ARTICLES
https://doi.org/10.1038/s41587-020-0677-y

nature biotechnology

Predicting the efficiency of prime editing guide RNAs in human cells

Hui Kwon Kim[1,2,3,4,10], Goosang Yu[1,2,10], Jinman Park[1,2], Seonwoo Min[5], Sungtae Lee[1], Sungroh Yoon[5,6,7] and Hyongbum Henry Kim[1,2,3,4,8,9] ✉

Prime editing enables the introduction of virtually any small-sized genetic change without requiring donor DNA or double-strand breaks. However, evaluation of prime editing efficiency requires time-consuming experiments, and the factors that affect efficiency have not been extensively investigated. In this study, we performed high-throughput evaluation of prime editor 2 (PE2) activities in human cells using 54,836 pairs of prime editing guide RNAs (pegRNAs) and their target sequences. The resulting data sets allowed us to identify factors affecting PE2 efficiency and to develop three computational models to predict pegRNA efficiency. For a given target sequence, the computational models predict efficiencies of pegRNAs with different lengths of primer binding sites and reverse transcriptase templates for edits of various types and positions. Testing the accuracy of the predictions using test data sets that were not used for training, we found Spearman's correlations between 0.47 and 0.81. Our computational models and information about factors affecting PE2 efficiency will facilitate practical application of prime editing.

The genetic changes that prime editing can introduce include insertions, deletions and all 12 possible point mutations, as well as combinations of these changes[1]. Prime editors are composed of a Cas9 nickase–reverse transcriptase (RT) fusion protein and a pegRNA. The pegRNA contains a guide sequence that recognizes the target sequence, a tracrRNA scaffold sequence, a primer binding site (PBS) required for the initiation of reverse transcription and an RT template that includes the desired genetic changes and sequences homologous to the targets[1]. Four types of prime editors have been developed: PE1, PE2, PE3 and PE3b[1]. Because PE1 is less efficient than the others, it is not expected to be widely used. Unlike PE2, PE3 and PE3b require a single guide RNA (sgRNA) in addition to a pegRNA. Furthermore, compared to PE2, PE3 results in more frequent unintended indels[1,2], and the use of PE3b is often restricted by the target sequence composition[1]. PE3 and PE3b usually, but not always, show higher efficiency than PE2 (refs. [1,2]). Thus, PE2, PE3 or PE3b will be chosen depending on the purpose and conditions of the experiments in question, as well as the target sequences of interest. Given that both PE3 and PE3b are composed of PE2 and an additional sgRNA[1], the efficiency of PE2 at a given target sequence will also affect the efficiencies of PE3 and PE3b at the target[1]. Thus, the evaluation and prediction of PE2 activity at a given target sequence should also assist in the prediction of PE3 and PE3b efficiencies.

Previously, high-throughput evaluation of the activities of Cas9, Cas12a and base editors at a large number of target sequences in human cells enabled the identification of factors associated with such activities and the development of computational models that predict Cas9 and Cas12a efficiencies at given target sequences, both of which have greatly assisted genome editing using these CRISPR nucleases[3–14]. Similarly, the identification of factors affecting prime editing efficiencies and the development of computational models predicting prime editing activities based on high-throughput evaluation would greatly facilitate prime editing, especially given that prime editing efficiencies have been tested at only a limited number of target sequences and that no computational models that predict prime editing efficiency are currently available[1,2]. In this study, we tested PE2 efficiencies using 54,836 pairs of pegRNA-encoding sequences and corresponding target sequences, which enabled the identification of factors associated with PE2 efficiency and the development of a computational model that predicts PE2 efficiency at given target sequences.

Results

High-throughput evaluation of PE2 efficiency. For high-throughput analysis of PE2 efficiencies, we adopted and modified the paired library approach that we and others previously used to evaluate the activities and outcomes of Cas12a and Cas9 at thousands of target sequences[6–11,15]. We prepared a lentiviral plasmid library, named library 1, from a pool of oligonucleotides that contained 48,000 pairs of pegRNA-encoding sequences and corresponding target sequences (= 2,000 target sequences × 24 combinations of PBS and RT templates per target sequence) (Supplementary Fig. 1a,b). The position numbering system used for the pegRNA and target sequence in this study is described in Supplementary Fig. 1c. To test the effect of changing the PBS and RT template lengths, the library included 24 different combinations of PBS and RT template lengths (six PBS lengths (7, 9, 11, 13, 15 and 17 nucleotides (nts) × four RT template lengths (10, 12, 15 and 20 nts) = 24 combinations) for 2,000 pairs of guide and target sequences to induce a transversion

[1]Department of Pharmacology, Yonsei University College of Medicine, Seoul, Republic of Korea. [2]Brain Korea 21 Plus Project for Medical Sciences, Yonsei University College of Medicine, Seoul, Republic of Korea. [3]Center for Nanomedicine, Institute for Basic Science (IBS), Seoul, Republic of Korea. [4]Graduate Program of Nano Biomedical Engineering (NanoBME), Advanced Science Institute, Yonsei University, Seoul, Republic of Korea. [5]Electrical and Computer Engineering, Seoul National University, Seoul, Republic of Korea. [6]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea. [7]Graduate School of Data Science, Seoul National University, Seoul, Republic of Korea. [8]Severance Biomedical Science Institute, Yonsei University College of Medicine, Seoul, Republic of Korea. [9]Graduate Program of NanoScience and Technology, Yonsei University, Seoul, Republic of Korea. [10]These authors contributed equally: Hui Kwon Kim, Goosang Yu. ✉e-mail: hkim1@yuhs.ac

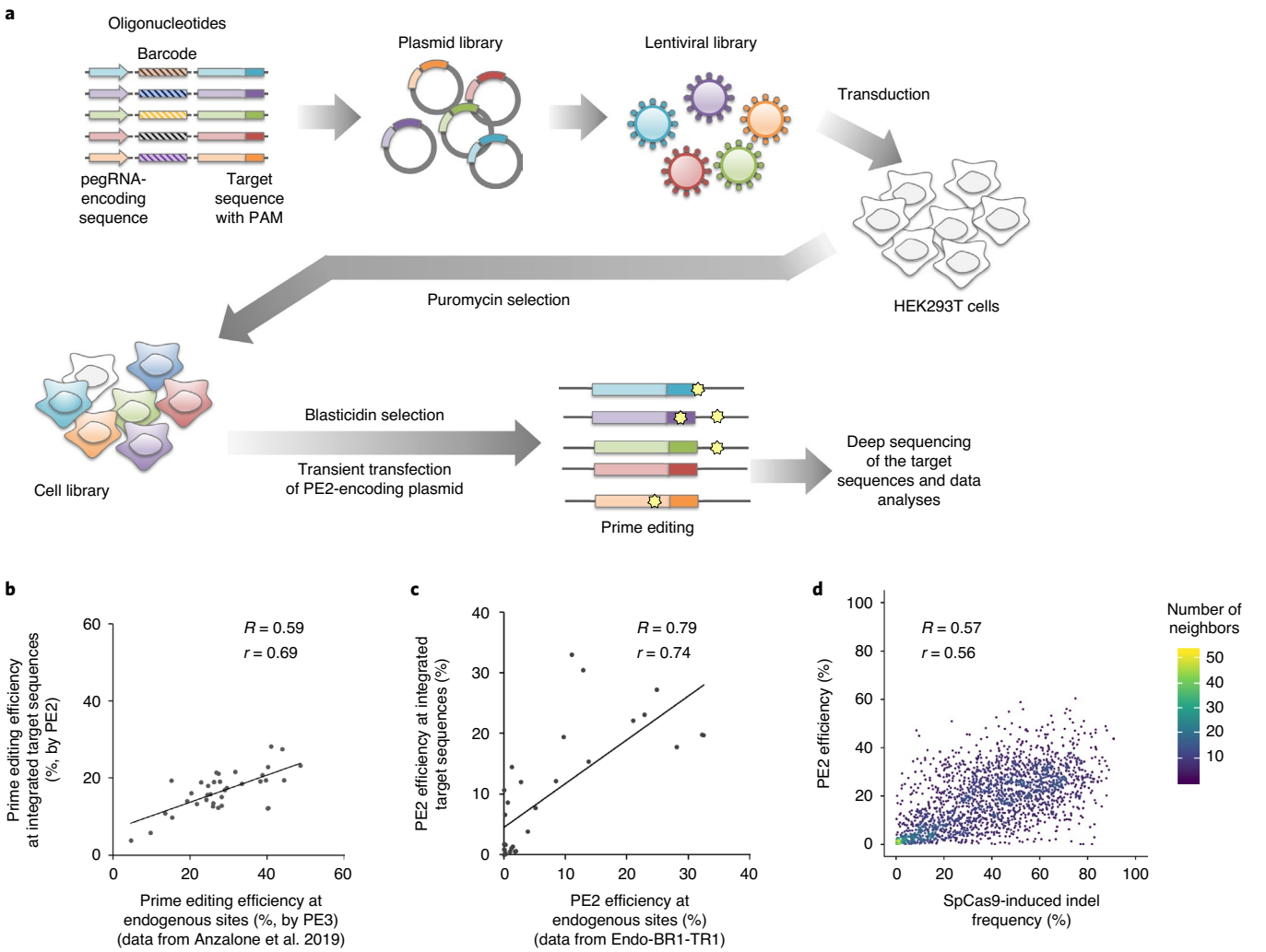

Fig. 1 | High-throughput evaluation of PE2 activity using libraries of pegRNA–target sequence pairs. a, Schematic representation of the experimental procedure. **b**, **c**, The correlations between PE2 efficiencies measured at endogenous sites and those at corresponding integrated target sequences. The data sets of PE2 efficiencies at endogenous sites; the data from the initial study (**b**, ref. [1], $n = 36$ pairs of pegRNAs and target sequences, HEK293T cells); or the data set, named PE2-Endo, that we newly generated in the current study (**c**, $n = 31$ pairs of pegRNAs and target sequences, HEK293T cells) were used. PE2 efficiency was determined as the proportion of sequence reads with specified edits among the total sequence reads. The Spearman's (R) and Pearson's (r) correlation coefficients are shown. **d**, The correlation between SpCas9-induced indel frequencies and PE2 efficiencies determined for identical target sequences. To minimize the effect of PBS and RT template lengths, the pegRNA that showed the highest efficiency among 24 pegRNAs with different PBS and RT template lengths was chosen per target sequence. The color of each dot was determined by the number of neighboring dots (that is, dots within a distance that is three times the radius of the dot). Similar graphs based on PBSs and RT templates with fixed lengths are shown in Supplementary Figs. 5 and 6. The number of pegRNA and target sequence pairs $n = 1{,}956$.

mutation from G to C at position +5 from the nicking site (position 22 within the wide target sequence), which comprises 48,000 (= 24 × 2,000) pairs of pegRNA and target sequences (Supplementary Fig. 1b). Furthermore, to evaluate the effect of factors other than PBS and RT template lengths on the PE2 efficiency, we generated one more paired library, named library 2, which contains 6,800 pairs of pegRNA-encoding sequences and corresponding target sequences. The factors tested using library 2 include editing positions, types of editing (for example, insertion, deletion or substitution) and locations of two-position editing (Supplementary Fig. 1b).

