Efficient C•G-to-G•C base editors developed using CRISPRi screens, target-library analysis, and machine learning

Development of C•G-to-G•C transversion base editors from CRISPRi screens, target-library analysis and machine learning

Luke W. Koblan[1,2,3,†], Mandana Arbab[1,2,3,†], Max W. Shen[1,2,3,4,†], Jeffrey A. Hussmann[5-7], Andrew V. Anzalone[1,2,3], Jordan L. Doman[1,2,3], Gregory A. Newby[1,2,3], Dian Yang[5-7], Beverly Mok[1,2,3], Joseph M. Replogle[5-7,8], Albert Xu[5-7,8], Tyler A. Sisley[2], Jonathan S. Weissman[5,7,9,10*], Britt Adamson[5-7,11,12*], David R. Liu[1,2,3*]

[1]*Merkin Institute of Transformative Technologies in Healthcare, Broad Institute of Harvard and MIT, Cambridge, MA, USA.*
[2]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA.*
[3]*Howard Hughes Medical Institute, Harvard University, Cambridge, MA, USA.*
[4]*Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, MA, USA.*
[5]*Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, CA 94158, USA*
[6]*Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA 94158, USA*
[7]*Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, CA 94158, USA*
[8]*Medical Scientist Training Program, University of California, San Francisco, San Francisco, CA 94158, USA*
[9]*Present address: Whitehead Institute for Biomedical Research, Cambridge, MA, USA.*
[10]*Present address: Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA.*
[11]*Present address: Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA*
[12]*Present address: Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA*

[†]*Denotes equal contribution*

*Correspondence should be addressed to Jonathan Weissman (weissman@wi.mit.edu), Britt Adamson (badamson@princeton.edu), and David R. Liu (drliu@fas.harvard.edu)

Abstract

Programmable C•G-to-G•C base editors (CGBEs) have broad scientific and therapeutic potential, but CGBE editing outcomes are unpredictable due to a high degree of target sequence dependence. We describe a suite of novel engineered CGBEs paired with machine learning models to enable efficient, high-purity C•G-to-G•C base editing. We performed a CRISPRi screen targeting 476 genes enriched for those with roles in DNA repair to identify factors that affect C•G-to-G•C editing outcomes and used these insights to develop CGBEs with diverse editing profiles. We characterized ten promising CGBEs on a library of 10,638 genomically integrated target sites in mammalian cells and trained machine learning models that accurately predict the purity and yield of edited outcomes (R=0.90) using these data. These CGBEs enable correction to wild-type coding sequence of 546 disease-related transversion SNVs with >90% precision and up to 70% efficiency. We demonstrate that computational prediction of optimal CGBE-sgRNA pairs enables high-purity transversion base editing at >4-fold more target sites than can be achieved using any single CGBE variant.

Single nucleotide variants (SNVs) represent approximately half of currently known human pathogenic gene variants[1]. Base editors, fusions of programmable DNA-binding proteins with base-modifying enzymes, enable conversion of individual target nucleotides in the genome[2-10]. The two major classes of base editors are cytosine base editors (CBEs), which convert C•G to T•A, and adenine base editors (ABEs), which convert A•T to G•C[2,3,8]. CBEs and ABEs can install transition mutations with high efficiency and product purity (the fraction of all edited alleles that contain only the desired edit), but in general cannot efficiently install transversion mutations including C•G to G•C[2,5,11,12].

We previously demonstrated that CBE editing byproducts, including C•G-to-G•C or C•G-to-A•T transversion outcomes, are inhibited by knockout of cellular uracil DNA N-glycosylase (UNG) or by fusion of uracil glycosylase inhibitor (UGI)[2,7,8,11,12], suggesting that transversion byproducts result from an abasic intermediate that is generated by UNG-catalyzed excision of deaminated target cytosines (**Fig. 1a**). Consistent with this model, first-generation C•G-to-G•C base editors (CGBEs) were CBE derivatives that lack UGI domains[11]. These CGBEs, including editors with fusions to UNG and other DNA-repair proteins[13-16], can provide efficient C•G-to-G•C editing but only at a minority of tested target sites with few criteria to identify sites amenable to CGBE editing[13-15].

Previously, we used libraries containing thousands of genomically integrated target sites and corresponding guide RNAs in mammalian cells to comprehensively characterize CBE and ABE base editing profiles. We used these data to train machine learning models (collectively named BE-Hive) that learned the sequence determinants driving CBE and ABE base editing outcomes[12,17]. We envisioned that broad characterization of the sequence determinants of CGBE editing outcomes could enable accurate prediction of editing efficiencies and product purities, and thus facilitate the broader use of CGBEs.

To reveal genetic determinants of cytosine transversion base editing, we performed a focused CRISPR interference (CRISPRi) screen to identify DNA repair genes that impact cytosine base editing efficiency and purity. Guided by these data, we constructed various fusions of DNA repair proteins, deaminases, and Cas proteins to engineer novel CGBEs with promising C•G-to-G•C editing activities. We characterized ten CGBEs with diverse editing profiles using a “comprehensive context library” of 10,638 genomically integrated, highly variable target sites in mouse embryonic stem cells (mESCs)[12]. We used the resulting data to train machine learning models that successfully predict CGBE editing efficiency, purity, and bystander editing patterns with high accuracy (CGBE-Hive), enabling reliable identification of

CGBE variants and target sites that together support high-purity C•G-to-G•C editing. We show that editing activity is predicted with substantially higher accuracy by deep learning models compared to simpler models, indicating that CGBE-Hive has learned complex sequence features that play important roles in determining C-to-G editing activity. Notably, 247 cytosines predicted by CGBE-Hive to be edited by a CGBE with >80% C•G-to-G•C editing purity were indeed edited in mammalian cell experiments with an average of 83% purity.

The panel of CGBEs in this study offer diverse editing profiles that collectively expand the sequence landscape amenable to high-quality C•G-to-G•C editing by up to 4.1-fold over the number predicted to be amenable to editing by any single CGBE. Finally, we demonstrate CGBE-mediated correction of 546 disease-associated single-nucleotide variants (SNVs) with >90% precision among the resulting edited amino acid sequences. These findings advance our understanding of transversion base editing outcomes and provide new CGBEs that improve the scope and utility of base editing.

Results

Exploring the activity of DNA glycosylases in C•G-to-G•C transversion outcomes

Previous work[2,11] suggested that excision of uracil from genomic DNA to form an abasic site is an important early step in transversion outcomes. These observations suggested that CBE-mediated transversions arise from uracil excision to generate an abasic lesion followed by error-prone polymerase activity on the strand opposite the abasic site (**Fig. 1a**)[2,11,16]. Motivated by this model, we sought to develop C•G-to-G•C base editors that enhanced uracil excision at CBE-edited nucleotides. We started with a CBE architecture lacking UGI (BE4B) (bpNLS–APOBEC1–Cas9 D10A–bpNLS; abbreviated AC), similar to other reported CGBEs[13-15].

We fused a variety of known uracil excising and binding enzymes to the C-terminus of the BE4B (AC) scaffold and assessed the frequency of C•G-to-G•C edits across five genomic loci in HEK293T cells (**Fig. 1b**). Several glycosylases (i.e., SMUG1, MBD4, and TDG2) did not alter editing outcomes, and fusion to UNG led to a reduction of C•G-to-G•C editing yield and purity at three out of five targeted sites, consistent with a recent report[13]. We found that fusion of a UNG orthologue from *B. smegmatis* (UdgX) moderately improved C•G-to-G•C product purity by 1.2-fold on average[18-20], with the largest improvement at the *RNF2* locus (56±0.8% with BE4B to 72±2.1% with AC–UdgX; p=0.0002, Student's two-sided t-test) and

significant changes observed at HEK site 2 C6, HEK site 3 C5, and EMX1 C6 ($p<0.01$, Student's two-sided t-test). However, we observed only modest changes to editing yield (1.1-fold relative to BE4B at the most efficiently edited C across the five tested genomic loci). These observations suggested that fusion partners may enhance C•G-to-G•C transversion base editing outcomes.

Next, we asked whether the orientation of the glycosylase fusion impacts editing outcomes. We constructed BE4B (AC) fusion variants with either UdgX (abbreviated X) or GFP in three orientations: at either the N- or C-terminus (e.g., XAC or ACX) or between the deaminase and Cas9 (e.g., AXC). We observed that C•G-to-G•C editing was similar or slightly improved for UdgX fusions compared to N- and C-terminal GFP fusions (**Fig. 1c**). However, the editing efficiency and purity of AXC was modestly higher than that of the best GFP fusion at a majority of sites (four out of five sites for efficiency; three out of five sites for purity). We chose to advance the AXC architecture since it offered similar or better performance than the XAC and ACX variants at these test loci.

CRISPRi screen for determinants of base editing outcomes

Next, we investigated whether other DNA repair or translesion synthesis factors impact C•G-to-G•C editing outcomes of AXC. We observed no significant changes in editing purity of AXC in individual *UNG*, *APE1/APEX1*, *MLH1*, *REV1* knockout cell lines and direct AXC fusions to mammalian polymerase domains did not consistently improve editing outcomes (**Supplementary Figs. 1-2; Supplementary Discussion 1**). We thus performed a much broader search for genetic modulators that impact cytosine transversion editing.

