nature research

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

Supplementary information

Predicting the efficiency of prime editing guide RNAs in human cells

In the format provided by the authors and unedited

Supplementary Information

Predicting the efficiency of prime-editing guide RNAs in human cells

Contents:

Supplementary Text 1

As described in the legends of Fig. 2e and Supplementary Figs. 12, 16, instances with high and low feature values are represented as red and blue dots, respectively. For example, in the second feature shown in Fig. 2e, target sequences with high and low GC counts in PBS are represented as red and blue dots, respectively. SHAP values for red dots are mostly higher than 0 and SHAP values for blue dots are mostly lower than 0. Given that high SHAP values are linked with high PE2 efficiencies, high and low GC counts in PBS are linked with high and low prime editing efficiencies, respectively. In the case of the fourth feature, target sequences with a high number of UUs in the PBS and RT template regions (thus, shown in red) have SHAP values lower than 0, suggesting that a high number of UUs in those regions is associated with low prime editing efficiencies.

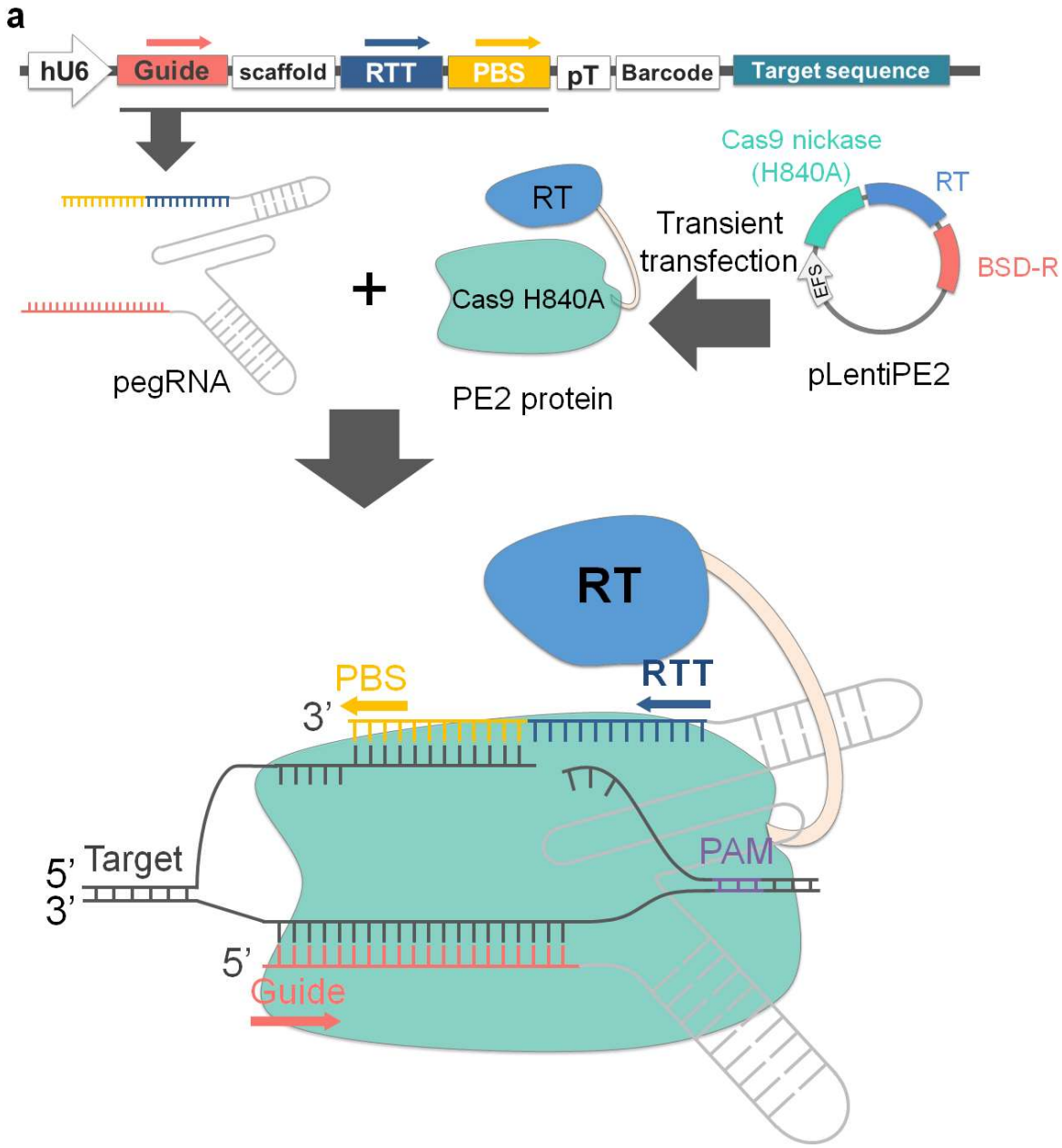
Supplementary Text 2

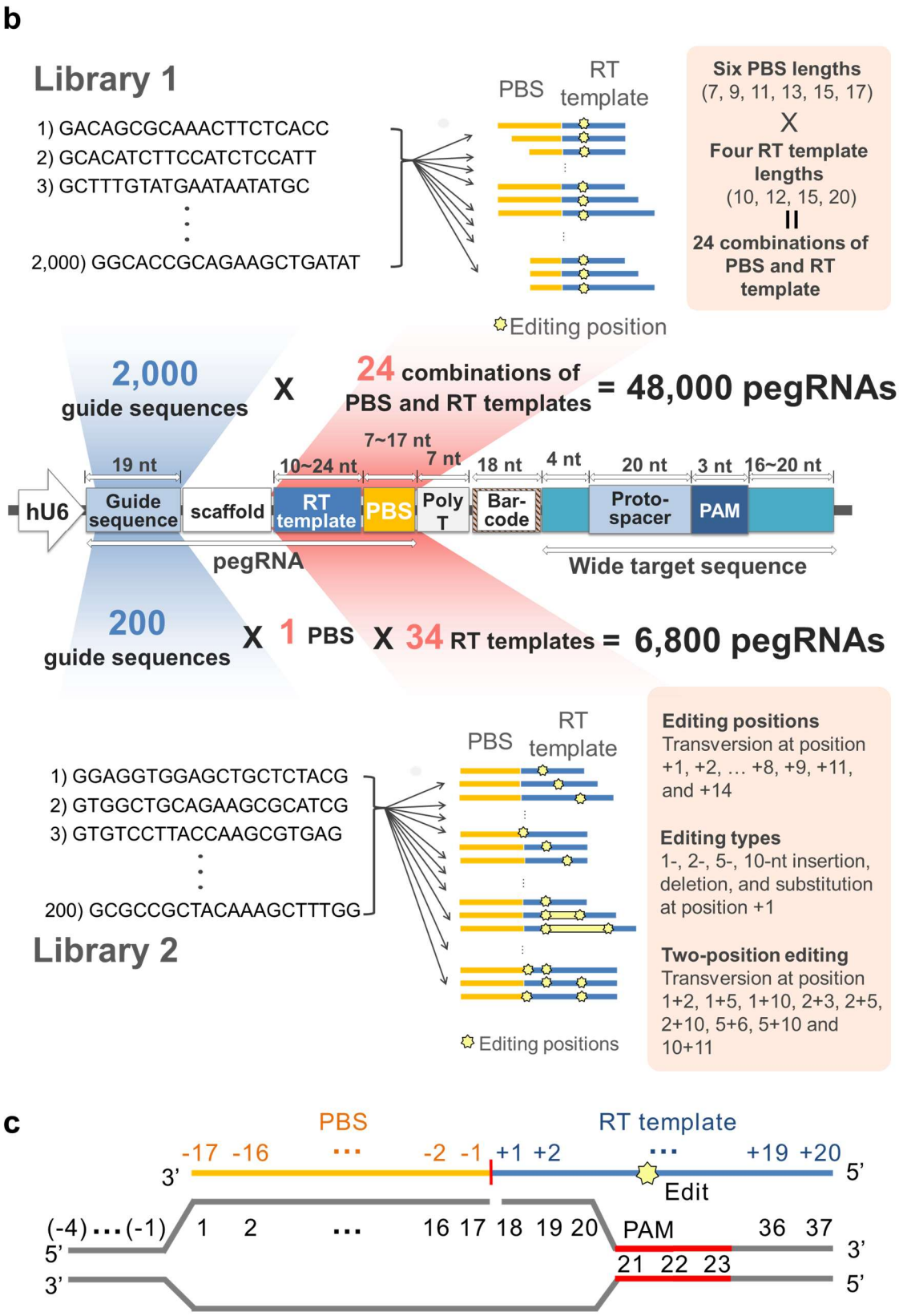
We also evaluated factors affecting PE2 activity when the DeepSpCas9 score was excluded (Supplementary Fig. 16). In this case, the most important feature was the GC counts in the PBS; its absolute SHAP value (0.0242) was basically similar to that from the analysis that included the DeepSpCas9 score as a feature (0.0188). All ten of the most important features when the DeepSpCas9 score was included as a feature showed basically comparable feature ranks and SHAP values when that score was excluded. However, some of the features that were less important in the evaluation that included the DeepSpCas9 score became more important in this evaluation without that score. As representative examples, G at position 20 (favored, average absolute SHAP value = 0.015 without the DeepSpCas9 score; the 94th most important feature, average absolute SHAP value = 0.001 with the DeepSpCas9 score), C at position 18 (favored, average absolute SHAP value = 0.011 without score; the 43th most important feature, average absolute SHAP value = 0.002 with score), and T at position 17 (disfavored, average absolute SHAP value = 0.010 without score; the 58th most important feature, average absolute SHAP value = 0.001 with score) became the second, fifth, and eighth most important features, respectively, all of which substantially affect SpCas9 activity[4, 5, 15].

Supplementary Text 3

Together with the use of DeepPE, PE_type, and PE_position, we recommend the following for efficient prime editing. In general, 13-nt PBS and 12-nt RT template lengths are recommended, which is basically in line with the recommendation of the initial study[1]. However, we recommend that the PBS length should be increased to 15 nt when the GC contents are lower than 40% or decreased to 9 nt when the GC contents are higher than 60%. High GC counts in PBS are generally associated with high PE2 efficiencies. The RT template length should also be modified according to the last templated nucleotide. When

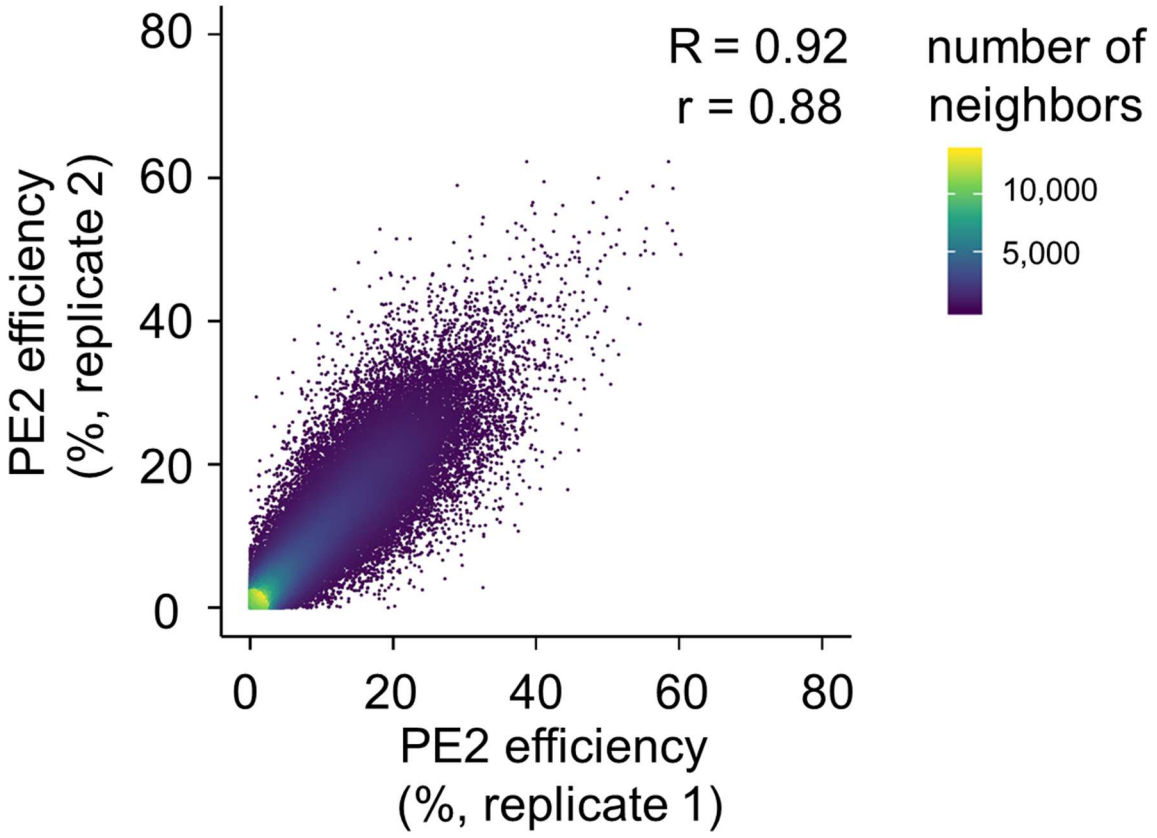
the last templated nucleotide is not a G and the RT template length is ≤12 nt, we recommend altering the RT template length to 11, 13, 14, or 15 nt so that the last templated nucleotide is a G if possible, which basically differs from the recommendation of the initial study[1]. However, when the RT template length is ≥20 nt, making the last templated nucleotide a C is recommended, which is compatible with the recommendation of the initial study[1]. When the intended editing does not disrupt the PAM (NGG), we recommend including PAM editing (e.g. editing G to C at position +5) in addition to the intended editing, which is also in line with the recommendation of the initial study[1].


a
hU6
Guide
scaffold
RTT
PBS
pT
Barcode
Target sequence
pegRNA
RT
Cas9 H840A
PE2 protein
Transient transfection
Cas9 nickase (H840A)
RT
BSD-R
EFS
pLentiPE2
RT
PBS
RTT
3'
5' Target
3'
PAM
5'
Guide

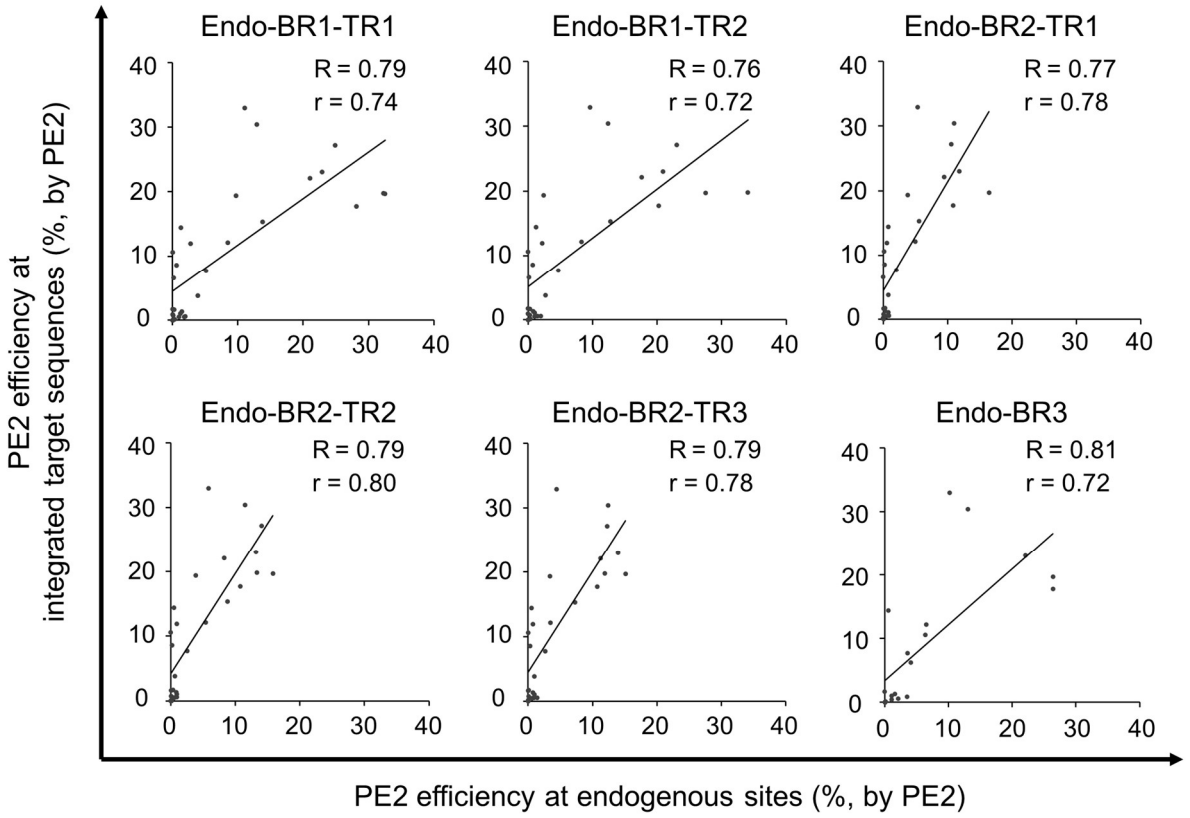


Supplementary Figure 1. Schematic representation of prime editing components and strategy used in this study. Each pegRNA was paired with a wide target sequence that includes a protospacer, a protospacer adjacent motif (PAM), and neighboring sequences.

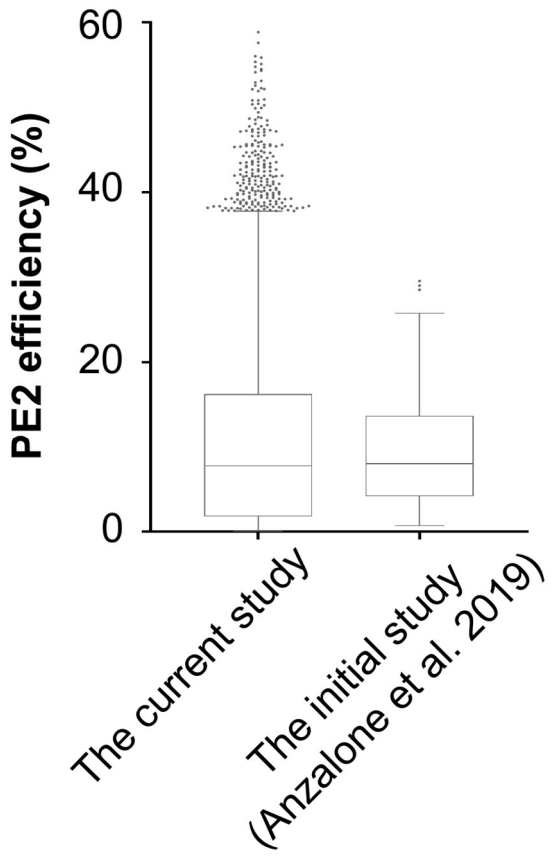
(**a**) Schematic of prime editing components and strategy used in this study. The PE2 protein was expressed by transient transfection. The human U6 promoter (hU6) was used for the expression of pegRNAs, which guide PE2 to the target sequence. Guide, guide sequence; RTT, reverse transcriptase template; PBS, primer binding site; RT, reverse transcriptase; BSD-R, blasticidin-resistant gene. (**b**) Construction of libraries 1 and 2. In library 1, 2,000 guide sequences were associated with 24 combinations of different primer binding site (PBS) and RT template lengths, resulting in 48,000 pegRNAs. In library 2, 200 guide sequences were linked with 34 different combinations of PBSs and RT templates, designed to generate various types of intended edits in different positions, resulting in 6,800 pegRNAs. (**c**) Schematic showing how positions are designated within the pegRNA, cDNA, and wide target sequence in this study. Positions within the pegRNA and the cDNA generated from the pegRNA are numbered starting at the nicking site of the Cas9 nickase. Positions within the wide target sequence are designated such that the 20^{th} nucleotide upstream from the PAM is position 1 and the nucleotides in the NGG PAM are positions 21 - 23.