HEK293T cells were transduced with lentivirus generated from the plasmid library to construct a cell library at 0.3 multiplicity of infection (MOI), and untransduced cells were removed by puromycin selection (Fig. 1a). Each cell in this library expresses a pegRNA and includes the corresponding integrated target sequence. This cell library was then transfected with a plasmid-encoding PE2, and untransfected cells were removed by blasticidin selection. Four and a half days after the transfection with the PE2 plasmid, genomic DNA was isolated from the cells and subjected to polymerase chain reaction (PCR) to amplify the target sequences. The amplicons were subjected to deep sequencing to measure the mutation frequencies induced by PE2. Sanger sequencing showed that 8.5% (= 12/142) of the copies in the plasmid library contained at least one mutation in the guide sequence, scaffold, PBS, RT template or target sequence regions (Supplementary Table 1a), which would be attributable to errors introduced during oligonucleotide synthesis and PCR-based amplification. Furthermore, when high-throughput evaluations are performed using lentiviral vectors, two distant elements can be shuffled[16–19]. When we measured the rate of uncoupling between the pegRNA-encoding and barcode target sequences in the cell library, it was 4.2% (Supplementary Table 1b), compatible with previously observed rates[14,16–19]. Given that almost no prime editing would

occur with these mutant or uncoupled sequences, the observed PE2 efficiency would be 87% (= 100% − 8.5% − 4.2%) of the true PE2 efficiency (that is, if the true PE2 efficiency is 25%, the observed PE2 efficiency would be 25% × 87% = 22%). We observed strong correlations between replicates independently transfected by two different experimentalists (Supplementary Fig. 2) and combined the data from the two replicates for subsequent analyses. This result is in line with the strong correlation between replicates in our previous similar experiments with Cas9 (refs. [10,11]).

We next determined the correlations between editing efficiencies measured at the integrated sequences using the high-throughput approach and those at endogenous sites evaluated by individual tests. When we evaluated this correlation using a published data set of PE3 efficiencies from the initial study[1], there was also a strong correlation (Fig. 1b; Spearman's correlation coefficient (R) = 0.59, Pearson's correlation coefficient (r) = 0.69). Furthermore, we also generated six new data sets of PE2 efficiencies at 20–31 endogenous sites randomly selected from the 54,836 pegRNAs of libraries 1 and 2 (Supplementary Tables 2 and 3). In these experiments, plasmids encoding PE2 and pegRNAs were transiently transfected. We observed high correlations between the PE2 efficiencies at the endogenous sites and the corresponding integrated target sequences in a reproducible manner (Fig. 1c and Supplementary Fig. 3). The average PE2 efficiency obtained using libraries 1 and 2 was 9.9%, which is similar to 9.5%, the efficiency observed in the initial study[1] (Supplementary Fig. 4).

The correlation between SpCas9 and PE2 activities. For prime editing, Cas9 needs to bind the target sequence and make a nick[1]. Thus, it is expected that the activities of PE2-pegRNA and Cas9-sgRNA would be highly correlated. We previously evaluated the indel frequencies associated with Cas9-sgRNA activity at 2,000 target sequences[10]. When we evaluated the association of the activities of PE2-pegRNA and Cas9-sgRNA at the same target sequences, we observed modest correlations (Fig. 1d and Supplementary Figs. 5 and 6) as expected. The reason for the modest, rather than strong, correlations would be that prime editing requires additional processes that are barely or not relevant to the indel-generating activity of Cas9; these processes include reverse transcription of the pegRNA, 5′ flap cleavage and DNA repair. Factors associated with these processes are described below. Both PE2 (when the optimal combination of PBS and RT template lengths are chosen) and Cas9 nuclease efficiencies showed basically uniform distributions, with the exception that those with high activities were relatively rare (Supplementary Fig. 7). Additionally, PE, but not Cas9, efficiencies showed a very weak tendency toward a bimodal distribution, with modes when the editing was almost nonexistent (lower than 2% efficiency) and when it was around 25%. However, when all pegRNAs with the 24 combinations of PBS and RT template lengths are considered, the relative frequency of pegRNAs generally decreased as the PE2 efficiency increased.

The effect of PBS and RT template lengths on PE2 efficiency. For prime editing at a given target sequence, various combinations of PBS and RT template lengths can be chosen, and the lengths of these two regions in the pegRNA significantly affect prime editing efficiency[1]. Thus, we next evaluated the effect of different PBS and RT template lengths on the PE2 efficiencies at 2,000 target sequences. When we calculated the average editing efficiencies for each combination of PBS and RT template lengths, they showed a unimodal distribution; the highest average efficiency (13.4%) was observed when pegRNAs with an 11- to 13-nt PBS and a 10- to 12-nt RT template were used (Fig. 2a and Supplementary Fig. 8). If we define poorly working pegRNAs as those associated with PE2 efficiencies lower than 5%, depending on the PBS and RT template lengths, 28~81% (average, 43%) of pegRNAs fell into this category (Supplementary Fig. 9); in other words, 19~72% (average, 57%) of pegRNAs led to PE2 efficiencies higher than 5% (Fig. 2b and Supplementary Fig. 9b). We found that the optimal combination of the PBS and RT template lengths is variable depending on the target sequences, which is compatible with previous observations using human cells[1] and plants[2]. Thus, we next evaluated how frequently each combination of PBS and RT template lengths induced the highest editing efficiencies per given target sequence. These values also showed a unimodal distribution; the highest editing efficiencies were the most frequently observed when a 9- to 13-nt PBS and a 10- to 12-nt RT template were used (Fig. 2c). In the past, when we chose the combination of PBS and RT template lengths that led to the highest editing efficiency at a given target sequence, it was not known whether the effects of the PBS and RT template lengths were independent of each other. Analysis of our large data set allowed us to determine that these two parameters are independent (P = 0.25 by a chi-square test; Supplementary Fig. 10).

We also compared the average editing efficiencies of each combination of PBS and RT template lengths when the most efficient pegRNA at each target was selected. Surprisingly, the average editing efficiencies under these optimal combinations of PBS and RT template lengths were the highest when the lengths of the PBS and RT template were short (for example, a 7-nt PBS and a 10- to 12-nt RT template) and decreased as the PBS and RT template lengths increased (Fig. 2d). Taken together, these results lead us to recommend using a 13-nt PBS and a 12-nt RT template for initial testing of PE2 efficiencies and expanding to a 9- to 15-nt PBS and a 10- to 15-nt RT template for the second round of testing, which is basically compatible with the lengths of the initial study recommendation (for an approximately 13-nt PBS and a 10- to 16-nt RT template), which is based on individual evaluations at five target sequences[1]. When we compared the efficiencies of pegRNAs with a 13-nt PBS and identical target sequences and intended edits but with different RT template lengths, we observed relatively high correlations between them (Supplementary Fig. 11), suggesting that other factors affect PE2 efficiencies.

Factors associated with PE2 efficiency. To evaluate other factors associated with PE2 efficiency in a more systematic manner, we next performed Tree SHAP (SHapley Additive exPlanations merged into XGBoost algorithm)[20] using 1,766 features that include melting temperature, GC counts, GC contents, the minimum self-folding free energy of various regions in the pegRNAs, the lengths of PBS and the RT template, the DeepSpCas9 score (computationally predicted Cas9 nuclease activities at a given target sequence[10]) and direct sequence information, such as all position-dependent and position-independent mononucleotides and dinucleotides. When high feature values were linked with high and low prime editing efficiencies, then the features were classified as favored and disfavored features, respectively. The most important feature was the DeepSpCas9 score (favored) at the corresponding target sequence (Fig. 2e), which is in line with the correlation between SpCas9-induced indel frequencies and PE2 efficiencies as shown above. GC counts in PBS (favored) was the second most important feature. In line with this result, GC contents in PBS (favored) was also the 11th most important feature (Supplementary Fig. 12). GC content can be calculated by dividing the GC count (the number of G or C nucleotides) with the length of the relevant DNA strand. The importance of these features can be understood given that a high GC count in PBS would result in strong binding of the pegRNA to the nicked strand of the target DNA, which is required for reverse transcription. When we systematically evaluated the effects of GC contents and GC counts in PBS, the RT template and the combination of PBS and RT template on PE2 efficiency, we clearly observed higher PE2 efficiencies as the GC contents and GC counts of PBS increased (Supplementary Fig. 13). When the GC contents of the