We performed a high-throughput CRISPRi-based screen designed to read out editing outcomes from BE1 (deaminase–dCas9) and BE4B (AC) editors by DNA sequencing (**Fig. 2a-b, Supplementary Fig. 3a**). We targeted 476 genes enriched for regulators of DNA repair using a custom sgRNA library comprising 1,513 targeting sgRNAs and 60 non-targeting sgRNA controls (**Supplementary Table 1**). We transduced this library into HeLa cells stably expressing the CRISPRi effector protein dSpCas9–KRAB[21]. After allowing 5 days for gene knockdown, we transfected the cells with plasmids encoding SaCas9-based CBEs (either SaCas9-BE1 or SaCas9-BE4B) and an SaCas9 sgRNA that targets a sequence adjacent to the genomically integrated SpCas9 sgRNA expression cassette. The proximity of the target site and CRISPRi sgRNA enabled these features to be read out together by paired-end DNA sequencing, linking editing outcomes to CRISPRi perturbation identities (**Fig. 2a**). Notably,

we used SaCas9-based CBEs to eliminate guide exchange between the base editors and CRISPRi machinery.

After base editing, we isolated genomic DNA from treated cells, affixed unique molecular identifiers (UMIs) to DNA fragments containing both the sgRNA expression cassettes and edited target sites, and sequenced the linked sgRNA, target sites, and UMI sequences. Comparing frequencies of editing outcomes from each CRISPRi sgRNA with those from non-targeting sgRNAs (**Fig. 2b, Supplementary Fig. 3a)** identified genes that promote or suppress various editing outcomes (**Supplementary Table 2**).

BE1 and BE4B editors showed strong baseline activity under screening conditions, enabling quantitation of editing differences driven by CRISPRi-sgRNAs (**Fig. 2, Supplementary Fig. 3, Supplementary Fig. 4**). To evaluate these outcomes, we calculated the effects of all CRISPRi sgRNAs on the frequencies of two major categories of edits: outcomes containing any C•G-to-T•A point mutation and outcomes containing any C•G-to-G•C point mutation (**Fig. 2c**). For both mutation classes, the effects of individual CRISPRi sgRNAs were consistent between replicates (**Fig. 2c**, upper left and lower right panels).

Some CRISPRi sgRNAs showed different effects on C•G-to-T•A versus C•G-to-G•C outcomes, indicating that specific genes influence partitioning between these outcomes (**Fig. 2c**, upper right panel). In the BE4B screen, the clearest differential effects resulted from sgRNAs targeting *UNG* (**Fig. 2b, c**). Consistent with the effects of UGI fusions and *UNG* loss[2,11], *UNG* knockdown increased frequencies of C•G-to-T•A editing while decreasing frequencies of C•G-to-G•C editing. Notably, the effects of *UNG* repression on BE1 editing were not as significant or straightforward (**Supplementary Fig. 3a,c**), perhaps reflecting differences in how nicked versus unnicked target substrates are processed (**Fig. 2b and Supplementary Fig. 3a)**.

Screens with sequencing-based readouts of editing outcomes can detect changes to a diverse range of editing products. For example, we observed that CRISPRi-mediated depletion of double-strand breaks (DSB) repair genes affect the frequency of rare indels that can result from base editing, though these pathway-phenotype relationships were not always straightforward (**Supplementary Fig. 4a, Supplementary Table 2**). For example, while knockdown of HDR factors *BRCA1*, *BRCA2*, and *PALB2* increased AC-generated deletions, depletion of the HDR gene *BLM* decreased them. Moreover, depletion of *BRCA2* was among the strongest reducers of C•G-to-T•A editing outcomes (**Supplementary Fig. 4b**). We also identified genes that affect the base editing window (**Supplementary Figs. 4c, 5;**

Supplementary Discussion 2). These phenotypes suggest a complex interplay between the function of these genes and the formation of DSBs by CBEs.

We also identified genes that specifically promote C•G-to-G•C editing. We calculated the relative fraction of sequencing reads containing any C•G-to-G•C edit among all reads containing any point mutation of the target for each CRISPRi sgRNA and identified genes whose knockdown significantly reduced the C•G-to-G•C editing fraction compared to non-targeting sgRNAs (**Fig. 2d** and **Supplementary Fig. 4d**). The strongest hit was *RFWD3*, an E3 ligase with multiple roles in DNA repair recently identified as required for successful translesion synthesis across a variety of genomic lesions [22]. In addition to UNG, other hits included multiple subunits of the replicative polymerase *POLD* and replicative clamp loader RFC; *EXO1*; translesion polymerases *REV1* and *REV3L*; and *RAD18*, an E3 ubiquitin ligase involved in translesion synthesis (**Supplementary Table 2**). The different outcomes for *REV1* knockdown versus our individual knockout cell line may arise from compensatory mechanisms that could alter DNA repair outcomes in cells lacking *REV1*. We also identified genes whose knockdown reduced frequencies of both C•G-to-T•A and C•G-to-G•C base editing for both BE1 and BE4B (**Supplementary Fig. 4e**), including *ASCC3*, which may act by affecting accessibility of the target locus, a known determinant of base editing efficiency[2,3,8]. Together, these screen results suggest important roles for DNA replication processes, especially translesion synthesis, in modulating C•G-to-G•C base editing outcomes.

CBE fusion proteins can alter C•G-to-G•C transversion outcomes

To further advance the development of CGBEs, we generated new CGBE candidates by fusing AXC, the prototype CGBE described above, to proteins encoded by 15 genes identified through the CRISPRi screen. These included genes that reduced C•G-to-G•C editing following knockdown, including *DDX1*, *EXO1*, *POLD1*, *POLD2*, *POLD3*, *RAD18*, *RBMX*, *REV1*, *RFWD3*, and *TIMELESS*, and several additional genes involved in DNA polymerization, some of which also affected editing outcomes in the CRISPRi screen (PCNA, POLH, POLK, UBE2I, and UBE2T, **Supplementary Table 2**). Notably, although replication clamp loader *RFC* components were observed among the genes whose knockdown most reduced C•G-to-G•C-specific editing, other members of the *RFC* ring complex were observed to affect C•G-to-T•A and indel outcomes (**Fig. 2d, Supplementary Fig. 4b, Supplementary Table 2**), leading us to exclude *RFC* factors from subsequent efforts.

We fused each of these proteins to the N- or C-terminus of AXC to assess their effect on C•G-to-G•C editing efficiency or purity and assessed their editing performance at five genomic loci in HEK293T cells. Three proteins increased C•G-to-G•C editing purity when fused to the N-terminus of AXC (**Supplementary Fig. 6a**): DNA polymerase D2 (POLD2), exonuclease 1 (EXO1), and RNA binding motif protein X-linked (RBMX). Editing improvements for fused constructs varied by site. The most pronounced effects were observed at the *RNF2* locus, where editing purity significantly improved from 54±1.4% with AXC to 73±0.4% with RBMX–AXC, 74±1.4% for EXO1–AXC, and 77±0.8% for POLD2–AXC ($p < 0.001$, Student's two-sided t-test). Marginal improvements in purity were also observed at HEK site 2, HEK site 3, and HEK site 4 loci. At *RNF2* we also observed a significant increase in editing yield from 43±2.4% with AXC to 50±5.2% with RBMX–AXC, 53±3.6% with EXO1–AXC, and 55± 5.5% for POLD2–AXC ($p < 0.05$, Student's two-sided t-test). C-terminal fusions typically did not perform as well as N-terminal fusions (**Supplementary Data 1**).

Motivated by these improvements, we developed additional candidate CGBEs containing RBMX, EXO1, POLD2, and UdgX as fusions to AXC. We compared single and dual pairwise fusion architectures for these components, testing N- and C-terminal dual fusions as well as tandem N terminal fusions (N-, N-) using 32-residue linkers identified in a linker-testing experiment for these constructs (**Supplementary Fig. 7**). From a total of 28 single- and dual-fusion proteins tested, the four dual fusion architectures POLD2–deaminase–UdgX–nCas9–RBMX, POLD2–deaminase–UdgX–nCas9–UdgX, UdgX–deaminase–UdgX–nCas9–UdgX, and UdgX–deaminase–UdgX–nCas9–RBMX further increased C•G-to-G•C editor yield and purity at some sites (on average, by +10% and +13%, respectively) compared to single fusion architectures across nine cytosines in five genomic loci (**Supplementary Fig. 6b**).

Collectively, these results indicate that CGBEs including fusions to DNA repair proteins identified in the CRISPRi screen can affect C•G-to-G•C editing outcomes in a site-dependent manner. Some base editing applications may prioritize protein size over other base editing characteristics. We explored the use of trans-splicing split-inteins as a means to reduce the size of large CGBEs into two smaller protein components[23], and observe no changes in editing outcomes of split-CGBEs compared to their full-length counterparts (**Supplementary Fig. 8**). When necessary, these split CGBE variants yield and purity benefits for cytosine transversion outcomes without requiring the expression of full-length proteins.

Base editor deaminase and Cas9 domains bias repair outcomes

We next sought to understand how different deaminase domains affect C•G-to-G•C editing in the AXC architecture. Since the base editing window may influence cytosine transversion outcomes[2,11,12], we examined a panel of catalytically impaired deaminases that support different CBE editing windows[24] and observed an increase in C•G-to-G•C editing purity at three of five tested loci (**Fig. 3a**). The APOBEC1 R126E R132E (EE)[24] deaminase showed the greatest improvement, averaging 1.2-fold higher product purity at HEK site 2, HEK site 3, and *RNF2*. Editing yield with these deaminase alternatives varied by locus. We observed similar or reduced editing yield compared to AXC at four out of five loci that is likely due to the lower catalytic activity of these deaminases, though reduced yield did not correlate with altered C•G-to-G•C purity. Editing yield by EE-AXC at the *RNF2* locus significantly improved (AXC=52±3.2% vs. EE-AXC=66±3.5%, p=0.0070, Student's two-sided t-test).