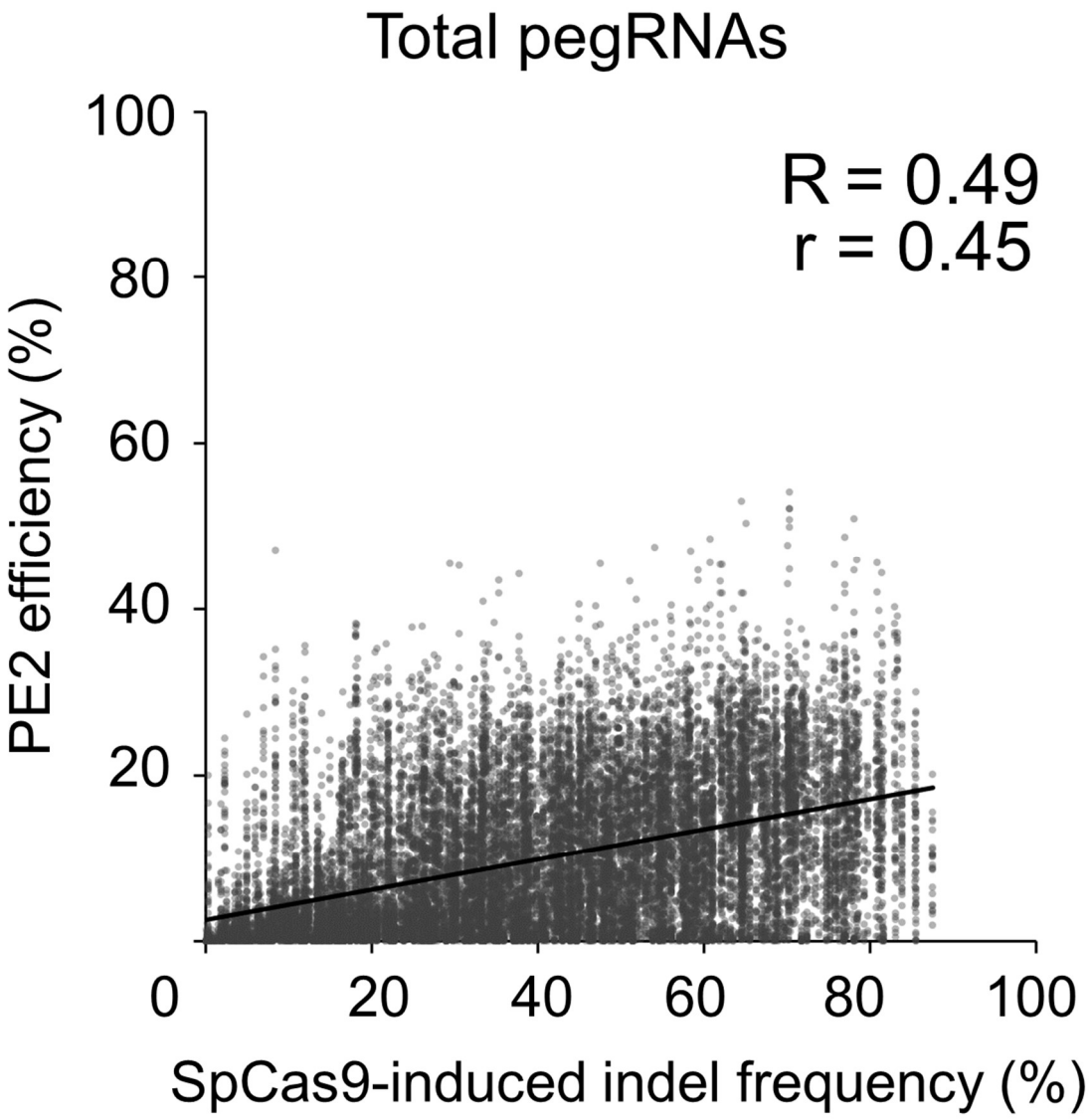
Supplementary Figure 2. Correlation between PE efficiencies in replicates independently transfected with the PE2-encoding plasmid. The combined results from libraries 1 and 2 are shown here. To increase the accuracy of the analyses, pegRNA and target sequence pairs for which the deep sequencing read counts were below 200 or the background prime editing frequencies were above 5% were filtered out. The color of each dot was determined by the number of neighboring dots (i.e., dots within a distance that is three times the radius of the dot). The number of pegRNA and target sequence pairs n = 49,301. The Spearman (R) and Pearson (r) correlation coefficients are shown.



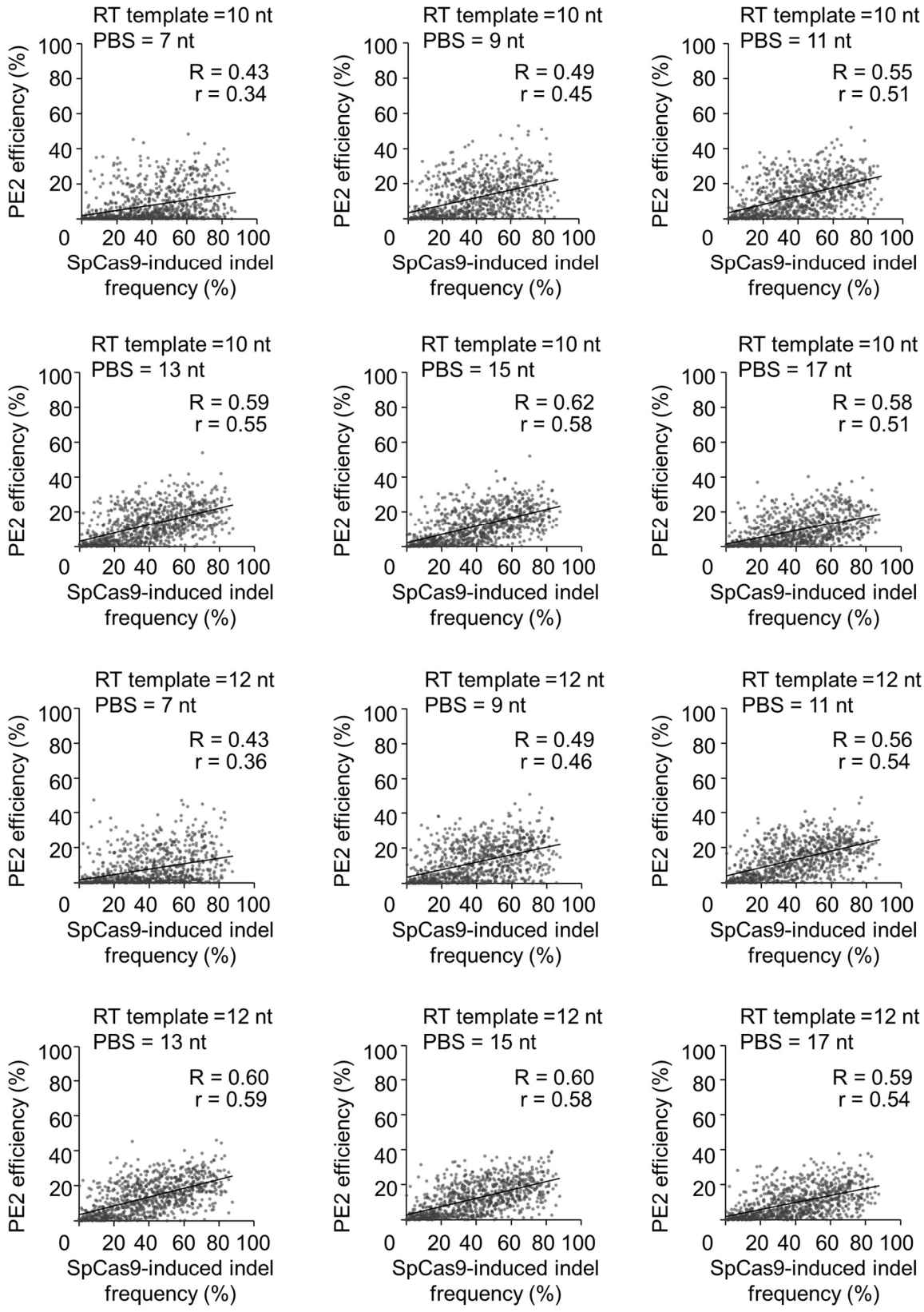
Supplementary Figure 3. Correlations between PE efficiencies measured at endogenous sites and those at corresponding integrated target sequences. Three biological replicates (BR1, BR2, and BR3) were evaluated and each biological replicate had one, two, or three technical replicates (TRs). Six datasets of PE2 efficiencies at endogenous sites in HEK293T cells were generated by the evaluations 4.5 (Endo-BR1-TR1, Endo-BR1-TR2, Endo-BR2-TR1, Endo-BR2-TR2, and Endo-BR2-TR3) or 7 (Endo-BR3) days after transient transfection of plasmids encoding PE2 and pegRNAs. The results using Endo-BR1-TR1 are also shown in Fig. 4b. The number of pegRNA and target sequence pairs n = 31, 30, 28, 28, 28, and 20 for Endo-BR1-TR1, Endo-BR1-TR2, Endo-BR2-TR-1, Endo-BR2-TR2, Endo-BR2-TR3, and Endo-BR3, respectively. In Endo-BR3, 18 pegRNAs randomly chosen from a total of 31 pegRNAs, and two pegRNAs that were not tested in the other replicates, were evaluated. In Endo-BR1-TR2, Endo-BR2-TR-1, Endo-BR2-TR2, and Endo-BR2-TR3, one or three pegRNAs were removed from the analysis due to transfection failure or an insufficient deep sequencing read count (less than 200). The Spearman (R) and Pearson (r) correlation coefficients are shown.

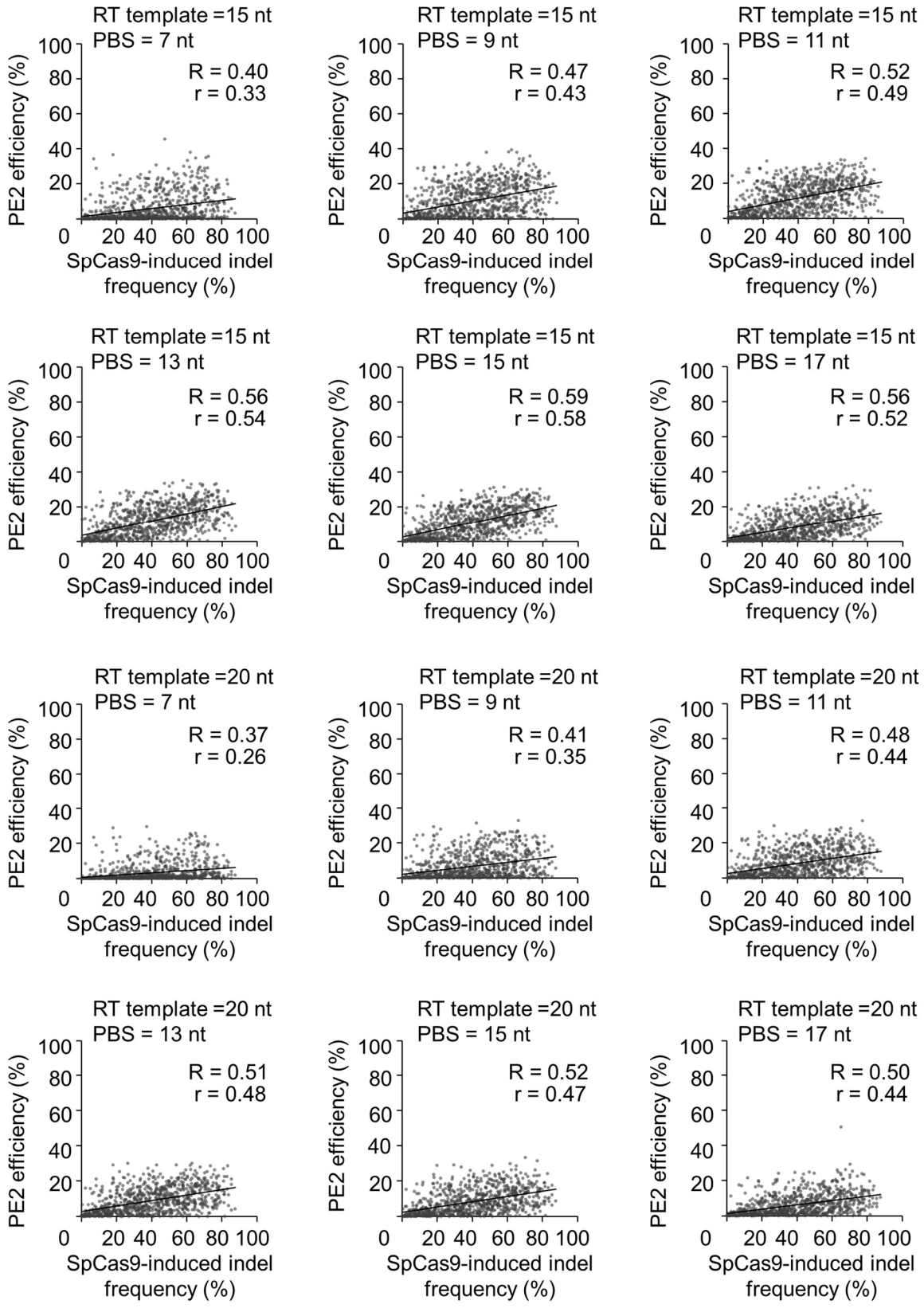


Supplementary Figure 4. The distribution of PE2 efficiencies in comparison with that of the initial study[1]. In the boxes, the top, middle, and bottom lines represent the 25th, 50th, and 75th percentiles, respectively, whiskers indicate the 10th and 90th percentiles, and outliers are shown as individual dots. The number of pegRNAs n = 49,301 (the current study using libraries 1 and 2) and 186 (the initial study[1]).



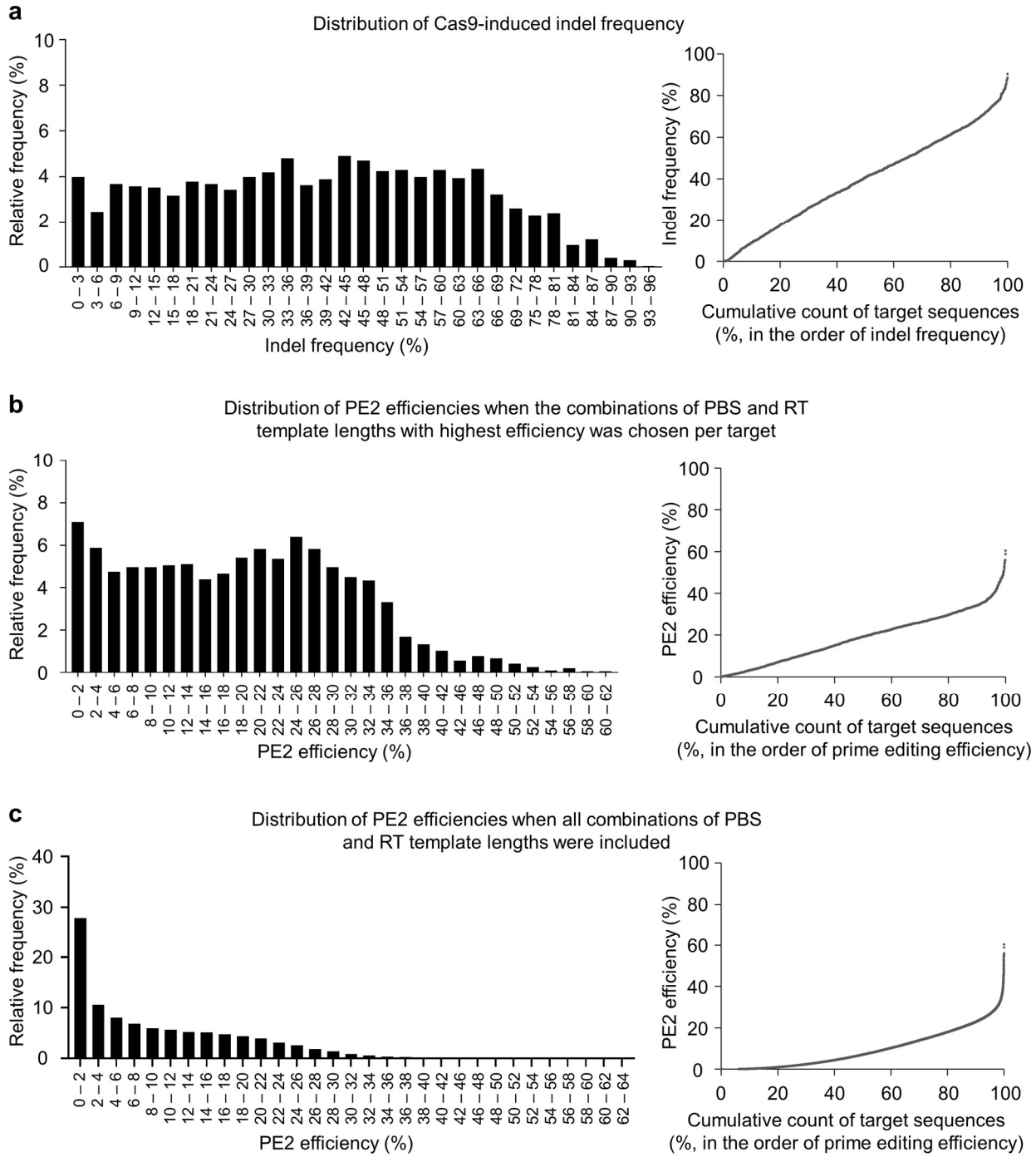
Supplementary Figure 5. Correlations between SpCas9-induced indel frequencies and prime editing efficiencies determined for identical target sequences using library 1. The correlation was evaluated when all 24 combinations of the PBS and RT template lengths were considered. To increase the accuracy of the analyses, 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 filtered out. If any of the 24 pegRNAs with different combinations of the PBS and RT template lengths for a given target sequence were filtered out, the target sequence and remaining pegRNAs for that target were also filtered out from the analysis. The number of pegRNA and target sequence pairs n = 21,288. The Spearman (R) and Pearson (r) correlation coefficients are shown.


