nature portfolio

https://doi.org/10.1038/s41587-022-01276-4

Supplementary information

High-throughput functional evaluation of human cancer-associated mutations using base editors

In the format provided by the authors and unedited

Supplementary Information

High-throughput functional evaluation of human cancer-associated mutations using base editors

Contents:

Supplementary note 1

When lentiviral vectors are used for high-throughput evaluations, barcodes and sgRNA-encoding sequences can be shuffled during lentivirus generation at a rate dependent on the distance between the two elements[33]. The distance between barcodes and sgRNA-encoding sequences in our constructs was only 100 bp. Given that the switching rates were 4.2~4.3% when the distances between barcodes and sgRNA guide sequences were 92~134 bp in similar lentiviral vectors[31, 32] (Supplementary Fig. 1), the switching rate for the current study would be expected to be about 4.3%. Because base editing at the integrated targets would not occur with the uncoupled sequences, the observed base editing would be about 95.7% (=100% - 4.3%) of the true base editing efficiency (that is, if the real base editing efficiency is 70%, the observed efficiency would be 70% x 95.7% = 67%). This low level of template switching would not substantially affect our high-throughput evaluations.

Supplementary note 2

Let us suppose that a protein variant called “variant a” was generated by the transduction of lentivirus encoding sgRNA A, an example sgRNA, at an efficiency E(a) and, to make the mathematical model simpler, that sgRNA A does not induce other variants (Supplementary Fig. 2a). To determine the relationship between the LFC (log fold change) in cell number and the editing efficiency E(a) for the generation of variant a, we first define the following five parameters.

N(a,x): the number of cells transduced with lentivirus that encodes sgRNA A, at day x

N(nt,x): the number of cells transduced with lentivirus that encodes nontargeting negative control sgRNA, at day x

E(a): the base editing efficiency for the generation of variant a by the activity of sgRNA A, $0 \leq E(a) \leq 1$

k: the proliferation rate of the negative control cell population (i.e., the cell population transduced with a negative control sgRNA-encoding lentivirus) (unit: /day) (for example, if the number of negative control cells increase by 20% in one day, then k = 1.2/day), $k > 1$

α(a): the increment in the proliferation rate in the cell population containing variant a generated by sgRNA A (unit: /day) (for example, if the population proliferation rate of variant a-containing cells were 20% higher than that of the negative control cell population, then $\alpha(a)/k = 0.2$. As another example, if the population proliferation rate of variant a-containing cells were 30% lower than that of the negative control cell population, then $\alpha(a)/k = -0.3$), $\alpha(a)/k \geq -1$

The fold-change in the cell number from day 0 to day 1 in cell populations transduced with nontargeting sgRNA can be calculated as follows.

$N(nt,1)/N(nt,0) = k$ ……equation 1

The fold-change in the cell number from day 0 to day 14 in cell populations transduced with nontargeting sgRNA can be calculated as follows.

$N(nt,14)/N(nt,0) = k^{14}$ …….. equation 2

Similarly, the fold change in the cell number from day 0 to day 1 in cell populations transduced with sgRNA A can be calculated. In this case, the cell population is composed of variant a-containing cells, in which base editing has occurred, and unperturbed cells, in which base editing has not occurred; the proportions of these populations would be E(a) and 1-E(a), respectively (Supplementary Fig. 2a). The fold change in the cell number from day 0 to day 1 in variant a-containing cells and unperturbed cells would be (k + α (a)) and k, respectively. Thus, the fold change in the cell number from day 0 to day 1 in cell populations transduced with sgRNA A can be calculated as follows.

$N(a,1)/N(a,0) = (1 – E(a)) \times k + E(a) \times (k + \alpha(a))$ …….equation 3

Similarly, the fold change in the cell number from day 0 to day 14 in cell populations transduced with sgRNA A can be calculated as follows.

$N(a,14)/N(a,0) = (1 - E(a)) \times k^{14} + E(a) \times (k + \alpha(a))^{14}$ ……. equation 4

If we divide equation 4 by equation 2, the fold-change in the proliferation rate of the cell population transduced with sgRNA A normalized to that of cells transduced with nontargeting control sgRNAs over 14 days can be calculated as follows.

$$(N(a,14)/N(a,0))/(N(nt,14)/N(nt,0)) = 1 - E(a) + E(a) \times (1 + \alpha(a)/k)^{14}$$

$$= 1 + E(a) \times ((1 + \alpha(a)/k)^{14} - 1) \quad \ldots\ldots \text{ equation 5}$$

If we take the logarithm of both sides of equation 5,

$$\text{Log}_2(N(a,14)/N(a,0))/(N(nt,14)/N(nt,0)) = LFC = \log_2(1 + E(a) \times ((1 + \alpha(a)/k)^{14} - 1)) \ldots.$$ equation 6

For example, if variant a-containing cells grow 10% faster than unperturbed cells including non-targeting sgRNA-expressing cells, then $\alpha(a)/k = 0.1$.

In this case, the correlation of LFC and E(a) is as follows.