We also hypothesized that changes to the Cas9 binding domain of CGBEs could alter editing windows and C•G-to-G•C editing outcomes by altering the competition between Cas9 and repair machinery for access to the target locus. We assessed AXC editors that use Cas9 variants with different binding kinetics, including new variants with combinations of previously reported Cas9 mutations (**Fig. 3b**)[25-28]. AX–HF-nCas9 substantially improved C•G-to-G•C editing at the C9 position of the HEK site 3 locus, increasing yield (AXC=34±1.9% vs. AX–HF-nCas9=52±1.7%,) and purity (AXC=49±2.2% vs. AX–HF-nCas9=60±1.2%) ($p < 0.005$ for both, Student's two-sided t-test) (**Fig. 3b**). AX-Hypa-nCas9 showed similar effects but AX-HF-nCas9 typically performed modestly better. These results suggest Cas protein binding parameters can affect C•G-to-G•C editing yield and purity of CGBEs at some target loci.

The balance of editing yield and purity among candidate CGBEs and the variability in these two measures across different loci suggests that different target sites will be best edited by different CGBEs. Therefore, a suite of CGBEs with different kinetics and substrate preferences would likely enable efficient and high-purity C•G-to-G•C editing across a broader range of diverse target sequences than could be achieved by any single CGBE variant alone.

Combining deaminase, Cas9 domain, and DNA repair fusion proteins into new CGBEs

We integrated the above findings from varying protein fusions, deaminases, and Cas domains into improved CGBEs. We evaluated the four most promising dual-fusion AXC editors (POLD2–AXC–RBMX, POLD2–AXC–UdgX, UdgX–AXC–RBMX, and UdgX–AXC–

UdgX), four single-fusion AXC editors (POLD2–AXC, RBMX–AXC, EXO1–AXC, and UdgX–AXC), AXCs with deaminase variants of those same editors, and direct deaminase–nCas9 CGBEs without additional fusion proteins. The five cytidine deaminases tested in these 10 CGBE architectures included rAPOBEC1, EE, Anc689 (ancestrally-reconstructed APOBEC1 node 689[29]), eA3A, and eA3A-T31A[12]. In addition, we tested both SpCas9 nickase and HF-Cas9 nickase variants. In total, we evaluated 95 candidate CGBEs at eight genomic loci in HEK293T cells.

No single CGBE outperformed all other candidates at all sites (**Fig. 4a**). To identify a set of the most promising CGBEs, we selected 32 editors that demonstrated improved C•G-to-G•C editing outcomes at some sites for testing at eight additional genomic loci (**Fig. 4b**). We used these data to identify ten CGBEs with high purity, yield, and maximally distinct activities at different endogenous loci using quadratic programming and hierarchical clustering (Supplementary Methods): 689–nCas9, UdgX–689–UdgX–nCas9–RBMX, eA3A–nCas9, RBMX–eA3A–UdgX–HFnCas9, RBMX–eA3A–UdgX–nCas9, EE–nCas9, UdgX–EE–UdgX–nCas9–UdgX, APOBEC1–nCas9, UdgX–APOBEC1–UdgX–HFnCas9, and POLD2–APOBEC1–UdgX–nCas9–UdgX.

To test how this set of CGBEs performed in human cell lines other than HEK293T cells, we assayed the ability of each of these CGBEs to edit five target genomic sites in K562, U2OS, and HeLa (**Supplementary Fig. 9**). We observed that while CGBE outcomes vary modestly by cell type, the top-performing CGBE variants for each tested site were generally the same in all three additional cell lines. These results indicate that deaminase, Cas protein, and DNA repair protein variants can improve C•G-to-G•C editing in across different cell types.

Target library characterization of CGBEs

We observed that different target loci were best edited by different CGBEs, indicating that diverse CGBE sequence preferences may be strong determinants of C•G-to-G•C editing efficiency and purity. Previously, we used high-throughput analysis of base editing outcomes at thousands of genomically integrated target sequences to better understand CBE and ABE sequence-activity relationships, and we used these data to train machine learning models that facilitate the selection of target sequences amenable to C•G-to-G•C conversion by CBEs[12]. We envisioned that comprehensive characterization of our top ten promising and diverse CGBEs could similarly aid in the selection of targets amenable to efficient and high-purity C•G-to-G•C editing by specific CGBEs.

We characterized each of the ten CGBEs using a high-throughput genome-integrated library assay of 10,638 matched sgRNA and target pairs in mESCs, previously referred to as the "comprehensive context library"[12]. The target sequences in this library cover all possible sequence contexts surrounding the edited C•G with minimal sequence bias (**Fig. 5a**, Online Methods). To detect editing outcomes with high sensitivity, we maintained an average coverage of ≥300x per library member throughout the course of the experiment and an average sequencing depth of ≥4,000x per target. We collected two biological replicates per CGBE characterization experiment. We previously validated that the library assay data has strong consistency between biological replicates and is concordant with data from base editing endogenous genomic loci[12,30].

We used the resulting library data to quantify editing windows and product purities for each CGBE (**Fig. 5b**, Online Methods). CGBE editing activity was generally centered around protospacer position 6 with editing window widths ranging from 3 nt (EE–nCas9; positions 5-7) to 8 nt (UdgX–APOBEC1–UdgX–HF-nCas9 nickase; positions 4-11). The editing windows of CGBEs with additional components beyond Cas and deaminase domains were shifted by up to 3 nt compared to direct deaminase–Cas fusions, indicating that CGBE protein fusions can affect editing window size and position.

Engineered CGBE architectures showed significant improvements in C•G-to-G•C product purity compared to simple deaminase–nCas9 fusions. Across the 10,638 target sites in the comprehensive context library, the fusion CGBEs POLD2–APOBEC1–UdgX–nCas9–UdgX, UdgX–EE–UdgX–nCas9–UdgX, and UdgX–689–UdgX–nCas9–RBMX showed 25% higher mean C•G-to-G•C purity than their corresponding deaminase–nCas9 counterparts within each editor's editing window ($P < 5.1\times10^{-9}$; Welch's t-test) (**Fig. 5c**). We observed large variation in CGBE editing efficiency, with mean efficiency ranging from 1.8% by UdgX–EE–UdgX–nCas9–UdgX to 23.0% by Anc689–nCas9 across the comprehensive context library within the same experimental batch. Notably, the protein fusion CGBEs exhibiting increased C•G-to-G•C purity also reduced editing yield by 1.4- to 1.6-fold on average.

C•G-to-G•C editing purity exceeded 90% for at least one of the tested CGBEs at 895 cytosines across the comprehensive context library. Some cytosines edited with purities as high as 90-100% by some CGBEs were edited with purity as low as 0-10% by other CGBEs, indicating that these CGBEs indeed offer complementary editing characteristics, and confirming that a panel of diverse CGBEs maximizes the utility of C•G-to-G•C base editing compared to using any single CGBE (**Fig. 5d**). We clustered CGBEs by C•G-to-G•C editing

purity across the comprehensive context library and observed that engineered CGBEs did not cluster by deaminase (**Fig. 5e**), indicating that protein fusion engineering of CGBE architectures resulted in distinct sequence preferences governing C•G-to-G•C editing.

Sequence determinants and machine learning modeling of CGBE activity

C•G-to-G•C product purity of CGBEs varies substantially by sequence context (**Fig. 5f**). We observed 24.7±26.3% average C•G-to-G•C purity across all tested CGBEs for cytosines positioned near the center of the editing window, with substantial variation across target sequences: the top 5% had >79.6% C•G-to-G•C purity while the bottom 5% had <1.0%. To decipher the sequence determinants that underly CGBE activity, we computed simple motifs for editing efficiency and transversion purity using a logistic regression model that considers each nucleotide independently (**Fig. 5g**, Online Methods)[12]. These motifs revealed that TC is strongly favored while GC is disfavored for editing efficiency across the tested CGBEs. We further trained gradient-boosted regression trees to predict CGBE editing efficiency sequence context, which achieved good accuracy with R=0.57-0.77 at held-out target sites. Consistent with our previous characterization of BE4 variants[12], we observed sequence motifs that associated RCTA with higher C•G-to-G•C purity (R=A or G) across all characterized CGBEs. Cytosines in an ACTA motif were edited with an average C•G-to-G•C purity of 68.7% (N=1,760) across CGBEs, substantially higher than the 24.7% average across all sequence contexts, indicating a major role for sequence context in determining C•G-to-G•C editing outcomes. These simple target sequence motifs predicted 27.0%-53.3% of the variation in C•G-to-G•C purity.

Next, we trained BE-Hive models for these ten CGBEs (termed CGBE-Hive) and evaluated the models' ability to predict C•G-to-G•C editing purity at held-out sequence contexts not seen during training. These models explained 58.3%-76.3% of the variance in C•G-to-G•C purity in the held-out dataset, a substantial improvement over logistic regression described above (27.0%-53.3%) (**Fig. 5h**). This performance improvement highlights that while C•G-to-G•C purity can be predicted using a simple motif such as RCTA that considers each nucleotide independently, higher-order interactions between nucleotides learned by deep neural networks substantially improve C•G-to-G•C editing purity predictions. Collectively, these observations establish that CGBE editing efficiency and purity can be accurately predicted by machine learning models.