RT template =10 nt
PBS = 7 nt
R = 0.43
r = 0.34
RT template =10 nt
PBS = 9 nt
R = 0.49
r = 0.45
RT template =10 nt
PBS = 11 nt
R = 0.55
r = 0.51
RT template =10 nt
PBS = 13 nt
R = 0.59
r = 0.55
RT template =10 nt
PBS = 15 nt
R = 0.62
r = 0.58
RT template =10 nt
PBS = 17 nt
R = 0.58
r = 0.51
RT template =12 nt
PBS = 7 nt
R = 0.43
r = 0.36
RT template =12 nt
PBS = 9 nt
R = 0.49
r = 0.46
RT template =12 nt
PBS = 11 nt
R = 0.56
r = 0.54
RT template =12 nt
PBS = 13 nt
R = 0.60
r = 0.59
RT template =12 nt
PBS = 15 nt
R = 0.60
r = 0.58
RT template =12 nt
PBS = 17 nt
R = 0.59
r = 0.54
PE2 efficiency (%)
SpCas9-induced indel frequency (%)
0
20
40
60
80
100

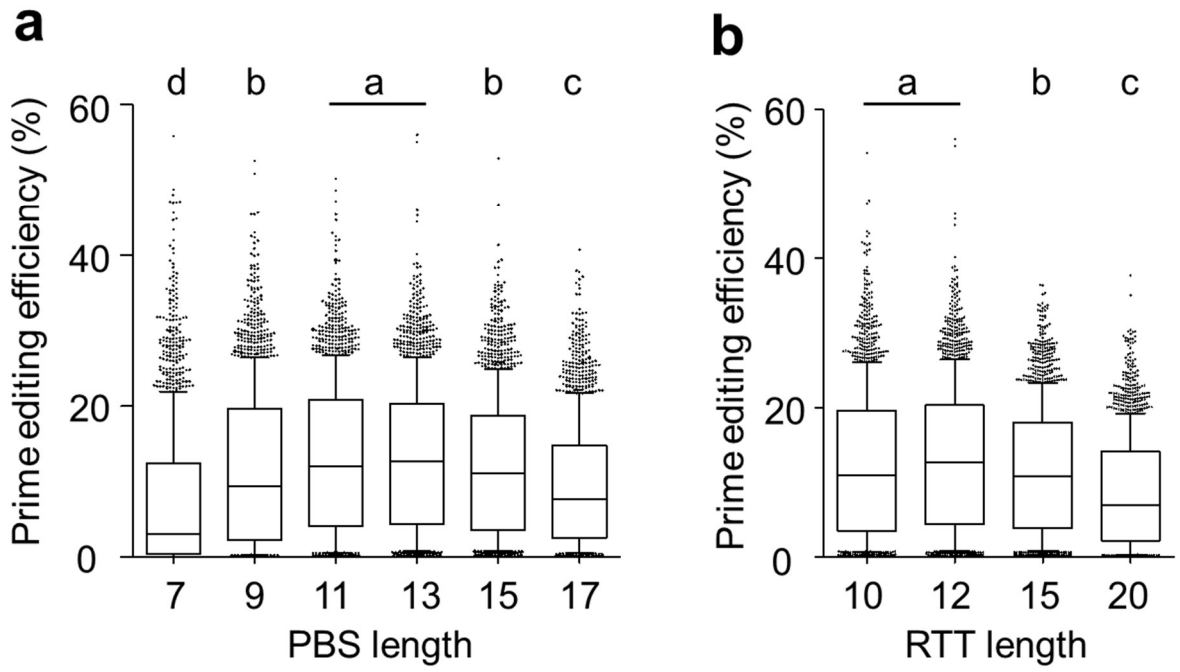


Supplementary Figure 6. Correlations between SpCas9-induced indel frequencies and prime editing efficiencies determined for identical target sequences using library 1. The correlations were evaluated when the specified lengths of the PBS and RT template,

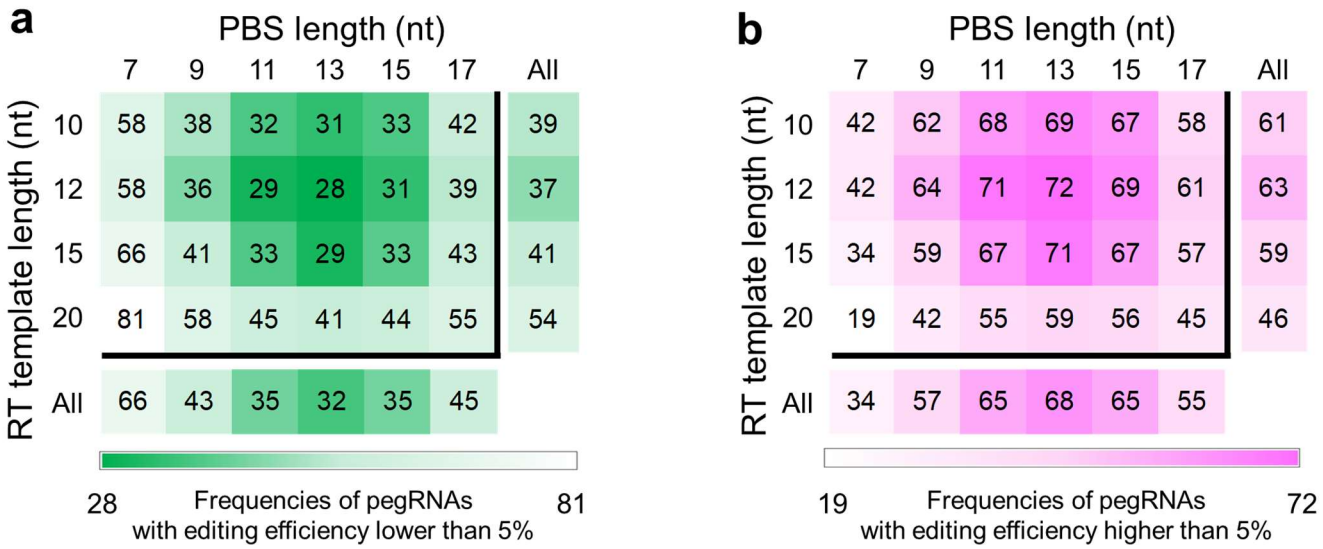
shown on the top of each graph, were chosen. To increase the accuracy of the analyses, pegRNA and target sequence pairs for which the deep sequencing read counts were below 200 or the background prime editing frequencies were above 5% were filtered out. If any of the 24 pegRNAs with different combinations of the PBS and RT template lengths for a given target sequence were filtered out, the target sequence and remaining pegRNAs for that target were also filtered out from the analysis. The number of pegRNA and target sequence pairs n = 887. The Spearman (R) and Pearson (r) correlation coefficients are shown.



Supplementary Figure 7. Distribution of Cas9-induced indel frequencies and PE2 efficiencies for the same set of target sequences. Histograms (left) and ranked scatter plots (right) show the distribution of SpCas9-induced indel frequencies (a) and PE2 efficiencies (b, c) measured at 1,956 target sequences. (b) The pegRNA that showed the highest efficiency among the 24 pegRNAs with different PBS and RT template lengths was chosen per target sequence. (c) All pegRNAs, with the 24 combinations of PBS and RT template lengths, were considered.



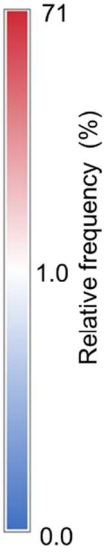
Supplementary Figure 8. Effect of PBS and RT template lengths on prime editing (PE) efficiency. PE efficiencies were evaluated with various lengths of the PBS (a) or the RT template (b) while the length of the RT template (a) and the PBS (b) were fixed to 12 nt and 13 nt, respectively. Subsets of experimental groups without statistically significant differences in PE efficiencies are represented with letters such as a, b, c, and d. P = 1.0, 0.98, 1.0, and 1.0 for subsets a, b, c and d, respectively, for groups with various PBS lengths; P = 0.08, 1.0, and 1.0 for subsets a, b, and c, respectively, for groups with various RT template lengths; ANOVA followed by two-sided Tukey's post hoc 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. The number of pegRNA and target sequence pairs per experimental group specified on the x-axis n = 1,809, 1,826, 1,824, 1,810, 1,799, and 1,760 for pegRNAs with a 7-, 9-, 11-, 13-, 15-, and 17-nt PBS, respectively (a) and n = 1,790, 1,810, 1,825, and 1,772 for pegRNAs with a 10-, 12-, 15-, and 20-nt RT template, respectively.

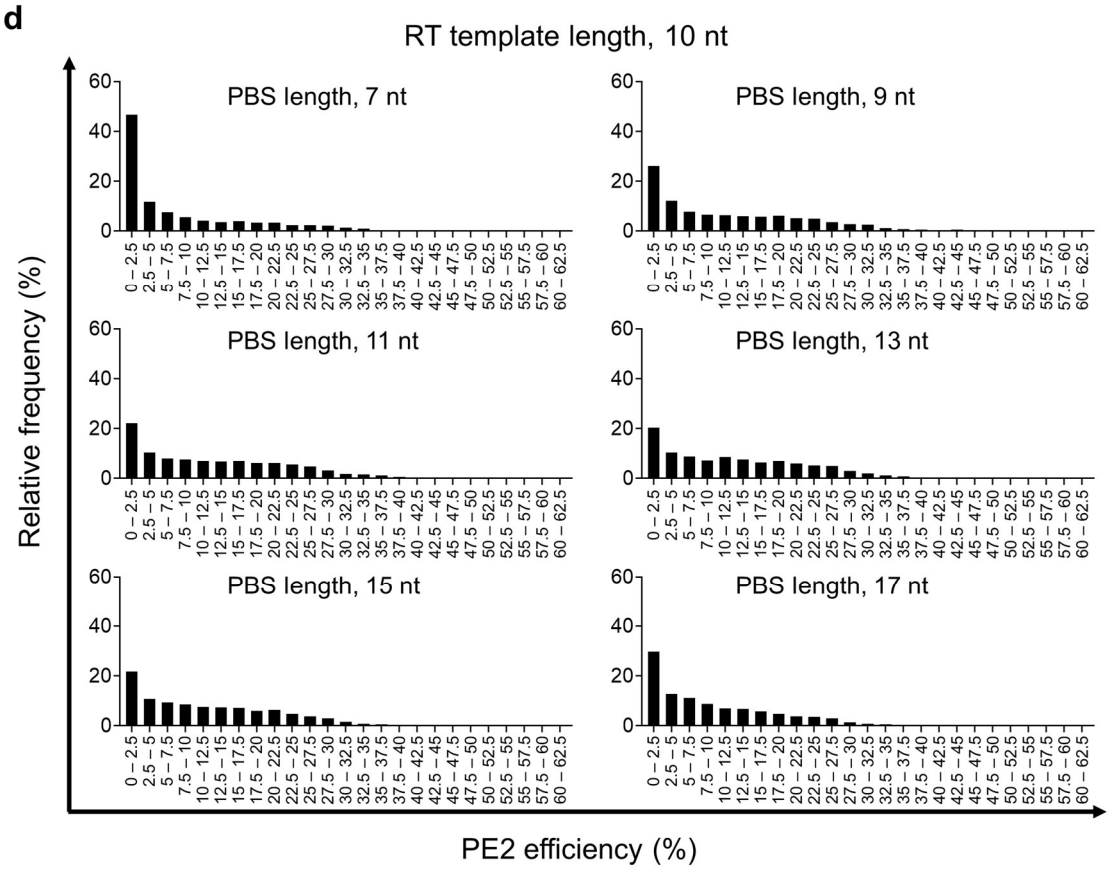


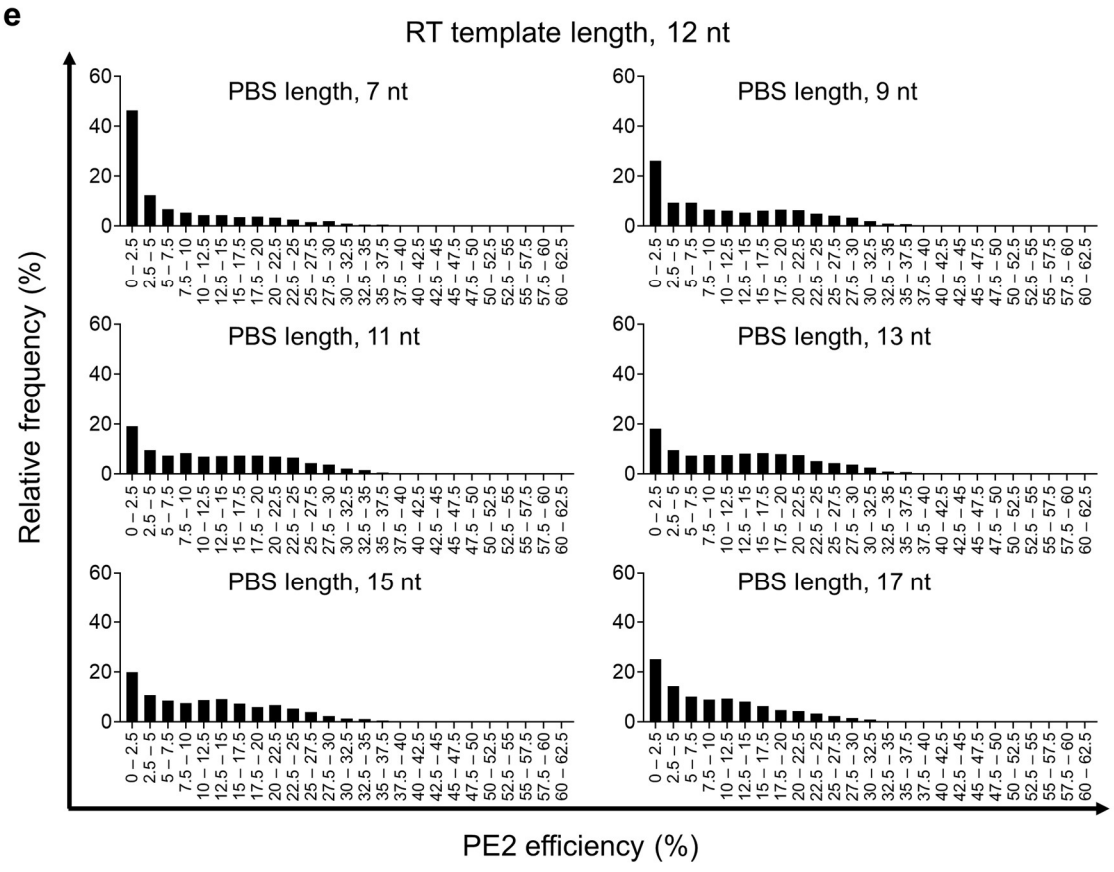
c

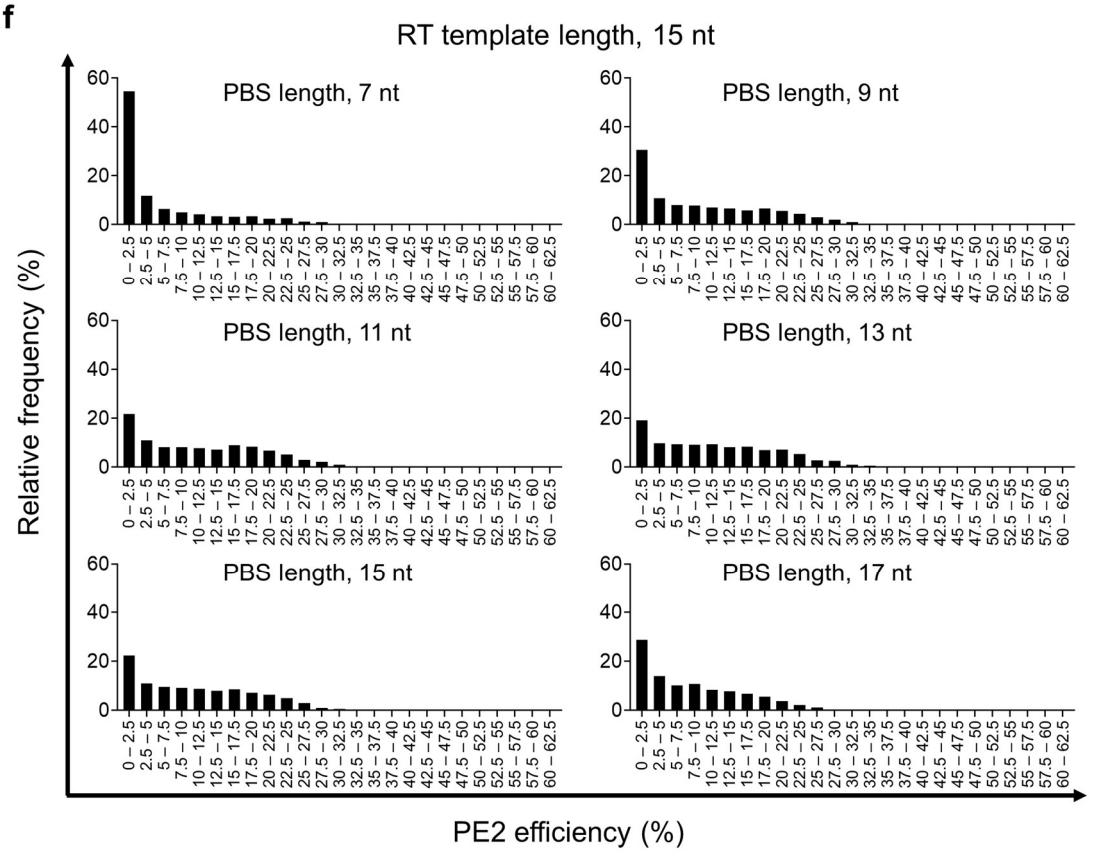
RT template length (nt)	10						12						15						20					
PBS length (nt)	7	9	11	13	15	17	7	9	11	13	15	17	7	9	11	13	15	17	7	9	11	13	15	17
0 – 2.5	47	26	22	20	22	30	46	26	19	18	20	25	55	31	22	19	22	29	71	45	32	27	28	38
2.5 – 5	12	12	10	10	11	13	12	9.4	9.5	9.5	11	14	12	11	11	9.8	11	14	10	13	14	14	15	17
5 – 7.5	7.5	7.8	7.9	8.8	9.4	11	6.8	9.4	7.4	7.3	8.5	10	6.4	7.9	8.2	9.4	9.6	10	5.7	9.0	11	12	13	12
7.5 – 10	5.6	6.5	7.5	7.2	8.5	8.8	5.4	6.6	8.3	7.5	7.6	8.9	4.9	7.7	8.2	9.2	9.1	11	3.1	6.4	8.4	9.1	9.5	10
10 – 12.5	4.1	6.4	7.0	8.5	7.5	7.0	4.4	6.3	7.0	7.5	8.7	9.3	4.1	7.0	7.7	9.3	8.7	8.4	2.5	5.9	8.0	8.2	7.9	7.0
12.5 – 15	3.5	6.0	6.8	7.5	7.4	6.7	4.3	5.5	7.2	8.1	9.2	8.1	3.4	6.5	7.2	8.2	8.0	7.8	2.6	6.0	6.4	7.8	7.3	5.1
15 – 17.5	4.0	5.7	6.9	6.4	7.2	5.7	3.6	6.2	7.3	8.4	7.3	6.4	3.3	5.7	8.9	8.4	8.5	6.7	1.6	5.0	6.2	7.6	6.3	3.9
17.5 – 20	3.3	6.1	6.2	6.9	6.0	4.8	3.7	6.5	7.4	7.9	6.0	4.8	3.4	6.6	8.3	7.0	7.2	5.6	1.6	3.2	5.9	5.8	5.4	2.9
20 – 22.5	3.3	5.1	6.1	6.0	6.4	3.8	3.4	6.4	6.9	7.6	6.8	4.4	2.4	5.6	6.7	7.2	6.3	3.7	0.9	3.2	4.4	4.3	3.8	1.7
22.5 – 25	2.3	5.0	5.5	5.2	4.7	3.5	2.7	5.0	6.5	5.2	5.3	3.3	2.5	4.4	5.2	5.4	4.9	2.2	0.8	1.6	2.3	2.2	1.7	1.4
25 – 27.5	2.3	3.6	4.7	4.9	3.8	3.0	1.5	4.2	4.4	4.3	4.0	2.3	1.2	3.1	3.0	2.7	3.0	1.1	0.2	0.8	1.1	1.0	0.9	0.4
27.5 – 30	2.1	2.7	3.1	2.9	2.9	1.4	2.0	3.4	3.8	3.8	2.4	1.5	0.9	2.0	2.2	2.5	0.9	0.4	0.2	0.3	0.8	0.8	0.5	0.2
30 – 32.5	1.3	2.6	1.7	2.0	1.5	0.7	1.1	1.9	2.2	2.5	1.4	1.0	0.3	1.0	1.0	0.9	0.5	0.4	0.0	0.2	0.1	0.2	0.2	0.0
32.5 – 35	1.0	1.2	1.6	1.1	0.8	0.5	0.7	0.9	1.5	1.0	1.1	0.2	0.4	0.4	0.4	0.6	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.1
35 – 37.5	0.4	0.8	1.1	0.7	0.5	0.3	0.6	0.8	0.5	0.8	0.6	0.3	0.1	0.2	0.3	0.2	0.2	0.1	0.0	0.0	0.0	0.1	0.1	0.0
37.5 – 40	0.3	0.6	0.5	0.3	0.4	0.3	0.3	0.4	0.2	0.2	0.2	0.1	0.1	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0
40 – 42.5	0.1	0.4	0.4	0.4	0.2	0.1	0.2	0.3	0.3	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
42.5 – 45	0.2	0.5	0.3	0.2	0.2	0.0	0.2	0.1	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
45 – 47.5	0.1	0.3	0.2	0.1	0.1	0.0	0.3	0.2	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
47.5 – 50	0.2	0.1	0.1	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50 – 52.5	0.1	0.1	0.1	0.0	0.1	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
52.5 – 55	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
55 – 57.5	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
57.5 – 60	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
60 – 62.5	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