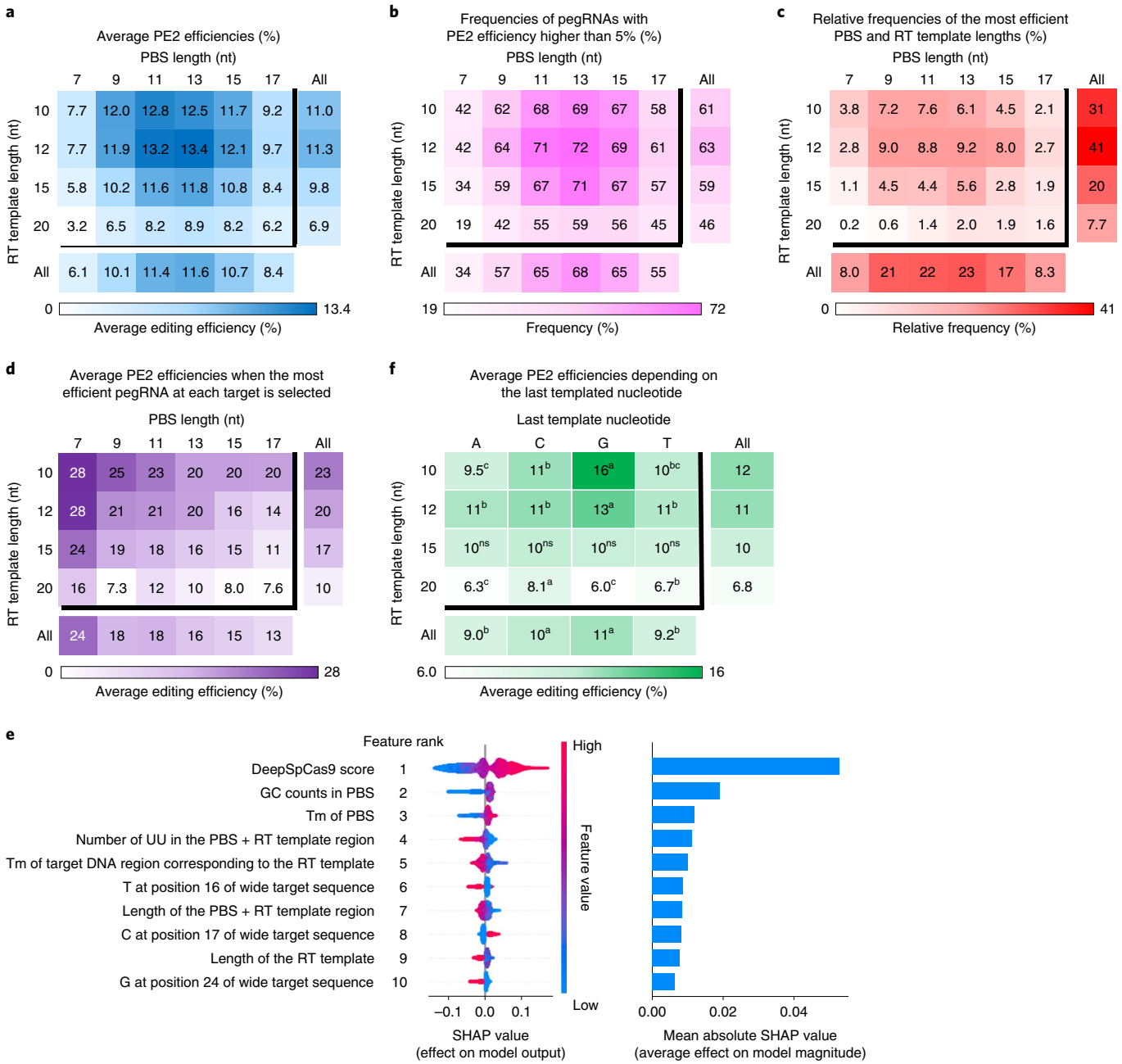

Fig. 2 | Factors affecting PE2 efficiency. a–d, The effect of PBS and RT template length on PE2 efficiency. The heat maps show the average editing efficiencies for given lengths of PBS and RT templates (**a**), frequencies of pegRNAs with PE2 efficiencies higher than 5% for given lengths of PBS and RT templates (**b**), the frequencies of PBS and RT template length combinations that induced the highest editing efficiencies per given target sequence (**c**) and the average editing efficiencies when the combination of PBS and RT template lengths that showed the highest editing efficiency at each target was selected (**d**). **e**, The ten most important features associated with PE2 efficiency determined by Tree SHAP (XGBoost classifier). On the summary violin plot (the left graph), each target sequence is represented by a dot; the position of the dot on the *x* axis shows its SHAP value. High and low SHAP values are linked with high and low prime editing efficiencies, respectively. The color of the dot indicates the value of the relevant feature for that particular target sequence; red and blue represent high and low values of the relevant feature as shown in the figure. Overlapping points are slightly separated in the *y*-axis direction so that the density is apparent. Examples of the summary plot interpretations are included as Supplementary Text 1. The 100 most important features are shown in Supplementary Fig. 12. Tm, melting temperature. The number of pegRNA and target sequence pairs (that is, the number of dots in the summary plot) $n = 38{,}692$. **f**, Average PE2 efficiencies of pegRNAs depending on the identity of the last templated nucleotide. In each category of RT template lengths, statistical analysis was conducted; subsets of experimental groups without statistically significant ($P < 0.05$, ANOVA followed by two-sided Tukey's post hoc test; exact *P* values are described in Supplementary Table 7) differences in PE2 efficiencies are represented with letters such as a, b and c in the order of the average PE2 efficiency. NS, not significant.

PBS were lower than 30%, PE2 efficiencies were poor for all tested PBS lengths, although longer lengths, such as 15 nts, resulted in relatively high editing efficiency. Conversely, when the GC contents of the PBS were higher than 60%, shortening the PBS to a length of 7–11 nts led to a relatively high PE2 efficiency. On the basis of these results, we recommend using a PBS that is 15 or 9 nts in length when

the GC contents are lower than 40% or higher than 60%, respectively (Supplementary Fig. 14). However, the GC contents and GC counts of the RT template only slightly affected PE2 efficiencies, and the PE2 efficiencies tended to be low when the GC-related parameters were extremely high or low. Compatible with these findings, neither the GC contents nor GC counts in the RT template were included in the 40 most important features.

The third and fifth most important features were, respectively, the melting temperature of PBS (favored) and that of the target DNA region that corresponds to the RT template (that is, between the strand containing the protospacer adjacent motif (PAM) and the opposite strand, here called the PAM-opposite strand; this feature was disfavored only when the melting temperature was higher than 35 °C). A high PBS melting temperature is likely to be associated with high GC counts in the PBS and would be linked with strong binding of the PBS region of the pegRNA to the target DNA, which would facilitate the reverse transcription reaction. When we examined the relationship between PE2 efficiency and the PBS melting temperature, we found that, as the PBS melting temperature increased, PE2 efficiency also increased (Supplementary Fig. 15a). If the melting temperature of the target DNA region that corresponds to the RT template is too high, the conversion of the 3′ flap into a 5′ flap, a process that is required for incorporation of the reverse transcribed DNA sequence into the genome[1], might be prevented. We analyzed the relationship between PE2 efficiencies and the melting temperature of this region and found that, when the melting temperature increased above 35 °C, the PE2 efficiency tended to decrease, although the difference was not statistically significant (Supplementary Fig. 15b). The fourth most important feature was the number of UUs in the RT + PBS region (disfavored). This feature would result from a large number of Ts in the pegRNA-encoding sequences, corresponding to a large number of Us in the pegRNAs, which could reduce the efficiency of transcription by RNA polymerase III[21,22], leading to a decrease in intracellular pegRNA concentrations.

The sixth and eighth most important features were the presence of a T at position 16 (disfavored) and a C at position 17 (favored) in the wide target sequence (position 1 is the 20th nucleotide from the NGG PAM). It has previously been shown that a T at position 16 is associated with decreased Cas9 nuclease activity[4,5,15]. Furthermore, a T at position 16 decreases GC counts in PBS, which is not favorable for reverse transcription, especially when the length of PBS is short. These two effects combined would result in a T at position 16 as the sixth most important feature. Similarly, previous work showed that, when a C is at position 17, the Cas9 nuclease activity slightly increased[4,5,15]. More importantly, a C at position 17 increases GC counts in PBS, facilitating reverse transcription. The combination of these two effects would render a C at position 17, a favored feature. The seventh, ninth and twelveth most important features were the RT and PBS length (generally disfavored), the RT template length (disfavored only when it is long) and the PBS length (generally disfavored), respectively, all of which were more deeply evaluated above. The tenth most important feature was a G at position 24 (disfavored) in the wide target sequence. The intended edit (+5 G to C) would replace a G at position 22, resulting in PAM editing, which would prevent re-binding of Cas9 to the target sequence. However, if the 24th nucleotide is a G, then a GG PAM sequence could be generated to span positions 23 and 24, a core PAM shift that would allow re-binding of Cas9 (refs. [11,23–25]), leading to the nicking of the reverse transcribed DNA strand before the repair of the complementary strand[1]. In addition, we evaluated factors affecting PE2 activity when the DeepSpCas9 score was excluded (Supplementary Text 2 and Supplementary Fig. 16).

For efficient prime editing, the initial study recommended that the last templated nucleotide should not be a G to avoid using RT templates that locate a C close to the 3′ hairpin of the sgRNA scaffold[1]. To examine the validity of this recommendation, which was based on observations at three target sequences (72 pegRNAs), we categorized the PE2 efficiencies at 887 target sequences (21,288 pegRNAs) depending on the last templated nucleotide. Contrary to the initial finding, PE2 efficiencies were overall the highest when the last templated nucleotide was a G. Interestingly, the preferred nucleotide at the last templated position varied depending on the RT template length. When the RT template was relatively short, such as 10 or 12 nts, a G was strongly preferred, whereas an A or a T was not preferred (Fig. 2f). However, when the RT template was relatively long, such as 20 nts, then a C was favored, whereas A and G were not, which is partially in line with the initial study recommendation. These preferences for the identity of the last templated nucleotide were similarly observed across six tested PBS lengths (Supplementary Fig. 17).