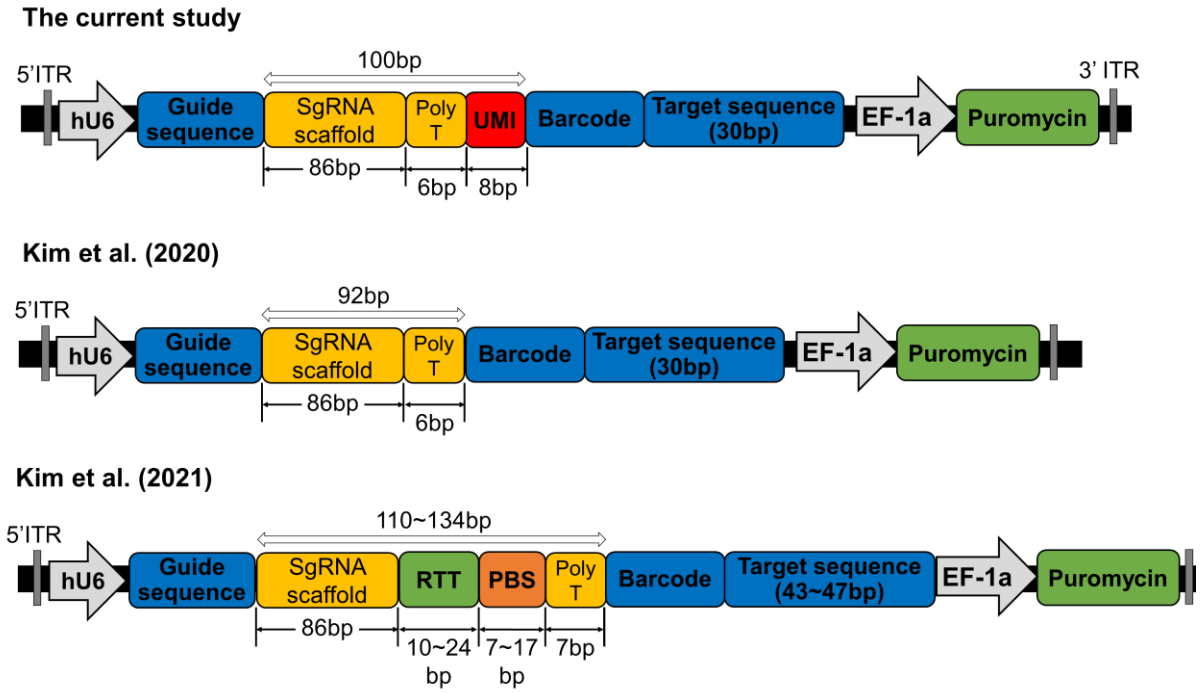
$$LFC = \log_2(1 + E(a) \times ((1.1)^{14} - 1) = \log_2(1 + E(a) \times 2.797) \ldots\ldots\ldots\ldots \text{equation 7}$$

When we plot the relationship between LFC and E(a), it was close to, albeit not exactly, linearly proportional when the growth phenotype was relatively modest, such as $-0.15 \leq \alpha(a)/k \leq 0.15$ (Supplementary Fig. 2b).