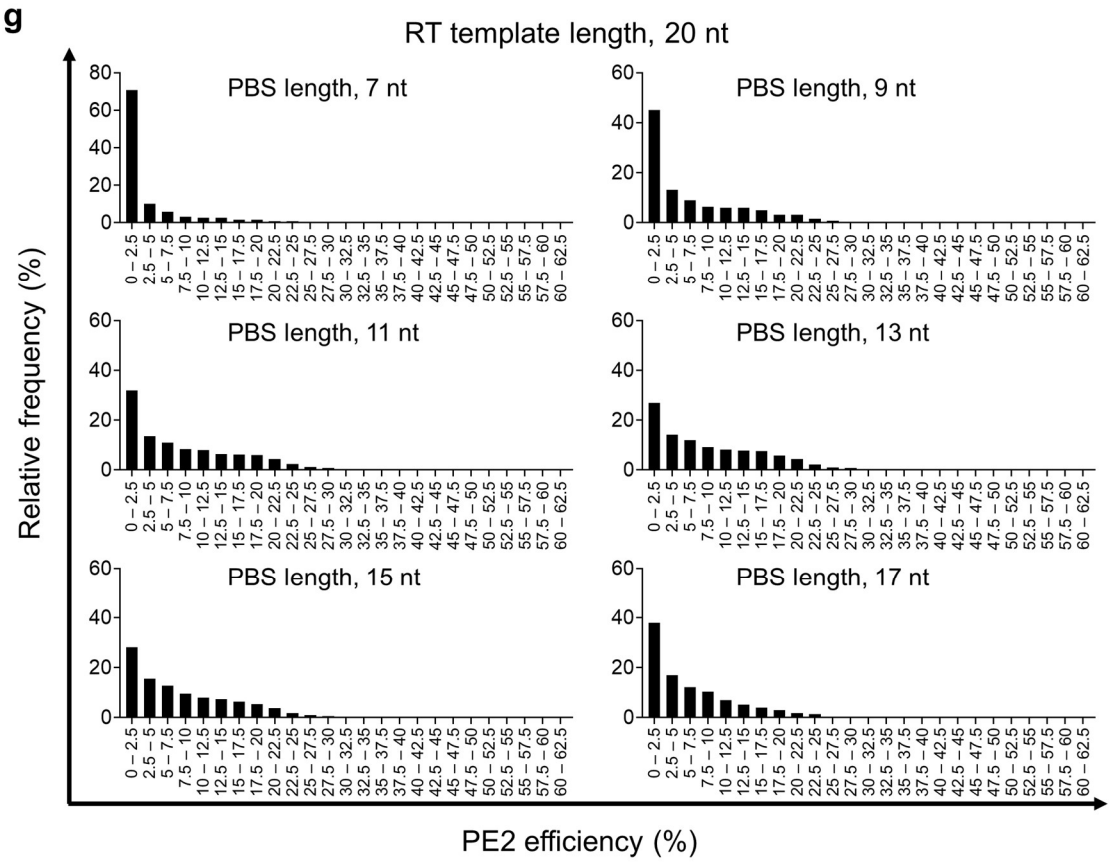
PE2 efficiency (%)



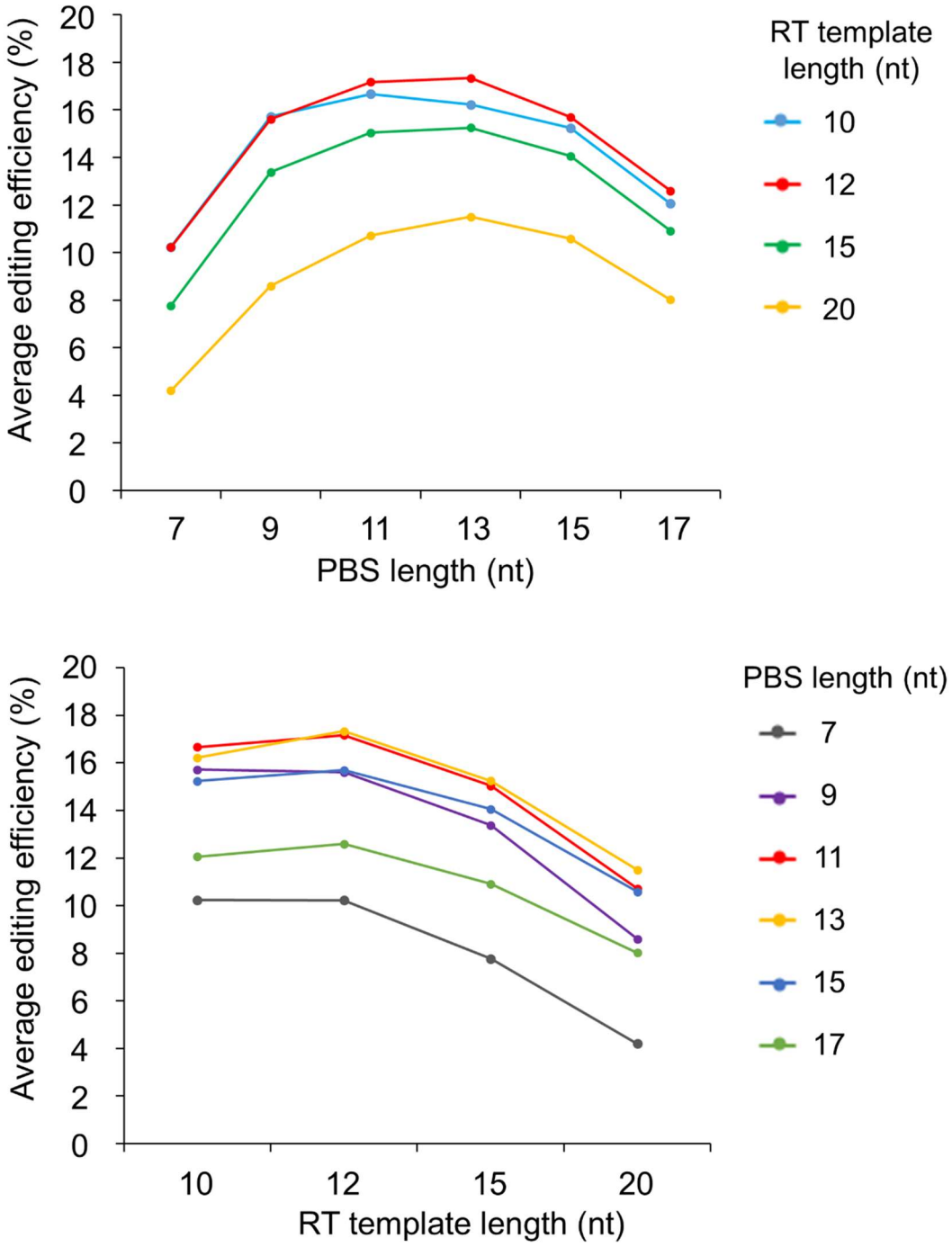

d
RT template length, 10 nt
PBS length, 7 nt
PBS length, 9 nt
PBS length, 11 nt
PBS length, 13 nt
PBS length, 15 nt
PBS length, 17 nt
Relative frequency (%)
PE2 efficiency (%)


e
RT template length, 12 nt
PBS length, 7 nt
PBS length, 9 nt
PBS length, 11 nt
PBS length, 13 nt
PBS length, 15 nt
PBS length, 17 nt
Relative frequency (%)
PE2 efficiency (%)

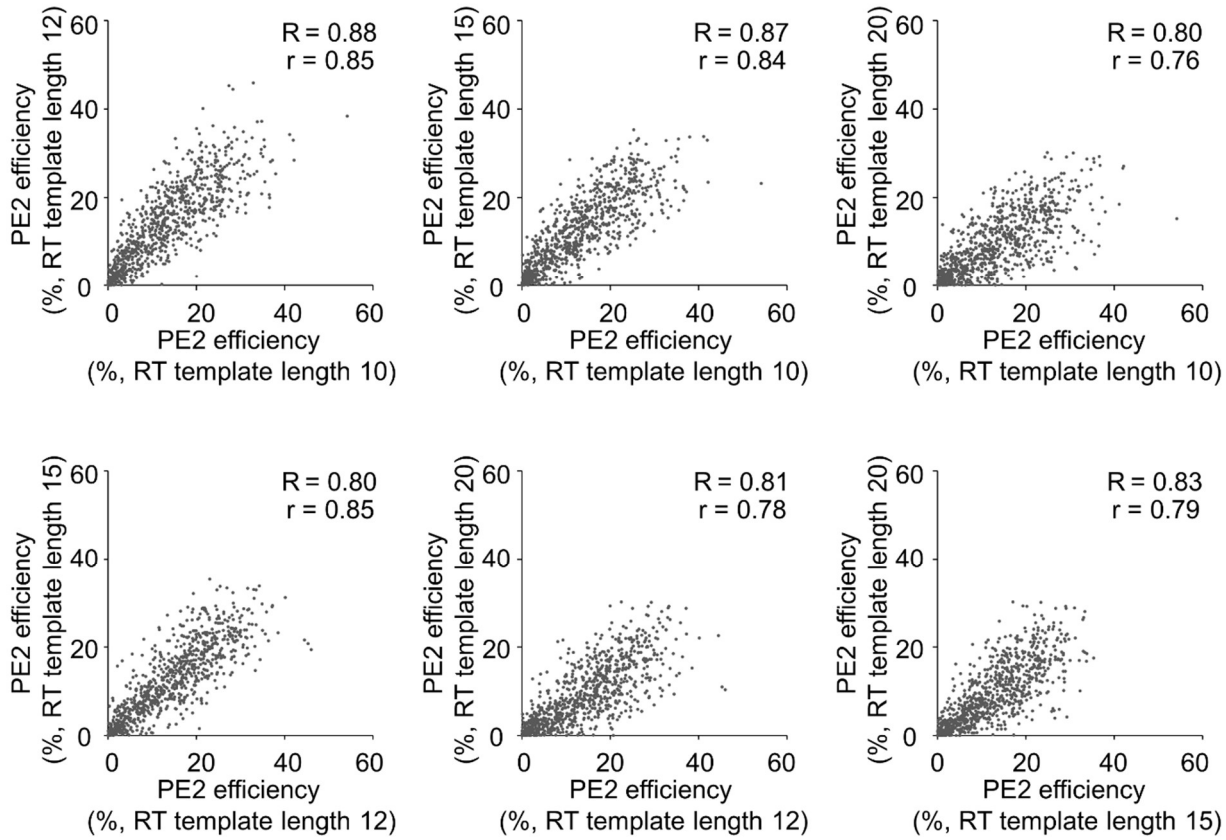

f
RT template length, 15 nt
PBS length, 7 nt
PBS length, 9 nt
PBS length, 11 nt
PBS length, 13 nt
PBS length, 15 nt
PBS length, 17 nt
Relative frequency (%)
PE2 efficiency (%)


g
RT template length, 20 nt
PBS length, 7 nt
PBS length, 9 nt
PBS length, 11 nt
PBS length, 13 nt
PBS length, 15 nt
PBS length, 17 nt
Relative frequency (%)
PE2 efficiency (%)

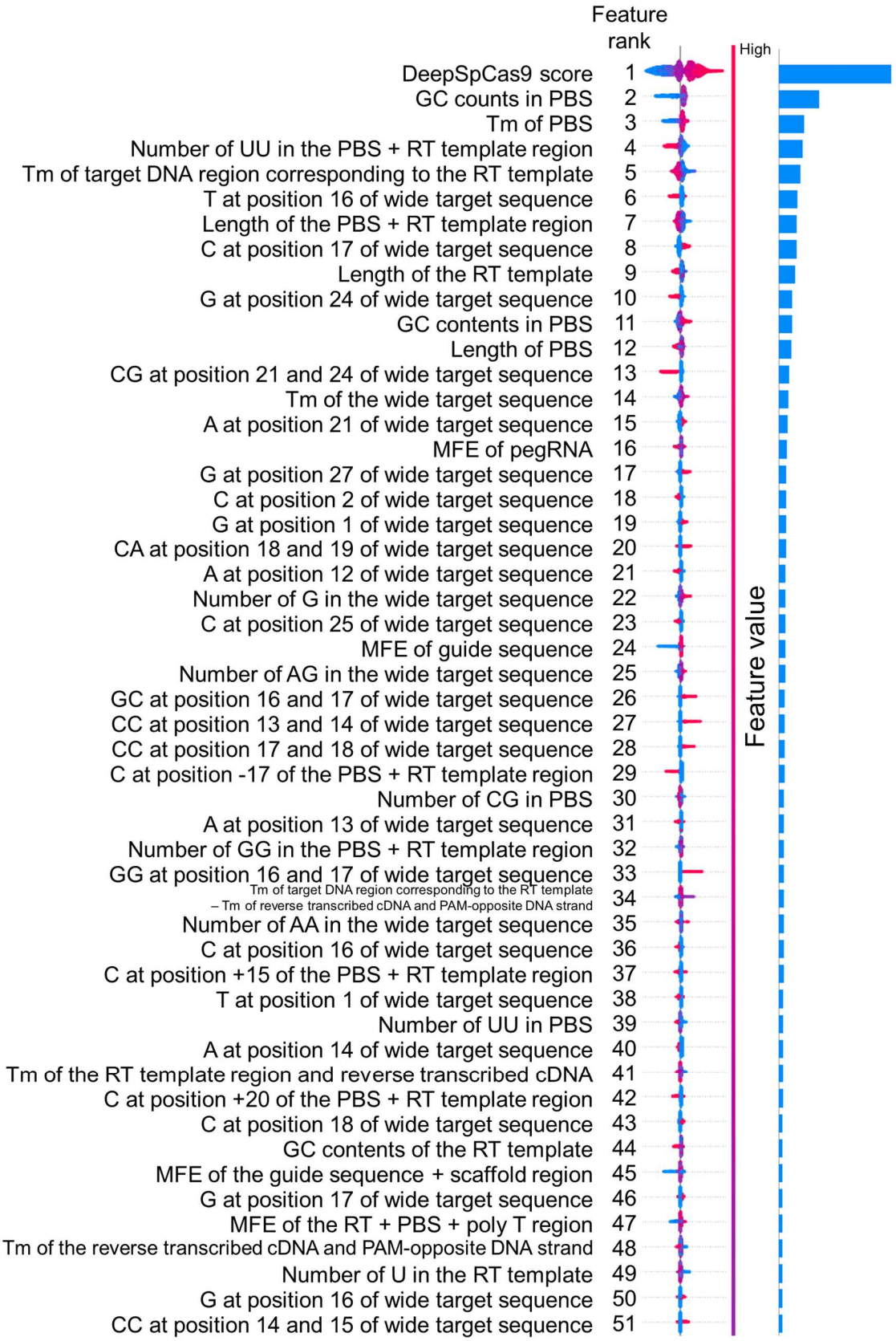
Supplementary Figure 9. Distribution of PE2 efficiencies depending on the PBS and RT template lengths. A total of 43,149 pegRNA and target sequence pairs in library 1 were used for this analysis. (**a, b**) Heat maps showing the frequencies of pegRNAs with editing efficiencies lower than 5% (a) and higher than 5% (b) for given PBS and RT template lengths. (b) is also shown in the main figure as Fig. 2b. (**c-g**) A heatmap (c) and histograms (d-g) showing the relative frequencies of pegRNAs with defined PBS lengths (7, 9, 11, 13, 15, and 17 nts) and RT template lengths (10, 12, 15, and 20 nts) associated with different PE2 efficiencies. With a 10-nt RT template, the number of pegRNA and target sequence pairs per group n = 1,809, 1,810, 1,804, 1,790, 1,779, and 1,754 for pegRNAs with a 7-, 9-, 11-, 13-, 15-, and 17-nt PBS, respectively; with a 12-nt RT template, n = 1,809, 1,826, 1,824, 1,810, 1,799, and 1,760 for pegRNAs with a 7-, 9-, 11-, 13-, 15-, and 17-nt PBS, respectively; with a 15-nt RT template, n = 1,811, 1,834, 1,840, 1,825, 1,822, and 1,789 for pegRNAs with a 7-, 9-, 11-, 13-, 15-, and 17-nt PBS, respectively; with a 20-nt RT template, n = 1,850, 1,831, 1,795, 1,772, 1,737, and 1,669 for pegRNAs with a 7-, 9-, 11-, 13-, 15-, and 17-nt PBS, respectively.

