Base editor-directed high-throughput functional evaluation of human cancer-associated transition mutations

Younggwang Kim

Department of Medical Science

The Graduate School, Yonsei University

Base editor-directed high-throughput functional evaluation of human cancer-associated transition mutations

Younggwang Kim

Department of Medical Science

The Graduate School, Yonsei University

Base editor-directed high-throughput functional evaluation of human cancer-associated transition mutations

Directed by Professor Hyongbum Kim

The Doctoral Dissertation
submitted to the Department of Medical Science,
the Graduate School of Yonsei University
in partial fulfillment of the requirements for the degree
of Doctor of Philosophy

Younggwang Kim

December 2020

This certifies that the Doctoral Dissertation of Younggwang Kim is approved.

Thesis Supervisor : Hyongbum Kim

Thesis Committee Member#1 : Hyunseok Kim

Thesis Committee Member#2 : Jae-Ho Cheong

Thesis Committee Member#3: Sangwoo Kim

Thesis Committee Member#4: Tae-Min Kim

The Graduate School
Yonsei University

December 2020

ACKNOWLEDGEMENTS

I am indebted to so many, and I acknowledge them with great pleasure. I first wish to express my deepest gratitude to Professor Hyongbum Kim, my dissertation advisor, for his guidance, encouragement, and continuing support. Without his intellectual energy and commitment, this dissertation would not have been the same. I would like to extend my sincere thanks to the members of my dissertation committee. Professors Hyunseok Kim and Sangwoo Kim provided extremely valuable insight into genetic screening and bioinformatic tools. Professors Jae-Ho Cheong and Tae-Min Kim offered cogent comments on cancer genomics. I am very grateful to Professor Phil Hyu Lee, my Master's thesis advisor. He first guided me to pursue my goals at the Physician-Scientist Program. His own innovative work and passion for both clinics and basic science continue to inspire me. The Department of Neurology at Severance Hospital, with its wonderful professors and colleagues, provided a wonderful home for my research. I am extremely thankful and indebted to Seungho Lee with whom I worked very closely for this project. Hui Kwon Kim, Myungjae Song, Jungmin Lim, Ramu Gopalappa and all other colleagues of the Department of Pharmacology shared their knowledge and expertise. I will always cherish the time we spent together. Finally, I wish to thank my parents and family members for their love and encouragement, without whom I would never have enjoyed so many opportunities. Above all, I owe heartfelt thanks to my loving wife and son for their company and moral support throughout my studies.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Base editor-directed high-throughput functional evaluation of human cancer-associated transition mutations

Younggwang Kim

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Hyongbum Kim)

Identifying causal driver mutations among the vast majority of other passenger mutations is key to understanding tumorigenesis and the development of cancer therapy. Here, we developed a novel high-throughput method using cytosine base editor (CBE) and adenine base editor (ABE) to evaluate the function of somatic cancer-associated mutations found in human cancer tissues. We designed a total of 83,731 and 23,613 sgRNAs for CBE and ABE, respectively, that would lead to the generation of 107,982 single nucleotide variants found in human somatic cancers. We performed screening to identify mutations with positive (outgrowing) or negative (depleting) effects on proliferation and survival followed by two sets of high-coverage small focused library evaluation with sgRNA libraries. We found that unique molecular identifier (UMI)-based analysis outperforms conventional one-sgRNA-one-mutation based analysis in identifying outgrowing or depleting mutations. Our screening platform using base editors should facilitate cancer genomics by identifying functional consequences of individual variants, which may contribute to the development of new therapeutic options.

Key words : CRISPR/Cas9, base editor, genome-wide screening, somatic mutations, Variants of uncertain significance

Base editor-directed high-throughput functional evaluation of human cancer-associated transition mutations

Younggwang Kim

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Hyongbum Kim)

I. INTRODUCTION

Since the development of high-throughput DNA sequencing, analysis of tens of thousands of pan-cancer tumor samples has revealed a landscape of cancer mutations[1-3] along with lists of putative cancer-related genes[4-6] that confer growth advantages on the cells carrying them, so-called driver genes[7]. Although these efforts greatly enhanced our understanding of the landscape of cancer genomics, current driver-detection methods mainly rely on statistical methods to compare the frequencies of mutations in an individual gene with those in other neutral genes[8,9], methods considering the computationally predicted effects of mutations on protein function[10,11], or both[5]. However, because the results of these retrospective driver discovery approaches could be affected by the background mutation rate[12], inherent[13,14] or targeted tumor therapy[15], it is not always straightforward to infer direct causal relationships between candidate cancer-related mutations and tumorigenesis in a specific cellular context. In addition, most cancer mutations are still annotated as variants of uncertain significance (VUS)[16], and recent subgene-resolution analysis showed that some of these mutations, not located in part of a canonical cancer driver gene, could be candidates of true drivers[17]. To overcome the gap between our current knowledge of cancer-related mutations and their role in a human cellular context, a need for direct functional

assessment of genetic variants has emerged[18]. Functional evaluation of cancer mutations including VUS in a cellular context would greatly facilitate the understanding carcinogenesis and tumor growth and subsequent development of cancer therapies.

As a method for high-throughput functional evaluation of VUS in cells, overexpression of transgenes containing VUS of *TP53*[19,20], *PPARG*[21], and *ERK2*[22] in cells in which the corresponding endogenous target genes had already been knocked out. However, functional consequences of overexpression of transgenes can be different from that of expression of endogenous genes containing variants. As a method for introducing VUS into the endogenous gene, homology-directed repair (HDR) can be used[23]. Although the functions of VUS of a gene, *BRCA1* have been well evaluated using the approach[24], the functional evaluation of VUS in other genes using HDR can be challenging mainly because the efficiency of HDR is usually limited in mammalian cells. Furthermore, these transgene or HDR-based approaches have not been used to evaluate VUS from a large number of genes.

About 95% of mutations found in cancer tissue are single nucleotide variants (SNV)[9]. Base editors, composed of Cas9, guide RNA, and deaminases, can induce transition mutations in a targeted manner[25,26]. Here, we developed a high-throughput method to evaluate the functional effects of 107,984 transition SNVs found in human cancer tissues on the proliferation and survival of non-cancer cells using cytosine base editor (CBE), adenine base editor (ABE) and associated single guide RNA (sgRNA) libraries. We increased the sensitivity and specificity of the functional evaluation by using unique molecular identifier (UMI) and closely monitored the outcomes of base editing using surrogate target sequences.

II. MATERIALS AND METHODS

1. Cell lines and cultures

HBEC30KT cells are normal human bronchial epithelial cells that were immortalized by stable expression of CDK4 and hTERT; they also exhibit trisomy of chromosome 5, which may have occurred during selection for immortalization. However, these cells exhibit intact contact inhibition of proliferation and lack tumorigenic potential[27]. HBEC30KT-shTP53 (HBEC^{P53}) was generated by lentiviral delivery of shRNA targeting TP53 into HBEC30KT cells, and the molecular features and properties of this cell line were previously described[28]. In brief, immunoblot analysis of the products of oncogenic genes that play important roles in lung cancer [TP53, KRAS, and LKB1 (STK1)] showed that HBEC^{P53} resembles its normal matched control HBEC30KT except for reduced expression of the TP53 protein. Additionally, HBEC^{P53} cells are nontumorigenic in immune-compromised mice.

HBEC^{P53} cells were cultured in ACL4 medium (RPMI 1640 (GIBCO, 2.05 mM L-glutamine) supplemented with 0.02 mg/ml insulin, 0.01 mg/ml transferrin, 25 nM sodium selenite, 50 nM hydrocortisone, 10 mM HEPES, 1 ng/ml epidermal growth factor, 0.01 mM ethanolamine, 0.01 mM O-phosphorylethanolamine, 0.1 nM triiodothyronine, 2 mg/ml bovine serum albumin, 0.5 mM sodium pyruvate) with 2% Tet system approved fetal bovine serum (FBS, Clontech) and 1% Penicillin-Streptomycin (GIBCO) at 37℃ with 5% CO_2. HEK293T cells (American Type Culture Collection) were cultured in Dulbecco's modified Eagle's Medium (DMEM, GIBCO) with 10% FBS (GIBCO) at 37℃ with 5% CO_2.

2. Lentivirus production

HEK293T cells were seeded in 100-mm culture dishes at a density of 5 X 10^6 cells per dish 24 hours before transfection. On the day of transfection, the growth medium was exchanged for 10 mL of DMEM containing 25 μM chloroquine diphosphate, after which cells were cultured for 5 hours. Transfer

plasmids containing the gene of interest, psPAX2, and pMD2.G were mixed at a molar ratio of 1.64:1.3:0.72 pmol and diluted into 500 μL of Opti-MEM (Life Technology). Polyethylenimine (PEI) was diluted into Opti-MEM in a total volume of 500 μL and added to the DNA mixture such that the ratio of μg DNA:μg PEI was 1:3, resulting in a total volume of 1000 μL. The mixture was incubated for 20 min and added to cells. To achieve a high viral titer, caffeine (Sigma-Aldrich, C0750) was added to the culture medium, at a final concentration of 4 mM, after treatment with the PEI:DNA mixture as previously described[29]. At 12 hours post-transfection, 10 mL of growth medium supplemented with 4 mM of caffeine was added to refresh the cells.

After 36 hours, the growth medium was harvested and centrifuged at 2000g for 10 min to pellet cell debris. The supernatant was filtered through a Millex-HV 0.45-μm low protein–binding membrane (Millipore), divided into aliquots, and kept frozen at -80°C until use.