To further investigate sequence determinants of CGBE editing outcomes, we calculated target sequence motifs for cytosines with the highest C•G-to-G•C efficiency for each CGBE (Online Methods). While most CGBEs shared sequence preferences favoring TC for overall editing efficiency and RCTA for purity, different CGBEs had distinct motifs that correlated with C•G-to-G•C yield. POLD2–APOBEC1–UdgX–nCas9–UdgX favored RCTA for C•G-to-G•C yield, while eA3A–nCas9 simply favored TC (**Fig. 5i**). Interestingly, RBMX–eA3A–UdgX–nCas9 favored CTC, while UdgX–EE–UdgX–nCas9–UdgX favored TCT, and 689–nCas9 favored CTA (**Fig. 5i**). These observations reveal that different CGBEs show distinct sequence preferences that influence the yield of C•G-to-G•C outcomes.

We provide machine learning models trained on up to 10,638 sgRNA-target pairs for these ten CGBEs in our online interactive web app (www.crisprbehive.design)[12]. Users can query sgRNAs and target sequences for data-driven predictions on editing outcomes of all CGBEs characterized in this study.

Model-guided correction of pathogenic transversion SNVs

To extend the applicability of these CGBEs, we assessed their compatibility with PAM-variant Cas9 proteins. We evaluated editing at eight loci by CGBEs using Cas9-NG, an engineered SpCas9 variant with broadened PAM compatibility[31], and observed similar editing purities to SpCas9 CGBEs at NGG PAM substrates (**Supplementary Fig. 10, 11**). The best performing NG-CGBEs at each locus retained >50% yield relative to SpCas9 CGBEs at targets with NGG PAMs (**Supplementary Fig. 10**).

Given the broadened targeting scope of NG-CGBEs we sought to characterize their performance on the "transversion-enriched SNV library"[12] in mESCs, which contains 3,400 sgRNA-target pairs selected by BE-Hive from 18,523 disease-related G•C-to-C•G and A•T-to-C•G SNVs from the ClinVar and HGMD databases that are targetable by Cas9-NG[1,32], predicted to be correctable by cytosine transversion base editing with high purity and yield. We generated the following NG-CGBEs based on their performance on the comprehensive context library: 689–nCas9-NG, APOBEC1–nCas9-NG, eA3A–nCas9-NG, UdgX–689–UdgX–nCas9-NG –RBMX, and UdgX–APOBEC1–UdgX–HFnCas9-NG. As Cas9-NG generally demonstrates reduced editing activity compared to wild-type SpCas9[31], similar to HF-Cas9, we included UdgX–APOBEC1–UdgX–nCas9-NG without the HF modifications as an alternative binding-impaired Cas9-fusion variant.

All six CGBEs tested on the transversion-enriched SNV library enabled high-purity C•G-to-G•C editing at disease-associated SNVs. At 247 cytosines predicted by CGBE-Hive to have >80% C•G-to-G•C editing purity, CGBEs demonstrated an average of 83% C•G-to-G•C editing purity (**Fig. 6a**). Each CGBE corrected > 200 SNVs to their wild-type coding sequence with >90% precision among edited amino acid sequences (amino acid correction precision; **Fig. 6b**), with a total of 546 unique SNVs across CGBEs. For example, in the genome-integrated library, eA3A–nCas9-NG corrected the G•C-to-C•G SNV in COL3A1 associated with Ehlers-Danlos syndrome[33] with 71.4% yield and 92.8% purity, and corrected an SNV in BRCA2 associated with familial breast and ovarian cancer[34] with 66.5% yield and 82.5% purity. The fusion CGBE UdgX–APOBEC1–UdgX–nCas9-NG corrected an SNV in NSD1 associated with Sotos syndrome[35] with 40.0% yield and 73.4% purity and corrected an SNV in NIPBL associated with Cornelia de Lange syndrome[36] with 38.8% yield and 76.9% purity. Collectively, these results reveal efficient and high-purity correction of hundreds of disease-related SNVs by CGBEs.

Notably, the UdgX–APOBEC1–UdgX–nCas9 CGBE maintained a similar high purity of C•G-to-G•C editing between HF-nCas9 and nCas9-NG variants. UdgX–APOBEC1–UdgX–nCas9-NG, however, offered substantially better yield of genotype and coding sequence corrected G•C-to-C•G SNVs (**Fig. 6a,b**). These results suggest that fusion of CGBEs to Cas9-NG variants may obviate the need to use HF-variant Cas9-proteins to alter their binding kinetics to promote C•G-to-G•C editing outcomes.

The best-edited targets in the transversion-enriched SNV library varied greatly by CGBE. Some SNVs edited with >90% purity by one CGBEs had purity below 5% for other CGBEs (**Supplementary Fig. 12**). CGBE-Hive models accurately accounted for this diversity in editing purity in the transversion-enriched SNV library, and accurately predicted the yield of exact genotype correction products and of alleles with corrected amino acid sequences (R=0.89-0.93 and R=0.91-0.94, respectively, **Fig. 6c**), as well as the DNA and amino acid correction precision (R=0.77-0.85 and R=0.82-0.90, respectively, **Fig. 6d**), including targets with multiple cytosines in the editing window. Since accurately predicting correction yield and precision requires accurate predictions for CGBE efficiency, C•G-to-G•C purity, and bystander editing patterns, these results establish that CGBE-Hive has learned important aspects of CGBE editing activity and can guide the use of CGBEs for high-purity correction of disease-related transversion SNVs.

Using CGBE-Hive to pick the best among the characterized CGBEs to correct each SNV should achieve greater C•G-to-G•C correction than applying any single CGBE to a set of targets. Indeed, we observed that using CGBE-Hive to choose the three CGBE variants predicted to best achieve the desired edit (top-3 performance) increased the number of targets corrected with ≥90% precision or to ≥40% efficiency by 4.1- and 5.0-fold, respectively, compared to the number of targets that are expected to be corrected with these precision and efficiency thresholds by picking any single CGBE (**Fig. 6e**). These improvements of 4.1- and 5.0-fold by using the top three CGBE-Hive choices were nearly identical to the performance from picking the best CGBE out of all six options in hindsight. CGBE-Hive also displayed strong top-1 performance: Using CGBE-Hive to choose just a single CGBE increased the number of targets corrected with ≥90% precision or to ≥40% efficiency to 1.7- and 4.0-fold, respectively, compared to picking a single CGBE in expectation.

For correction precision, CGBE-Hive recovered the best performing CGBE variant in its top choice in 43.3% of targets and in its top three choices in 84.2% of target sequences. For correction yield, CGBE-Hive recovered the best-performing CGBE variant in its top choice in 67.5% of targets and in its top three choices in 97.2% of targets. These results collectively demonstrate that this panel of CGBEs have diverse editing activities that CGBE-Hive has learned to predict, to optimize selection of the most promising CGBE variant to use for a desired edit. These improvements were also observed at endogenous loci in HEK293T cells (**Fig. 6f, Supplementary Discussion 3**). Thus, CGBE-Hive enables researchers to reap the benefits of the diversity of CGBEs developed in this study without the need to test all CGBE variants.

Comparisons with recently reported CGBEs, prime editing, and off-target profiling

Next, we determined whether the CGBE variants described in this work extend the scope of C•G-to-G•C base editing beyond those accessible with recently described CGBEs or PE. We were encouraged to find that the CGBEs developed in this study extend the scope of C•G-to-G•C genome editing by enabling higher yields and product purities at a wider array of target sequences compared to the use of previously described CGBEs alone except at loci already edited with high yield and purity by deaminase–nCas9 constructs (**Supplementary Fig. 13; Supplementary Discussion 4**). Furthermore, we observed that these novel CGBEs complement prime editing (PE) technology[37]. We found PE typically offers higher product purities while editing with CGBEs offers higher editing yields at some loci (**Supplementary**

Fig. 14; Supplementary Discussion 5), consistent with recent reports[13-15,37]. Notably, prime editing currently requires extensive optimization of pegRNA features to achieve high-efficiency edits, while CGBE-Hive prediction obviates CGBE editor selection. CGBEs complement prime editing for efficient C•G-to-G•C editing, although additional optimization of both technologies may further improve their properties.

We also sought to characterize potential off-target editing outcomes of CGBEs. Since the genome-wide off-targets of base editors that use cytosine deaminase enzymes are known to be predominantly sgRNA dependent, we characterized Cas9-dependent off-target editing profiles of CGBEs by examining the activity of CGBEs at previously confirmed off-target loci of corresponding Cas9:sgRNA complexes[8]. The architectural changes and protein fusions used to develop the CGBEs in this study resulted in lower Cas9-dependent off-target editing compared to corresponding CGBEs lacking protein fusions (**Supplementary Fig. 11, 15**), despite their generally higher on-target editing, perhaps because the more complex fusions or architectural changes introduce additional conformational requirements in editor:DNA complexes that are not met by some off-target loci (see **Supplementary Discussion 6**). While DNA repair protein CGBE components may result in additional Cas-independent off-target effects, these are likely to differ by cell type and delivery method, and therefore are best assessed for each application.