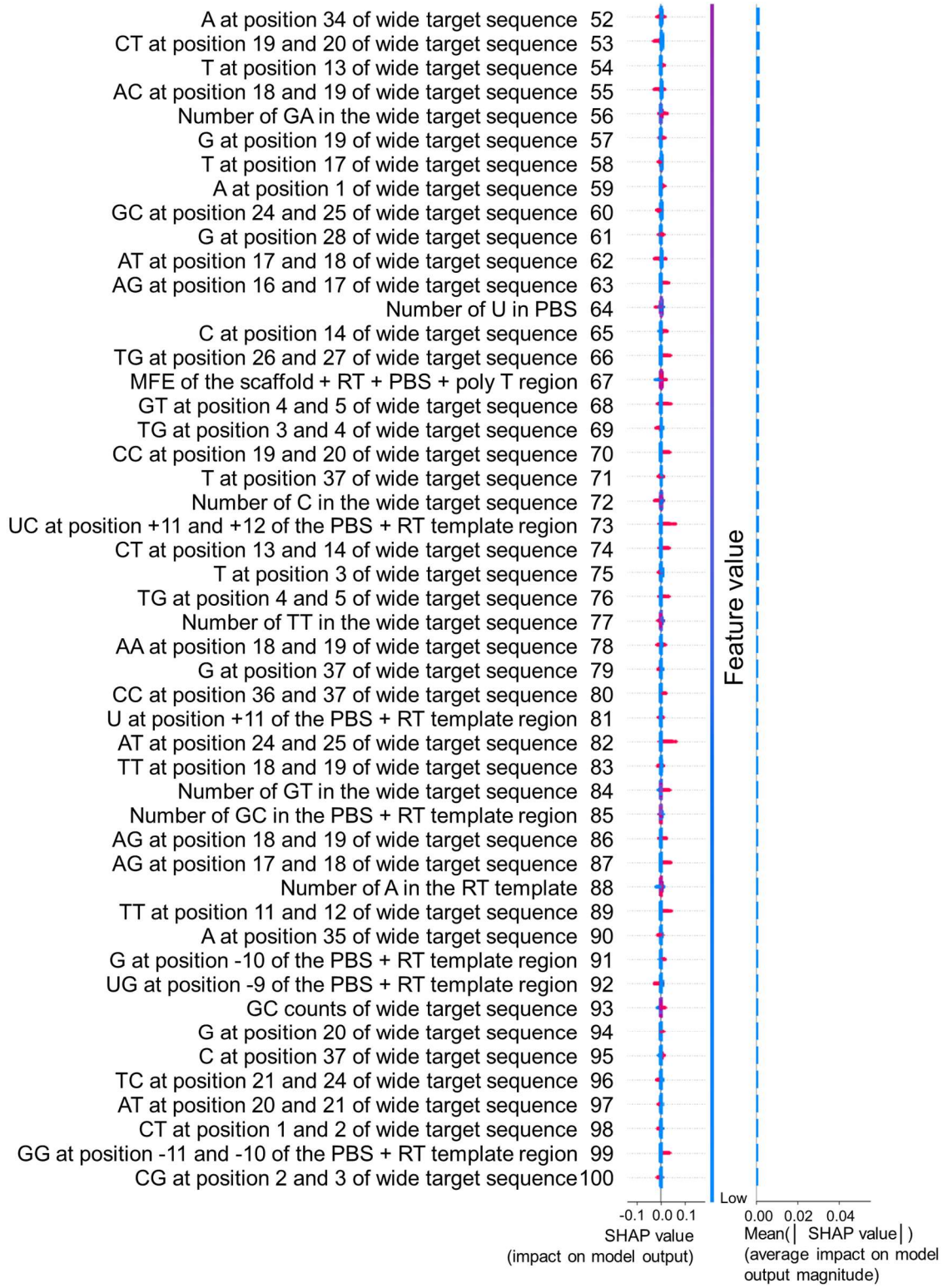


Supplementary Figure 10. The relationship between PBS lengths and RT template lengths when the most efficient combination of these two parameters per target sequence is selected. When all 24 PE2 efficiencies at a target sequence (obtained using the 24 combinations of PBS and RT template lengths) were lower than 10%, the target sequence was filtered out to minimize random errors and increase the accuracy of the comparison. These two parameters (i.e. the PBS lengths and the RT template lengths) were independent ($p = 0.25$ by a Chi-square test). The number of target sequences $n = 651$ per dot.



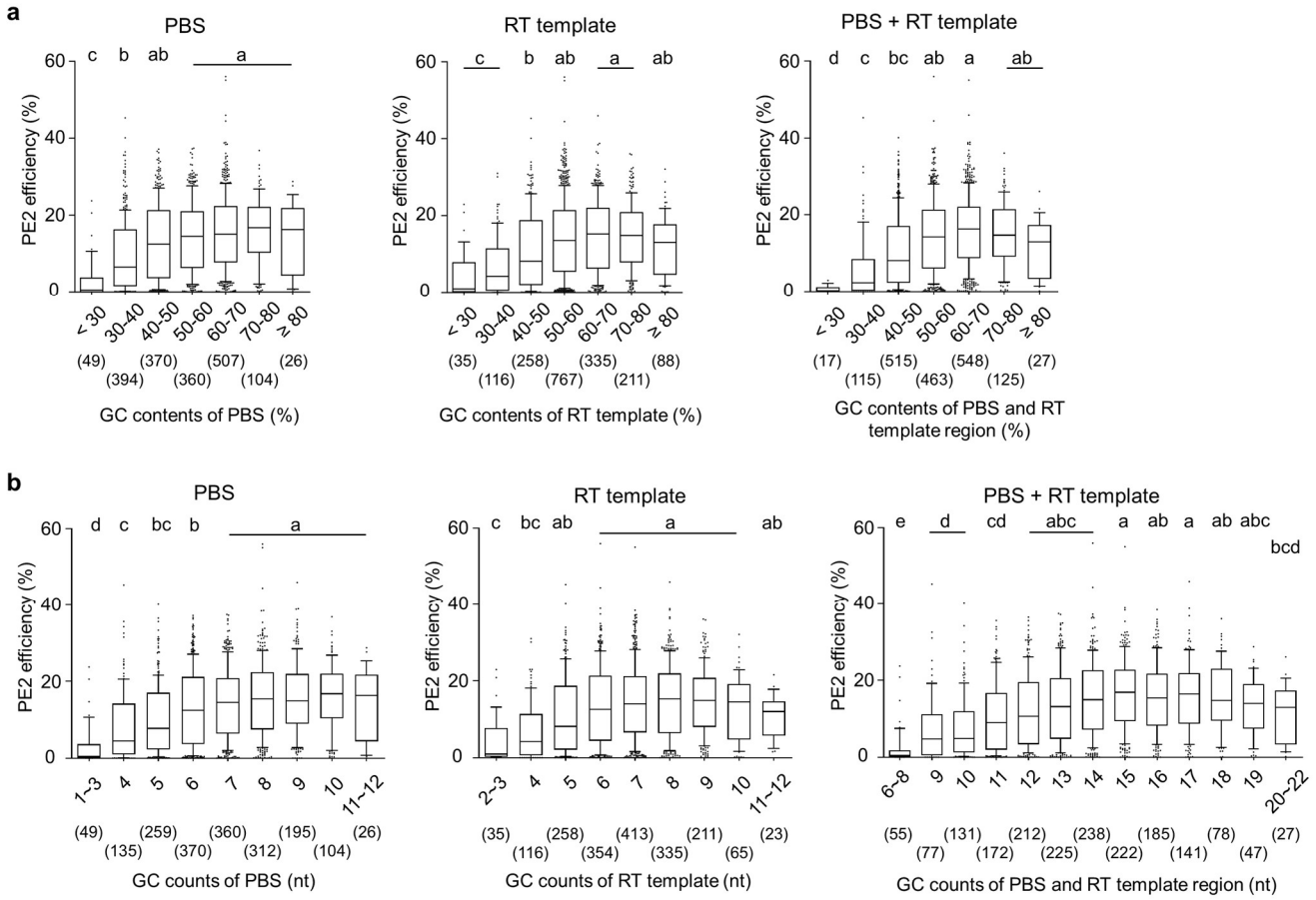
Supplementary Figure 11. Correlations between PE2 efficiencies of pegRNAs with a 13-nt PBS and identical target sequences and intended edits but with different RT template lengths (10, 12, 15, and 20 nt). The number of pegRNA and target sequence pairs n = 887. The RT template lengths are shown on the x and y axes. The Spearman (R) and Pearson (r) correlation coefficients are shown.


Feature rank
High
Feature value
DeepSpCas9 score 1
GC counts in PBS 2
Tm of PBS 3
Number of UU in the PBS + RT template region 4
Tm of target DNA region corresponding to the RT template 5
T at position 16 of wide target sequence 6
Length of the PBS + RT template region 7
C at position 17 of wide target sequence 8
Length of the RT template 9
G at position 24 of wide target sequence 10
GC contents in PBS 11
Length of PBS 12
CG at position 21 and 24 of wide target sequence 13
Tm of the wide target sequence 14
A at position 21 of wide target sequence 15
MFE of pegRNA 16
G at position 27 of wide target sequence 17
C at position 2 of wide target sequence 18
G at position 1 of wide target sequence 19
CA at position 18 and 19 of wide target sequence 20
A at position 12 of wide target sequence 21
Number of G in the wide target sequence 22
C at position 25 of wide target sequence 23
MFE of guide sequence 24
Number of AG in the wide target sequence 25
GC at position 16 and 17 of wide target sequence 26
CC at position 13 and 14 of wide target sequence 27
CC at position 17 and 18 of wide target sequence 28
C at position -17 of the PBS + RT template region 29
Number of CG in PBS 30
A at position 13 of wide target sequence 31
Number of GG in the PBS + RT template region 32
GG at position 16 and 17 of wide target sequence 33
Tm of target DNA region corresponding to the RT template – Tm of reverse transcribed cDNA and PAM-opposite DNA strand 34
Number of AA in the wide target sequence 35
C at position 16 of wide target sequence 36
C at position +15 of the PBS + RT template region 37
T at position 1 of wide target sequence 38
Number of UU in PBS 39
A at position 14 of wide target sequence 40
Tm of the RT template region and reverse transcribed cDNA 41
C at position +20 of the PBS + RT template region 42
C at position 18 of wide target sequence 43
GC contents of the RT template 44
MFE of the guide sequence + scaffold region 45
G at position 17 of wide target sequence 46
MFE of the RT + PBS + poly T region 47
Tm of the reverse transcribed cDNA and PAM-opposite DNA strand 48
Number of U in the RT template 49
G at position 16 of wide target sequence 50
CC at position 14 and 15 of wide target sequence 51

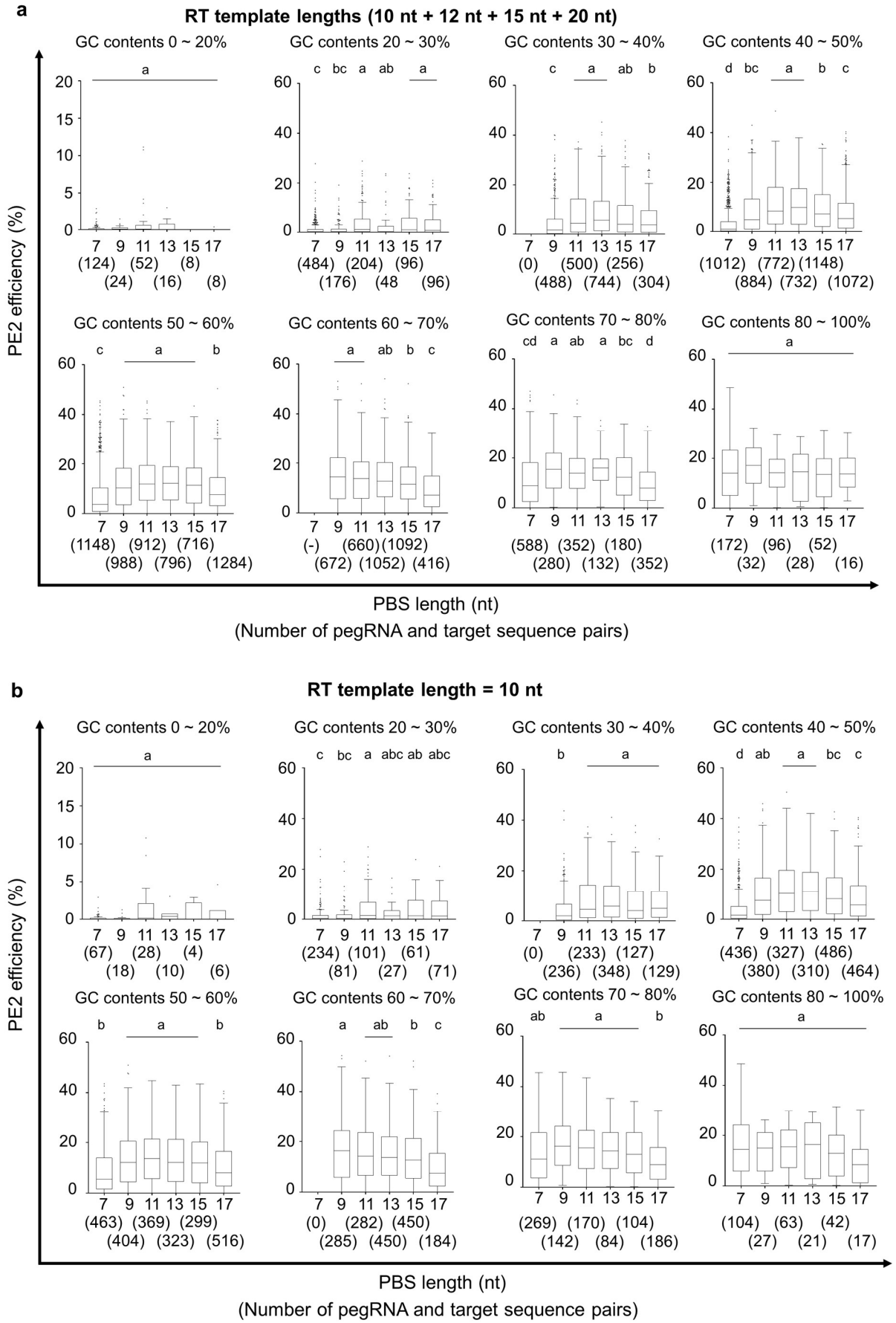


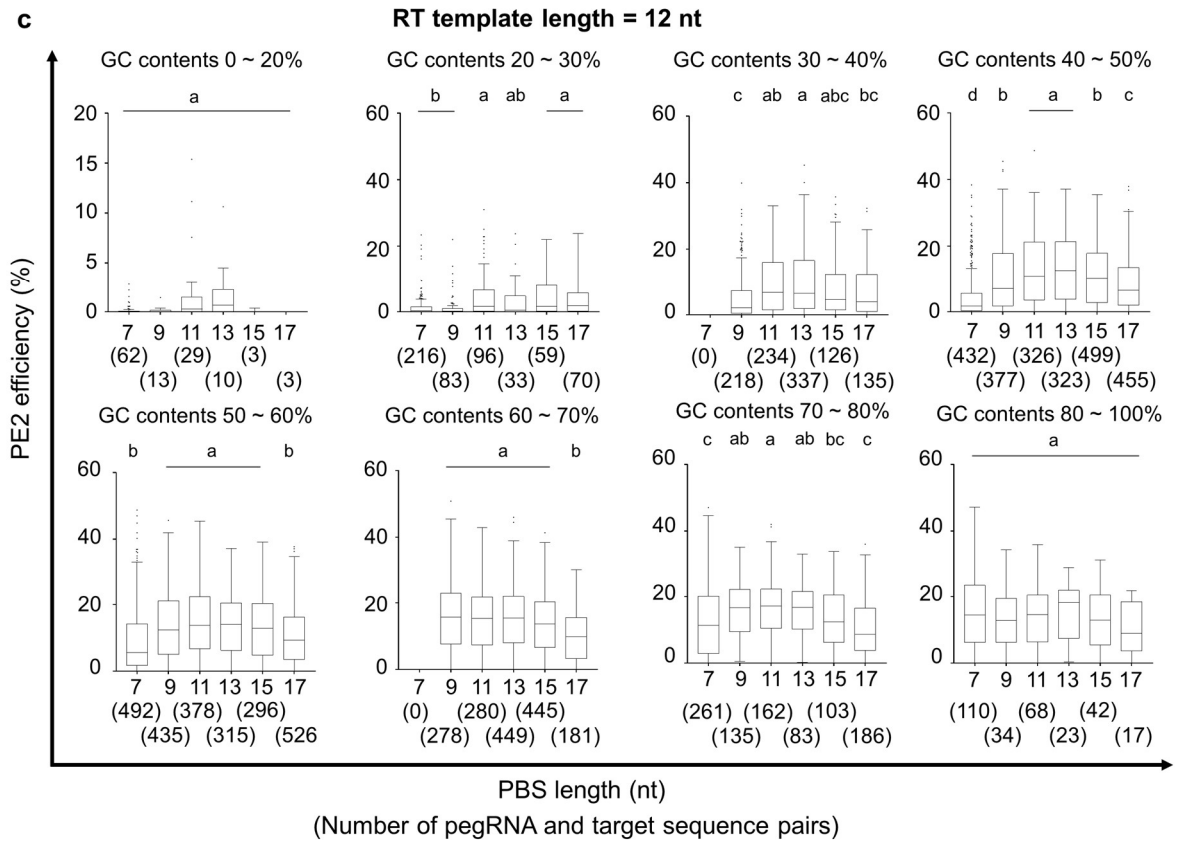
Supplementary Figure 12. The 100 most important features associated with prime editing efficiency determined by Tree SHAP (XGBoost classifier). On the summary violin plot (the left graph), each target sequence is represented by a dot; the position of the dot on the x-axis shows its SHAP value. High and 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

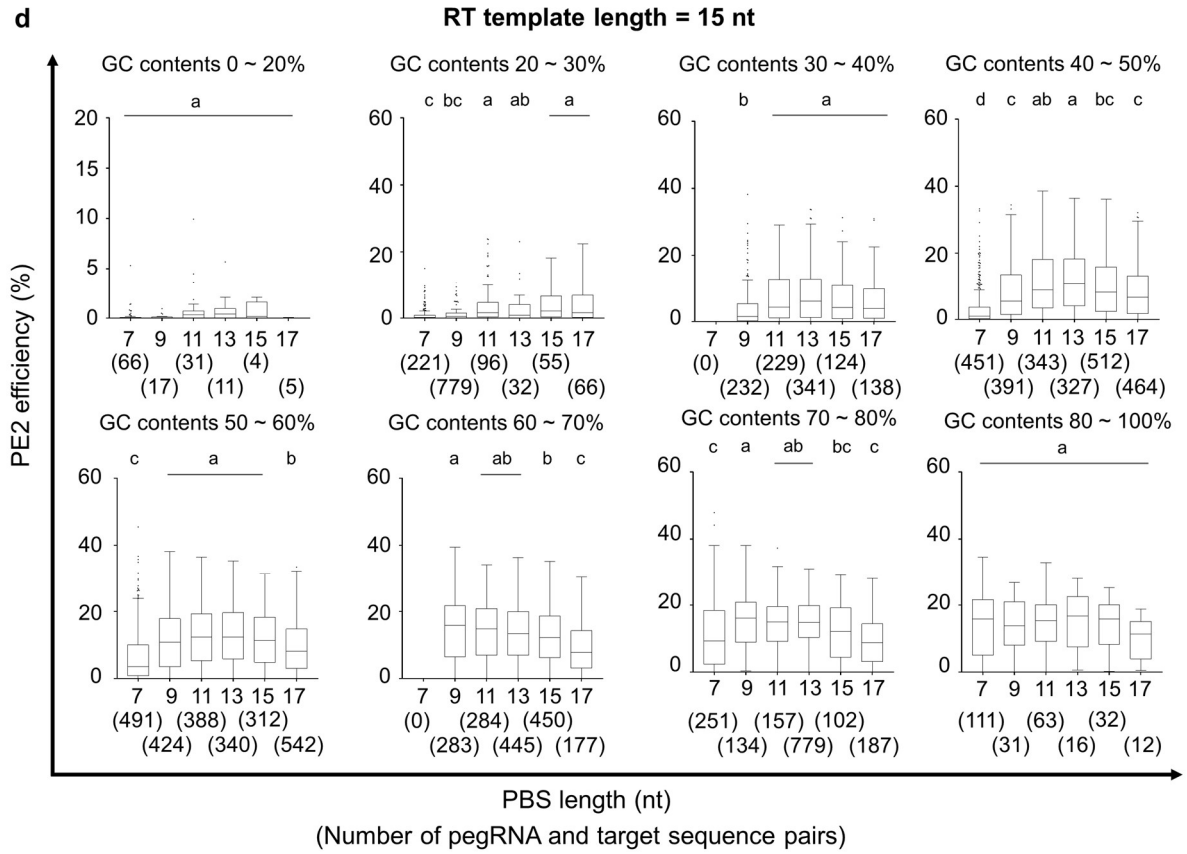
values of the relevant feature as shown in the figure. Overlapping points are slightly separated in the y-axis direction so that the density is apparent. Examples of the summary plot interpretations are included as Supplementary text 1. The 10 most important features are also shown in Fig. 2e. Tm, melting temperature; MFE, minimum self-folding free energy.