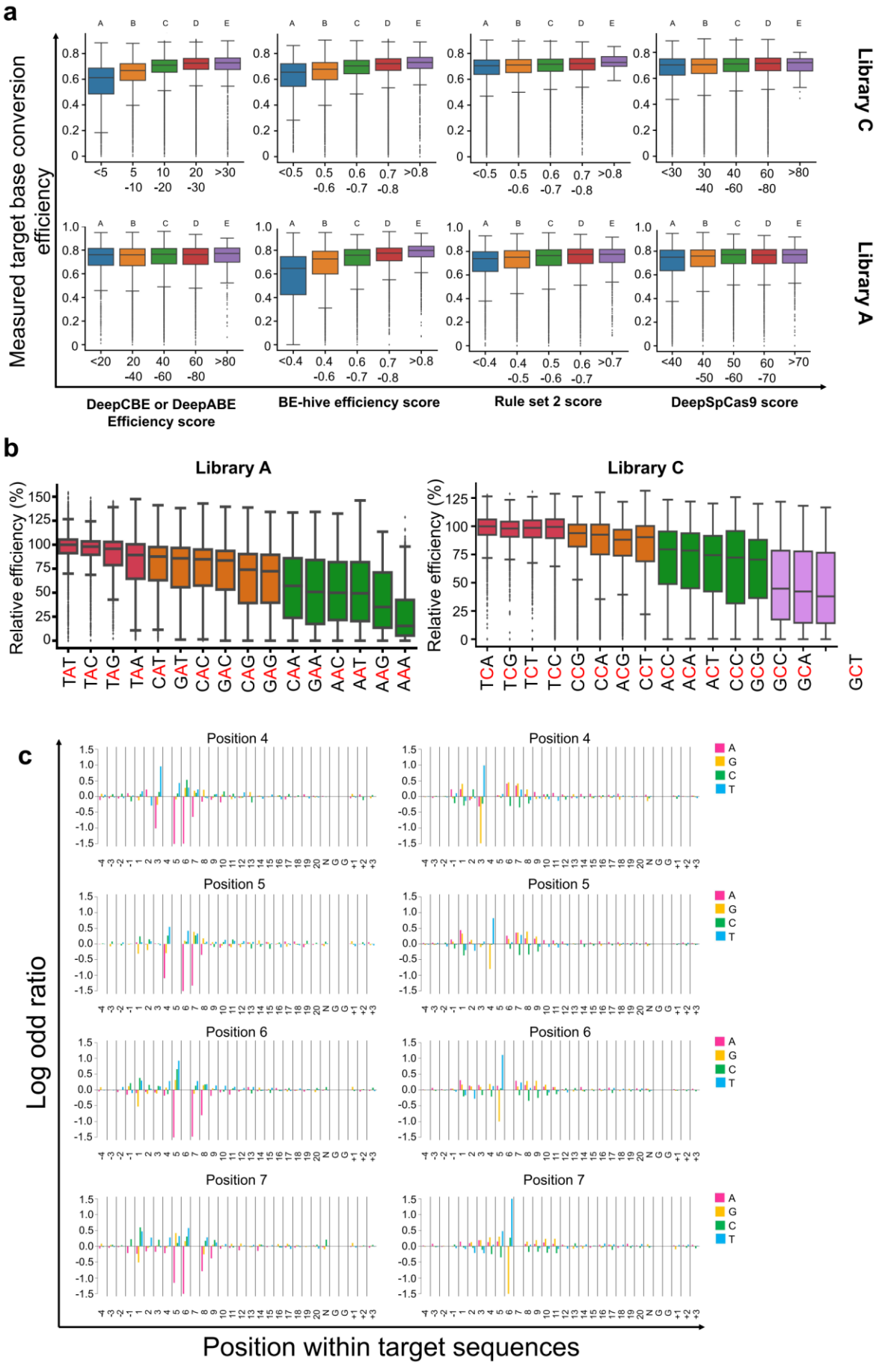
In other words, to reach LFCs that are higher than noise-level LFCs in the case of an outgrowing phenotype (or in the case of a depleting phenotype, LFCs that are lower than noise-level LFCs), base editing efficiencies are almost as important as the phenotypes induced by the variants generated by base editing when the phenotypes are modest. However, if we assume that the phenotypes are extreme, such as $\alpha(a)/k = 1, 2, -0.8, -0.9$, the increment or decrement in LFC does not exhibit a linear relationship with that in the base editing efficiency (Supplementary Fig. 2c). Interestingly, if a strong outgrowing phenotype is assumed (e.g. $\alpha(a)/k = 1$ or 2), LFCs substantially higher than the noise levels would be observed even when very low levels of base editing efficiencies (e.g. 10%) are assumed. However, in contrast, even if a strong depleting phenotype is

assumed (e.g. α (a)/k = -0.8 or -0.9), LFCs substantially lower than the noise levels are not expected to be observed when the levels of base editing are low.

The highest and lowest LFCs observed in our study were 2.6 and -2.8, respectively. In these cases, if we assume that the base editing efficiencies were 0.6 (60%) or 0.7, the α (a)/k values for LFC 2.6 would be 0.17 or 0.16, respectively, and those for LFC -2.8 cannot be calculated (the base editing efficiency should be at least 86% to reach LFC -2.8). When base editing efficiencies were assumed to be 90%, then α (a)/k values for LFCs 2.6 and -2.8 would be 0.14 and -0.19, respectively. These results suggest that all observed phenotypes, especially outgrowing phenotypes, in our experimental settings would be relatively modest.