Effects of editing type and position on PE2 efficiency. So far, we have described our evaluation of PE2 efficiencies for G-to-C conversions at a fixed position (+5 from the nicking site) at 2,000 target sequences using library 1. We next evaluated PE2 efficiencies for more diverse kinds of genome editing using the 6,800 pegRNA and target sequence pairs (= 200 target sequences × 1 PBS per target sequence × 34 RT templates per target sequence) in library 2 to determine the effect of the type of genome editing (that is, the generation of indels versus substitutions), the position edited and the number of inserted or deleted nucleotides on the efficiency. We first evaluated the efficiencies of generating 1-bp insertions, 1-bp deletions and 1-bp substitutions and found that the general efficiencies could be ranked as insertion ≥ deletion ≥ substitution and that the difference between the insertion and substitution efficiencies was statistically significant (Fig. 3a). Then, we assessed the effect of the type and number of inserted nucleotides on prime editing-induced insertions and found that the identity of the inserted nucleotide did not affect the 1-bp insertion efficiency. When we increased the number of inserted nucleotides from 1 bp to 2, 5 and 10 bp, the insertion efficiencies were similar for 1- and 2-bp insertions, decreased for 5-bp insertions and drastically decreased for 10-bp insertions (Fig. 3b). In parallel, we also evaluated the PE2 efficiency for 1-, 2-, 5- and 10-bp deletions; we found that the PE2 efficiencies were similar for 1-, 2- and 5-bp deletions and drastically decreased for 10-bp deletions (Fig. 3c).

We next examined the effect of the substituted nucleotide identity on the PE2 efficiency. We tested all 12 possible types of 1-bp substitutions at position +1 from the nicking site, which is between positions 17 and 18 in the wide target sequence, and found that the PE2 efficiencies differed slightly depending on the type of substitution; C-to-T and T-to-G conversions showed the highest and the lowest PE2 efficiencies, respectively (Fig. 3d). To gain mechanistic insights into these effects, we considered the temporary base pairing between the nucleotide in the complementary DNA (cDNA) generated from the RT template and the corresponding nucleotide in the PAM-opposite strand. Interestingly, the PE2 efficiencies could be ranked as follows: T (cDNA) – G (corresponding nucleotide in the PAM-opposite strand) and G – T pairings ≥ C – T and T – C pairings ≥ C – A and A – C pairings ≥ A – G and G – A pairings. The differences between the T – G and G – T pairing groups and the A – G and G – A pairing groups were statistically significant, implying the possibility that temporary base pairing between the cDNA and PAM-opposite strands might affect PE2 efficiency. When the temporary base pairings were formed between the same nucleotides, such as T (cDNA) – T (corresponding nucleotide in the PAM-opposite strand), G – G, C – C and A – A, which correspond to A-to-T, C-to-G, G-to-C and T-to-A conversions, respectively, the PE2 efficiencies were all similar (Fig. 3d). In addition, when we analyzed the PE2 efficiencies for these four conversions mediated by the temporary base pairing between the same nucleotides at different

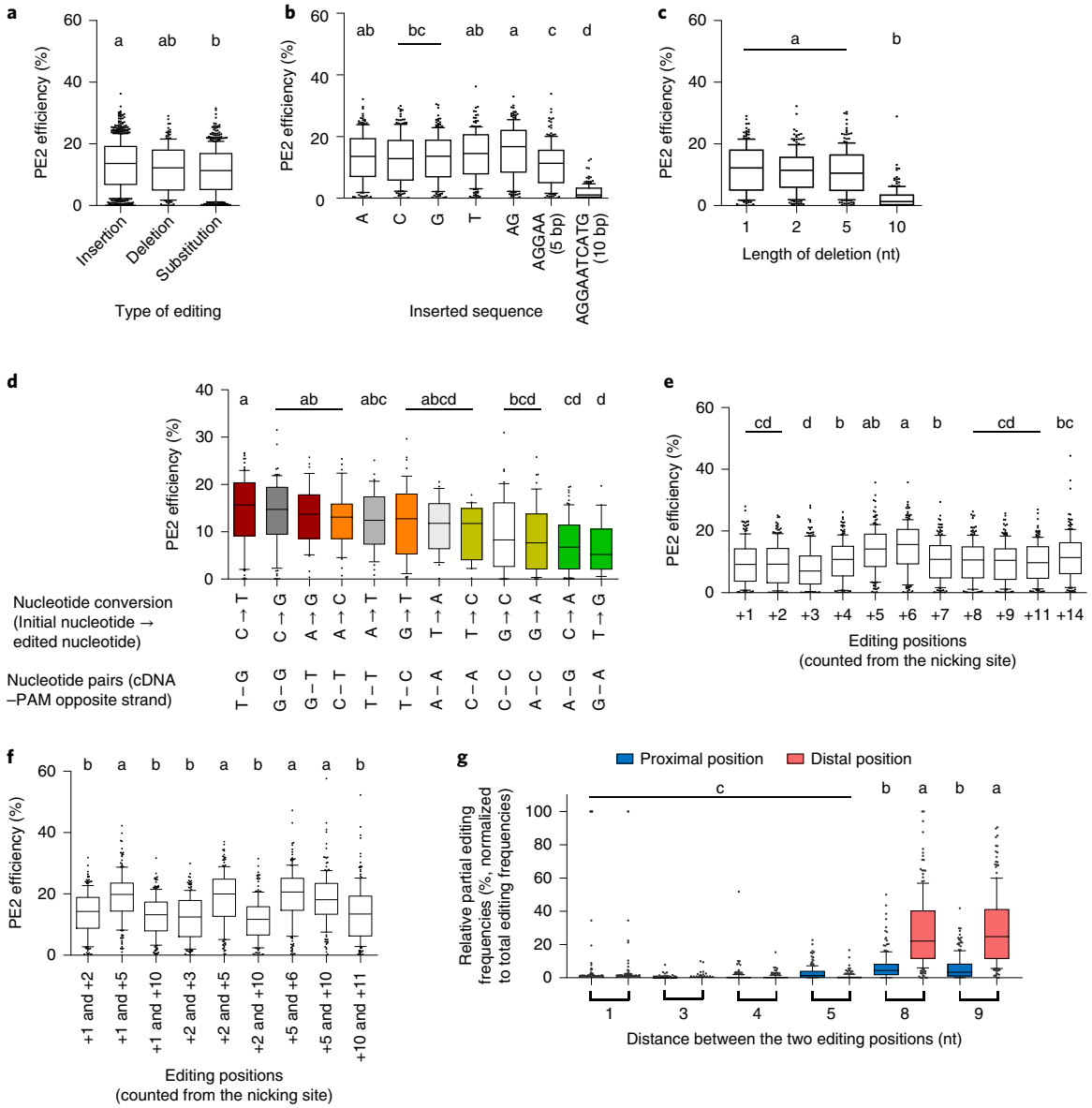

Fig. 3 | Effects of editing type and position on PE2 efficiency. Prime editing efficiency for insertions, deletions and substitutions in 170 target sequences. Editing occurred at position +1 from the nicking site of the Cas9 nickase, and the length of PBS was 13 nt. The combined length of the left and right homology arms of the RT template was 14 nt (that is, in the cases of 1-bp substitutions, the length of the RT templates was 15 nt). Subsets of experimental groups without statistically significant ($P < 0.05$, ANOVA followed by two-sided Tukey's post hoc test; exact P values are described in Supplementary Table 7) differences in PE2 efficiencies are represented with letters such as a–d in the order of the average PE2 efficiency (for example, if four hypothetical groups 1, 2, 3 and 4 are designated as b, c, ab and a, then the statistically significant differences in the prime editor efficiency are 4 > 1 > 2, 4 = 3 and 3 = 1). In the boxes, the top, middle and bottom lines represent the 25th, 50th, and 75th percentiles, respectively; whiskers indicate the 10th and 90th percentiles; and outliers are shown as individual dots. **a**, PE2 efficiencies for 1-bp insertions, deletions and substitutions. The number of pegRNA and target sequence pairs $n = 739$ for insertions, 178 for deletions and 566 for substitutions. **b**, Effect of the inserted nucleotide type and number on PE2 efficiency. The number of pegRNA and target sequence pairs $n = 183$, 183, 188, 185, 184, 179 and 163 for the insertion of A, C, G, T, AG, AGGAA (5 bp) and AGGGAATCATG (10 bp), respectively. **c**, Effect of the length of deletion on PE2 efficiency. The number of pegRNA and target sequence pairs $n = 178$, 189, 185 and 169 for 1-, 2-, 5- and 10-bp deletions, respectively. **d**, Effect of the type of substitution on PE2 efficiency. The number of pegRNA and target sequence pairs $n = 88$, 87, 36, 35, 34, 44, 21, 20, 45, 45, 90 and 21 for C-to-T, C-to-G, A-to-G, A-to-C, A-to-T, G-to-T, T-to-A, T-to-C, G-to-C, G-to-A, C-to-A and T-to-G conversions, respectively. **e**, Effect of the editing position on the PE2 efficiency for 1-bp transversion substitutions. Editing positions shown on the x axis were counted from the nicking site. The number of pegRNA and target sequence pairs $n = 179$, 186, 184, 180, 173, 184, 182, 178, 177, 178 and 173 for position +1, +2, +3, +4, +5, +6, +7, +8, +9, +11 and +14, respectively. **f**, Effect of the editing position on the prime editing efficiency for 1-bp transversion substitutions at two positions. The number of pegRNA and target sequence pairs $n = 190$, 181, 186, 190, 177, 180, 183, 170, and 169 for position +1 and +2, +1 and +5, +1 and +10, +2 and +3, +2 and +5, +2 and +10, +5 and +6, +5 and +10 and +10 and +11, respectively. **g**, Relative frequency of partial editing depending on the distance between two editing positions described in **f**. Absolute frequencies of partial editing are shown in Supplementary Fig. 19. The total editing frequency includes the sum of the efficiencies of the intended edit at only the proximal position (the position closer to the nicking site than the distal position), only the distal position and both positions.