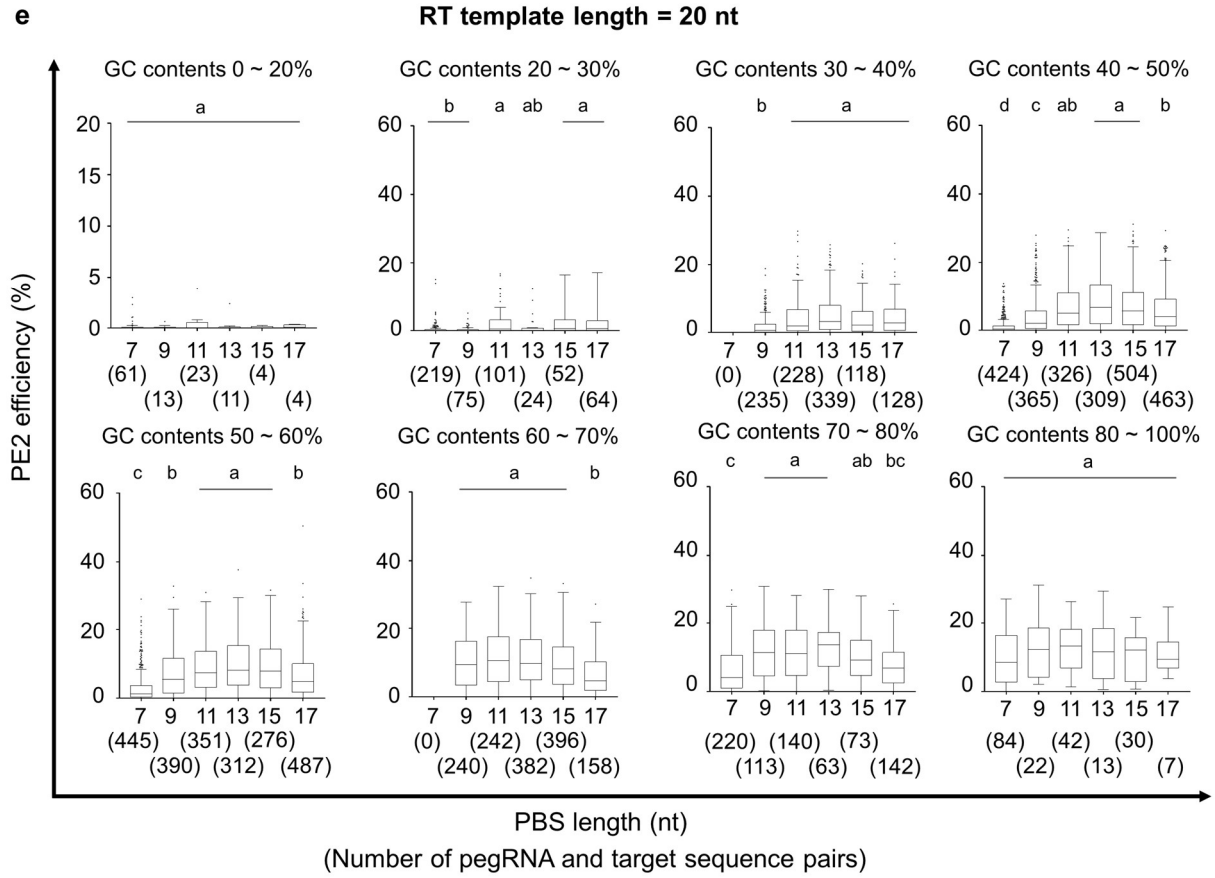


Supplementary Figure 13. Effect of GC contents and GC counts in PBS and the RT template on prime editing (PE) efficiency. The pegRNAs were classified into groups depending on the GC contents (a) and GC counts (b) in the PBS and RT template regions. The PE efficiencies of these groups were compared. The lengths of PBS and the RT template were 13 nt and 12 nt, respectively, across groups. Subsets of experimental groups without statistically significant differences in PE efficiencies are represented with letters such as a, b, c, d, and e. P = 0.41, 0.09, and 1.0 for subsets a, b, and c, respectively, for groups with various GC contents in the PBS; P = 0.35, 0.16, and 0.33 for subsets a, b, and c, respectively, for groups with various GC contents in the RT template; P = 0.08, 0.15, 0.10, and 1.0 for subsets a, b, c, and d, respectively, for groups with various GC contents in the PBS and RT template region; P = 0.64, 0.07, 0.64, and 1.0 for subsets a, b, c, and d, respectively, for groups with various GC counts in the PBS; P = 0.09, 0.09, and 0.69 for subsets a, b, and c, respectively, for groups with various GC counts in the RT template; P = 0.08, 0.08, 0.06, 0.14, and 1.0 for subsets a, b, c, d, and e, respectively, for groups with various GC counts in the PBS and RT template region; ANOVA followed by two-sided Tukey's post hoc 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. The numbers of pegRNA and target sequence pairs per experimental group (n) are specified on the x-axis within the parentheses.

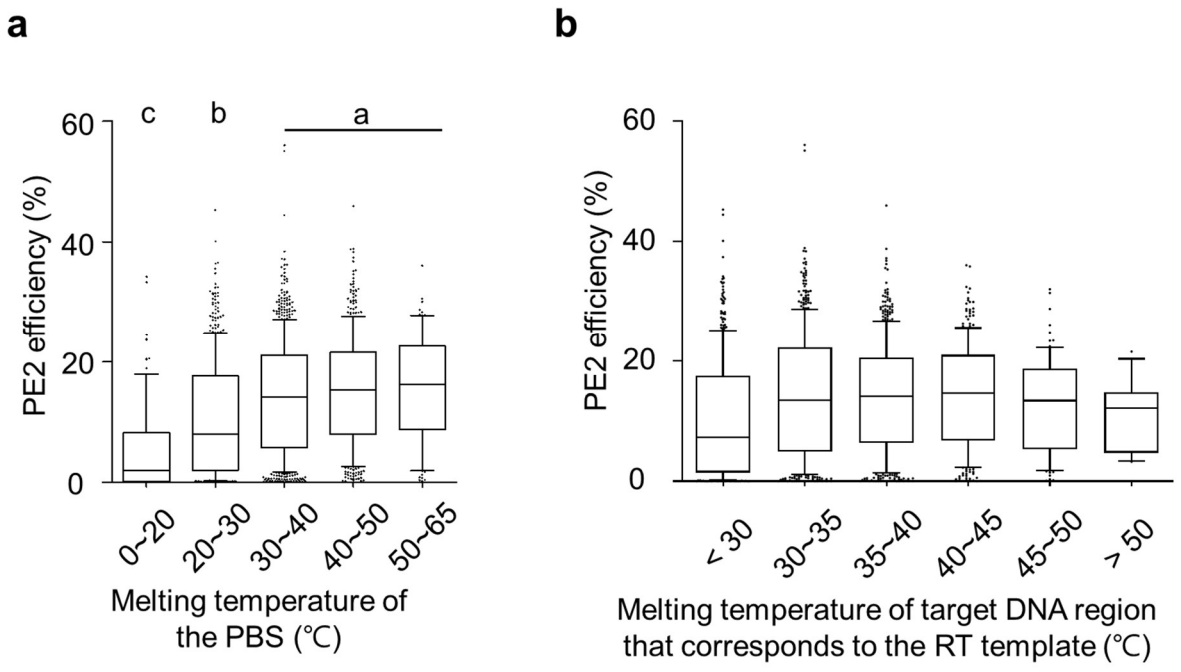

a
RT template lengths (10 nt + 12 nt + 15 nt + 20 nt)
GC contents 0 ~ 20%
GC contents 20 ~ 30%
GC contents 30 ~ 40%
GC contents 40 ~ 50%
GC contents 50 ~ 60%
GC contents 60 ~ 70%
GC contents 70 ~ 80%
GC contents 80 ~ 100%
PE2 efficiency (%)
PBS length (nt)
(Number of pegRNA and target sequence pairs)
b
RT template length = 10 nt
GC contents 0 ~ 20%
GC contents 20 ~ 30%
GC contents 30 ~ 40%
GC contents 40 ~ 50%
GC contents 50 ~ 60%
GC contents 60 ~ 70%
GC contents 70 ~ 80%
GC contents 80 ~ 100%
PE2 efficiency (%)
PBS length (nt)
(Number of pegRNA and target sequence pairs)


c
RT template length = 12 nt
GC contents 0 ~ 20%
GC contents 20 ~ 30%
GC contents 30 ~ 40%
GC contents 40 ~ 50%
GC contents 50 ~ 60%
GC contents 60 ~ 70%
GC contents 70 ~ 80%
GC contents 80 ~ 100%
PE2 efficiency (%)
PBS length (nt)
(Number of pegRNA and target sequence pairs)

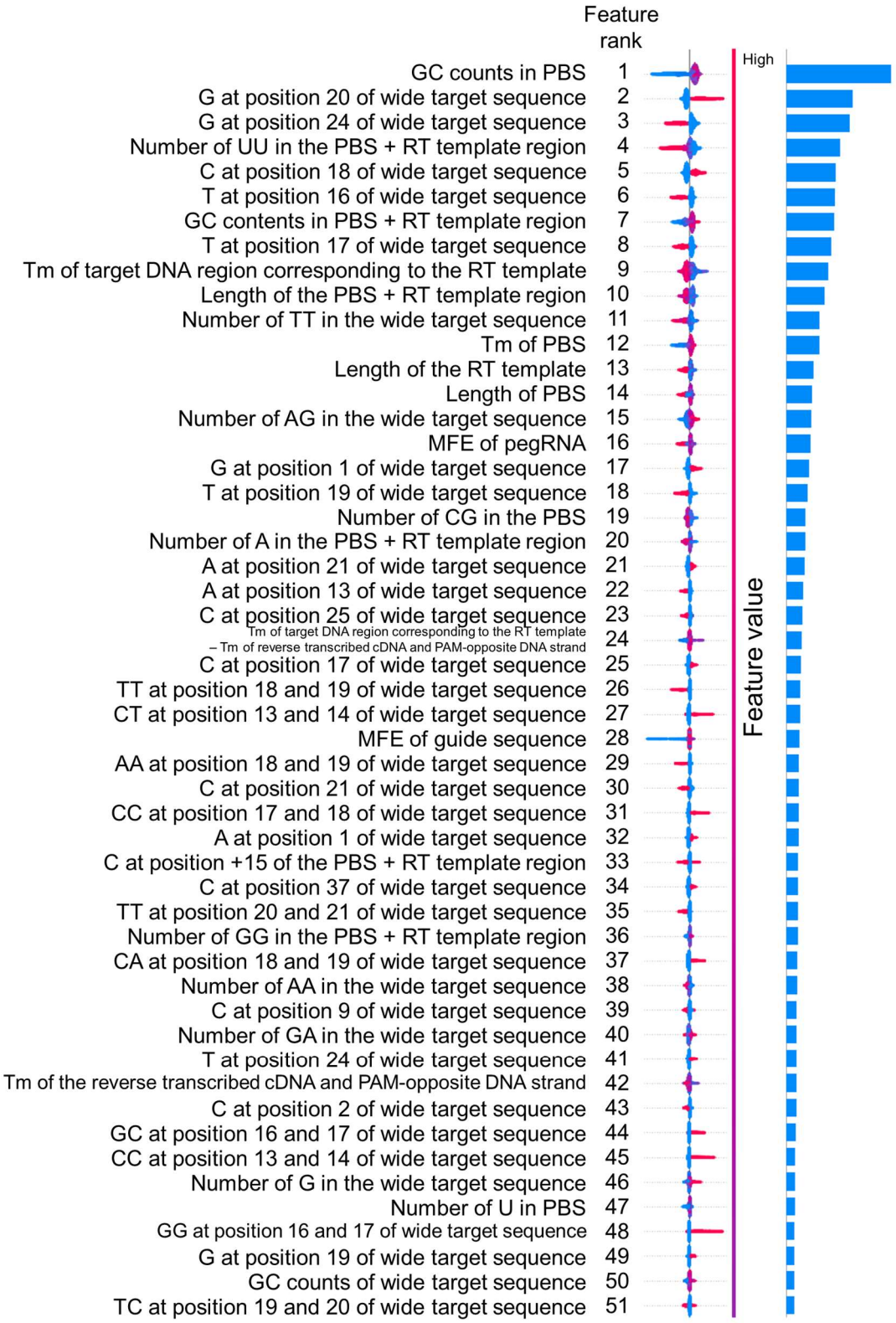

d
RT template length = 15 nt
GC contents 0 ~ 20%
GC contents 20 ~ 30%
GC contents 30 ~ 40%
GC contents 40 ~ 50%
GC contents 50 ~ 60%
GC contents 60 ~ 70%
GC contents 70 ~ 80%
GC contents 80 ~ 100%
PE2 efficiency (%)
PBS length (nt)
(Number of pegRNA and target sequence pairs)



Supplementary Figure 14. Effect of GC contents in PBS and PBS lengths on PE2 efficiency. pegRNAs were classified into groups depending on their GC contents and PBS lengths. The PE2 efficiencies of these groups were compared under all tested RT template lengths (a) or the same RT template lengths (b – e). Subsets of experimental groups without statistically significant (P < 0.05, ANOVA followed by Tukey's post hoc test) differences in PE efficiencies are represented with letters such as a, b, and c in the order of the average PE2 efficiency (exact p-values are listed in Supplementary Table 7). 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. The numbers of pegRNA and target sequence pairs per experimental group (n) are specified on the x-axis within the parentheses.

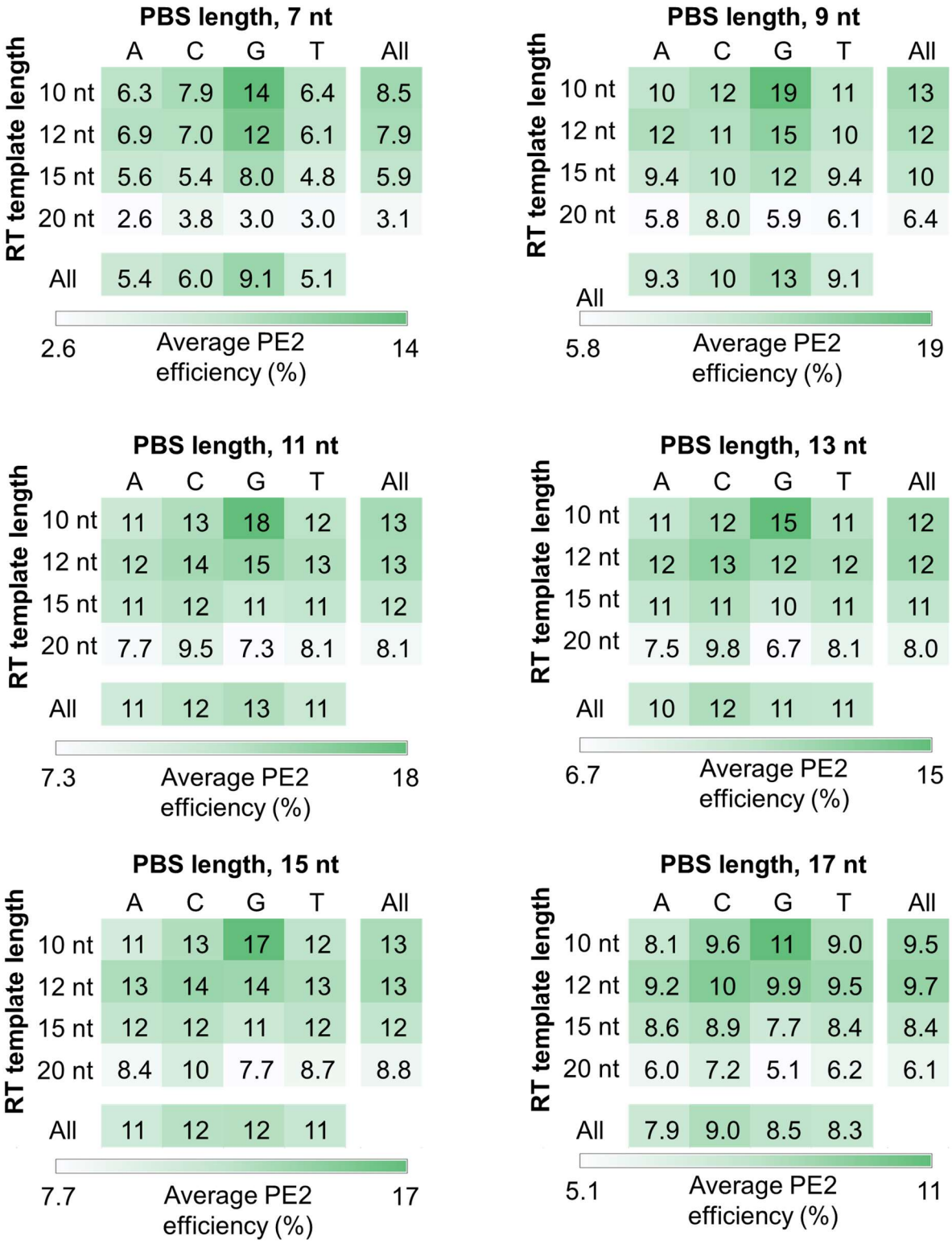


Supplementary Figure 15. Effect of the melting temperatures of PBS and the target DNA region that corresponds to the RT template on prime editing (PE) efficiency. The pegRNAs were classified into groups depending the melting temperatures of the PBS (a) and the target sequence region corresponding to the RT template (b). The PE efficiencies of these groups were compared. The lengths of PBS and the RT template were 13 nt and 12 nt, respectively, across groups. Subsets of experimental groups without statistically significant differences in PE efficiencies are represented with letters such as a, b, and c. P = 0.65, 1.0, and 1.0 for subsets a, b, and c, respectively, for groups with various PBS melting temperatures; ANOVA followed by two-sided Tukey's post hoc 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. The number of pegRNA and target sequence pairs per experimental group n = 76, 498, 736, 420, and 80 for pegRNAs with PBS melting temperatures of 0 – 20, 20 – 30, 30 – 40, 40 – 50, and 50 – 65 ˚C, respectively; n = 429, 497, 518, 271, 82, and 13 for target sequences with melting temperatures of the target DNA region that corresponds to the RT template of < 30, 30 – 35, 35 – 40, 40 – 45, 45 – 50, and > 50 ˚C, respectively.