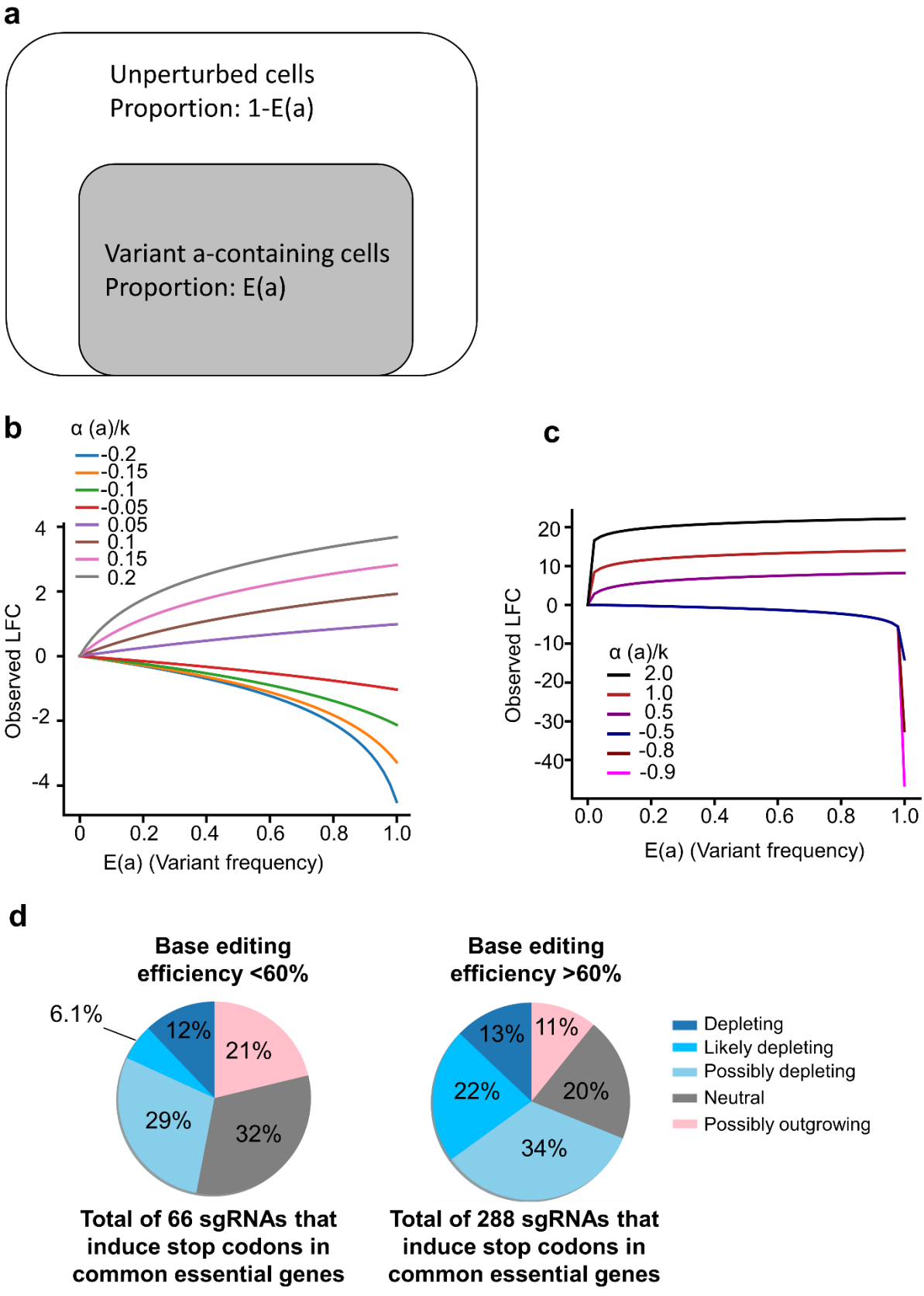


Supplementary Fig. 1. A map of the lentiviral vector containing the library of sgRNA-encoding sequence and surrogate target sequence pairs used in the current study and previous studies[45,46]. UMI, 8-nt unique molecular identifier.

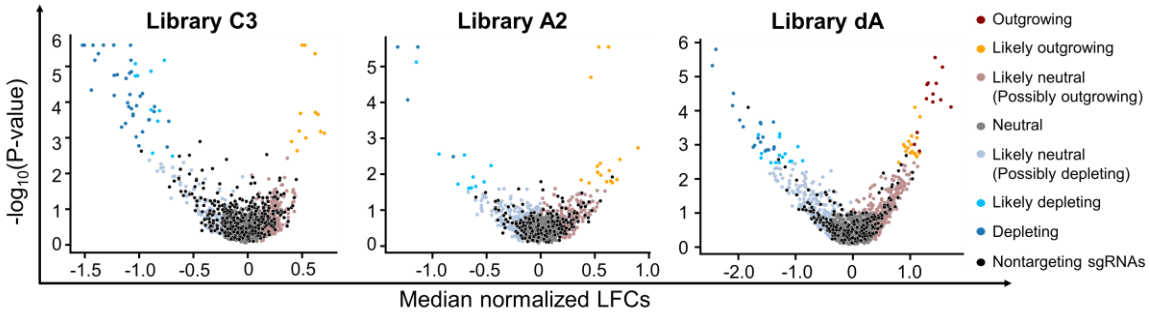

a
Measured target base conversion efficiency
Library C
Library A
DeepCBE or DeepABE Efficiency score
BE-hive efficiency score
Rule set 2 score
DeepSpCas9 score
b
Library A
Library C
Relative efficiency (%)
TAT TAC TAG TAA CAT GAT CAC GAC CAG GAG CAA GAA AAC AAT AAG AAA
TCA TCG TCT TCC CCG CCA ACG CCT ACC ACA ACT CCC GCG GCC GCA GCT
c
Position 4
Position 5
Position 6
Position 7
A
G
C
T
Log odd ratio
Position within target sequences