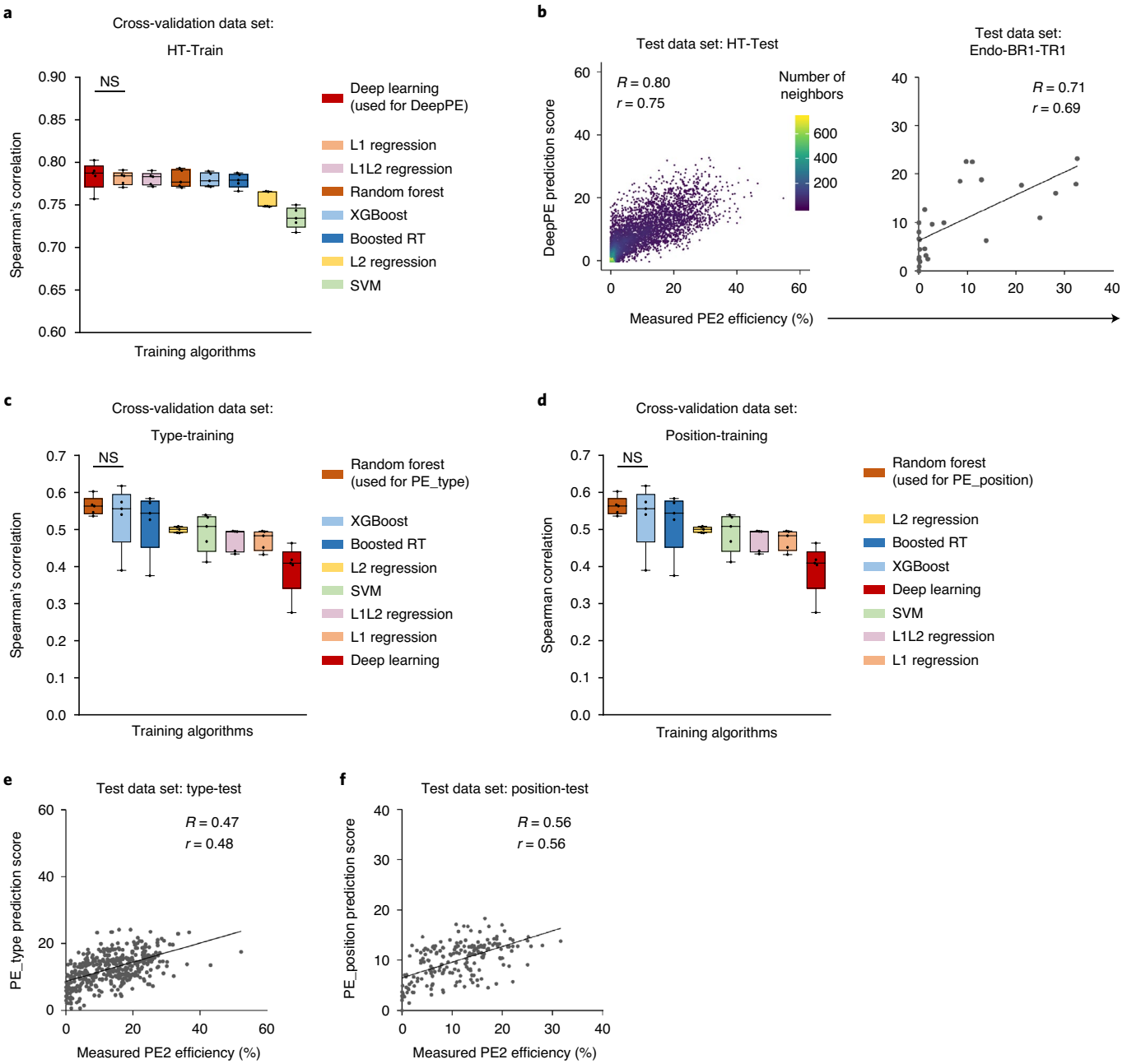

Fig. 4 | Development of computational models for predicting PE2 efficiencies. a–d, Cross-validation of prediction models depending on the machine learning framework used. Each dot represents the Spearman's correlation coefficient between the measured prime editor efficiency and the predicted activity from five-fold cross-validation (total, $n = 5$ correlation coefficients). For clarity, results from statistical testing are shown only for the best model versus the next best model; NS, not significant; two-sided Steiger's test. In the boxes, the top, middle and bottom lines represent the 25th, 50th and 75th percentiles, respectively; whiskers indicate the 10th and 90th percentiles; and outliers are shown as individual dots. **b**, **e**, **f**, Evaluation of DeepPE (**b**), PE_type (**e**) and PE_position (**f**). The Spearman's (R) and Pearson's (r) correlation coefficients are shown. **b**, Evaluation of DeepPE using data set HT-Test (the number of pegRNA and target sequence pairs $n = 4,457$) and Endo-BR1-TR1 ($n = 26$). The color of each dot was determined by the number of neighboring dots (that is, dots within a distance that is three times the radius of the dot). **e**, **f**, Evaluations of PE_type (**e**) and PE_position (**f**) using Type-test ($n = 403$) and Position-test ($n = 200$), respectively.

positions, such as +9, +11 and +14 from the nicking site, we also observed that the efficiencies were similar for the four tested conversions at all three tested positions (Supplementary Fig. 18), which is in line with the analyses at position +1 from the nicking site.

We also investigated the effect of the editing position on 1-bp substitution efficiencies. We found that editing efficiencies were generally similar at all tested positions, which ranged from +1 to +14 from the nicking site, except at positions +3, +5 and +6 (Fig. 3e). The lowest editing efficiency was observed at position +3, although the underlying mechanism for this effect is not clear. The highest editing efficiencies were observed at positions +5 and +6, the position of the GG PAM; as stated above, if the PAM is not edited, Cas9 can re-bind to the target sequence and nick the reverse-transcribed DNA strand before the repair of the complementary strand[1],

resulting in a decrease in the PE2 efficiency. This effect of PAM editing on PE2 efficiency can be also observed when 2-bp substitution efficiencies are evaluated. We generated 2-bp substitutions at various positions and found that the editing efficiencies were consistently higher when one or both nucleotides in the PAM (positions +5 and +6) were edited (for example, positions +1 and +5, positions +2 and +5, positions +5 and +6 and positions +5 and +10) than when the PAM was left intact (positions +1 and +2, positions +1 and +10, positions +2 and +3, positions +2 and +10 or positions +10 and +11 were edited) (Fig. 3f). Given that the editing position affects PE2 efficiency, the use of SpCas9 variants that recognize different PAMs[11,26–32] instead of wild-type SpCas9 could improve PE2 efficiencies at some target sequences. Interestingly, up to a median of 20% of sequences in which at least one of two intended edits were introduced have only one edit (Fig. 3g and Supplementary Fig. 19). Such partial editing rates were higher at the positions distal to the nicking site than at the proximal positions and showed a tendency to increase as the distance between the two positions increased.

Computational models that predict PE2 efficiencies. We next attempted to develop a computational model that predicts PE2 efficiencies at a given target sequence paired with 24 different pegRNAs with variable PBS and RT template lengths. We previously used deep learning to develop accurate computational models that predict the efficiencies of Cas12a[8] and Cas9 (refs. [10,11]) at given target sequences. The PE2 efficiencies obtained using library 1, with 48,000 pairs of pegRNA and target sequences, were split into two data sets, named HT-training ($n=38,692$) and HT-test ($n=4,457$), by random sampling (the same target sequences were never shared between the two data sets) (Supplementary Tables 2 and 4). Using HT-training as the training data, we generated computational models that predict PE2 efficiencies at a given target sequence paired with 24 pegRNAs with different combinations of PBS and RT template lengths when prime editing is designed for G-to-C conversion at position +5. Cross-validation showed that the deep learning framework has the highest performance, although the difference with L1 regression, the second best framework, was not statistically significant (Fig. 4a). When evaluated using HT-test as the test data set, we found that DeepPE, a deep learning-based model, significantly, albeit slightly, outperformed other models based on conventional machine learning (Fig. 4b and Supplementary Fig. 20), which is in line with the results of deep learning models of Cas12a[8] and Cas9 (ref. [10]). When tested using six replicates of PE2 efficiencies at endogenous sites as the testing data sets, the Spearman's and Pearson's correlation coefficients (R and r) were $R=0.67$~0.77 (average, 0.73) and $r=0.63$~0.74 (average 0.69), respectively (Fig. 4b and Supplementary Fig. 21), suggesting good performance of DeepPE in predicting PE2 efficiencies at endogenous sites. Evaluation of DeepPE in two additional cell types, HCT116 (a colorectal carcinoma cell line) and MDA-MB-231 (a human breast adenocarcinoma cell line), at target sequences that were never used for DeepPE training also revealed its good performance across biologicial and technical replicates (HCT116, $R=0.70$~0.77 (average, 0.74), $r=0.57$~0.61 (average, 0.59); MDA-MB-231, $R=0.76$~0.81 (average, 0.79), $r=0.62$~0.65 (average, 0.64)) (Supplementary Fig. 22 and Supplementary Table 5). We determined the usefulness of DeepPE for choosing the most efficient combination of PBS and RT template lengths (out of 24 possible combinations) for a given target sequence. When DeepPE was used, the average absolute and relative PE2 efficiencies were 1.2% and 8.3%, respectively, which was significantly higher than those obtained using recommendations based on the initial study (that is, 13-nt PBS and 12-nt RT template, avoiding a G at the last templated nucleotide) (Supplementary Fig. 23). Furthermore, for an intended edit, there could be multiple target sequences; in this case, DeepPE would be useful for choosing the target sequence that could be edited with the highest efficiency.

We also used the data set obtained using library 2 to develop two more computational models that predict PE2 efficiencies for various other editing types and positions than were evaluated above. The data obtained using library 2 was split into Type-training, Type-test, Position-training and Position-test such that target sequences were never shared between the training and test data sets (Methods and Supplementary Tables 2 and 4). Cross-validation using Type-training and Position-training revealed that random forest had the highest performances, although the differences with the second best frameworks were not statistically significant (Fig. 4c,d). In both cases, deep learning showed limited performance, possibly owing to the relatively small number of target sequences and pegRNAs. When we evaluated using Type-test and Position-test, we observed that PE_type and PE_position, random forest-based models, showed useful performance (PE_type, $R=0.47$, $r=0.48$; PE_position, $R=0.56$, $r=0.56$) (Fig. 4e,f). Evaluation of prime editing efficiencies at a larger number of target sequences using pegRNAs with every possible PBS and RT template length and more diverse intended edits could yield more informative models.