3. Plasmid library construction

Pooled, 150-nt oligonucleotides for plasmid construction were array-synthesized by Twist Bioscience. Each plasmid in our library was designed to include the following elements (**Figure 2**): (i) a 19-nt homology arm with a U6 promoter at the 3' terminus, (ii) a 20-nt sequence with a G at the 5' terminus followed by a 19-nt sgRNA guide sequence, (iii) a random 20-nt sequence flanked by a BsmBI cut site on either side (11 nt each), (iv) a 20-nt unique barcode sequence corresponding to each sgRNA, (v) a 30-nt surrogate target sequence containing a PAM (a 4+23-nt target sequence plus a 3-nt PAM+3-nt), which is identical to an endogenous genomic target locus, and (iv) a 20-nt homology arm.

The methods used to generate the plasmid library were previously described in detail[30]. In brief, the pooled oligonucleotides were amplified using primer pair 17/18 and Phusion High-Fidelity DNA Polymerase (NEB), after which they were size-selected by electrophoresis on a 2% agarose gel.

The amplicons were assembled into linearized Lenti-gRNA-Puro (Addgene, 84752) after digestion with BsmBI using NEBuilder HiFi DNA Assembly Master Mix (NEB). 200 ng of linearized vector and 120 ng of purified oligonucleotides were used in one Gibson assembly reaction (with a total volume of 20 μL). A total of 16 and 8 reactions were performed for the CBE and ABE libraries (designated the C and A libraries), respectively. After the assembly reactions, the mixtures were pooled and concentrated using a MEGAquick-spin Total Fragment DNA Purification kit (iNtRON Biotechnology, South Korea) and used in up to 12 and 8 electroporation reactions to maximize the library complexity.

An improved form of sgRNA scaffold[31] and a unique molecular identifier (UMI) were synthesized (IDT, Primer 19) and amplified using primer pair 20/21. The resulting amplicon was digested with BsmBI and purified. A ligation reaction was then performed using 60 ng of the sticky-ended sgRNA scaffold-UMI fragments and 250 ng of the scaffoldless plasmid library generated above, also digested with BsmBI. A total of 16 and 8 reactions were performed for the CBE and ABE libraries, respectively. The reaction mixtures were pooled, concentrated, and used in up to 12 and 8 electroporation reactions.

4. Screening experiments

Twenty-four hours before transduction, 168 million HBEC^{P53}-rtTA-CBE (P-C) cells and 48 million HBEC^{P53}-rtTA-ABE (P-A) cells were seeded in duplicate, resulting in 2000-fold coverage of the sgRNA libraries in each replicate. The cells in each replicate were infected with the lentiviral screening library (library C or A) with 10 μg/ml of polybrene at an MOI of approximately 0.3, such that every sgRNA was represented in approximately 600 cells. After 24 hours of infection, the medium was replaced with fresh medium containing 20 μg/ml of puromycin (Invitrogen) and 2 μg/ml of doxycycline hyclate (Sigma) to induce expression of the base editor; cells

were cultured under these conditions for an additional 9 days, and were harvested at day 10 post-infection with approximately 1000~1500-fold coverage of the sgRNA libraries. The remaining cells were maintained without doxycycline treatment with 2000-fold coverage of the sgRNA libraries for an additional 14 days. These cells, which were passaged every 2 to 3 days, were seeded at a relatively high confluency (40%, or 3 million cells per 150-mm culture dish) to mimic the level of proliferation in the bulk cell population. At day 24 post-infection, the cells were collected for genomic DNA extraction.

Similarly, for the two focused C libraries and two focused A libraries (C1, C2 and A1, A2), 24 hours before transduction, 42 million P-C cells and 24 million P-A cells were seeded in duplicate, resulting in ~10,000X coverage of the sgRNA library in each replicate. The remaining steps were the same as those used for the screening library, except that the cells were maintained with 10,000-fold coverage of the sgRNA libraries during screening.

5. Genomic DNA preparation and deep sequencing.

Genomic DNA was extracted from harvested cell pellets with a Wizard Genomic DNA Purification Kit (Promega) according to the manufacturer's protocol.

The integrated barcode and target sequences were amplified and prepared for deep sequencing through two PCR steps using 2X Pfu PCR Smart mix (Solgent). The first step was performed using genomic DNA, which was divided into multiple 50-μl reactions containing 2.5 μg of genomic DNA, 20 pmol of forward primer, 20 pmol of reverse primer, and 25 μL of PCR premix. The PCR cycling parameters were as follows: an initial 2 minutes at 95℃; followed by 30 seconds at 95℃, 30 seconds at 60℃, and 40 seconds at 72℃, for 24 cycles; and a final 5 minutes extension at 72℃. The total amount of genomic DNA for each experiment represented more than 1,000X coverage of the library, assuming 6.6 μg of genomic DNA per 10^6 cells[32].

In the second PCR step, which was performed to attach sequencing adaptors and barcodes, a total of 250 ng of purified PCR product from the first step was used in eight separate 50-μL reactions for the screening libraries and a total of 40 ng of purified PCR product from the first step was used in two separate 50-μL reactions for the focused libraries, with 20 pm of Illumina indexing primers in each reaction. The PCR cycling parameters were as follows: an initial 2 minutes at 95℃; followed by 30 seconds at 95℃, 30 seconds at 60℃, 40 seconds at 72℃, for 8 cycles; and a final 5 minutes extension at 72℃. Amplicons for each experiment were size-selected with agarose gel electrophoresis and sequenced using a HiSeq 2500 System (Illumina).

6. Validation experiments

We individually cloned sgRNA-encoding sequences in the Lenti-Guide-Puro vector (Addgene, #52963). 1.2 million base editor knock-in cells per sgRNA were seeded in 100 mm culture dishes 24 hours before transduction. The cells were infected with lentivirus harboring sequences encoding individual sgRNAs at a low MOI (~0.4). In addition, base editor knock-in cells were seeded as above for a GFP positive control. In this case, lentivirus harboring an empty sgRNA cassette, the puromycin resistance gene-p2A-GFP fusion gene was used to infect cells at a low MOI (~0.4). The day after transduction, the medium was replaced with fresh medium containing 20 μg/ml of puromycin (Invitrogen) and 2 μg/ml of doxycycline hyclate (Sigma) to induce expression of the base editor; these conditions were maintained for 48 hours. After removal of puromycin the cells were maintained for an additional 7 days with doxycycline treatment.

Competitive growth assay. Ten days after infection, cells transduced with lentivirus encoding candidate hit sgRNAs (GFP-) and cells transduced with the positive control lentivirus (GFP+) were mixed and grown together. The cells were sampled every 3 or 4 days and the ratio of GFP positive cells in the mixture was quantified via longitudinal flow cytometry. By assuming that the

cells exhibit an exponential growth rate, the number of cells (N) at times t1 and t2 can be described by the following equation, where f_0 is the absolute fitness of the reference cells and Δf_{gRNA} is the fitness change caused by the transduced sgRNA.

$$N_{t2} = N_{t1} \times 2^{(f_0+\Delta f_{gRNA})(t_2-t_1)}$$

The $\Delta f_{gRNA,ti}$ between a certain timepoint t_i and the reference timepoint t_0 was obtained according to the equation:

$$\frac{N_{gRNA,ti}}{N_{c,ti}} = \frac{N_{gRNA,t0} \times 2^{(f_0+\Delta f_{gRNA,ti})(t_i-t_0)}}{N_{c,t0} \times 2^{(f_0)(t_i-t_0)}}$$

$$\frac{\frac{N_{gRNA,ti}}{N_{c,ti}}}{\frac{N_{gRNA,t0}}{N_{c,t0}}} = 2^{\Delta f_{gRNA,ti}}$$

The ratio between the number of GFP- cells (N_{gRNA}) and the number of GFP+ cells (N_c) was obtained from the competitive growth assay, and we assumed that the relative fitness of the GFP+ cells was equal to the fitness of the reference cells (f_0). The relative enrichment ($E_{gRNA,ti}$) between a certain timepoint t_i and the reference timepoint t_0 (**Figure 17E**) was determined as follows:

$$E_{gRNA,ti} = \frac{\frac{N_{gRNA,ti}}{N_{c,ti}}}{\frac{N_{gRNA,t0}}{N_{c,t0}}} \times 100\,(\%)$$

Allele frequency tracking. The cells harboring individual sgRNA and base editors seeded duplicate after removal of doxycycline at 10 days post-infection. These cells were cultured for an additional 2 weeks, and harvested at 10, 17, 24 days post-infection. Each sgRNA-targeted genomic site was amplified using site-specific primers and analyzed by deep sequencing.

We assumed that the fold change in the frequency of each allele could be explained by the sum of the effect of the allele frequency change induced by the activity of the base editor and the effect of competition between each allele. However, the surrogate target context sequence was randomly integrated via lentivirus in the genome, so fold change in the frequency of each allele in surrogate sequence is only affected by the activity of base editors. Therefore, we normalized the log fold change (LFC) of each allele in its endogenous genomic site with that of allele in surrogate sequence to estimate the effect proliferation of each allele excluding the influence of the base editor by adjusted LFC (aLFC) as follows:

$$aLFC = \log_2 \frac{Proportion\ in\ endogenous\ sequence\ in\ D24}{Proportion\ in\ endogenous\ sequence\ in\ D10} - \log_2 \frac{Proportion\ in\ surrogate\ sequen}{Proportion\ in\ surrogate\ sequen}$$

7. Screening data analysis

For UMI analysis, 8-nt UMI sequences were counted and analyzed according to the sorting barcode by deep sequencing with in-house Python scriptsClement, 2019 #246}. To minimize UMI sequencing errors, we used directional network integrating UMIs within threefold read count difference between edit distance 1 UMIs[33]. For UMI MAGeCK analysis, we generated a UMI MAGeCK count file using the reads per million of each UMI. We eliminated UMIs with less than a total of eleven read counts from the combined treated and control samples and sorting barcodes with less than six types of UMIs. We computed LFCs, MAGeCK positive and negative scores, and the false discovery rate (FDR) of the sgRNA using MAGeCK 0.5.9.3[34]. Plotting the LFC (x-axis) and negative logarithm of the MAGeCK score (y-axis) produced a volcano plot. The MAGeCK score used in the volcano plot was the selected lower value between the negative and positive MAGeCK scores. To calculate the integrated performance of an sgRNA

between replicates in each dataset, every UMI was assigned a new UMI barcode that did not overlap between different replicates, after which it was used as input for MAGeCK to calculate the performance of an sgRNA from all UMIs.