Supplementary Fig. 2. Base editing efficiencies at integrated target sequences and effects of sequence context surrounding target sequences. a, Base conversion efficiency of the target base of sgRNAs at increasing thresholds based on scores from the indicated computational models in library C (top) and library A (bottom). The number of analyzed target in each subset are as follows, from left to right: Library C: 3,242, 6,611, 22,782, 24,715, 1,557 (DeepCBE score); 1,575, 57,01, 23,645, 26,357, 1,629 (BE-Hive score); 24,747, 20,701, 11,555, 1,854, 50 (Rule set 2 score); and 7,157, 9,631, 28,929, 13,110, 80 (DeepSpCas9 score); Library A: 1,232, 8,346, 9,904, 2,871, 306 (DeepABE score); 1,296, 5,217, 5,581, 6,640, 3,925 (BE-Hive score); 1,774, 5,644, 8,812, 5,543, 886 (Rule Set 2 score); 5,873, 5,405, 6,309, 4,154 and 918 (DeepSpCas9 score). P-values of one-way analysis of variance followed by Dunn's post hoc test are as follows: Library C: A-B: $1.99x10^{-66}$, A-C: 0, A-D: 0, A-E: $4.25x10^{-249}$, B-C: $7.98x10^{-230}$, B-D: 0, B-E: $3.96x10^{-125}$, C-D: $1.74x10^{-105}$, C-E: $8.42x10^{-17}$, D-E: 0.498 (DeepCBE score), A-B: $1.61x10^{-9}$, A-C: $9.27x10^{-68}$, A-D: $9.58x10^{-146}$, A-E: $4.54x10^{-124}$, B-C: $8.35x10^{-81}$, B-D: $7.51x10^{-252}$, B-E: $5.05x10^{-124}$, C-D: $2.12x10^{-126}$, C-E: $6.13x10^{-51}$, D-E: $2.53x10^{-11}$ (BE-Hive score), A-B: $9.78x10^{-19}$, A-C: $4.61x10^{-46}$, A-D: $4.64x10^{-24}$, A-E: $4.04x10^{-4}$, B-C: $2.81x10^{-11}$, B-D: $3.71x10^{-11}$, B-E: 0.003, C-D: $9.0x10^{-4}$, C-E: 0.0164, D-E: 0.0727 (Rule set 2 score), A-B: 0.0239, A-C: $1.45x10^{-18}$, A-D: $1.21x10^{-31}$, A-E: 0.0992, B-C: $6.41x10^{-12}$, B-D: $2.19x10^{-24}$, B-E: 0.181, C-D: $1.08x10^{-7}$, C-E: 0.536, D-E: 0.905 (DeepSpCas9); Library A: A-B: 0.588, A-C: 0.110, A-D: 0.364, A-E: 0.077, B-C: $1.27x10^{-5}$, B-D: 0.028, B-E: 0.027, C-D: 0.412, C-E: 0.270, D-E: 0.176 (DeepABE score), A-B: $1.16x10^{-41}$, A-C: $2.12x10^{-109}$, A-D: $2.58x10^{-194}$, A-E: $2.08x10^{-301}$, B-C: $3.02x10^{-43}$, B-D: $1.82x10^{-150}$, B-E: $3.91x10^{-290}$, C-D: $3.74x10^{-33}$, C-E: $3.9x10^{-129}$, D-E: $9.85x10^{-46}$ (BE-Hive score), A-B: $8.74x10^{-6}$, A-C: $1.12x10^{-21}$, A-D: $8.63x10^{-38}$, A-E: $1.03x10^{-18}$, B-C: $6.28x10^{-14}$, B-D: $6.77x10^{-34}$, B-E: $2.01x10^{-11}$, C-D: $3.1x10^{-9}$, C-E: 0.001, D-E: 0.723 (Rule set 2 score), and A-B: $3.94x10^{-6}$, A-C: $1.58x10^{-29}$, A-D: $4.8x10^{-20}$,

A-E: 5.88×10^{-9}, B-C: 2.21×10^{-10}, B-D: 1.64×10^{-6}, B-E: 8.09×10^{-4}, C-D: 0.349, C-E: 0.956, D-E: 0.571 (DeepSpCas9 score). Boxplots are represented as follows: center line of box indicating the median, box limits indicating the upper and lower quartile; whiskers show the 1.5 times interquartile range. **b,** Effect of the sequence context surrounding the target base (red letters) on the ABE- (left, library A) and CBE- (right, library C) directed base-editing frequency at protospacer positions 4 to 8. The target base conversion frequency was normalized to the median frequency of the sequence motif that showed the highest median editing frequency at each position, yielding the relative frequency. Boxplots are represented the same as Supplementary Fig.2a. The number of analyzed target motifs (*n*) are as follows: Library A; *n*= 2,097 (TAT), 2,387 (TAC), 736 (TAG), 1,339 (TAA), 3,119 (CAT), 1,954 (GAT), 2,637 (CAC), 2,683 (GAC), 3,762 (CAG), 2,908 (GAG), 3,975 (CAA), 3,607 (GAA), 2,960 (AAC), 2,138 (AAT), 3,714 (AAG), 3,408 (AAA); Library C; n= 4,586 (TCA), 5,801 (TCG), 3,950 (TCT), 13,557 (TCC), 12,127 (CCG), 11,224 (CCA), 5,811 (ACG), 11,141 (CCT), 7,298 (ACC), 2,639 (ACA), 2,182 (ACT), 18,969 (CCC), 9,918 (GCG), 12,503 (GCC), 5,755 (GCA) and 5,823 (GCT). **c**, Sequence preferences at each position in efficient (top 20%) vs. inefficient (bottom 20%) target sequences. Sequence preferences for ABE (left) and CBE (right) are shown. The log odds ratios of nucleotide frequencies between efficient and inefficient target sequences are represented on the y axis. A nucleotide is preferred in efficient targets at a position if the log odds ratio has a positive value, but is preferred in inefficient targets at the position if the log odds ratio has a negative value.