We provide a web tool that provides the results of DeepPE, PE_type and PE_position for a given target sequence at http://deepcrispr.info/DeepPE. When a sequence containing a target sequence is entered, this web tool identifies candidate target sequences and provides the expected PE2 efficiencies for a total of 57 pegRNAs (24 pegRNAs from DeepPE, 23 pegRNAs from PE_type and ten pegRNAs from PE_position) per target sequence. The 23 pegRNAs from PE_type are designed to generate four types of deletions (1-, 2-, 5- and 10-bp deletions at position +1), seven types of insertions (insertions of A, C, G, T (1-bp), AG (2-bp), AGGAA (5-bp) and AGGAATCATG (10-bp) at position +1), all three possible 1-bp substitutions at position +1 and nine types of substitutions (A to T, C to G, G to C and T to A) at two positions (positions +1 and +2, +1 and +5, +1 and +10, +2 and +3, +2 and +5, +2 and +10, +5 and +6, +5 and +10 or +10 and +11). The ten pegRNAs from PE_position are designed to generate 1-bp substitutions (A to T, C to G, G to C and T to A) at positions +1, +2, +3, +4, +6, +7, +8, +9, +11 or +14.

Discussion

Prime editing is revolutionary in that it enables the introduction of any small genetic mutation in a fairly efficient manner without the use of donor DNAs. Together with the use of computational models (DeepPE, PE_type and PE_position), based on the results obtained in this study we recommend: 1) using a 13-nt PBS and a 12-nt RT template 2) with a high GC count in the PBS region if possible; 3) using a G at the last templated nucleotide when the RT template length is ≤12 nt; and 4) including PAM editing (detailed recommendations are provided as Supplementary Text 3). We expect that, together with the computational models, the information about factors that affect PE2 efficiency identified in the current study based on high-throughput analyses will facilitate prime editing.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0677-y.

Received: 14 April 2020; Accepted: 17 August 2020;
Published online: 21 September 2020

References

1. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

2. Lin, Q. et al. Prime genome editing in rice and wheat. *Nat. Biotechnol.* **38**, 582–585 (2020).
3. Chari, R., Mali, P., Moosburner, M. & Church, G. M. Unraveling CRISPR–Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods* **12**, 823–826 (2015).
4. Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR–Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
5. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR–Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
6. Kim, H. K. et al. In vivo high-throughput profiling of CRISPR–Cpf1 activity. *Nat. Methods* **14**, 153–159 (2017).
7. Allen, F. et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* **37**, 64–72 (2018).
8. Kim, H. K. et al. Deep learning improves prediction of CRISPR–Cpf1 guide RNA activity. *Nat. Biotechnol.* **36**, 239–241 (2018).
9. Shen, M. W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
10. Kim, H. K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci. Adv.* **5**, eaax9249 (2019).
11. Kim, H. K. et al. High-throughput analysis of the activities of xCas9, SpCas9-NG and SpCas9 at matched and mismatched target sequences in human cells. *Nat. Biomed. Eng.* **4**, 111–124 (2020).
12. Song, M. et al. Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0453-z (2020).
13. Arbab, M. et al. Determinants of base editing outcomes from target library analysis and machine learning. *Cell* **182**, 463–480 (2020).
14. Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-020-0537-9 (2020).
15. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).
16. Schlub, T. E., Smyth, R. P., Grimm, A. J., Mak, J. & Davenport, M. P. Accurately measuring recombination between closely related HIV-1 genomes. *PLoS Comput. Biol.* **6**, e1000766 (2010).
17. Sack, L. M., Davoli, T., Xu, Q., Li, M. Z. & Elledge, S. J. Sources of error in mammalian genetic screens. *G3 (Bethesda)* **6**, 2781–2790 (2016).
18. Feldman, D., Singh, A., Garrity, A. J. & Blainey, P. C. Lentiviral co-packaging mitigates the effects of intermolecular recombination and multiple integrations in pooled genetic screens. Preprint at https://doi.org/10.1101/262121 (2018).
19. Hill, A. J. et al. On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).
20. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
21. Dang, Y. et al. Optimizing sgRNA structure to improve CRISPR–Cas9 knockout efficiency. *Genome Biol.* **16**, 280 (2015).
22. Nielsen, S., Yuzenkova, Y. & Zenkin, N. Mechanism of eukaryotic RNA polymerase III transcription termination. *Science* **340**, 1577–1580 (2013).
23. Lin, Y. et al. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473–7485 (2014).
24. Chen, H., Choi, J. & Bailey, S. Cut site selection by the two nuclease domains of the Cas9 RNA-guided endonuclease. *J. Biol. Chem.* **289**, 13284–13294 (2014).
25. Zeng, Y. et al. The initiation, propagation and dynamics of CRISPR–SpyCas9 R-loop complex. *Nucleic Acids Res.* **46**, 350–361 (2018).
26. Kleinstiver, B. P. et al. Engineered CRISPR–Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
27. Kleinstiver, B. P. et al. High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
28. Anders, C., Bargsten, K. & Jinek, M. Structural plasticity of PAM recognition by engineered variants of the RNA-guided endonuclease Cas9. *Mol. Cell.* **61**, 895–902 (2016).
29. Hu, J. H. et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).
30. Nishimasu, H. et al. Engineered CRISPR–Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).
31. Miller, S. M. et al. Continuous evolution of SpCas9 variants compatible with non-G PAMs. *Nat. Biotechnol.* **38**, 471–481 (2020).
32. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR–Cas9 variants. *Science* **368**, 290–296 (2020).

Methods

Construction of PE2-expressing vector. The LentiCas9-Blast plasmid (Addgene no. 52962) was digested with AgeI and BamHI restriction enzymes (NEB) at 37 °C for 4 h and treated with 1 µl of Quick-CIP (NEB) at 37 °C for 10 min. Next, the linearized plasmid was gel purified using a MEGAquick-spin Total Fragment DNA Purification Kit (iNtRON Biotechnology). The PE2-encoding sequence from pCMV-PE2 (Addgene no. 132775) was amplified by PCR using Sol 2× pfu PCR Smart Mix (SolGent). The amplicons were assembled with the linearized LentiCas9-Blast plasmid using an NEBuilder HiFi DNA Assembly Kit (NEB). The assembled plasmid is referred to as pLenti-PE2-BSD.

Oligonucleotide library design. An oligonucleotide pool containing 54,836 pairs of pegRNAs and target sequences was synthesized by Twist Bioscience. Each oligonucleotide contained the following elements: a 19-nt guide sequence, BsmBI restriction site #1, a 15-nt barcode stuffer sequence, BsmBI restriction site #2, the RT template sequence, the PBS sequence, a poly T sequence, an 18-nt barcode sequence (identification barcode) and a corresponding 43~47-nt wide target sequence that included a PAM and an RT template binding region. The barcode stuffer was later removed by cleavage with BsmBI, whereas the identification barcode (located upstream of the target sequence) allowed individual pegRNA and target sequence pairs to be identified after deep sequencing[10,11]. Oligonucleotides that included unintended BsmBI restriction sites in their sequences were excluded.

To test the effect of PBS and RT template length on PE2 efficiency, we prepared pegRNAs with 24 combinations of PBS and RT template lengths (six PBS lengths (7, 9, 11, 13, 15, 17 nts × four RT template lengths (10, 12, 15, 20 nts) = 24 different possibilities) for 2,000 pairs of guide and target sequences, resulting in a total of 48,000 (= 24 × 2,000) pairs of pegRNA and target sequences (library 1). The pegRNAs were designed to generate a G-to-C transversion mutation at position +5 from the nicking site. The 2,000 target sequences were randomly selected from human protein-coding genes, at which SpCas9-induced indel frequencies were previously measured[10], to allow the correlation between SpCas9 and PE2 efficiencies at identical target sequences to be determined.

We also prepared another library, named library 2, to evaluate the effects of editing position, type and length on PE2 efficiencies. We randomly selected 200 target sequences from the 2,000 target sequences used for library 1 and designed 34 different templates for each target sequence as follows.

i. The effect of editing position (11 RT templates): The RT templates were designed to introduce transversion mutations at positions +1, +2, …, +8, +9, +11 and +14 from the nicking site. The lengths of PBS and the RT template were fixed at 13 and 20 nts, respectively.
ii. The effect of editing type and length (14 RT templates): The RT templates were designed to introduce insertions (inserted sequences = A, G, C, T, AG, AGGAA and AGGAATCATG), deletions (1-, 2-, 5- and 10-nt) and single base substitutions (all possible 1-nt substitutions) at position +1 from the nicking site. The lengths of PBS and the right homology arm of the RT template were fixed at 13 and 14 nts, respectively.
iii. The effect of PAM editing (nine RT templates): The RT templates were designed to introduce 2-bp transversion mutations at positions +1 and +2, +1 and +5, +1 and +10, +2 and +3, +2 and +5, +2 and +10, +5 and +6, +5 and +10 and +10 and +11. The lengths of PBS and the RT template were fixed at 13 and 16 nts, respectively.

Furthermore, we included 36 pairs of pegRNAs and target sequences used in the initial prime editing study[1] with five unique barcodes per target sequence. This set was used to determine the correlation between prime editing efficiencies at integrated sequences and endogenous sites. All together, a total of 54,836 pairs of pegRNAs and target sequences—consisting of 48,000 (2,000 × 24, for library 1) + 6,800 (200 × 34, for library 2) + 36 (from the initial prime editing study)—were used to create libraries 1 and 2.

Plasmid library preparation. The plasmid library containing pairs of pegRNA-encoding and corresponding target sequences was prepared using a two-step cloning process: (Step I) Gibson assembly and (Step II) restriction enzyme-induced cutting and ligation. Uncoupling between paired guide RNA and target sequences during oligonucleotide amplification via PCR is effectively prevented by this two-step process[33]. The multi-step procedure was adapted and modified from a previously reported method[34].