Discussion

Understanding and controlling the outcomes of genome editing experiments are important challenges for achieving targeted, precise genome manipulation. We investigated molecular determinants of transversion base editing, including the effects of the deaminase, Cas effector domain, and many DNA repair proteins identified in targeted CRISPRi screens, and used these insights to engineer novel CGBEs. We characterized the editing outcomes and performance of these reagents using a high-throughput genome-integrated library assay in mammalian cells and identified sequence features that affect base editing outcomes of ten diverse CGBEs. We showed that C-to-G editing activity is predicted with substantially higher accuracy by deep learning models compared to simpler models, indicating that complex sequence features drive C•G-to-G•C editing activity.

We provide trained CGBE-Hive machine learning models which accurately predict CGBE efficiency, C•G-to-G•C editing purity, and bystander editing patterns (R=0.90) to enable predictable and consistently pure CGBE editing. We demonstrate a machine learning

workflow using CGBE-Hive to identify optimal CGBE and sgRNA editing strategies to install a desired edit and show that this workflow expands high-efficiency and high-purity C•G-to-G•C editing to more loci than using any single CGBE by 5.0-fold and 4.1-fold with the top three CGBE-nominated choices. We demonstrate CGBE-mediated correction of the amino acid sequences of 546 disease-associated single nucleotide variants (SNVs) with >90% precision. Furthermore, we demonstrated efficient and pure installation of four disease-relevant SNPs and tested the performance of these tools in other mammalian cell lines. Collectively, the base editor and computational tools presented in this work substantially improve the targeting scope, effectiveness, and utility of CGBE-mediated transversion base editing.

Data and code availability

The target library sequencing data generated during this study are available at the NCBI Sequence Read Archive database under PRJNA631290. Processed target library data used for training machine learning models have been deposited under the following DOIs: 10.6084/m9.figshare.12275645 and 10.6084/m9.figshare.12275654. Code used for analyzing CRISPRi screens is available at github.com/jeffhussmann/repair-seq. Code used for target library data processing and analysis are available at https://github.com/maxwshen/lib–dataprocessing and https://github.com/maxwshen/lib–analysis. The machine learning models for CGBEs trained on target library data are available as a part of the BE-Hive interactive web application at https://crisprbehive.design and the BE-Hive Python package at https://github.com/maxwshen/be_predict_efficiency and https://github.com/maxwshen/be_predict_bystander.

Acknowledgements

This work was supported by U.S. NIH U01AI142756, UG3AI150551, RM1HG009490, and HHMI. Research reported in this publication was supported by NIGMS of the National Institutes of Health under award number R35GM138167-01 and the Searle Scholars Program (B.A.). The authors acknowledge NSF Graduate Research Fellowships to L.W.K., M.W.S., and T.A.S.; a NWO Rubicon Fellowship to M.A.; a Jane Coffin Childs postdoctoral fellowship to A.V.A.; fellowship support from the NSF and Hertz Foundation to J.L.D.; a Helen Hay Whitney postdoctoral fellowship to G.A.N.; a Damon Runyon Postdoctoral Fellowship to D.Y.; a Singapore A*STAR NSS fellowship to B.M.; and an NIH/NINDS Ruth L. Kirschstein National Research Service Award to J.M.R. J.A.H. was the Rebecca Ridley Kry Fellow of the Damon Runyon Cancer Research Foundation.

Author contributions

L.W.K, M.A., M.W.S., J.A.H., A.V.A., J.S.W., B.A., D.R.L. designed the research. L.W.K., M.A., M.W.S., J.A.H., A.V.A., J.L.D., G.A.N., D.Y., B.M., J.M.R., A.X., T.A.S., B.A. performed experiments. J.S.W., B.A., and D.R.L. supervised the project. L.W.K. and D.R.L. wrote the manuscript with input from all other authors.

Author information

The authors declare competing financial interests: J.A.H. is a consultant for Tessera Therapeutics. J.M.R. is a consultant for Maze Therapeutics. JSW is a consultant for and holds equity in Maze Therapeutics, Chroma Medicine, and KSQ Therapeutics. B.A. was a member of a ThinkLab Advisory Board for Celsius Therapeutics. D.R.L. is a consultant for and holds equity in Beam Therapeutics, Prime Medicine, Pairwise Plants, and Chroma Medicine.

References

1. Landrum, M.J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research* **44**, D862-D868 (2016).
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. Gaudelli, N.M. et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464-471 (2017).
4. Gehrke, J.M. et al. An APOBEC3A-Cas9 base editor with minimized bystander and off-target activities. *Nature Biotechnology* **36**, 977-982 (2018).
5. Nishida, K. et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. *Science* **353**, aaf8729-aaf8729 (2016).
6. Richter, M.F. et al. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nature Biotechnology* **38**, 883-891 (2020).
7. Rees, H.A. & Liu, D.R. Base editing: precision chemistry on the genome and transcriptome of living cells. *Nature Reviews Genetics* **19**, 770-788 (2018).
8. Anzalone, A.V., Koblan, L.W. & Liu, D.R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology* **38**, 824-844 (2020).
9. Gaudelli, N.M. et al. Directed evolution of adenine base editors with increased activity and therapeutic application. *Nature Biotechnology* **38**, 892-900 (2020).
10. Mok, B.Y. et al. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature* **583**, 631-637 (2020).
11. Komor, A.C. et al. Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Science Advances* **3**, eaao4774 (2017).
12. Arbab, M. et al. Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning. *Cell* **182**, 463-480.e430 (2020).
13. Kurt, I.C. et al. CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. *Nature Biotechnology* (2020).
14. Zhao, D. et al. Glycosylase base editors enable C-to-A and C-to-G base changes. *Nature Biotechnology* (2020).
15. Chen, L. et al. Precise and programmable C: G to G: C base editing in genomic DNA. *bioRxiv* (2020).
16. Liu, D.R.a.K., L.W. Cytosine to Guanine Base Editor. *World Intellectual Property Organization* (2018).
17. Marquart, K.F. et al. Predicting base editing outcomes with an attention-based deep learning algorithm trained on high-throughput target library screens. *bioRxiv* (2020).
18. Sang, P.B., Srinath, T., Patil, A.G., Woo, E.-J. & Varshney, U. A unique uracil-DNA binding protein of the uracil DNA glycosylase superfamily. *Nucleic Acids Res* **43**, 8452-8463 (2015).
19. Ahn, W.-C. et al. Covalent binding of uracil DNA glycosylase UdgX to abasic DNA upon uracil excision. *Nat Chem Biol* **15**, 607-614 (2019).
20. Tu, J., Chen, R., Yang, Y., Cao, W. & Xie, W. Suicide inactivation of the uracil DNA glycosylase UdgX by covalent complex formation. *Nat Chem Biol* **15**, 615-622 (2019).
21. Gilbert, L.A. et al. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442-451 (2013).
22. Gallina, I., Hendriks, I.A., Hoffmann, S., Larsen, N.B., Johansen, J., Colding-Christensen, C.S., Schubert, L., Sellés-Baiget, S., Fábián, Z., Kühbacher, U., Gao,

A.O., Räschle, M., Rasmussen, S., Nielsen, M.L., Mailand, N., Duxin, J.P. The ubiquitin ligase RFWD3 is required for translesion DNA synthesis. *Molecular Cell* **81**, 1-17 (2020).
23. Levy, J.M. et al. 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).
24. Kim, Y.B. et al. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nature Biotechnology* **35**, 371-376 (2017).
25. Kleinstiver, B.P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490-495 (2016).
26. Slaymaker, I.M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84-88 (2015).
27. Chen, J.S. et al. Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407-410 (2017).
28. Lee, J.K. et al. Directed evolution of CRISPR-Cas9 to increase its specificity. *Nature Communications* **9**, 3048 (2018).
29. Koblan, L.W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nature Biotechnology* **36**, 843-846 (2018).
30. Shen, M.W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646-651 (2018).
31. Nishimasu, H. et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259-1262 (2018).
32. Stenson, P.D. et al. Human Gene Mutation Database: towards a comprehensive central mutation database. *Journal of Medical Genetics* **45**, 124-126 (2007).
33. Frank, M. et al. The type of variants at the COL3A1 gene associates with the phenotype and severity of vascular Ehlers–Danlos syndrome. *European Journal of Human Genetics* **23**, 1657-1664 (2015).
34. Petrucelli, N., Daly, M.B. & Feldman, G.L. Hereditary breast and ovarian cancer due to mutations in BRCA1 and BRCA2. *Genetics in Medicine* **12**, 245-259 (2010).
35. Douglas, J., Hanks, S., Temple, I.K. & of …, D.-S. NSD1 mutations are the major cause of Sotos syndrome and occur in some cases of Weaver syndrome but are rare in other overgrowth phenotypes. *The American Journal of …* (2003).
36. Luna-Peláez, N. et al. The Cornelia de Lange Syndrome-associated factor NIPBL interacts with BRD4 ET domain for transcription control of a common set of genes. *Cell Death Dis* **10** (2019).
37. Anzalone, A.V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149-157 (2019).