Conventional MAGeCK analysis was performed with sgRNAs for which there were at least thirty raw read counts (non-normalized) in the control sample[35]. Conventional MAGeCK scores were selected in the same manner as for UMI MAGeCK analysis, after which volcano plots were generated. The integrated performance of the sgRNA in each dataset was calculated using conventional MAGeCK analysis from the deep sequencing results of two replicates.

8. Functional scoring of sgRNAs

First, we eliminated UMIs with less than a total of eleven read counts from treated and control samples and sgRNAs with less than six types of UMIs. Second, we excluded sgRNAs associated with nonsynonymous editing rates of <40% in the surrogate sequence. We used a two-component Gaussian mixture model (GMM) to calculate the probability that mutations belong to each cluster. The GMM was performed using the GaussianMixture function of the sklearn.mixture package of Python library. To calculate Gaussian Mixture probability, the MAGeCK score and LFC were used as variables[36]. sgRNAs with a >90% probability of belonging to a neutral cluster were defined as ‘Neutral’ and scored as 0. In addition, sgRNAs with a >99% probability of belonging to a pathogenic cluster were defined as ‘Depleting’ or ‘Outgrowing’ according to their fold change and scored as -2 or 2, respectively. sgRNAs that were not included in any cluster were defined as ‘Likely depleting’ or ‘Likely outgrowing’ according to their fold changes and scored as -1 or 1, respectively.

Next, to integrate scores from each replicate, we averaged every score for each sgRNA from each replicate to calculate a ‘functional score’ ranging from

+2 to -2. We classified sgRNAs as 'Depleting' (Score ≤ -1.5), 'Likely depleting' (- 1.5<Score<-0.5), 'Neutral' (-0.5 ≤ Score ≤ 0.5), 'Likely outgrowing' (0.5 <Score< 1.5), and 'Outgrowing' (1.5 ≤ Score) **(Figure 12).**

9. Functional classification of cancer mutations

First, we defined a 'coverage score' for each mutation, which was set based on how many datasets were screened for targeted mutations. If a mutation was only screened in a large screening library, it received a score of 1. Each mutation was given an additional 2 points for each focused library that was screened; therefore, 'coverage scores' range from 0 to 5 points. Mutations that were not classified in any dataset because of low UMI counts, read counts, or base editing rates were annotated as 'Not assessed'. We classified mutations that were only assessed in a screening library (with a coverage score of 1) into neutral, intermediate, depleting, and outgrowing clusters according to their functional score. Only neutral clusters with a coverage score of 1 were considered likely to be neutral; other mutations were annotated as having an uncertain role, considering that there were high frequencies of neutral mutations in outgrowing or depleting clusters in the screening library (**Figure 16**). For mutations assessed in a focused library (with a coverage score of 2 or more), we annotated each cluster according to the functional scores of sgRNAs obtained from the focused library. If a mutation was screened in both the 1st and 2nd focused libraries, the mean functional score from the two libraries was used for clustering.

10. Statistical significance

To compare the number of UMIs between the day 10 and day 24 samples and the enrichment values between the target sgRNA and nonessential sgRNA control (**Figure 17E**), we used the two-tailed Student's t-test. Statistical significance was calculated using PASW Statistics (version 18.0, IBM).

III. RESULTS

1. Generation of cancer-associated SNVs using CBE and ABE

To introduce the selected cancer-associated SNVs in endogenous target sequences using CBE and ABE, we first generated cell lines that express CBE or ABE in a doxycycline-responsive manner. For the target cell type, we searched cell lines derived from normal or close-to-normal cells rather than those derived from malignant cancer cells so that we could evaluate the function of the variant proteins encoded by the SNV-containing mutant genes in non-cancer (i.e., normal or precancerous) cells. We initially attempted to use HBEC30KT cells, an immortalized cell line derived from normal lung cells[28]. However, the cell proliferation was so slow that it would have taken an unreasonably long time to obtain enough cells for high-throughput evaluations. Thus, we used HBEC30KT cells that lentivirally express a shRNA targeting TP53 (hereafter, for brevity, P cells), which showed faster proliferation than HBEC30KT cells although the proliferation rate of P cells was still much lower than that of the matched transformed cancer cells (HCC4017)[28]. We sequentially transduced lentiviral vectors expressing reverse tetracycline-controlled transactivator (rtTA) and a base editor (CBE or ABE) into P cells, after which untransduced cells were removed by G418 and hygromycin selection. The resulting cell lines, which express CBE or ABE in a doxycycline-inducible manner, were named P-C or P-A cells, respectively.

2. Selection of cancer-associated SNVs that can be generated using CBE and ABE

To identify target sequences that can be modified to contain SNVs observed in human cancer tissues using CBE and ABE, we first extracted 2,825,363 C>T and G>A SNVs as potential candidates for generation by CBE and 624,931 A>G and T>C SNVs as potential candidates for generation by ABE from the Catalogue of Somatic Mutations in Cancer (COSMIC)[37], one of the most extensive datasets of manually curated somatic mutations from the

scientific literature. From these lists, we identified 153,425 and 35,163 SNVs that can be generated using CBE and ABE, respectively, in the highly active 4 bp activity window spanning protospacer positions 4 to 7, numbered such that the end distal to the protospacer-adjacent-motif (PAM) is designated as position 1 with an NGG PAM (**Figure 1**). After filtering out synonymous SNVs, we attempted to remove SNVs that cannot be generated at high efficiency. For this purpose, we removed the 10% of the target sequences with the lowest DeepSpCas9 scores, which represent computationally predicted SpCas9 activities[38], given that the base editing efficiency at a given target sequence is usually low when the Cas9 nuclease activity at the same target sequence is low[39]. Then, we added two negative control groups of target sequence and sgRNA pairs: the first group contained synthetic target sequences for sgRNAs that do not target any sequences in the human genome (hereafter, nontargeting control sgRNAs), and the second group included target sequences that can be modified with CBE or ABE to contain synonymous SNVs that do not change the amino acid sequence [40,41]. As a result of this process, we obtained 83,731 and 23,613 pairs of target sequences and sgRNAs for CBE and ABE, respectively.

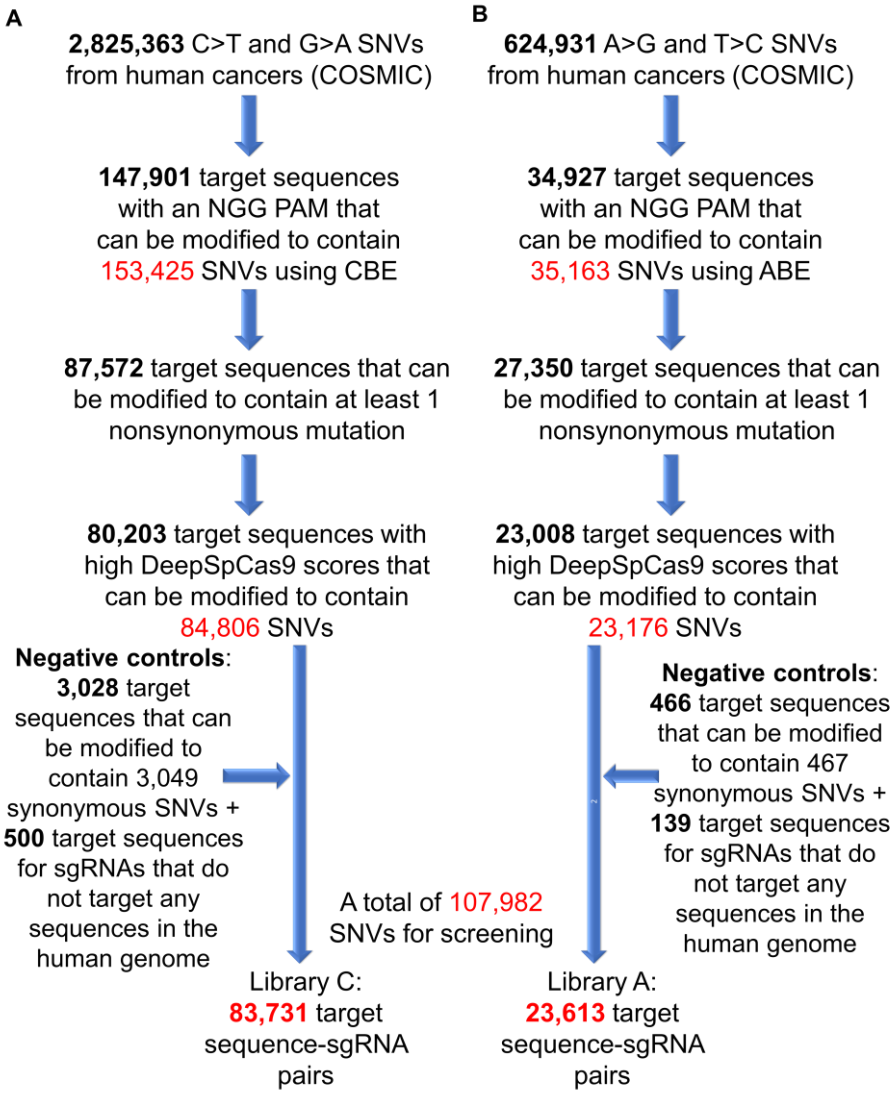


Figure 1. Flow chart of sgRNA design process. Selection of single nucleotide variants (SNVs) that can be generated using CBE (library C) and ABE (library A) in a high-throughput manner. (**A**) Determination of SNVs used to make library C. (**B**) Determination of SNVs used to generate library A.