a
Unperturbed cells
Proportion: 1-E(a)
Variant a-containing cells
Proportion: E(a)
b
α (a)/k
-0.2
-0.15
-0.1
-0.05
0.05
0.1
0.15
0.2
Observed LFC
E(a) (Variant frequency)
c
α (a)/k
2.0
1.0
0.5
-0.5
-0.8
-0.9
Observed LFC
E(a) (Variant frequency)
d
Base editing efficiency <60%
Base editing efficiency >60%
6.1%
12%
21%
29%
32%
13%
11%
22%
20%
34%
Depleting
Likely depleting
Possibly depleting
Neutral
Possibly outgrowing
Total of 66 sgRNAs that induce stop codons in common essential genes
Total of 288 sgRNAs that induce stop codons in common essential genes

Supplementary Fig. 3. Effect of base editing efficiencies in the high-throughput evaluations. a, Schematics of cell populations in which base editing has occurred. E(a), the base editing efficiency for the generation of variant a by sgRNA A. **b,c,** Theoretical relationship between LFC and E(a) (base editing efficiency) with modest (**b**) or extreme (**c**) changes in the proliferation rate (α (a)/k). See Supplementary text for details. k, the proliferation rate of the negative control cell population. α(a): the increment in the proliferation rate in the cell population containing variant a generated by sgRNA A. **d**, the distribution of sgRNAs predicted to induce nonsense mutations in common essential genes in each classification in library C.

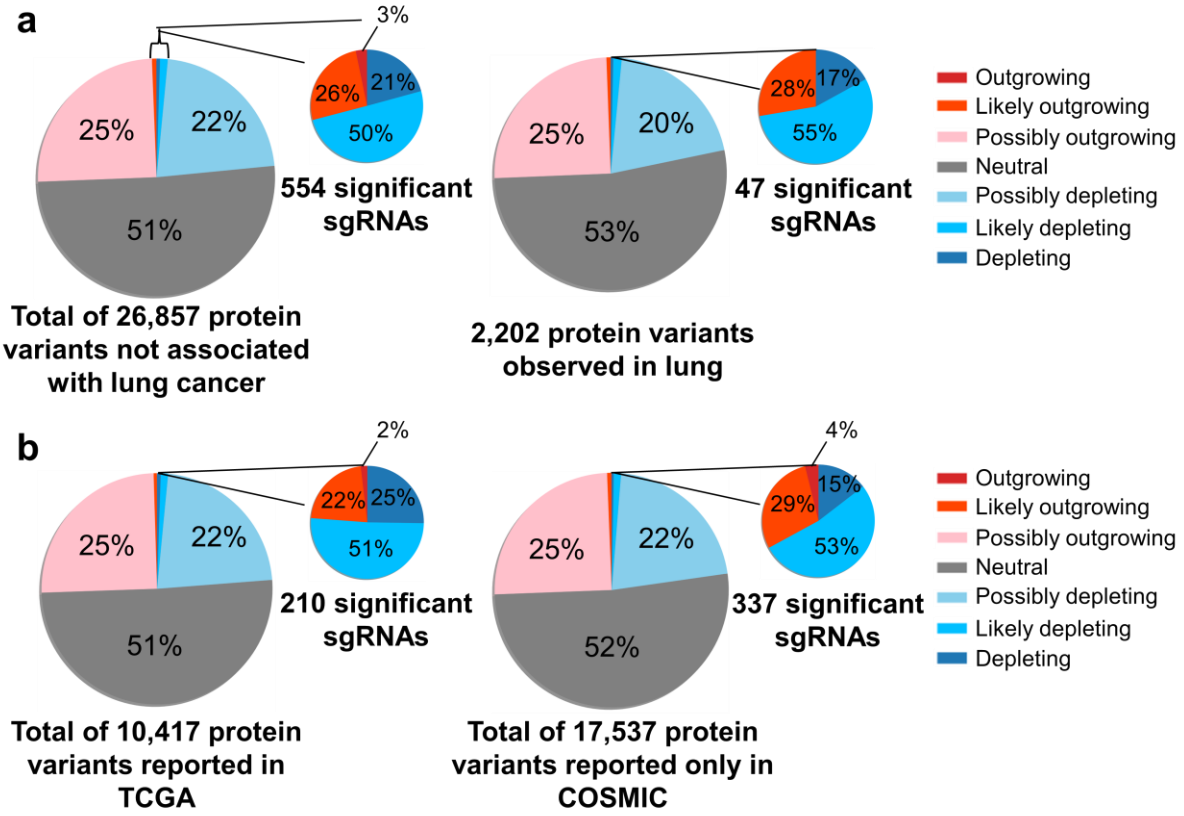


Supplementary Fig. 4. High-throughput classifications in three small libraries. Volcano plots of nLFCs and negative logarithm of RRA *P*-value of sgRNAs in libraries C3, A2, and dA. The colors of the dots (sgRNAs) represent their functional classifications (using the same color code shown in Fig. 2b). Nontargeting sgRNAs are shown in black.