Step I: Construction of the initial plasmid library containing pairs of pegRNA-encoding and target sequences. The oligonucleotide pool was amplified via PCR for 15 cycles using Phusion Polymerase (NEB), after which the amplicons were gel purified. The Lenti_gRNA-Puro vector (Addgene no. 84752) was digested with BsmBI enzyme (NEB) at 55 °C for 6 h. The linearized vector was then treated with 1 µl of Quick CIP at 37 °C for 10 min, followed by gel purification. Gibson assembly was used to assemble the amplified pool of oligonucleotides with the linearized Lenti_gRNA-Puro vector. After column purification, the assembled products were transformed into electrocompetent cells (Lucigen) using a MicroPulser (Bio-Rad). SOC media (2 ml) was then added to the transformation mixture, which was incubated at 37 °C for 1 h. The cells were then spread on Luria–Bertani agar plates containing 50 µg ml^{-1} of carbenicillin and incubated. Small fractions of the culture (0.1, 0.01 and 0.001 µl) were separately spread to allow determination of the library coverage. Plasmids were extracted from the total harvested colonies. The calculated coverage of this initial plasmid library was 113× the number of oligonucleotides.

Step II: sgRNA scaffold insertion. The initial plasmid library produced in Step I was digested with BsmBI for 8 h, followed by treatment with 1 µl of Quick CIP at 37 °C for 10 min. The digested product was gel purified after size selection on a 0.6% agarose gel. The sgRNA scaffold sequence in the pRG2 plasmid (Addgene no. 104174) was PCR amplified for 30 cycles using Phusion Polymerase and a primer pair with a BsmBI restriction site in each member of the pair. The resulting amplicon was digested with BsmBI for at least 12 h and gel purified on a 2% agarose gel. The purified insert (10 ng) was ligated with the digested initial plasmid library vector (200 ng) using T4 ligase (Enzynomics) at 16 °C for 16 h. The ligation products were column purified and electroporated into Endura electrocompetent cells (Lucigen). Colonies were harvested, and the final plasmid library was extracted. The calculated coverage of the final plasmid library was 785×.

Production of lentivirus. HEK293T cells (4.0×10^6 or 8.0×10^6) were seeded on 100-mm or 150-mm cell culture dishes containing DMEM. Fifteen hours later, the DMEM was exchanged with fresh medium containing 25 µM chloroquine diphosphate, after which the cells were incubated for another 5 h. The plasmid, psPAX2 (Addgene no. 12260), was mixed with pMD2.G (Addgene no. 12259) at a molar ratio of 1.3:0.72:1.64 and co-transfected into HEK293T cells using polyethyleneimine. At 15 h after transfection, cells were refreshed with maintaining medium. At 48 h after transfection, the lentivirus-containing supernatant was collected, filtered through a Millex-HV 0.45-µm low protein-binding membrane (Millipore), aliquoted and stored at −80°C. To determine the virus titer, serial dilutions of a viral aliquot were transduced into HEK293T cells in the presence of polybrene (8 µg ml^{-1}). Both untransduced cells and cells treated with the serially diluted virus were cultured in the presence of 2 µg ml^{-1} of puromycin (Invitrogen). When virtually all of the untransduced cells had died, we counted the number of living cells in the virus-treated population to estimate the viral titer as previously described[35].

Generation of the cell library. In preparation for lentivirus transduction, HEK293T cells were seeded on nine 150-mm dishes (at a density of 1.6×10^7 cells per dish) and incubated overnight. The lentiviral library was transduced into the cells at an MOI of 0.3 to achieve >500× coverage relative to the initial number of oligonucleotides. The cells were then incubated overnight, after which they were maintained in 2 µg ml^{-1} of puromycin for the next 5 d to remove untransduced cells. To preserve its diversity, the cell library was maintained at a count of at least 3.0×10^7 cells throughout the study.

PE2 delivery into the cell library. A total of 3.0×10^7 cells (from three 150-mm culture dishes, each containing 1.0×10^7 cells) were transfected with pLenti-PE2-BSD plasmid (80 µg per dish) using 80 µl of Lipofectamine 2000 (Thermo Fisher Scientific) according to the manufacturer's instructions. The culture medium was replaced with DMEM supplemented with 10% fetal bovine serum and 20 µg ml^{-1} of blasticidin S (InvivoGen) at 6 h after transfection. At 4.8 d after transfection, the cells were harvested.

Measurement of PE2 efficiencies at endogenous sites. To validate the results of the high-throughput experiment, 33 individual pegRNA-encoding plasmids were randomly selected from the plasmid library. In preparation for transfection, HEK293T cells were seeded on 48-well plates at a density of 5.0×10^4 or 1.0×10^5 cells per well 16–18 h in advance. Cells were transfected with a mixture of the plasmid-encoding PE2 (pLenti-PE2-BSD, 75 ng per 1.0×10^4 cells) and the pegRNA-encoding plasmid (25 ng per 1.0×10^4 cells) using 1 µl of Lipofectamine 2000 or TransIT-2020 transfection reagent per 1,000 ng of DNA according to the manufacturer's instructions. After incubation overnight, the culture medium was replaced with DMEM containing puromycin (2 µg ml^{-1}). The cells were harvested 4.5 d (for Endo-BR1 and Endo-BR2) or 7 d (for Endo-BR3) after the transfection.

Measurement of PE2 efficiencies in HCT116 and MDA-MB-231 cell lines. HCT116 and MDA-MB-231 cells were cultured and passaged in DMEM and RPMI, respectively, each supplemented with 10% (vol/vol) fetal bovine serum at 37 °C in the presence of 5% CO_2. To generate PE2-expressing cell lines, the PE2-encoding lentiviral vector was transduced into HCT116 and MDA-MB-231 cells at an MOI of 0.3 in culture medium containing 8 µg ml^{-1} of polybrene. After an overnight incubation, the cells were cultured in the presence of 10 µg ml^{-1} of blasticidin S for 7 d to remove untransduced cells.

Seventy-five plasmids, each containing a pair of pegRNA-encoding and corresponding target sequences, were randomly selected from plasmid library 1; plasmid identity was determined by Sanger sequencing (Supplementary Table 5). A small lentiviral library was then generated from this pool of plasmids as described above. The PE2-expressing HCT116 and MDA-MB-231 cells were seeded on six-well plates at a density of 2.0×10^5 cells per well, incubated overnight

and transduced with the lentiviral library. After incubation overnight, the culture medium was replaced with DMEM containing 1 μg ml^{-1} of puromycin and 10 μg ml^{-1} of blasticidin S or RPMI containing 2 μg ml^{-1} of puromycin and 10 μg ml^{-1} of blasticidin S for the HCT116 and MDA-MB-231 cell lines, respectively. At 4.5 d after the transduction, the cells were harvested and analyzed.

Deep sequencing. Genomic DNA was extracted from harvested cells with a Wizard Genomic DNA Purification Kit (Promega). For our high-throughput experiment, integrated barcodes and target sequences were PCR amplified using 2× Taq PCR Smart Mix (SolGent). For each cell library, the first PCR included a total of 400 μg of genomic DNA; given an assumption of 10 μg of genomic DNA per 10^6 cells, coverage would be more than 700× over the library. Eighty independent 50-μl PCR reactions were performed with an initial genomic DNA concentration of 5 μg per reaction, after which the products were pooled and gel purified with a MEGAquick-spin Total Fragment DNA Purification Kit (iNtRON Biotechnology). Then, 100 ng of purified DNA was amplified by PCR using primers that included both Illumina adaptor and barcode sequences (Supplementary Table 6). For measuring PE2 efficiencies at endogenous sites, the independent first PCR was performed in a 40-μl reaction volume that contained 200 ng of the initial genomic DNA template per sample. The second PCR to attach the Illumina adaptor and barcode sequences was then performed using 20 ng of the purified product from the first PCR in a 30-μl reaction volume. After gel purification, the resulting amplicons were analyzed using HiSeq or MiniSeq (Illumina). The PCR primers are shown in Supplementary Table 6.

Analysis of prime editing efficiencies. For analysis of deep sequencing data, we used in-house Python scripts (Supplementary Code 1) that were derived from previously used code[6]. Each pegRNA and target sequence pair was identified via a 22-nt sequence (the 18-nt barcode and 4-nt sequence located upstream of the barcode). The reads containing the specified edits without unintended mutations within the wide target sequence were considered to represent PE2-induced mutations. To exclude the background prime editing frequency originating from array synthesis and PCR amplification procedures, we subtracted the background prime editing frequency determined in the absence of PE2 from the observed prime editing frequencies as shown below.

$$= \frac{\begin{array}{c}\textit{Read counts with intended edit and specified barcode} \\ -(\textit{Total read counts with specified barcode} \\ \times \textit{background prime editing frequency}) \div 100\end{array}}{\begin{array}{c}\textit{Total read counts with specified barcode} \\ -(\textit{Total read counts with specified barcode} \\ \times \textit{background prime editing frequency}) \div 100\end{array}} \times 100$$

Deep sequencing data were filtered to improve the accuracy of our analysis. pegRNA and target sequence pairs for which the deep sequencing read counts were below 200 or the background prime editing frequencies were above 5% were excluded as we similarly performed previously[6,8,10,11].

Generation of data subsets for machine learning. PE2 efficiencies data obtained using library 1 were split into HT-training and HT-test by stratified random sampling such that the same target sequences were never shared between the two data sets (Supplementary Table 4). Similarly, PE2 efficiencies data obtained using library 2 were split into Type-training, Type-test, Position-training and Position-test such that the same target sequences were never shared between the training and test data sets (Supplementary Table 5). The target sequences used for the generation of data sets Endo-BR1, Endo-BR2, Endo-BR3, HCT-BR1, HCT-BR2, MDA-BR1 and MDA-BR2 were included in the corresponding test data sets to prevent the target sequences from being shared between the training and test data sets.