3. Lentiviral libraries of sgRNA-encoding and target sequence pairs with unique molecular identifiers

We next considered the lentiviral delivery of sgRNAs that would induce the formation of cancer-associated SNVs in P-C and P-A cells. However, as compared to the efficiency of SpCas9, which is frequently used for high-throughput screening[32,42,43], the efficiency of base editing is often limited[25,26,39,44], which can decrease the accuracy or robustness of high-throughput evaluations. Furthermore, base editing can result in multiple editing outcomes in addition to the intended cancer-associated SNVs[25,26,39,44]. Thus, base editing efficiencies and outcomes must be monitored for accurate functional evaluation of the cancer-associated SNVs that can be generated using CBE and ABE. However, it is practically almost impossible to monitor the efficiencies and editing outcomes at thousands of endogenous target sites. Thus, to the sgRNA-encoding lentiviral vector, we added corresponding surrogate target sequences, which can be integrated into the cell genome; we previously used a similar approach involving lentiviral libraries of target sequence and guide RNA pairs to evaluate the efficiencies of CRISPR nucleases[38,45-47], base editors[39], and prime editor 2[48]. We and others previously observed strong correlations between base editing efficiencies and the relative frequencies of base editing outcomes as well as Cas9, Cas12a, and prime editor efficiencies in the lentiviral surrogate target sequences and those at endogenous target sites[30,38,39,45-50].

We generated lentiviral libraries, respectively named libraries C and A, of the 83,731 (for CBE) and 23,613 (for ABE) pairs of sgRNA-encoding and target sequence pairs described above (**Figure 2**). In addition, unlike in our previous lentiviral libraries[30,38,39,45-47], we added an eight nucleotide-long UMI at the end of the sgRNA scaffold sequence in both libraries for tracking of transduced cells and subsequent analyses[51]. We transduced libraries C and A into P-C and P-A cells in duplicate, respectively, at a multiplicity of infection

(MOI) of 0.3, such that every sgRNA was represented in approximately 600 cells in each biological replicate; the cell libraries were maintained with 2000-fold coverage of the sgRNA libraries. Twenty-four hours after the transduction, we supplemented the culture medium with doxycycline (2 μg/ml) to induce CBE and ABE expression and then incubated the cultures for 9 days. When the base editing efficiencies at the integrated target sequences were measured at day 10 after the initial transduction, we observed high base editing efficiencies (**Figure 3**); the median efficiencies at positions 4, 5, 6, and 7 were 49%, 66%, 68%, and 63% for CBE and 16%, 72%, 73%, and 61% for ABE. To evaluate the reproducibility of base editing between replicates, we measured the nonsynonymous editing rate in both biological replicates by counting the proportion of reads that caused amino acid changes in the surrogate reporter sequence. Both libraries C and A showed very high correlations between replicates, with correlation coefficients of 0.94 and 0.91, respectively (**Figure 4**).

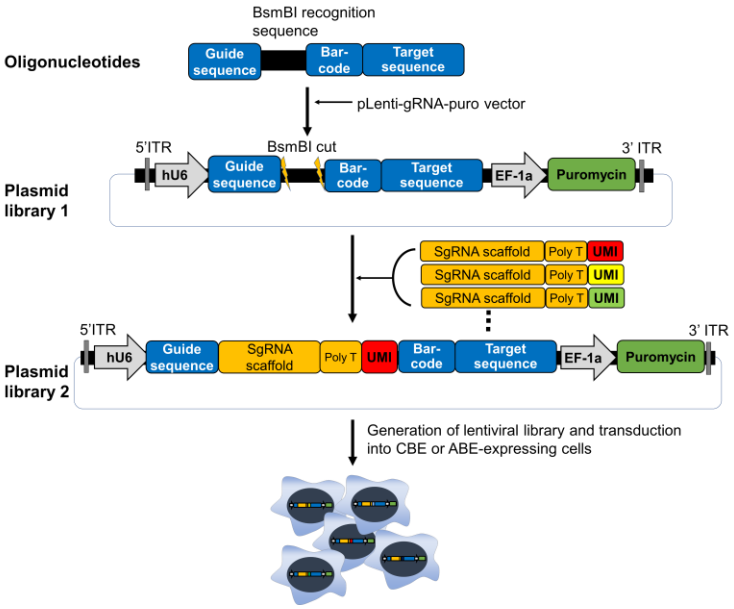


Figure 2. Generation of lentiviral libraries of sgRNA-encoding and target sequence pairs with unique molecular identifiers (UMIs). Oligonucleotides containing a 20-nt guide sequence and the corresponding target sequence were synthesized and cloned into the pLenti-gRNA-puro vector to create plasmid library 1. The plasmids were then digested with BsmBI restriction enzyme and ligated with fragments containing the sgRNA scaffold sequences and UMIs to create plasmid library 2. Lentiviral libraries generated from plasmid library 2 were then transduced into CBE- or ABE-expressing cells.

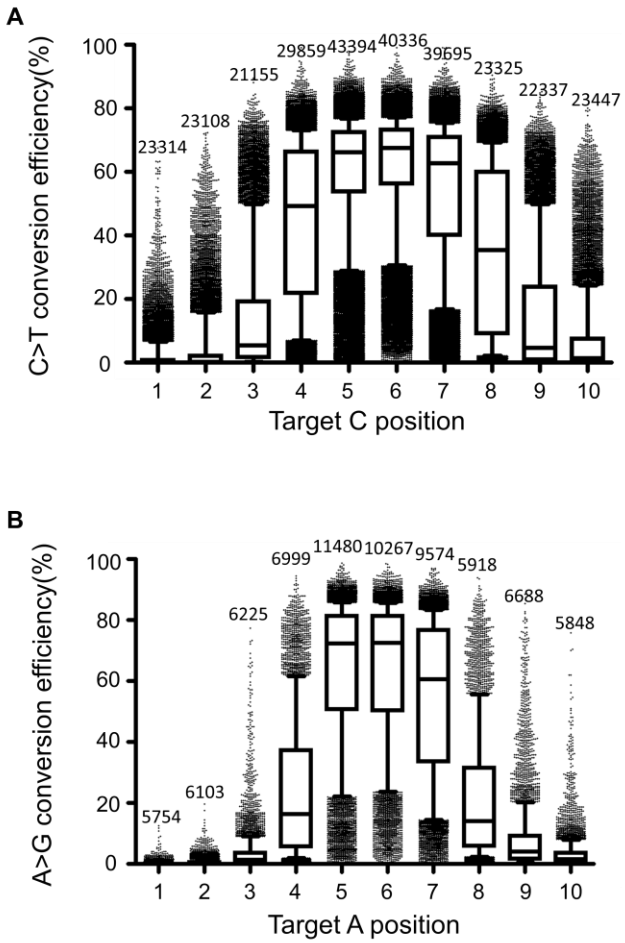


Figure 3. Analysis of base editing in surrogate sequences. Dependence of the C>T conversion efficiency on the relative position of the target C (**A**) and of the A>G conversion efficiency on the relative position of the target A (**B**) in the sgRNA. The position of the target A or C was counted with the end distal to the PAM defined as position 1. The number of analyzed adenines or cytosines is shown above each position. Data from replicate experiments were merged and sgRNAs with less than 100 raw read counts were excluded from the analysis. The boxes represent the 25th, 50th, and 75th percentiles; whiskers show the 10th and 90th percentiles.