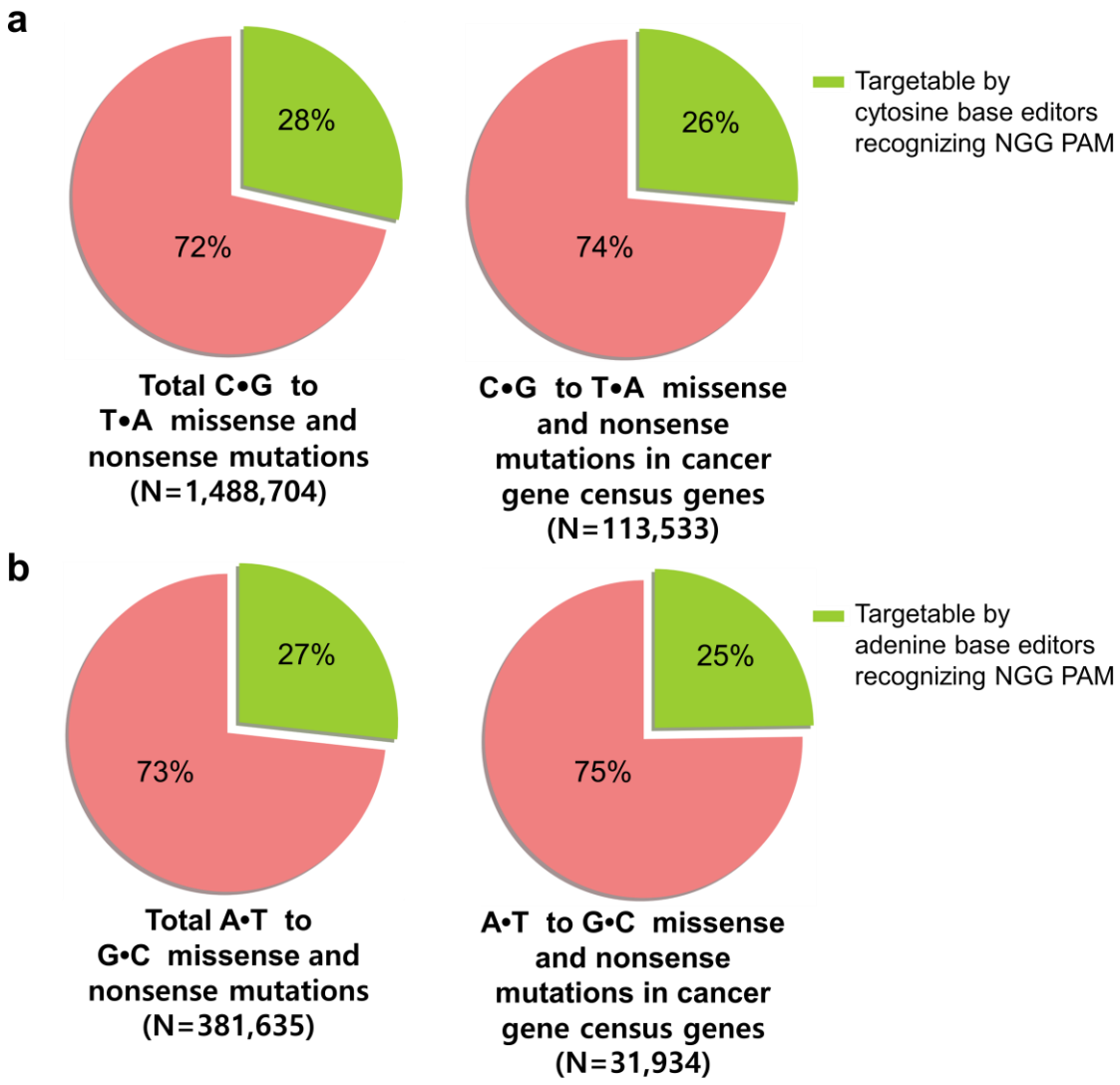
No	Sg.ID	Classification	Target gene
1	**Cg.TP53_p.Q192***	Likely outgrowing	Tumor suppressor gene
2	**Cg.TP53_p.T155I,p.T155=**	Likely outgrowing	Tumor suppressor gene
3	**Cg.TP53_p.Q100*,p.S99F**	Likely outgrowing	Tumor suppressor gene
4	**Cg.ACOX3_p.Q145***	Neutral	Enzyme gene
5	**Cg.KMT2C_p.R1906***	Likely neutral (Possibly outgrowing)	Tumor suppressor gene
6	**Cg.CDC23_p.T381M**	Likely outgrowing	Cell Cycle gene
7	**Cg.PTPN14_p.Q110***	Likely Outgrowing	Tumor suppressor gene
8	**Cg.POLR1C_p.A6V**	Depleting	Common essential gene
9	**Cg.MMS22L_p.R661***	Depleting	Common essential gene
10	**Cg.POLR2B_p.P714L**	Depleting	Common essential gene
11	**Cg.POLG_p.Q1029***	Likely depleting	DNA polymerase
12	**Ag.TP53_p.R280G**	Likely outgrowing	Tumor suppressor gene
13	**Ag.TP53_p.N239D**	Outgrowing	Tumor suppressor gene
14	**Ag.TP53_p.K120E,p.K120R**	Likely outgrowing	Tumor suppressor gene
15	**Ag.TP53_p.K351E**	Likely neutral (Possibly outgrowing)	Tumor suppressor gene
16	**Ag.TP53_p.T125A**	Likely outgrowing	Tumor suppressor gene
17	**Ag.CTCF_p.H312R**	Depleting	Common essential gene
18	**Ag.SRSF1_p.D139G**	Depleting	Common essential gene
19	**Ag.POLE_p.Y1889C**	Neutral	Common essential gene
20	**Ag.ACTL6A_p.T405A**	Neutral	Common essential gene
21	**Cg.GNA13_p.Q27***	Outgrowing	Oncogene
22	**Cg.CASP8_p.S158F**	Outgrowing	Tumor suppressor gene
23	**Cg.PSMB5_p.S261F**	Likely neutral (Possibly outgrowing)	Proteasome subunit
24	**Cg.PHLDA1_p.Q201***	Outgrowing	Apoptosis regulation
25	**Ag.PHLDA1_p.Y249C**	Likely outgrowing	Apoptosis regulation
26	**Ag.IRF6_p.Y97C**	Outgrowing	Transcription activator
27	**Ag.EGFR_p.Y727C**	Outgrowing	Oncogene
28	**Ag.SIK1_p.F26L**	Likely outgrowing	Tumor suppressor gene