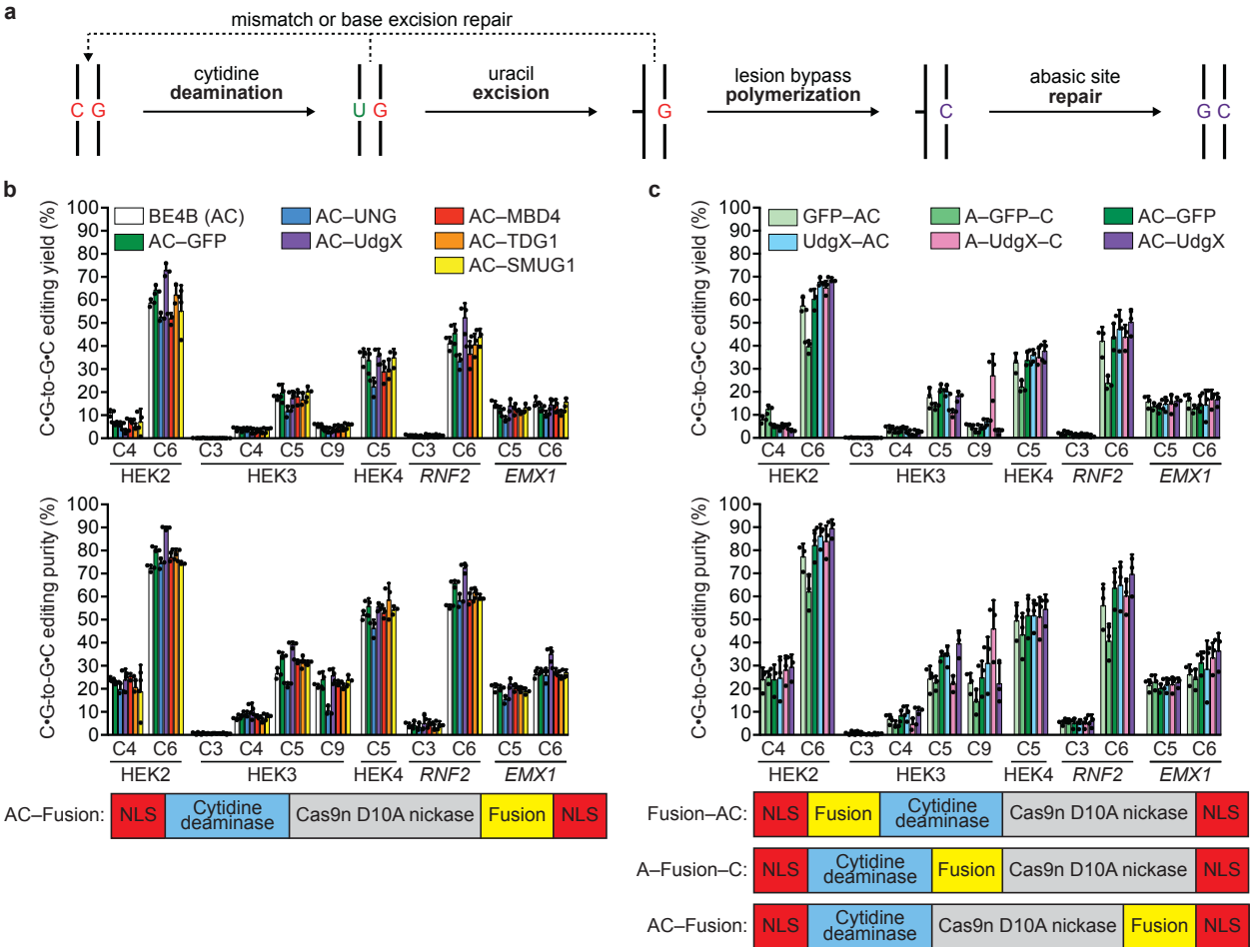

Figure 1. Development of prototype C•G-to-G•C base editors. (**a**) Potential pathway for C•G-to-G•C conversion. (**b**) C•G-to-G•C editing outcomes in HEK293T cells for C-terminal fusions of DNA glycosylases to BE4B (AC, APOBEC1 cytidine deaminase–Cas9 nickase). (**c**) Different fusion protein architectures lead to different C•G-to-G•C editing properties in HEK293T cells at the HEK3 locus for the Apo-UdgX-Cas9n (AXC) architecture. Values and error bars reflect the mean and standard deviation of three biological replicates, shown as individual data points. HEK2=HEK site 2; HEK3=HEK site 3; HEK4=HEK site 4. C4, C6, and similar annotations indicate the in-window target nucleotides where the SpCas9 PAM is at positions 21-23.

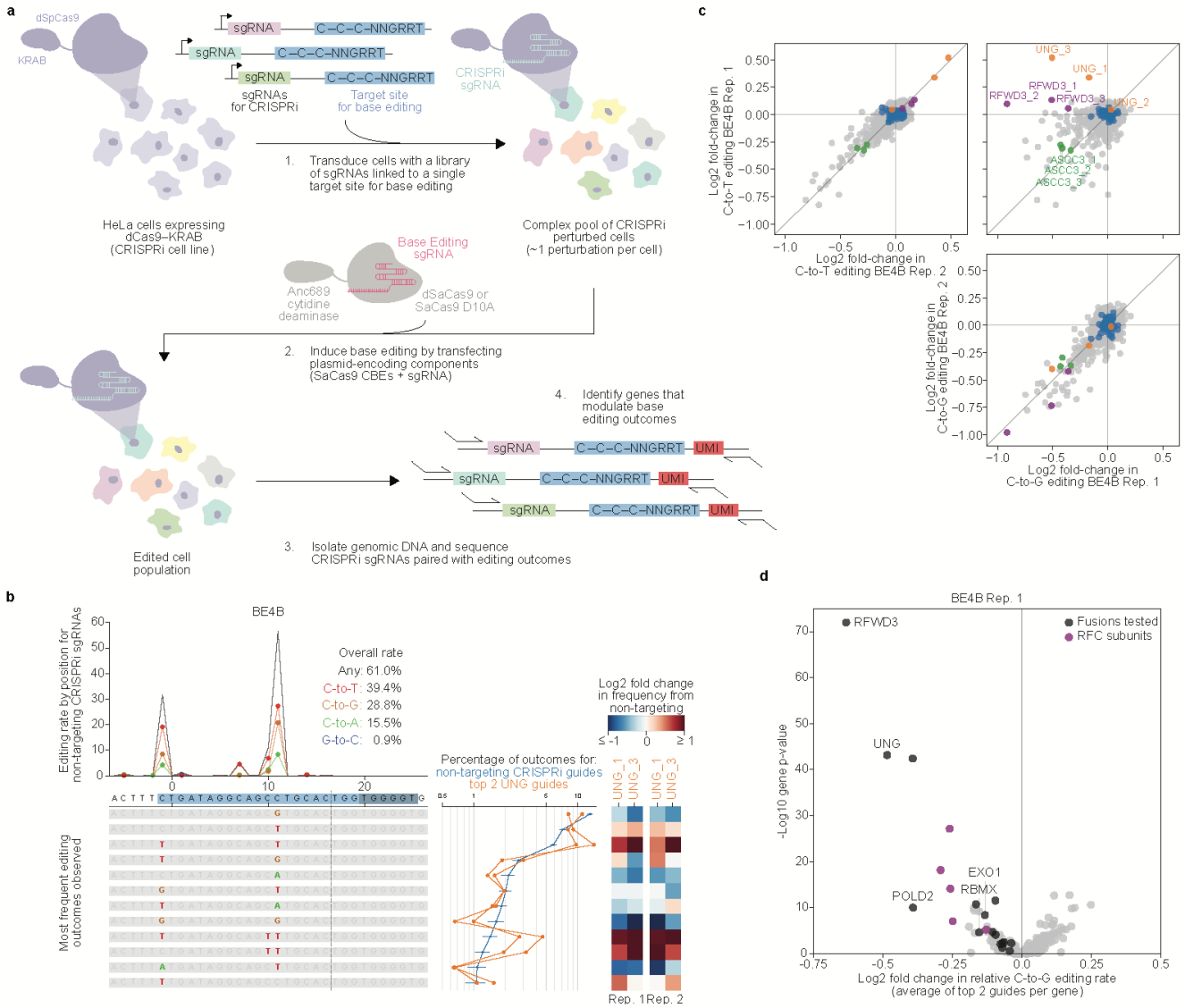

Figure 2. CRISPRi knockdown screen across 476 genes enriched for those with roles in DNA repair identifies candidate regulators of C•G-to-G•C editing. (**a**) Schematic of screen design. (**b**). Summary of base editing outcomes in BE4B (also AC) screen. Bottom left – all editing outcomes containing only point mutations present at >=1% frequency for non-targeting CRISPRi guides. Line plots above the individual outcomes show the total editing frequency (black line) and the frequencies of each single base edit (C-to-T=red, C-to-G=brown, C-to-A=green, and G-to-C=blue lines) at each position. Line plots to the right show frequencies of outcomes for specific CRISPRi guides (blue - average of all non-targeting guide +/- standard deviation across individual non-targeting guides; orange - top 2 most active *UNG* guides). Heatmaps show Log_2 fold changes in outcome frequencies for top 2 *UNG* guides relative to non-targeting guides. (**c**) Log_2 fold changes in frequency of outcomes containing C-to-T or C-to-G edits for each CRISPRi guide compared to non-targeting guides. Upper left - comparison of changes in C-to-T editing between two biological replicates. Lower right - comparison of changes in C-to-G editing between replicates. Upper right - comparison of changes in C-to-G editing to changes in C-to-T editing in replicate 1. All guides with at least 500 recovered UMIs in each replicate are plotted. Blue dots: individual non-targeting guides, orange dots: *UNG* guides, green dots: *ASCC3* guides, red dots: *RFWD3* guides, grey

dots: all other guides. (**d**) Effects of gene knockdown on relative C-to-G editing frequencies in BE4B screen. Each dot represents a gene, with the x-value representing the average of the two strongest Log_2 fold changes in normalized C-to-G editing for guides targeting the gene from the average of all non-targeting guides, and the y-value representing a gene-level p-value summarizing the combined statistical significance of all guides targeting each gene. Rep.=replicate.