Conventional machine learning-based model training. Seven models were trained based on conventional machine learning algorithms—that is, XGBoost, gradient-boosted regression tree, random forest, L1-regularized linear regression, L2-regularized linear regression, L1L2-regularized linear regression and support vector machine (SVM). We used the XGBoost Python package (version 0.90)[36] and scikit-learn (version 0.19.1)[37] for all other models. A total of 1,766 features were extracted from the wide target sequences and the PBS and RT template sequences. The features included position-independent and position-dependent nucleotides and dinucleotides, melting temperature, GC counts, the minimum self-folding free energy[5,10] of the wide target sequence, the PBS and RT template sequences and the DeepSpCas9 score[10]. The melting temperature was calculated by a program (https://biopython.org/docs/1.74/api/Bio.SeqUtils.MeltingTemp.html) using a default setting without considering the cellular nuclei milieu. For model selection among the regularization parameters and hyperparameter configurations, we performed five-fold cross-validation. For XGBoost and gradient-boosted regression tree, we searched over 144 models chosen from the following hyperparameter configurations: the number of base estimators (chosen from [5, 10, 50, 100]), the maximum depth of the individual regression estimators (chosen from [5, 10, 50, 100]), the minimum number of samples to be at a leaf node (chosen from [1, 2, 4]) and learning rate (chosen from [0.05, 0.1, 0.2]). For random forest, we searched over 144 models chosen from the same hyperparameter configurations listed above for XGBoost, except for the learning rate; we searched over the maximum number of features to consider when looking for the best split (chosen from [all features, the square root of all features, the binary logarithm of all features]). For L1-, L2-, and L1L2-regularized linear regression, to optimize the regularization parameter, over 144 points that were evenly spaced between 10^{-6} and 10^6 in log space were searched. For SVM, we searched over 144 models from the following hyperparameters: penalty parameter C and kernel parameter γ, 12 points that were evenly spaced between 10^{-3} and 10^3.

Evaluation of feature importance. To measure feature importance for predicting PE2 efficiencies, the Tree SHAP method[20] was used. We extracted features and trained XGBoost models with the best hyperparameter configurations determined from five-fold cross-validation as described above. In the Tree SHAP method, a per-sample importance score is assigned to each feature from the trained XGBoost models. The importance score, which represents the effect of the feature on the base value in the model output, is computed on the basis of a game theoretic Shapley value for optimal credit allocation[20]. We show SHAP value distributions for the entire data set or provide the mean absolute value to give a general overview of feature importance in our model.

Development of deep learning-based algorithms. DeepPE is a deep learning-based computational model that predicts the optimal combination of PBS and RT template lengths to introduce a G-to-C transversion mutation at position +5 from the nicking site. We used the training data set that consists of the prime editing efficiencies induced by PE2 and 38,692 pegRNAs; these training data contained the 47-nt-wide target sequences, the 17~37-nt RT template plus PBS sequences and 20 additional features, including melting temperature, GC counts, GC contents and minimum self-folding free energy. The nucleotide sequences were converted into four-dimensional binary matrixes by one-hot encoding.

DeepPE was developed using a convolutional layer and a fully connected layer. The convolution layer obtained two embedding vectors from the wide target sequences and RT template plus PBS sequences using ten filters at 3 nt in length. Then, the embedding vectors were concatenated with the 20 biological features. The pooling layer was excluded as the deep reinforcement learning algorithm was implemented to maintain local information[38]. The fully connected layer with 1,000 units multiplied the vectors with the rectified linear unit activation function. The regression output layer performed a linear transformation of the outputs and calculated the prediction scores for PE2 efficiency. After testing nine different models (hyperparameters; number of filters (10, 20, 40) and units (200, 500, 1,000) for the convolutional layer and fully connected layer, respectively), we chose the model that resulted in the highest Spearman's correlation coefficients between the experimentally measured and predicted activity levels during the five-fold cross-validation. Dropout was used to avoid overfitting with a rate of 0.3. The mean squared error, as the objective function, and an Adam optimizer with a learning rate of 10^{-3} were used. DeepPE was implemented using TensorFlow[39]. DeepPE is provided as Supplementary Code 2.

For the development of deep learning-based algorithms to predict PE2 efficiencies for various editing types and positions, we used multilayer perceptron (MLP) instead of a convolutional neural network, because an initial trial using a convolutional neural network showed poor performance. We have performed cross-validations to select among 18 MLP models that have similar architectures and number of parameters as DeepPE but lack the convolutions. The considered hyperparameter configurations were as follows: the number of layers (chosen from [2, 3]), the number of units in each hidden layer (chosen from [1,000, 200, 50] for the first hidden layer and [50] for the second hidden layer), the dropout regularization parameter, the learning rate (chosen from [0.01, 0.001, 0.0001]) and the ReLU activation function.

Statistics and reproducibility. To compare prime editing efficiencies between experiments using different pegRNAs, we used a one-way analysis of variance (ANOVA) followed by two-sided Tukey's post hoc test. To compare the Spearman's correlation between prediction scores from prediction models (Fig. 3 and Supplementary Fig. 21), we used a two-sided Steiger's test, which is a method for testing two dependent correlation coefficients from exactly the same data set. A chi-square test was performed to determine the relationship between PBS lengths and RT template lengths when the most efficient combination of these two parameters per target sequence was selected. For increased accuracy of the chi-square analysis, target sequences that showed a prime editing efficiency lower than 10%, even when the most efficient combination of the two parameters was selected, were filtered out from the analysis. To compare the PE2 efficiencies for pegRNAs with PBS and RT template lengths that were chosen using DeepPE versus the initial study's recommendation at given target sequences, we used a two-tailed paired *t*-test. To determine statistical significance, we used GraphPad Prism 8, PASW Statistics (version 18.0, IBM) and Microsoft Excel (version 16.0, Microsoft Corporation). For high-throughput evaluation of PE2 efficiencies using libraries 1 and 2, we combined the data from the two replicates independently transfected by two different experimentalists.

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

Data availability

The deep sequencing data from this study have been submitted to the National Center for Biotechnology Information Sequence Read Archive under accession number PRJNA624815. The data sets used in this study are provided as Supplementary Tables 3, 4 and 5.

Code availability

Source codes for DeepPE and the custom Python script used for the prime editing efficiency calculations are provided as Supplementary Codes 1 and 2 and are also available at https://github.com/hkimlab-PE/PE_SupplementaryCode.

References

33. Du, D. et al. Genetic interaction mapping in mammalian cells using CRISPR interference. *Nat. Methods* **14**, 577–580 (2017).
34. Shen, J. P. et al. Combinatorial CRISPR–Cas9 screens for de novo mapping of genetic interactions. *Nat. Methods* **14**, 573–576 (2017).
35. Shalem, O. et al. Genome-scale CRISPR–Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
36. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. Preprint at https://arxiv.org/abs/1603.02754 (2016).
37. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
38. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
39. Abadi, M. et al. In *Proc. of the 12th USENIX Conference on Operating Systems Design and Implementation* 265–283 (USENIX Association, 2016).

Acknowledgements

We would like to thank D. Kim, S. Park and Y. Kim for assisting with the experiments. This work was supported, in part, by the National Research Foundation of Korea (grants 2017R1A2B3004198 (H.H.K.), 2017M3A9B4062403 (H.H.K.), 2020R1C1C1003284 (H.K.K) and 2018R1A5A2025079 (H.H.K)), the Brain Korea 21 Plus Project (Yonsei University College of Medicine) and the Korean Health Technology R&D Project, Ministry of Health and Welfare, Republic of Korea (grants HI17C0676 (H.H.K.) and HI16C1012 (H.H.K.)).

Author contributions

G.Y. and H.K.K. performed the wet experiments, including high-throughput evaluation of PE2 efficiencies. S.M., S.L., S.Y. and H.K.K. developed DeepPE and the related web tools. J.P. substantially contributed to bioinformatics analyses and DeepPE development. H.K.K. and H.H.K. conceived of and designed the study. H.K.K., G.Y. and H.H.K. analyzed the data and wrote the manuscript.

Competing interests

Yonsei University has filed a patent application based on this work, in which H.K.K., G.Y. and H.H.K. are listed as inventors.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0677-y.

Correspondence and requests for materials should be addressed to H.H.K.

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

nature research

Corresponding author(s): Hyongbum Henry Kim

Last updated by author(s): Aug 11, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Targeted deep sequencing data were collected using HiSeq2500, HiSeq 4000, and MiniSeq (Illumina).
Data analysis	GraphPad Prism 8, PASW Statistics (version 18.0, IBM), Microsoft Excel (version, 16.0, Microsoft Corporation), XGBoost Python package (version 0.90), scikit-learn (version 0.19.1), and TensorFlow were used. Source codes for DeepPE and custom python script used for the prime editing efficiency calculation are provided as Supplementary codes 1 and 2 and also available at https://github.com/hkimlab-PE/PE_SupplementaryCode.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The deep sequencing data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number PRJNA624815.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Sample sizes were chosen after deep sequencing depending on the read number and quality. All sample sizes were sufficient for the following statistical test.
Data exclusions	To increase the accuracy of the analysis for PE efficiency, deep sequencing read counts were below 200 or the background PE frequencies were above 5% were excluded.
Replication	We replicated experiments as described in the manuscript. We performed high-throughput evaluation in duplicate by two independent researchers. We evaluated PE2 efficiencies at endogenous sites of HEK293T cells in sextuplicate and those of HCT116 and MDA-MB-231 cells in quadruplicate. All replcates were performed successfully and showed strong correlation between replicates.
Randomization	We selected target sequences for the development of DeepPE and other conventional machine learning based models by stratified random sampling.
Blinding	The investigators were not blinded to group allocation. This study does not involve animals or human research participants.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The source of the cell line, HEK293T, is American Type Culture Collection (ATCC). The source of the cell line, HCT116 and MDA-MB-231, are Korean Cell Line Bank (KCLB).
Authentication	Not been authenticated.
Mycoplasma contamination	Not been tested.
Commonly misidentified lines (See ICLAC register)	HEK293T, HCT116, and MDA-MB-231 are not listed in the ICLAC.