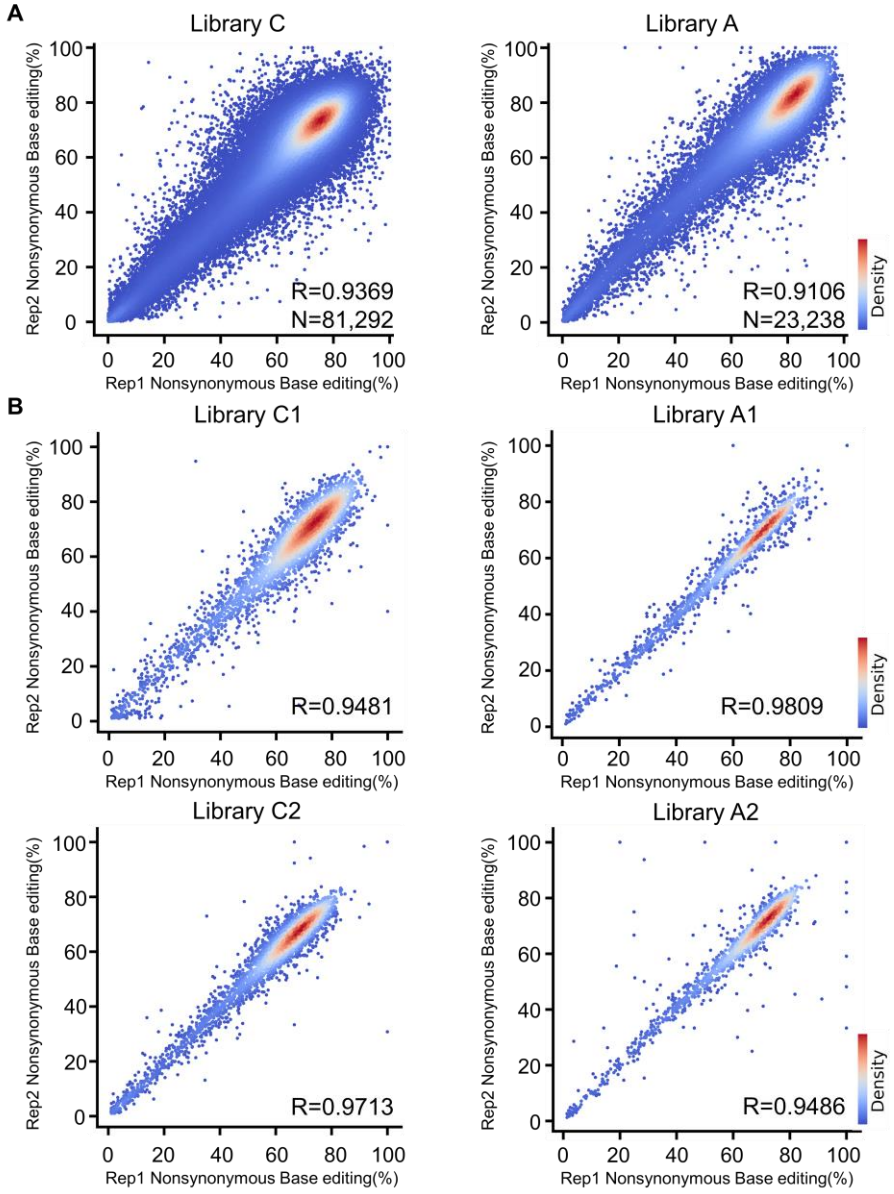


Figure 4. High reproducibility of nonsynonymous mutation rates in integrated target sequences. **(A)** Correlations between nonsynonymous mutation rates in integrated target sequences in replicates from library C (Left, N=81,292) and library A (Right, N=23,238). Pearson correlations for each

experiment are indicated. Dots are colored according to the sgRNA density. **(B)** Correlations between nonsynonymous mutation rates in integrated target sequences in replicates from focused libraries.

4. Comparison of CRISPR-UMI based metrics and conventional analysis

To evaluate the effect of the CBE- and ABE-generated variants on P-C and P-A cell proliferation and survival, we cultured these variant-containing cell populations in the absence of puromycin and doxycycline for 14 days. Genomic DNA was isolated from the cell populations at day 10 (baseline) and day 24 after the initial transduction of libraries C and A, and subjected to deep sequencing to evaluate the relative frequencies of sgRNA and target sequence pairs and UMIs (**Figure 5A**). In contrast to conventional CRISPR knockout screening that can involve several sgRNAs targeting a single gene, only one sgRNA can be used to induce a given SNV in most cases of base editor-based screening. Thus, conventional analysis that uses a single sgRNA as the input for calculating the log-fold change (LFC) could easily be distorted by stochastic noise caused by the underlying cell heterogeneity[51,52], and could be biased by differences in base editing efficiencies and diverse editing outcomes[39,53]. Previously, it was shown that the application of UMI-based lineage tracing in a Cas9-based loss-of-function genetic screen resulted in higher sensitivity and robustness than did conventional screening[51]. However, it has not been investigated whether the application of UMIs can increase the accuracy of high-throughput functional evaluation of CBE- and ABE-generated variants.

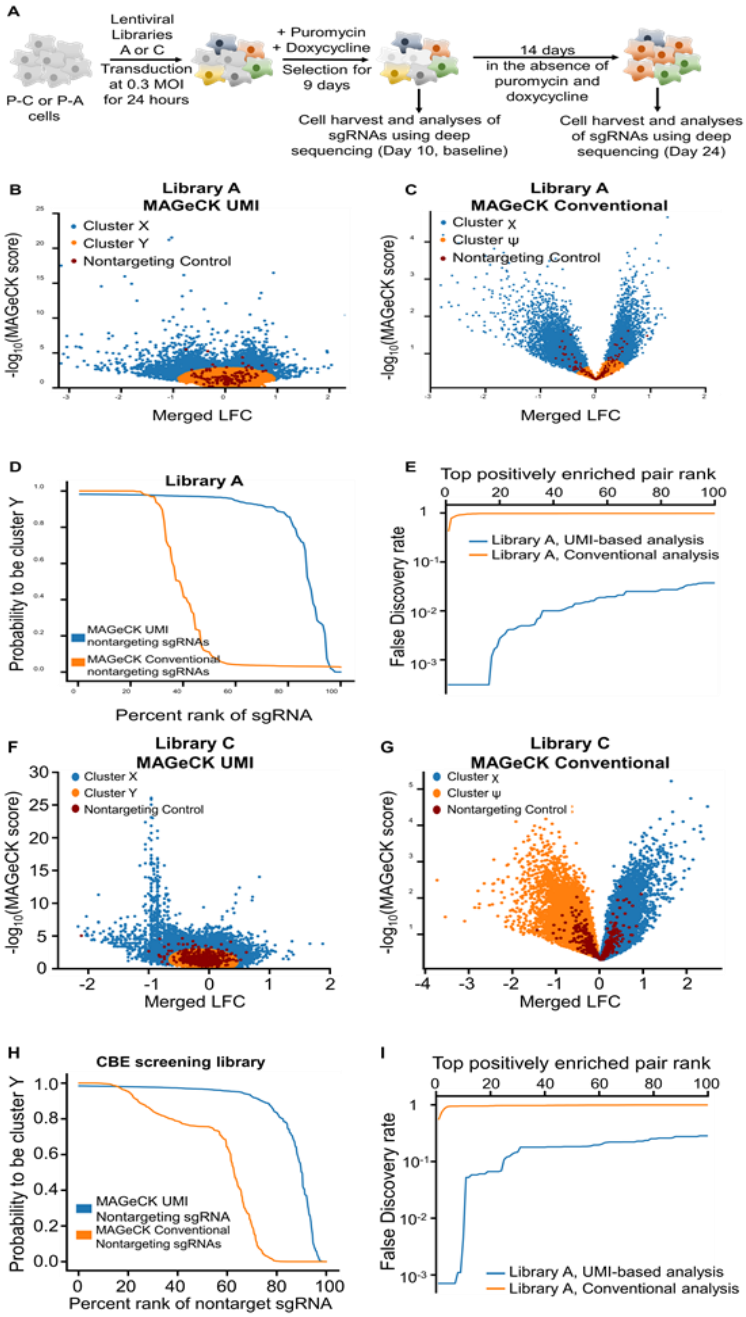


Figure 5. Base editing-mediated high throughput generation and evaluation of cancer-associated variants using lentiviral libraries containing UMI. **(A)**

Schematic for CBE and ABE-mediated high throughput evaluations of variants. P-C and P-A cells are untransformed human bronchial epithelial cells that express CBE and ABE, respectively, in a doxycycline-dependent manner. P-C or P-A cells are transduced with lentiviral libraries of sgRNAs, named libraries A or C, respectively. Untransduced cells were removed by puromycin selection and the expression of CBE or ABE were induced by adding doxycycline for nine days. Ten days after the transduction, half of the cells were harvested for analyses and the remaining cells were cultured in the absence of puromycin and doxycycline for 14 days. **(B-E)** Classification of target sequence and sgRNA pairs in library A. Using a two component GMM, the pairs were classified into two clusters according to the probability to be cluster X or Y. (B-C) Volcano plots using UMI-based MAGeCK (B) and conventional MAGeCK (C) analyses. Nontargeting control were shown as red dots. **(D)** The probability that nontargeting sgRNAs are classified in cluster Y. The results from UMI-based (blue) and conventional (orange) analyses are shown. **(E)** False discovery rate (FDR) from UMI-based and conventional analyses of library A. The results for the top 100 positively ranked variants are shown. **(F-G)** Volcano plot of CBE screening library using LFC and MAGeCK score calculated by UMI based analysis **(F)** and conventional analysis **(G).** Nontargeting control were shown as red dots. In contrast to ABE screening library, conventional analysis of CBE screening library failed to generate clusters of neutral mutations (Cluster B). **(H)** Rank distribution of probability that CBE screening library sgRNA was in cluster B, produced by Gaussian mixture model in nontargeting sgRNA between UMI based analysis (blue) and conventional analysis (orange). 90% probability that nontargeting sgRNA in cluster B was 74% in UMI analysis and 28% in conventional analysis, respectively. **(I)** False discovery rate (FDR) from UMI-based and conventional analyses of library C. The results for the top 100 positively ranked variants are shown.

We compared the accuracies of UMI-based MAGeCK[34] (hereafter, for brevity, MAGeCK-UMI) vs conventional MAGeCK[34] analyses in our experimental setting. We first compared LFCs between normalized read count ratios at the initial and final timepoints from two replicates, and found a lack of correlation between replicates by both UMI-based and conventional metrics despite similar base editing results (**Figure 4A and Figure 6**). This outcome suggests that most mutations in the library had no or minimal effects on proliferation, which is in line with previous findings[7,8]. Next, we used both UMI-based and conventional analysis to cluster neutral mutations that did not affect cell proliferation.