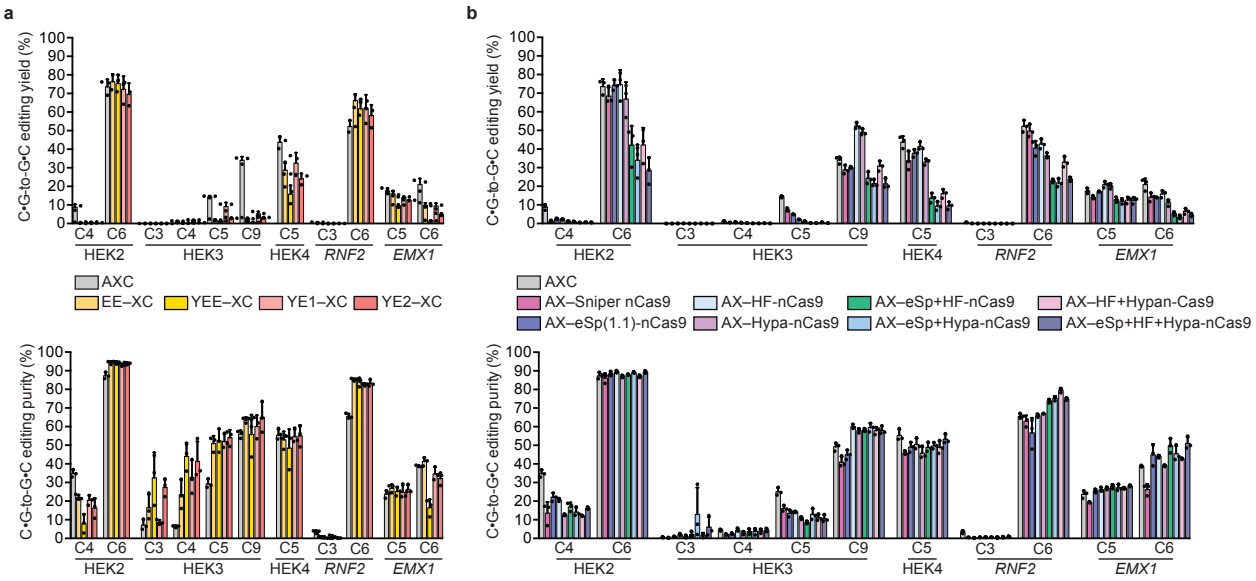

Figure 3. Effect of varying the cytidine deaminase and Cas9 components of CGBEs on C•G-to-G•C editing outcomes in HEK293T cells. (**a**) C•G-to-G•C editing outcomes for catalytically impaired, narrow-window cytidine deaminases show higher editing purity at HEK2 and *RNF2*. (**b**) C•G-to-G•C editing outcomes for high-fidelity Cas9 variants show altered editing windows and improved CGBE performance at some positions. “Cas9” represents the Cas9 D10A nickase variant of each Cas effector. Values and error bars reflect the mean and standard deviation of three biological replicates, shown as individual data points. HEK2=HEK site 2; HEK3=HEK site 3; HEK4=HEK site 4. C4, C6, and similar annotations indicate the in-window target nucleotides where the SpCas9 PAM is at positions 21-23.

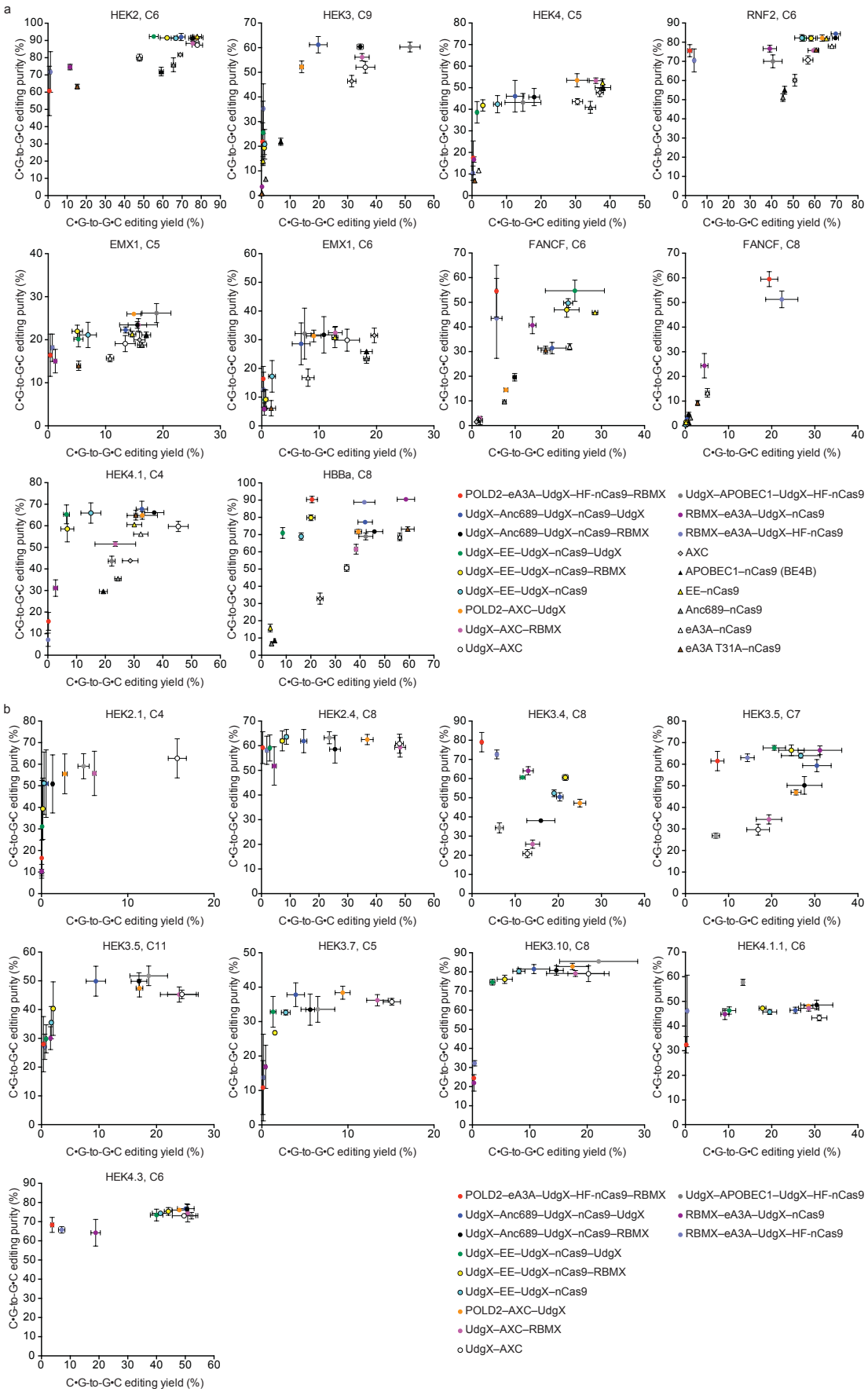

Figure 4. Novel engineered CGBEs with various DNA repair proteins, deaminases, Cas proteins, and architectures offer diverse editing performance on different target sites.

(**a**) C•G-to-G•C editing performance of CGBEs at eight genomic loci in HEK293T cells. (**b**) Further characterization of C•G-to-G•C editing outcomes for 12 variants from (a) at various genomic loci in HEK293T cells. Values and error bars reflect the mean and standard deviation of three biological replicates. HEK2=HEK293T cells site 2; HEK3=HEK293T cells site 3; HEK4=HEK293T cells site 4. C nucleotide annotations indicate the target nucleotide positions in the protospacer, where the SpCas9 PAM is at positions 21-23. Editing efficiencies, product purities, and indel frequencies for constructs that were tested but not shown in this figure can be found in Supplementary Data 1.