Feature rank
GC counts in PBS 1
G at position 20 of wide target sequence 2
G at position 24 of wide target sequence 3
Number of UU in the PBS + RT template region 4
C at position 18 of wide target sequence 5
T at position 16 of wide target sequence 6
GC contents in PBS + RT template region 7
T at position 17 of wide target sequence 8
Tm of target DNA region corresponding to the RT template 9
Length of the PBS + RT template region 10
Number of TT in the wide target sequence 11
Tm of PBS 12
Length of the RT template 13
Length of PBS 14
Number of AG in the wide target sequence 15
MFE of pegRNA 16
G at position 1 of wide target sequence 17
T at position 19 of wide target sequence 18
Number of CG in the PBS 19
Number of A in the PBS + RT template region 20
A at position 21 of wide target sequence 21
A at position 13 of wide target sequence 22
C at position 25 of wide target sequence 23
Tm of target DNA region corresponding to the RT template – Tm of reverse transcribed cDNA and PAM-opposite DNA strand 24
C at position 17 of wide target sequence 25
TT at position 18 and 19 of wide target sequence 26
CT at position 13 and 14 of wide target sequence 27
MFE of guide sequence 28
AA at position 18 and 19 of wide target sequence 29
C at position 21 of wide target sequence 30
CC at position 17 and 18 of wide target sequence 31
A at position 1 of wide target sequence 32
C at position +15 of the PBS + RT template region 33
C at position 37 of wide target sequence 34
TT at position 20 and 21 of wide target sequence 35
Number of GG in the PBS + RT template region 36
CA at position 18 and 19 of wide target sequence 37
Number of AA in the wide target sequence 38
C at position 9 of wide target sequence 39
Number of GA in the wide target sequence 40
T at position 24 of wide target sequence 41
Tm of the reverse transcribed cDNA and PAM-opposite DNA strand 42
C at position 2 of wide target sequence 43
GC at position 16 and 17 of wide target sequence 44
CC at position 13 and 14 of wide target sequence 45
Number of G in the wide target sequence 46
Number of U in PBS 47
GG at position 16 and 17 of wide target sequence 48
G at position 19 of wide target sequence 49
GC counts of wide target sequence 50
TC at position 19 and 20 of wide target sequence 51
High
Feature value

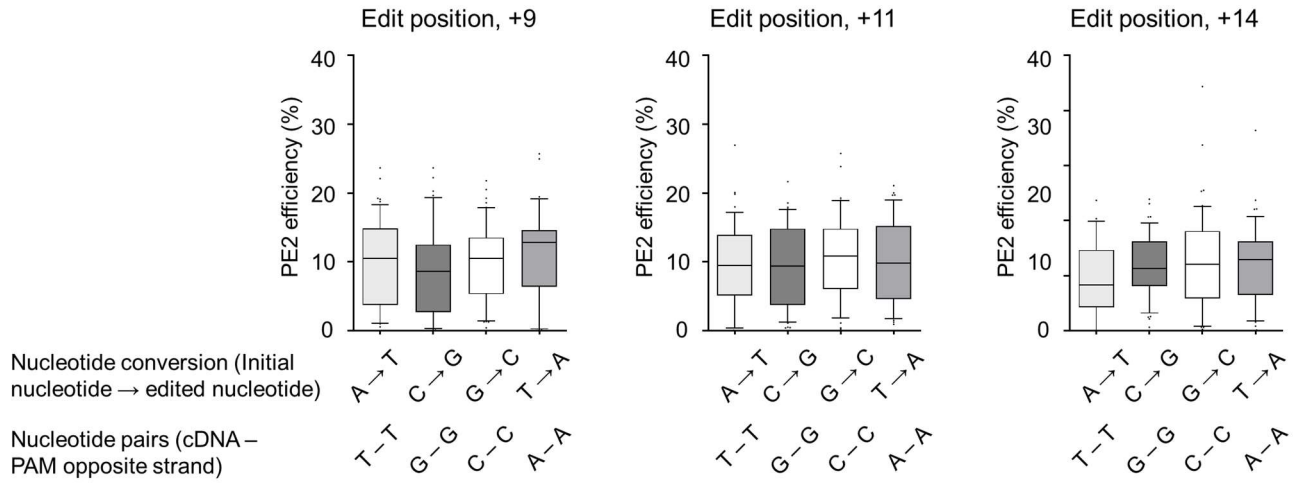
Supplementary Figure 16. The 100 most important features associated with prime editing efficiency determined by Tree SHAP (XGBoost classifier trained without the DeepSpCas9 score). On the summary violin plot (the left graph), each target sequence is represented by a dot; the position of the dot on the 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. Tm, melting temperature; MFE, minimum self-folding free energy.

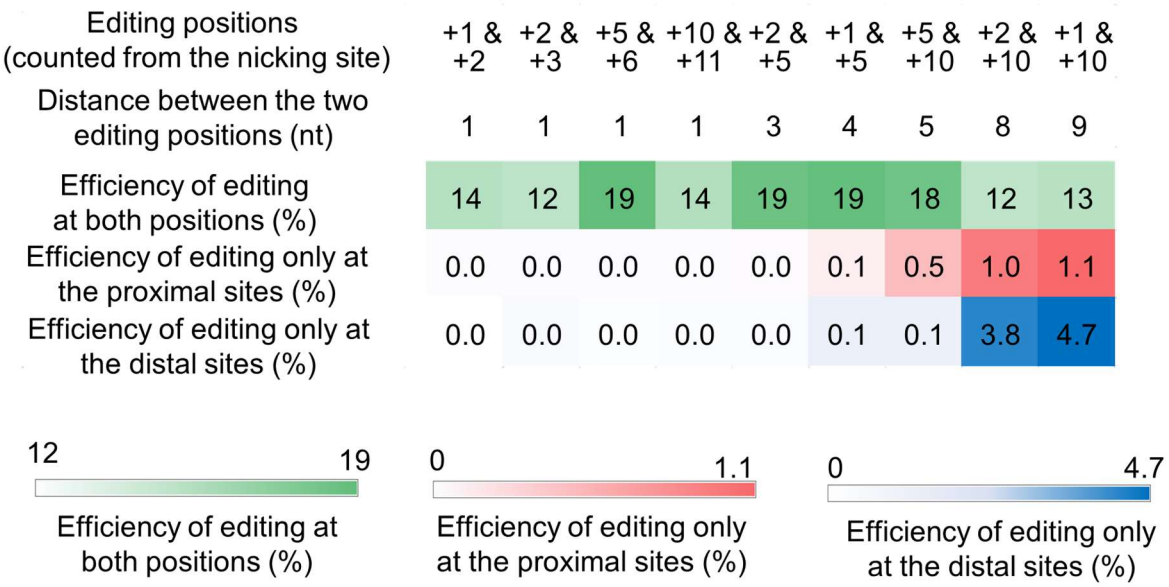


Supplementary Figure 17. Effect of the last templated nucleotide on PE2 efficiency. A total of 21,288 pegRNAs (887 target sequences x 24 combinations of PBS and RT template lengths) were grouped depending on their last templated nucleotide. The average editing efficiencies for groups with various last templated nucleotides and combinations of PBS and RT template lengths are shown as heat maps. When the last

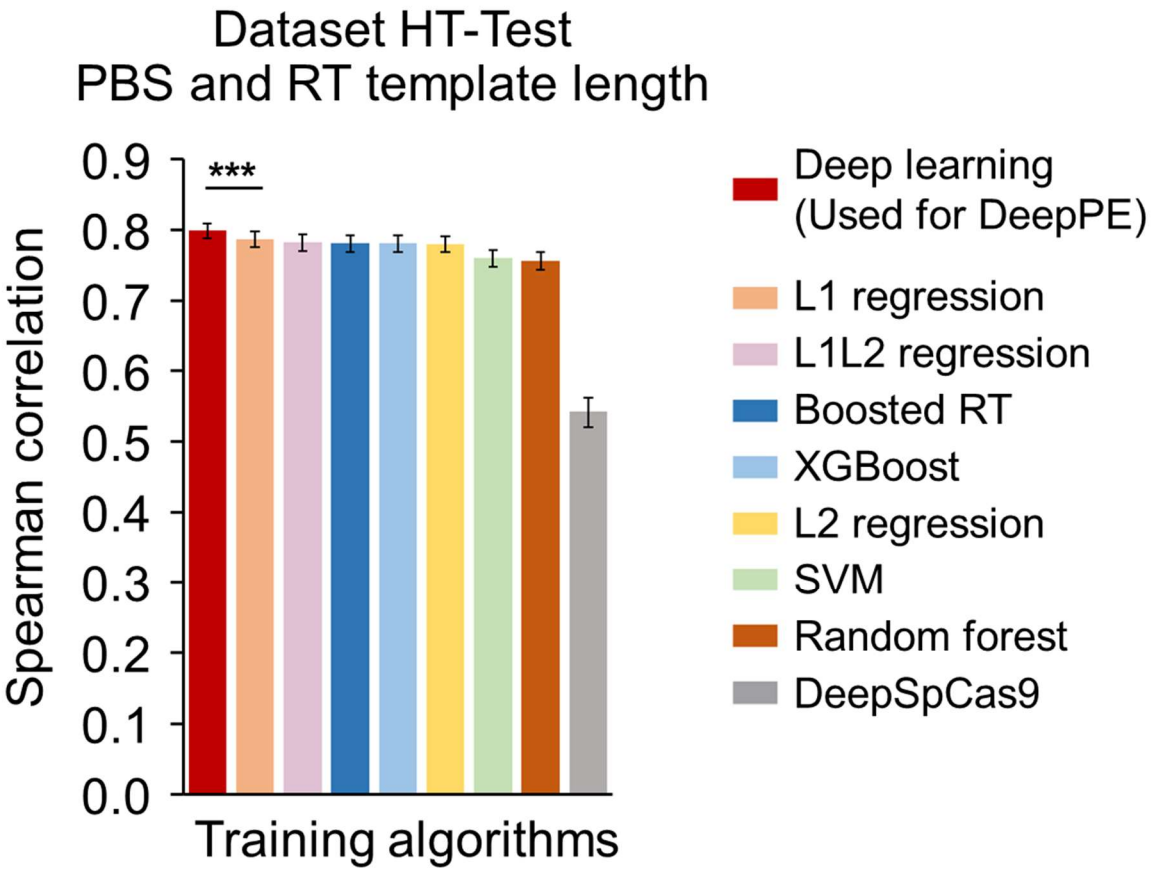
templated nucleotide was an A, C, G, or T, respectively, with a 10-nt RT template, the number of pegRNA and target sequence pairs per group n = 1,500, 1,452, 708, and 1,662; with a 12-nt RT template, n = 1,476, 1,494, 1,110, and 1,242; with a 15-nt RT template, n = 1,224, 1,536, 1,146, and 1,416; with a 20-nt RT template, n = 1,290, 1,560, 1,128, and 1,344.



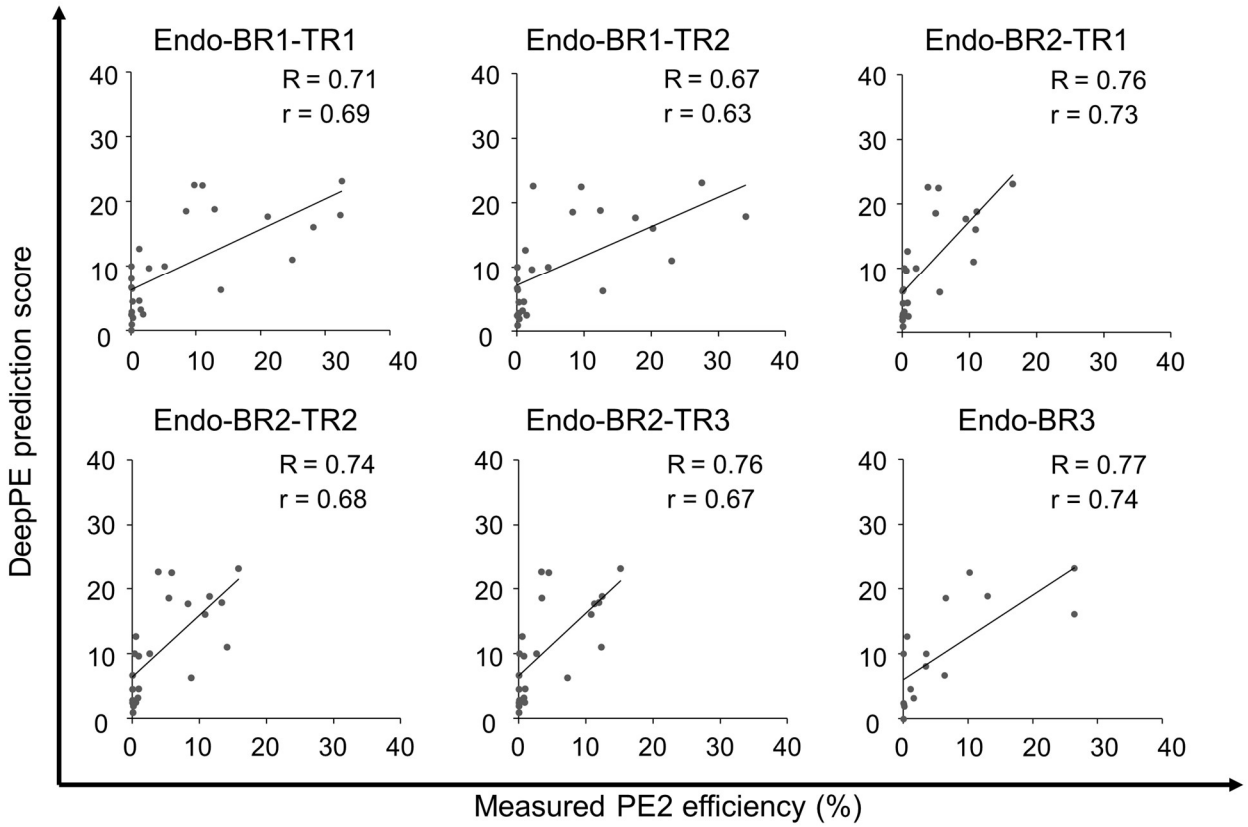
Supplementary Figure 18. Effect of the type of substitutions on prime editing (PE) efficiency. Prime editing was targeted to positions +9 (left), +11 (middle), and +14 (right) from the nicking site. In the boxes, the top, middle, and bottom lines represent the 25th, 50th, and 75th percentiles, respectively, whiskers indicate the 10th and 90th percentiles, and outliers are shown as individual dots. The number of pegRNA and target sequence pairs n = 52, 40, 50, and 35 for A to T, C to G, G to C, and T to A conversions (left), n = 49, 44, 43, and 42 for A to T, C to G, G to C, and T to A conversions (middle), and n = 29, 46, 51, and 47 for A to T, C to G, G to C, and T to A conversions (right).



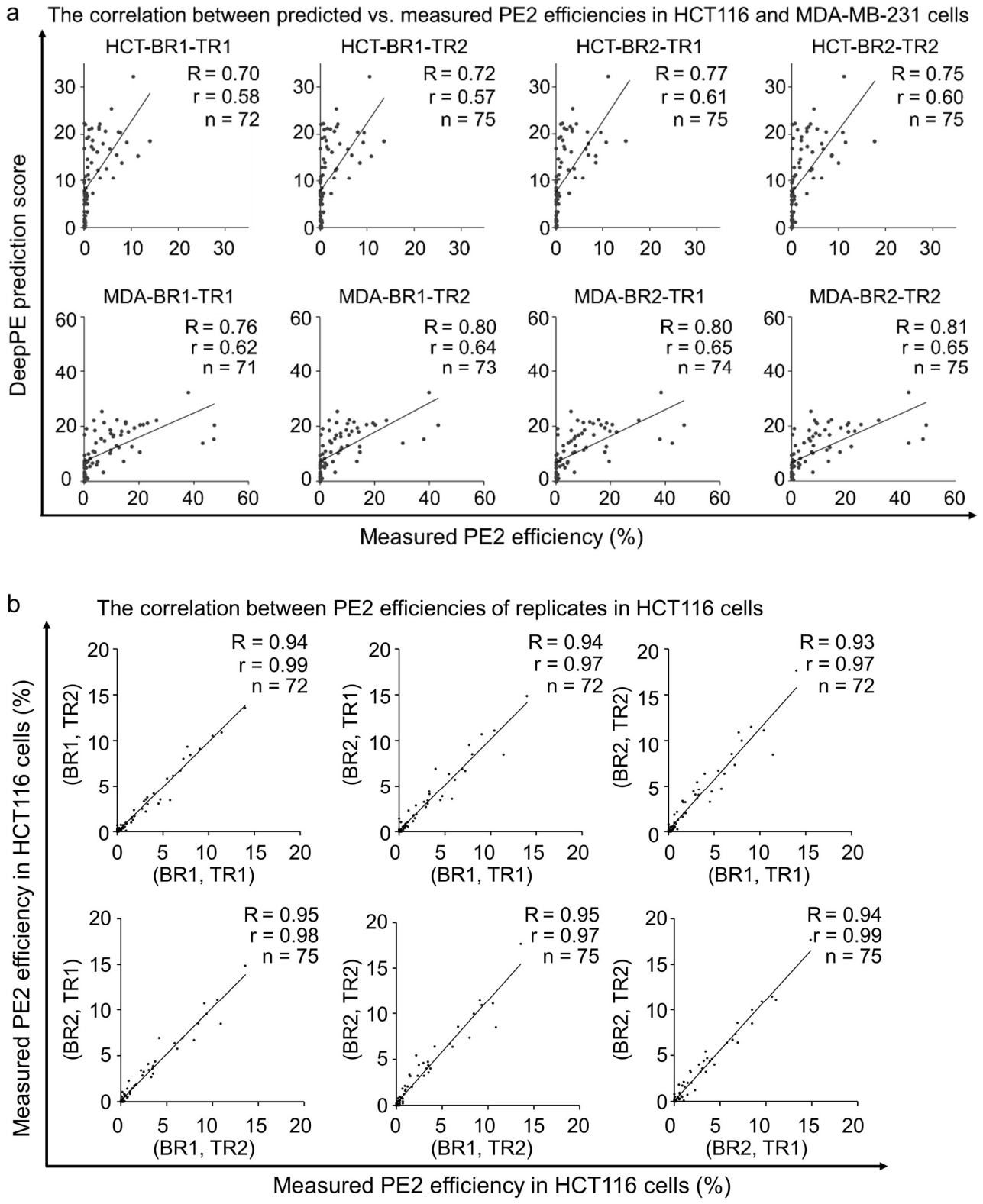
Supplementary Figure 19. Analysis of prime editing when two nucleotides are targeted for substitutions. Heatmap showing the average frequencies of partial (1 nt) and complete (2 nts) editing. The number of pegRNA and target sequence pairs n = 190, 181, 186, 190, 177, 180, 183, 170, and 169 for position +1 & +2, +1 & +5, +1 & +10, +2 & +3, +2 & +5, +2 & +10, +5 & +6, +5 & +10, and +10 & +11, respectively.