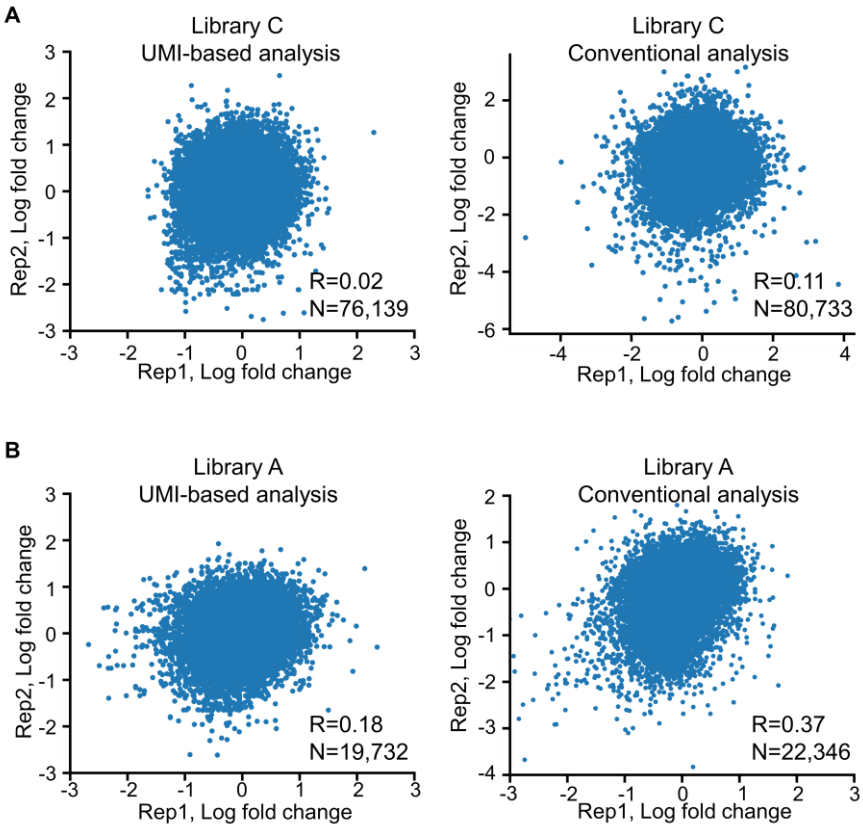


Figure 6. Log fold change of each sgRNA between the initial and final timepoints showed minimal correlations despite similar base editing results. **(A)** Correlations between log fold changes (LFCs) in two replicates of library C. LFCs calculated using CRISPR-UMI (Left, N=76,139) and conventional (Right, N=80,733) analyses are shown. For UMI-based analysis, only sgRNAs that passed the UMI and read count conditions (six types of UMIs and 11 read counts for the combined treated and control samples) in both replicates were used for analysis. For conventional analysis, sgRNAs that had more than 30 raw read counts in both replicates were used for analysis. **(B)** Correlations between LFCs in two replicates of library A (UMI, N=19,732; conventional, N=22,346).

To differentiate neutral mutations from those affecting cell proliferation, we classified target sequence-sgRNA pairs (hereafter, for brevity, pairs) into two groups by fitting a two-component Gaussian mixture model (GMM) using the $-\log_{10}$(MAGeCK score[34]) and LFC as variables (**Figure 5B, Figure 5C, Figure 5F and Figure 5G**). We named clusters in which cell proliferation seemed to be affected (indicated by a high absolute LFC value and a significant MAGeCK score) as cluster X or X, and the remaining targets as cluster Y or ψ, using the first and second letterforms to indicate MAGeCK-UMI or MAGeCK-conventional analyses, respectively. To compare the accuracies of MAGeCK-UMI and MAGeCK-conventional, we compared the probability of nontargeting pairs of belonging to cluster Y or ψ. The fraction of nontargeting pairs in library A that were classified as cluster Y or ψ was 75% using MAGeCK-UMI and 31% using MAGeCK-conventional and the fraction of nontargeting sgRNAs in library C that were so classified was 74% using MAGeCK-UMI and 28% using MAGeCK-conventional when a 90% probability of belonging to cluster Y or ψ was used as a threshold, suggesting that MAGeCK-UMI is more accurate than MAGeCK-conventional. (**Figure 5D and Figure 5H**). In addition, conventional analysis failed to generate clusters corresponding to a neutral cluster (cluster ψ) (**Figure 5G**). Next, we compared false discovery rates (FDRs) between MAGeCK-UMI and MAGeCK-conventional for the variants that were the most positively enriched over the 14 days (**Figure 5E and Figure 5I**). Even at FDR < 0.1, conventional analysis of libraries C and A failed to discover any hits. In contrast, at FDR < 0.05, UMI analyses found 11 and 135 variants from libraries C and A, respectively. These results suggest that MAGeCK-UMI is a more robust method for identifying pathogenic variants, a finding that is compatible with previous results based on Cas9 nuclease-based high-throughput evaluations[51].

To screen and validate more feasible candidate driver mutations, we screened two separate focused libraries. We prepared four smaller-sized focused libraries (containing nearly 3,261 and 3,173 unique sgRNAs for CBE and 1,595 and 1,582 unique sgRNAs for ABE), named C1, C2, A1, and A2, respectively. Pairs in libraries C1 and C2 were selected from library C and those in libraries A1 and A2 were from library A. The C2 and A2 libraries are respectively more enriched than C1 and A1 with positively ranked pairs. We performed screening experiments with 5-fold higher coverage than was used for library C or A such that every sgRNA was represented in more than 3000 cells at the time of transduction, and cell libraries were maintained with 10,000X coverage of the sgRNA libraries. We again evaluated the screening performance of MAGeCK-UMI and MAGeCK-conventional with the four focused libraries. Each screened guide-target pair was clustered by two-component GMM, and UMI analyses classified more nontargeting sgRNAs as cluster Y or ψ for the four focused libraries than for the original libraries (93% (C1), 88% (C2), 94% (A1) and 92% (A2) compared to 74% (C) and 75% (A); **Figure 7A and Figure 7B**). Furthermore, when we determined FDRs for all variants that were positively enriched over the 14 days, many more hits were obtained from UMI vs. conventional analyses for all four libraries (**Figure 7C and Figure 7D**). At FDR < 0.1, conventional analysis failed to discover any hits, with the exception of a small number of top ranking sgRNAs (26 in C1, 11 in C2, 130 in A1, and 175 in A2). UMI-based analysis discovered more hits compared to conventional analysis at FDR < 0.1 (65 in C1, 332 in C2, 286 in A1, and 876 in A2). These results corroborate that UMI-based analyses are more sensitive, robust, and accurate than conventional analyses.

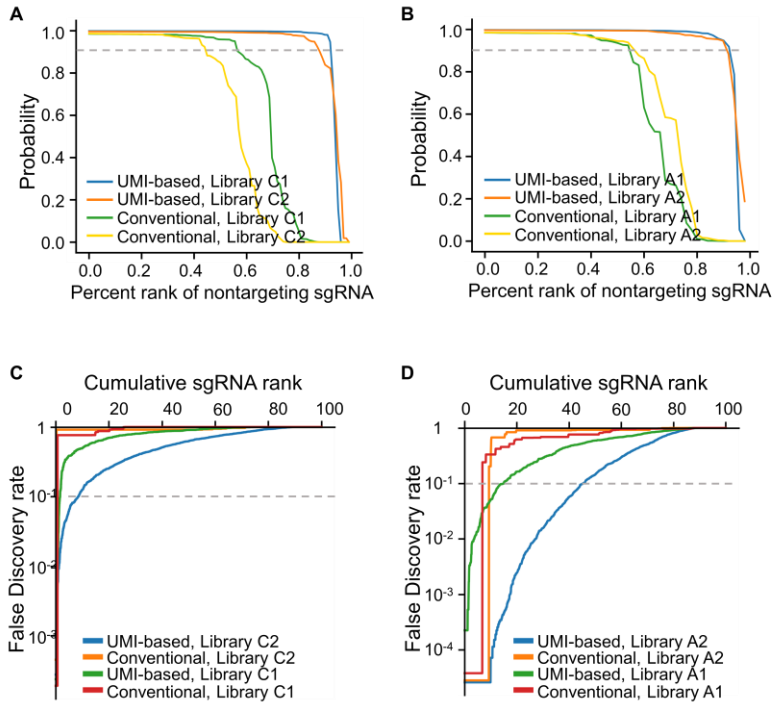


Figure 7. UMI-based analysis outperforms conventional analysis in the focused libraries. **(A and B)** Rank distributions of probabilities that nontargeting sgRNAs in the CBE focused libraries **(A)** and ABE focused libraries **(B)** were in cluster Y or ψ, produced by a Gaussian mixture model. The gray dashed lines indicate a probability value of 0.9. **(C and D)** Comparison between the false discovery rates (FDRs) of UMI-mediated analysis and conventional analysis for all sgRNAs in the CBE focused libraries **(C)** and ABE focused libraries **(D)**. The gray dashed lines were plotted at an FDR value of 0.1.

5. Functional classification of the sgRNA pairs in the focused libraries

We next compared the performance of high coverage focused library screening to that of library C or A screening by testing three sets of negative control target sequence and sgRNA pairs: nontargeting, synonymous, and nonessential. In addition to the aforementioned two negative control groups, we picked "nonessential control" sgRNAs from the original library that targeted a canonical nonessential gene set[54]. The focused libraries showed better performance than libraries C or A for discriminating the negative controls to cluster Y (**Figure 8 and 9**). For example, library C only recovered 64% of 92 overlapping nontargeting sgRNAs in C, C1, and C2, in contrast to C1 (95%) and C2 (88%) at a threshold of a 90% probability of belonging to cluster Y (**Figure 9B**), suggesting that screening of high coverage libraries could result in a higher power for detecting errors.