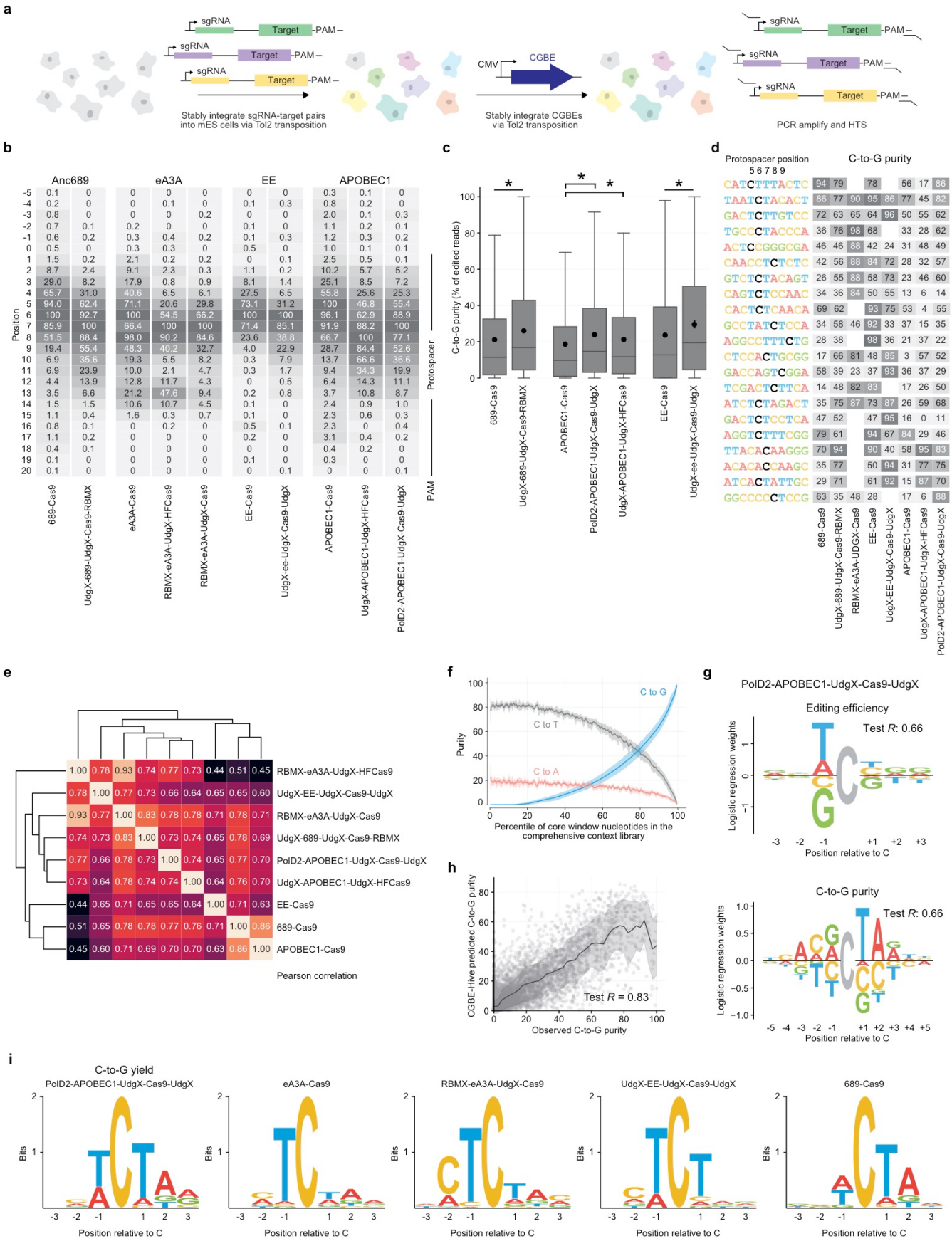

Figure 5. Target library characterization and machine learning modeling of 10 CGBE variants. (**a**) Overview of genome-integrated target library assay. Libraries of 12,000 or 4,000

pairs of sgRNAs and corresponding target sites are integrated into the genomes of mammalian cells using Tol2 transposase and treated with base editors. Edited cells are enriched by antibiotic selection, and library cassettes are amplified for high-throughput sequencing. (**b**) Base editing windows. Values are C•G-to-G•C editing efficiencies normalized to a maximum of 100. The protospacer is at positions 1-20, with the SpCas9 PAM at positions 21-23. All data are in mES cells except for eA3A-nCas9, which is in HEK293T cells. (**c**) C•G-to-G•C editing purity in the comprehensive context library in mES cells. Box plots indicate interquartile range, and black dots indicate mean. Welch's *T*-test * $P<5.1\times10^{-9}$. (**d**) Heatmap of observed C•G-to-G•C purities by CGBE in target contexts from the comprehensive context library in mES cells. Black nucleotides indicate the cytosine for which purity is calculated. Target sites were sorted by outcome variance and manually selected. (**e**) Clustering of CGBEs based on measured C•G-to-G•C purity in core window cytosines across the comprehensive context library in mESCs. Values are Pearson correlation. (**f**) Purity of editing outcomes across core window nucleotides in the comprehensive context library, ranked by C•G-to-G•C purity, averaged across CGBEs in mESCs. Trend lines and shading show the rolling mean and standard deviation across 1% intervals. (**g**) Representative sequence motifs for editing efficiency and C•G-to-G•C purity from logistic regression models. The sign of each learned weight indicates a contribution above (positive sign) or below (negative sign) the mean activity. Logo opacity is proportional to the motif's Pearson's *R* on held-out sequence contexts. (**h**) Observed C•G-to-G•C purity across CGBEs in mESCs compared to CGBE-Hive predictions. Trend lines and shading show the rolling mean and standard deviation. (**i**) Sequence motifs for C•G-to-G•C editing yield.

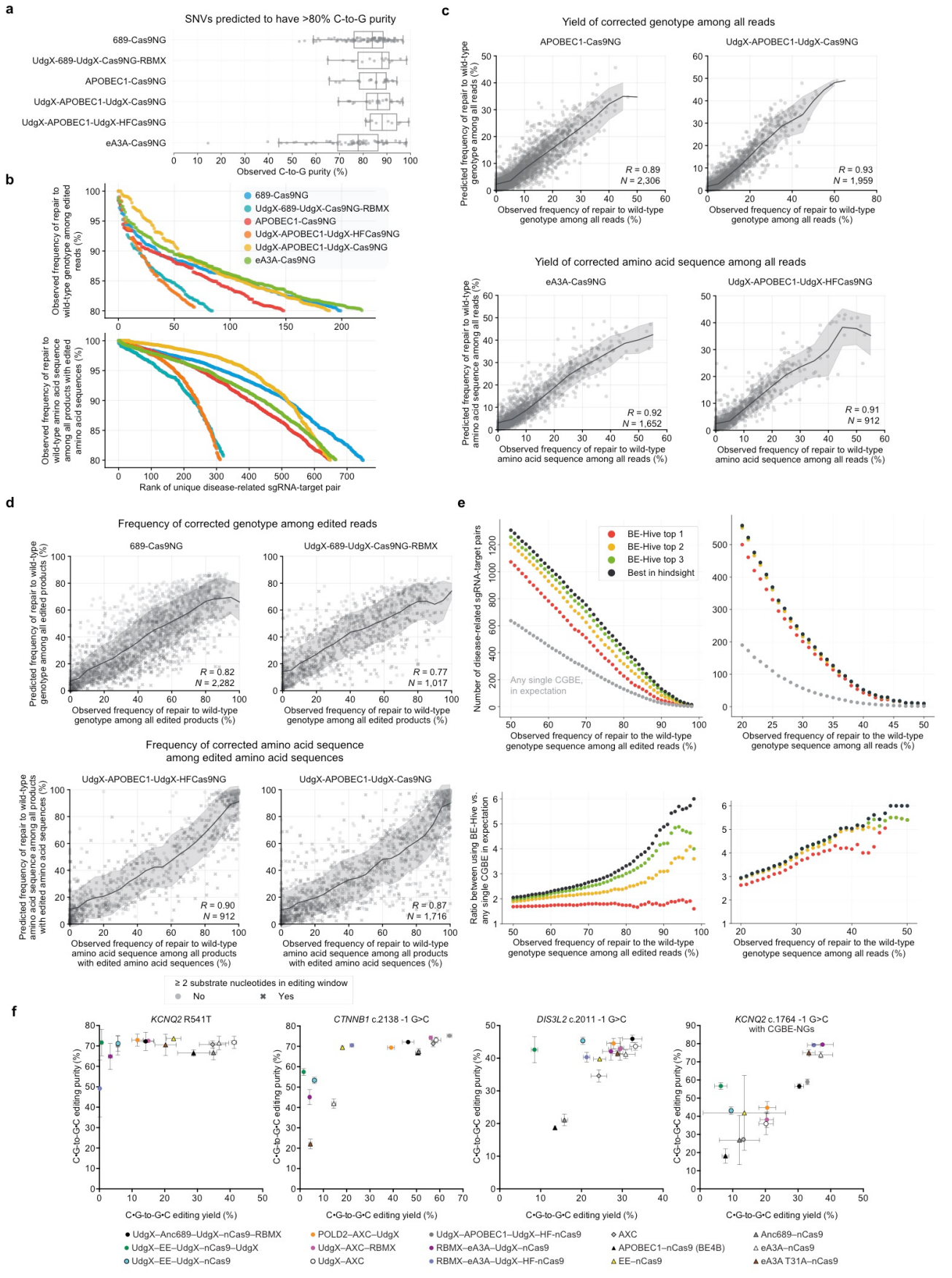

Figure 6. Target library characterization and machine learning modeling of CGBE variants. (**a**) Observed C-to-G purity by CGBE at SNVs predicted to have >80% C-to-G

purity. Box plot indicates median and interquartile range. (**b**) Observed number of disease-related sgRNA-target pairs corrected at varying genotype precision and amino acid precision thresholds by various strategies for selecting CGBEs. See Supplementary Table 3. (**c**) Comparison of predicted versus observed correction yield of disease-related transversion SNVs in mES cells. Trend lines and shading show the rolling mean and standard deviation. (**d**) Comparison of predicted versus observed correction precision of disease-related transversion SNVs in mES cells. Trend lines and shading show the rolling mean and standard deviation. (**e**) Observed number of sgRNA-target pairs containing disease-related transversion SNVs corrected at various thresholds for genotype and amino acid precision. (**f**) Installation of disease-associated SNPs using CGBEs.