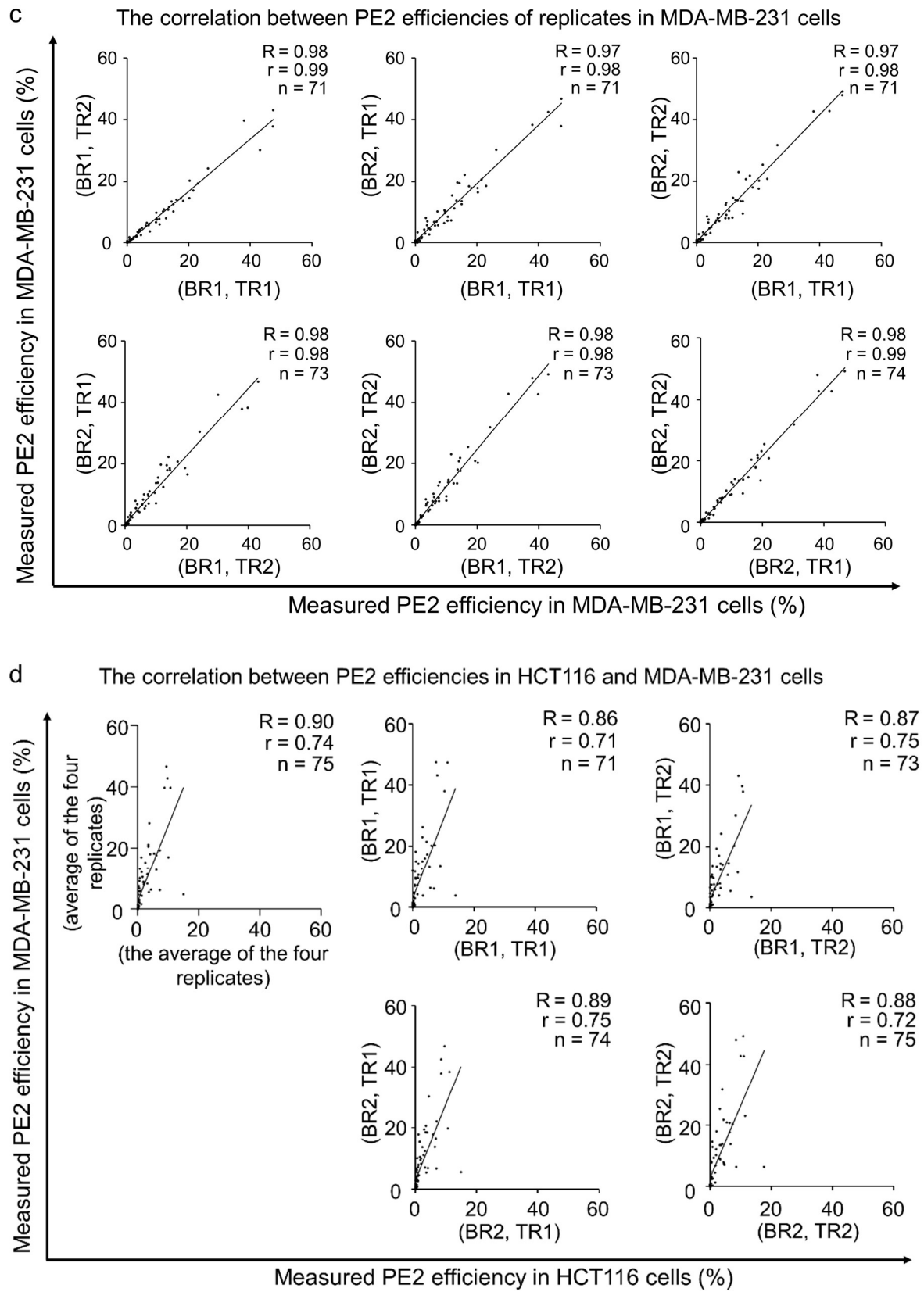


Supplementary Figure 20. Performance comparison of DeepPE with other prediction models using dataset HT-Test. The bar graph shows Spearman correlations between measured PE2 efficiencies and predicted activity scores. For the sake of clarity, results from statistical significance testing are shown only for DeepPE versus the next-best model (***P = 9.4 x 10^{-4}; two-sided Steiger's test). Error bars represent 95% confidence intervals and the number of pegRNA and target sequence pairs used for the analyses n = 4,457.



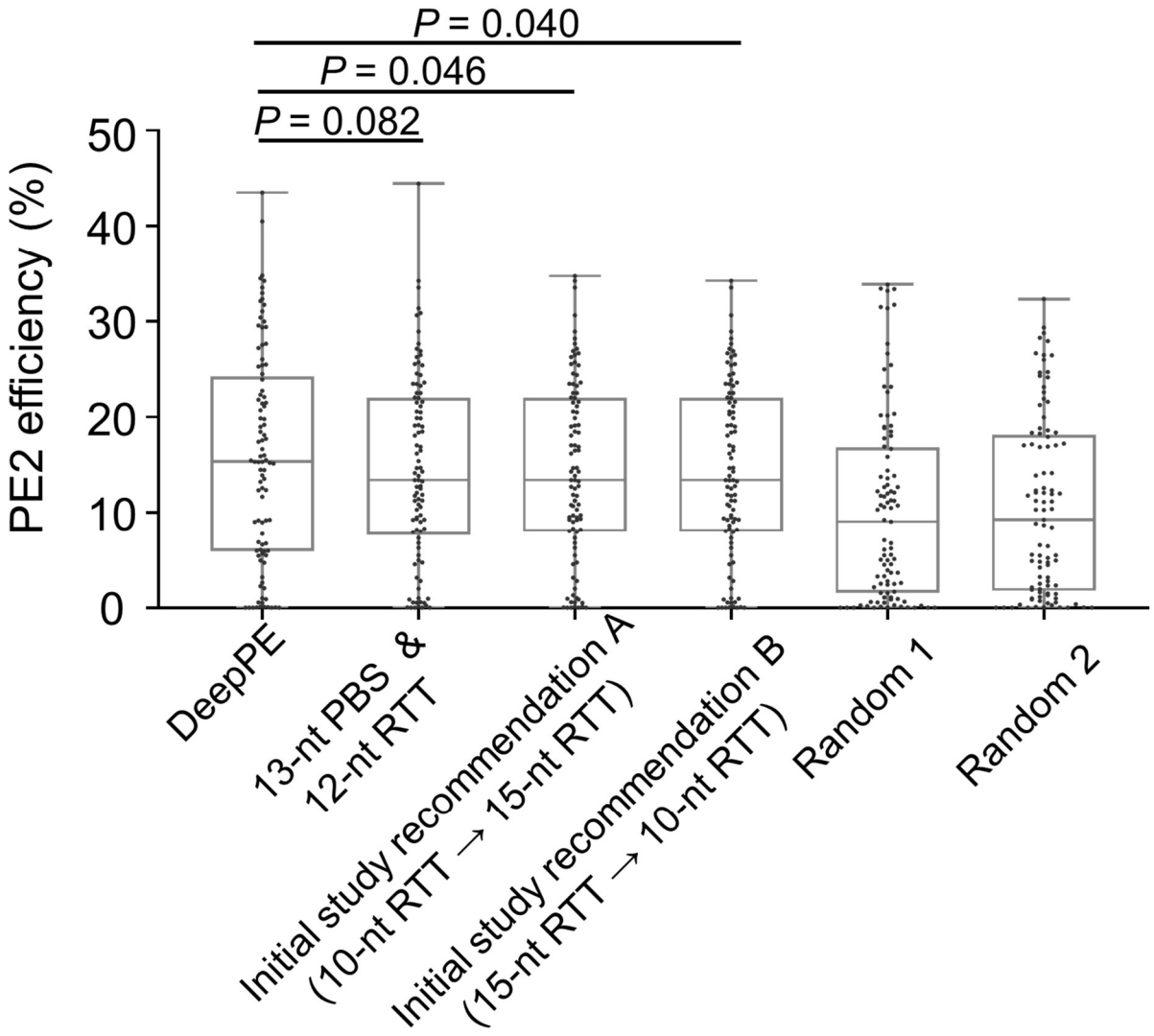
Supplementary Figure 21. Evaluation of DeepPE using six datasets obtained by measuring the PE2 efficiencies at endogenous sites after transient transfection of plasmids encoding PE2 and pegRNA into HEK293T cells. Three biological replicates (BR1, BR2, and BR3) were evaluated and each biological replicate had one, two, or three technical replicates (TRs). The number of target sequences n = 26, 25, 23, 23, 23, and 17 for datasets Endo-BR1-TR1, Endo-BR1-TR2, Endo-BR2-TR1, Endo-BR2-TR2, Endo-BR2-TR3, and Endo-BR3, respectively. In Endo-BR3, 15 pegRNAs randomly chosen from a total of 26 pegRNAs, and two pegRNAs that were not tested in the other replicates, were evaluated. In Endo-BR1-TR2, Endo-BR2-TR-1, Endo-BR2-TR2, and Endo-BR2-TR3, one or three pegRNAs were removed from the analysis due to transfection failure or an insufficient deep sequencing read count (less than 200). The results using Endo-BR1-TR1 are also shown in Fig. 4b. The Spearman (R) and Pearson (r) correlation coefficients are shown. Three biological replicates (BR1, BR2, and BR3) were evaluated and each biological replicate has one, two, or three technical replicates (TRs).


a
The correlation between predicted vs. measured PE2 efficiencies in HCT116 and MDA-MB-231 cells
HCT-BR1-TR1
R = 0.70
r = 0.58
n = 72
HCT-BR1-TR2
R = 0.72
r = 0.57
n = 75
HCT-BR2-TR1
R = 0.77
r = 0.61
n = 75
HCT-BR2-TR2
R = 0.75
r = 0.60
n = 75
MDA-BR1-TR1
R = 0.76
r = 0.62
n = 71
MDA-BR1-TR2
R = 0.80
r = 0.64
n = 73
MDA-BR2-TR1
R = 0.80
r = 0.65
n = 74
MDA-BR2-TR2
R = 0.81
r = 0.65
n = 75
DeepPE prediction score
Measured PE2 efficiency (%)
b
The correlation between PE2 efficiencies of replicates in HCT116 cells
R = 0.94
r = 0.99
n = 72
(BR1, TR2)
(BR1, TR1)
R = 0.94
r = 0.97
n = 72
(BR2, TR1)
(BR1, TR1)
R = 0.93
r = 0.97
n = 72
(BR2, TR2)
(BR1, TR1)
R = 0.95
r = 0.98
n = 75
(BR2, TR1)
(BR1, TR2)
R = 0.95
r = 0.97
n = 75
(BR2, TR2)
(BR1, TR2)
R = 0.94
r = 0.99
n = 75
(BR2, TR2)
(BR2, TR1)
Measured PE2 efficiency in HCT116 cells (%)
Measured PE2 efficiency in HCT116 cells (%)



Supplementary Figure 22. Evaluation of DeepPE using HCT116 and MDA-MB-231 cells. Eight datasets of PE2 efficiencies were generated using HCT116 (abbreviated as HCT) and MDA-MB-231 (abbreviated as MDA) cell lines at lentivirally integrated target

sequences that were never used for the training of DeepPE. Two biological replicates (BR1 and BR2) per cell line were evaluated and each biological replicate had two technical replicates (TR1 and TR2). The number of pegRNA and target sequence pairs (n) and the Spearman (R) and Pearson (r) correlation coefficients are shown in each graph. (a) The correlation of predicted vs. measured PE2 efficiencies in HCT116 and MDA-MB-231 cells. (b-c) The correlation between measured PE2 efficiencies of replicates in HCT116 (b) and MDA-MB-231 cells (c). (d) The correlation between measured PE2 efficiencies in HCT116 (b) and MDA-MB-231 cells.



Supplementary Figure 23. Performance comparison of DeepPE and other approaches for selecting the most efficient combination, out of 24 possibilities, of PBS and RT template lengths at a given target sequence. “13-nt PBS & 12-nt RT template” refers to choosing this combination of lengths regardless of the target sequence. Initial study recommendations A and B are based on using a 13-nt PBS and a 12-nt RT template (RTT) and avoiding a G as the last templated nucleotide by changing the RTT length as necessary. In recommendation A, if the last templated nucleotide is a G, then a 10-nt, rather than a 12-nt, RTT is chosen. If after this change the last templated nucleotide is again a G, then a 15-nt RTT is chosen. In recommendation B, if the last templated nucleotide is a G, then a 15-nt, rather than a 12-nt, RTT is chosen. If after this change the last templated nucleotide is again a G, then a 10-nt RTT is chosen. As controls, we also selected pegRNAs randomly (Random 1 and Random 2). Statistical significances determined by using the two-sided paired *t*-test are shown. In the boxes, the top, middle, and bottom lines represent the 25th, 50th, and 75th percentiles, respectively, and whiskers indicate the minimum and maximum values. The number of target sequences n = 97 per group.

Supplementary Table 1. Error rates in the plasmid and cell library. (a) The error rates in the copies in the plasmid library were evaluated by Sanger sequencing. (b) The lentiviral vector-induced shuffling frequency was evaluated using deep sequencing. Given that the PCR used for the deep sequencing sample preparation induces shuffling, the shuffling efficiency observed in the plasmid library using deep sequencing was subtracted from the total observed shuffling frequencies observed in the cell library[14].

(**a**) Error rate in the plasmid library	
Copies without any errors	130
Copies containing any error in the guide sequence, scaffold, RT template, PBS, or target sequence regions	12
Copies with shuffling	0
Total number of analyzed copies	142
Error rate in the plasmid library	12/142 = 8.5%

(**b**) Lentiviral vector-induced shuffling efficiency	
Initial template	Observed shuffling efficiency by deep sequencing
Plasmid library	11 / 1,101 (1.00%)
Genomic DNA from cell library	27 / 520 (5.19%)
Lentiviral vector-induced shuffling efficiency	(5.19% - 1.00%) = 4.2%

Supplementary Table 2. Datasets used for this study fall into three categories: PE2 efficiencies (i) at integrated target sites in HEK293T cells, (ii) at integrated target sequences in HCT116 and MDA-MD-231 cells, and (iii) at endogenous sites in HEK293T cells.

(i) Datasets HT-training, HT-test, Type-training, Type-test, Position-training, and Position-test contain PE2 efficiencies at integrated target sequences and were obtained from high-throughput experiments conducted in HEK293T cells. The datasets from the high-throughput experiments were divided into training and test datasets by random sampling.

(ii) Eight datasets named HCT-BR1-TR1, HCT-BR1-TR2, HCT-BR2-TR1, HCT-BR2-TR2, MDA-BR1-TR1, MDA-BR1-TR2, MDA-BR2-TR1, and MDA-BR2-TR2 contain the prime editing efficiencies at integrated target sequences obtained by using the high-throughput approach in HCT116 or MDA-MD-231 cells.

(iii) PE2 efficiencies at endogenous sites measured after transient transfection of plasmids encoding PE2 and pegRNA in HEK293T cells included six datasets named Endo-BR1-TR1, Endo-BR1-TR2, Endo-BR2-TR1, Endo-BR2-TR2, Endo-BR2-TR3, and Endo-BR3.

BR, Biological replicate TR, Technical replicate.

Dataset	Method/ Experiment	Cell line	Data size	Target information	Usage
HT-training	High-throughput experiments 1&2	HEK 293 T	38,692	Integrated target sequences (sequences derived from human coding regions)	Activity profiling of PE2, development of DeepPE
HT-test	High-throughput experiments 1&2	HEK 293 T	4,457	Integrated target sequences (sequences derived from human coding regions)	Activity profiling of PE2, validation of DeepPE
Type-training	High-throughput experiments 1&2	HEK 293 T	3,775	Integrated target sequences (sequences derived from human coding regions)	Activity profiling of PE2, development of PE_type
Type-test	High-throughput experiments 1&2	HEK 293 T	403	Integrated target sequences (sequences derived from human coding regions)	Activity profiling of PE2, validation of PE_type
Position-training	High-throughput experiments 1&2	HEK 293 T	1,774	Integrated target sequences (sequences derived from human coding regions)	Activity profiling of PE2, development of PE_position
Position-test	High-throughput experiments 1&2	HEK 293 T	200	Integrated target sequences (sequences derived from human coding regions)	Activity profiling of PE2, validation of PE_position
HCT-BR1-TR1	High-throughput experiment, BR1, TR1	HCT 116	72	Integrated target sequences (sequences derived from human coding regions)	Evaluation of DeepPE
HCT-BR1-TR2	High-throughput experiment, BR1, TR2	HCT 116	75	Integrated target sequences (sequences derived from human coding regions)	Evaluation of DeepPE
HCT-BR2-TR1	High-throughput experiment, BR2, TR1	HCT 116	75	Integrated target sequences (sequences derived from human coding regions)	Evaluation of DeepPE
HCT-BR2-TR2	High-throughput experiment, BR2, TR2	HCT 116	75	Integrated target sequences (sequences derived from human coding regions)	Evaluation of DeepPE

MDA-BR1-TR1	High-throughput experiment, BR1, TR1	MDA-MB-231	71	Integrated target sequences (sequences derived from human coding regions)	Evaluation of DeepPE
MDA-BR1-TR2	High-throughput experiment, BR1, TR2	MDA-MB-231	73	Integrated target sequences (sequences derived from human coding regions)	Evaluation of DeepPE
MDA-BR2-TR1	High-throughput experiment, BR2, TR1	MDA-MB-231	74	Integrated target sequences (sequences derived from human coding regions)	Evaluation of DeepPE
MDA-BR2-TR2	High-throughput experiment, BR2, TR2	MDA-MB-231	75	Integrated target sequences (sequences derived from human coding regions)	Evaluation of DeepPE
Endo-BR1-TR1	Transient transfection BR1, TR1	HEK293T	31	Endogenous target sites (sequences derived from human coding regions)	Validation of high-throughput approach, evaluation of DeepPE
Endo-BR1-TR2	Transient transfection BR1, TR2	HEK293T	30	Endogenous target sites (sequences derived from human coding regions)	Validation of high-throughput approach, evaluation of DeepPE
Endo-BR2-TR1	Transient transfection BR2, TR1	HEK293T	28	Endogenous target sites (sequences derived from human coding regions)	Validation of high-throughput approach, evaluation of DeepPE
Endo-BR2-TR2	Transient transfection BR2, TR2	HEK293T	28	Endogenous target sites (sequences derived from human coding regions)	Validation of high-throughput approach, evaluation of DeepPE
Endo-BR2-TR3	Transient transfection BR2, TR3	HEK293T	28	Endogenous target sites (sequences derived from human coding regions)	Validation of high-throughput approach, evaluation of DeepPE
Endo-BR3	Transient transfection BR3	HEK293T	20	Endogenous target sites (sequences derived from human coding regions)	Validation of high-throughput approach,

					evaluation of DeepPE

Supplementary Table 3. Datasets of PE2 efficiencies at endogenous sites (provided as a separate file). “BR” represents biological replicates, in which different experimentalists conducted the experiments using independently maintained cells. “TR” indicates technical replicates, in which the transfections of PE2- and pegRNA-encoding plasmids were independently performed.

Supplementary Table 4. Datasets HT-training, HT-test, Type-training, Type-test, Position-training, and Position-test (provided as a separate file).

Supplementary Table 5. Datasets of PE2 efficiencies generated using HCT116 and MDA-MB-231 cells (provided as a separate file).

Supplementary Table 6. Oligonucleotides used in this study (provided as a separate file).

Supplementary Table 7. Exact p-values for Figures 2, 3, and Supplementary Figure 15 (provided as a separate file).

Supplementary Code 1. Codes relevant to the PE efficiency analysis (provided as a separate file).

Supplementary Code 2. Codes relevant to DeepPE (provided as a separate file).