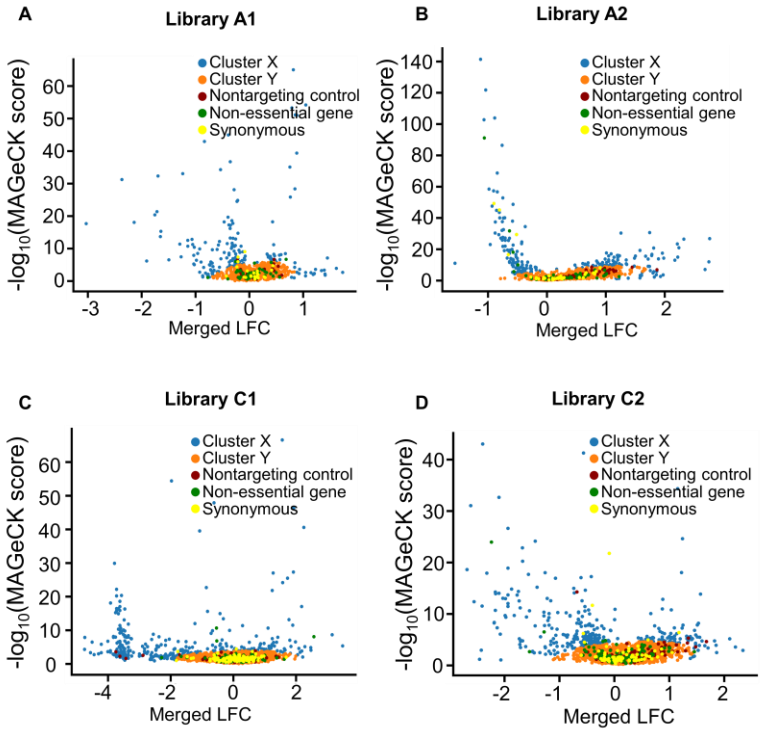
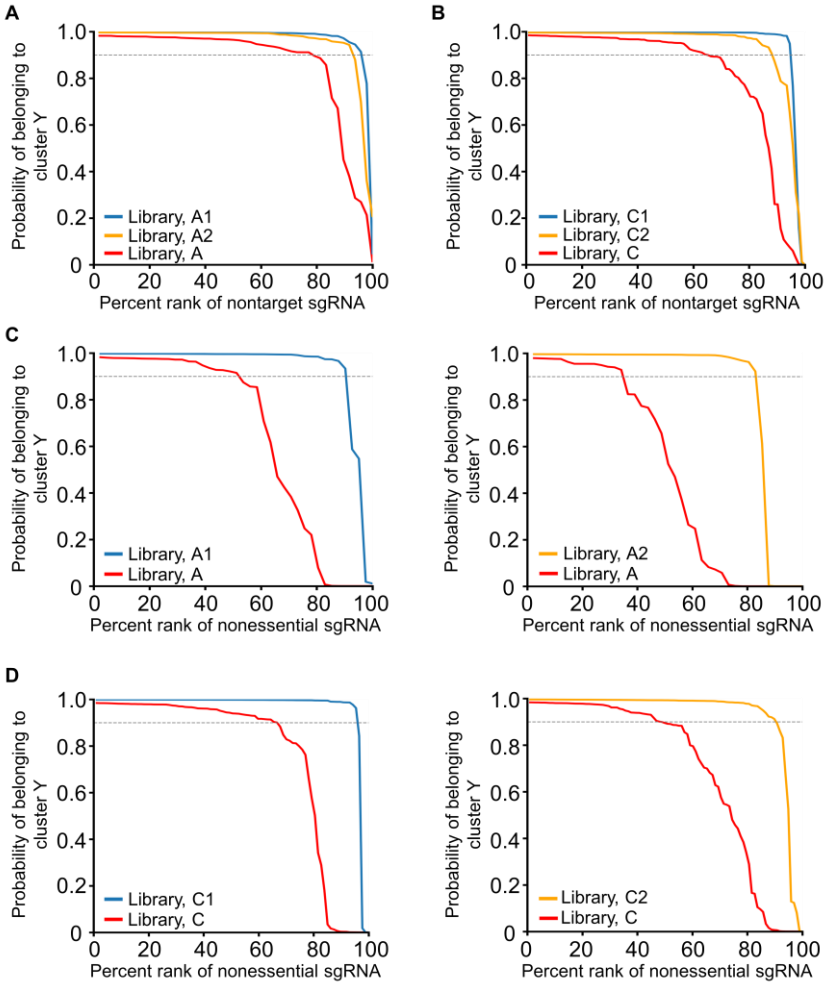


Figure 8. Clustering of cancer mutations in focused library. Volcano plots of UMI based analysis in focused library. Classification of target sequence and sgRNA pairs using a two component Gaussian mixture model in library A1(**A**), library A2 (**B**), library C1 (**C**), and library C2 (**D**). Along with nontargeting controls, (red dots) sgRNAs targeting non-essential gene sets (green) and synonymous sgRNAs (yellow) were plotted.



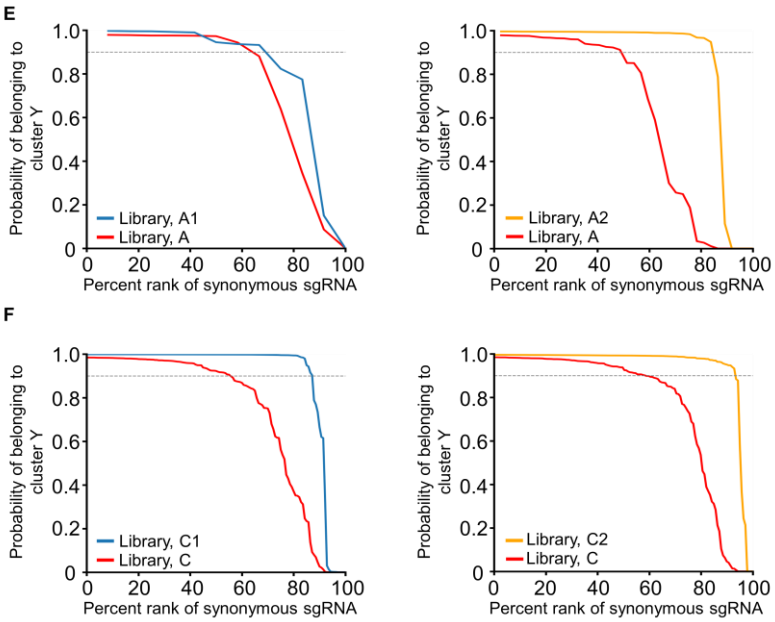


Figure 9. Focused libraries showed better performance than screening libraries for assigning negative controls to cluster Y. (**A-B**) Rank distributions of probabilities that overlapping nontargeting sgRNAs in the ABE focused libraries **(**N=49**, A)** and CBE focused libraries **(**N=92, **B)** were in cluster Y compared to those in libraries A and C, respectively. The gray dashed lines indicate a probability value of 0.9. **(C)** Rank distributions of probabilities that nonessential sgRNAs in libraries A1 (left, N=41) and A2 (right, N=41) were in cluster Y compared to those in library A. **(D)** Rank distributions of probabilities that nonessential sgRNAs in libraries C1 (left, N=87) and C2 (right, N=98) were in cluster Y compared to those in library C. (**E**) Rank distributions of probabilities that synonymous sgRNAs, which are predicted to cause mutations leading to synonymous amino acid changes, in libraries A1 (left, N=12) and A2 (right, N=37) were in cluster Y compared to those in library A. **(F)** Rank distributions of probabilities that synonymous sgRNAs in libraries C1 (left, N=140) and C2 (right, N=139) were in cluster Y compared to those in library C.

Next, we investigated the relationship between the base editing efficiency at integrated target sequences and phenotypic changes. As noted previously, base editing efficiencies at integrated target sequences show good correlations with those at endogenous target sequences[39]. Using 220 unique sgRNAs targeting 65 curated essential genes in the C2 library (see **Materials and Methods** for details), we plotted density plots of LFC vs. the nonsynonymous base editing efficiency (**Figure 10A**). We found robust depletion of cells harboring sgRNAs when the nonsynonymous base editing efficiency in the surrogate sequences was over 40%. Therefore, we assumed that sgRNAs associated with a less than 40% base editing efficiency in surrogate sequences resulted in incomplete base editing at the endogenous target sites. Thus, we filtered out those sgRNAs from further functional annotations. As noted previously, we found that most nontargeting sgRNAs, nonessential sgRNAs, and synonymous sgRNAs were included in cluster Y. In alignment with this result, we thought that sgRNAs with a high probability of belonging to cluster Y would have a minimal impact on proliferation. We tried to differentiate neutral mutations according to the probability of the sgRNAs belonging to cluster Y; 95%, 88%, 94%, and 92% of nontargeting pairs in libraries C1, C2, A1, and A2, respectively, had a probability of 90% or more of belonging to cluster Y (**Figures. 10A and 10B**). Therefore, we classified sgRNAs with a probability over 90% of belonging in cluster Y as ‘Neutral’. Considering that most mutations in the library were predicted to have minimal or no effects on proliferation, we tried to select pathogenic mutations, which have robust effects on proliferation, by using strict filtering conditions to exclude false positives. We examined the relationship between the probability of each mutation belonging in cluster X and the FDR (**Figure 10B, Figure 11A, 11D, 11G**) and found that sgRNAs with a probability over 99% of being in cluster X showed a much lower FDR value in all four focused libraries. Therefore, we classified sgRNAs with a probability of over 99% of being in cluster X as

'Pathogenic,' a category that we further divided into 'Outgrowing' and 'Depleting' according to the LFC. Finally, sgRNAs that were not classified as either 'Neutral' or 'Pathogenic' were classified as 'Likely Outgrowing' or 'Likely Depleting' according to the LFC.