Supplementary Fig. 5. sgRNAs selected for individual validations. 28 sgRNAs and their classifications in this study, and the putative functions of the genes targeted by the

sgRNAs, are shown.



Supplementary Fig. 6. The distribution of primary protein variants in each classification. a, The distribution of primary protein variants in each classification for lung cancer-related variants (right) and variants not related to lung cancer (left). Only sgRNA-induced single amino acid changes that represent a major proportion (>75%) of the edited alleles were included. **b**, The distribution of primary protein variants in each classification for variants reported in TCGA and COSMIC (left) and COSMIC only (right). Only sgRNA-induced single amino acid changes that represent a major proportion (>75%) of the edited alleles were included.



Supplementary Fig. 7. The proportion of SNVs listed in the COSMIC database that are targetable by cytosine base editors (**a**) and adenine base editors (**b**) recognizing an NGG PAM. Mutations in cancer-related genes (cancer gene census) are indicated in the right panel.

Supplementary Table 1. Composition of sgRNA-encoding libraries C, A, C1, C2, C3, A1, A2, dA, and eC. Barcode sequences used for sorting, sgRNA sequences, target sequences including neighboring sequences (5'-neighboring sequence (4 bp) + target sequence (20 bp + 3-bp PAM = 23 bp) + 3'-neighboring sequence (3 bp) = 30 bp of genomic DNA sequence). Information about intended mutations and DeepCBE or DeepABE efficiency scores are also included (provided as a separate Excel file).

Supplementary Table 2. The results of MAGeCK analyses. Raw reads per million (RPM) of 4 replicate^{UMI}, log fold changes (LFCs), median LFCs (mLFCs), positive or negative MAGeCK robust rank aggregation (RRA) *P*-values, and LFCs of UMI CPM are shown for each sgRNA (provided as a separate Excel file).

Supplementary Table 3. Functional classifications of sgRNAs and protein variants. **a**, Functional classification of sgRNAs based on the proliferation and survival (sheet 1). **b**, Functional classification of sgRNAs in library eC (sheet 2). **c**, Base editing outcomes and allele frequencies at the integrated target sequences (dependency on EGF signaling) (sheet 3). **d**, Potential classification of sgRNAs with low base editing efficiencies (sheet 4). (provided as a separate Excel file).

Supplementary Table 4. Results of allele frequency tracking after delivery of an individual sgRNA for 20 selected sgRNAs. After lentiviral transduction of the specified individual sgRNA, protein variant frequencies were calculated from DNA sequence analysis. Endogenous DNA sequence variants encoding the same amino acid change were

combined into one protein variant (provided as a separate Excel file).

Supplementary Table 5. Oligonucleotides used in this study (provided in a separate Excel file).