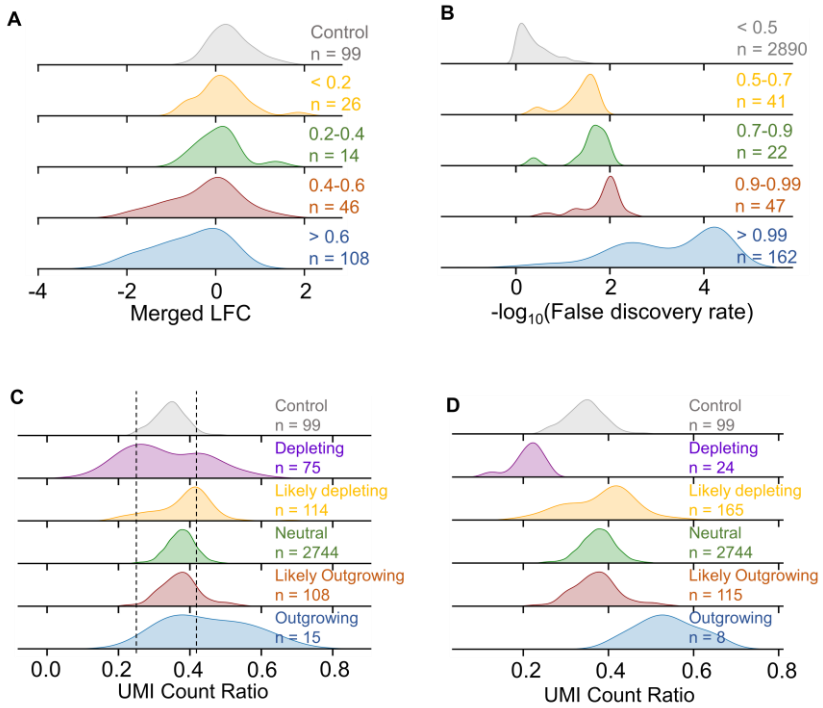


Figure 10. Functional annotations of mutations by UMI based clustering and surrogate sequence edit proportion. **(A)** LFC distributions of sgRNAs targeting essential genes according the percentage of nonsynonymous mutations in the integrated target sequence. **(B)** Density plots of -$\log_{10}$(FDR) of sgRNAs grouped by their probability of belonging to cluster Y in library C2. **(C)** Density plots of UMI count ratios grouped by functional classification in library C2. The dashed lines indicate the thresholds that were derived from two standard deviations of the UMI count ratio for nontargeting sgRNAs. **(D)** Density plots of UMI count ratios according to the final classifications.

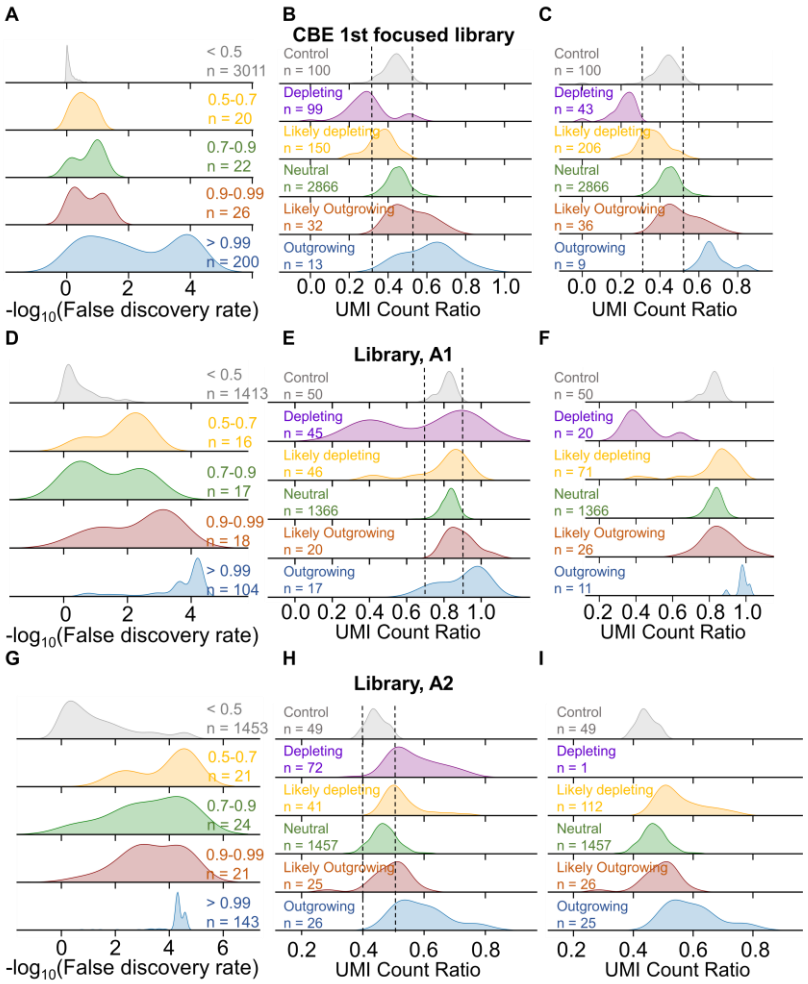


Figure 11. Functional annotations of mutations by UMI-based clustering and the proportion of edited surrogate sequences. **(A, D, G)** Density plots of $\log_{10}$(FDR) of sgRNAs grouped by the probability of belonging to cluster Y in libraries C1 (**A**), A1 (**D**), and A2 (**G**). (**B, E, H**) Density plots of the UMI count ratio grouped by functional classification in libraries C1 (**B**), A1 (**E**), and A2 (**H**). The dashed lines indicate the thresholds, which were derived from two standard deviations of the UMI count ratio for nontargeting sgRNAs in each library. (**C, F, I**) Density plots of the UMI count ratio according to the final classifications in libraries C1, (**C**) A1 (**F**), and A2 (**I**).

To further reduce false positives, we considered the UMI count ratio: the number of UMIs in the day 24 sample divided by the number of UMIs in the day 10 sample. Generally, the number of UMIs decreased over time due to dilution by the subculture process or selection pressure in the cell population If mutations that promote cell proliferation arise from base editing, UMIs from cells associated with such mutations would have a high probability of surviving as selection pressure occurs, and many clones would be maintained in the late phase of selection. As a result, there would be a high UMI count ratio for 'Outgrowing' compared to neutral mutations. Conversely, we presume that 'Depleting' mutations would have a low UMI count ratio. When comparing the UMI count between day 10 and day 24 (**Figure 10C, Figures. 11B, 11E, 11H**) in each group, we found that the UMI count ratio of nontarget controls was similar to that of Neutral sgRNAs. For example, in library C2, the mean UMI count ratios were 0.35 and 0.38 for nontargeting and neutral sgRNAs, respectively, suggesting that many sgRNA clones were eliminated during the screening period. However, we identified a number of sgRNAs in the 'Depleting' or 'Outgrowing' groups that showed a distinct pattern. For 32% of the sgRNAs in the 'Depleting' group, the UMI count ratio was less than the 5% rank of the UMI count ratio in the nontargeting control sgRNAs. In addition, for 53% of the sgRNAs in the 'Outgrowing' group, the UMI count ratio was greater than the 95% rank of the UMI count ratio in the nontargeting control sgRNAs (**Figure 10D**). This result indicated that sgRNAs inducing proliferative effects tended increase the number of clones that survived; in contrast, sgRNAs inducing anti-proliferative effects tended to increase the number that died. Similar results were found in the other three focused libraries (**Figures. 11B, 11C, 11E, 11F, 11H and 11I)**. Therefore, only sgRNAs for which the UMI counts were two standard deviations outside the range of nontargeting controls were considered as true hits and annotated as 'Outgrowing' or 'Depleting'.

Finally, we classified sgRNAs by scoring them according to the nonsynonymous editing rates in the surrogate target sequence, the GMM results, and the UMI count ratios (**Figure 12**). First, as previously noted, we eliminated those sgRNAs associated with a nonsynonymous editing efficiency of less than 40% in the surrogate target sequence. Next, we scored sgRNAs according to the GMM results and then calculated an average score for sgRNAs in two replicates (**Figure 12**). We classified sgRNAs with a score of -2 to -1.5 as 'Depleting', those with a score of -1.5 to -0.75 as 'Likely Depleting', those with a score of -0.5 to 0.5 as 'Neutral', those with a score of 0.75 or 1.5 as 'Likely Outgrowing', and those with a score of 1.5 to 2 as 'Outgrowing'. Finally, sgRNAs with scores that suggested that they were Depleting were finally defined as Depleting if their UMI count ratio was less than the UMI count ratio of 5% of the nontargeting sgRNAs; if this condition was not met, sgRNAs were re-classified as Likely Depleting. Similarly, sgRNAs with scores that suggested that they were Outgrowing were actually defined as Outgrowing only if their UMI count ratio was greater than the UMI count ratio of 95% of the nontarget sgRNAs; if not, the sgRNA was re-classified as Likely Outgrowing.

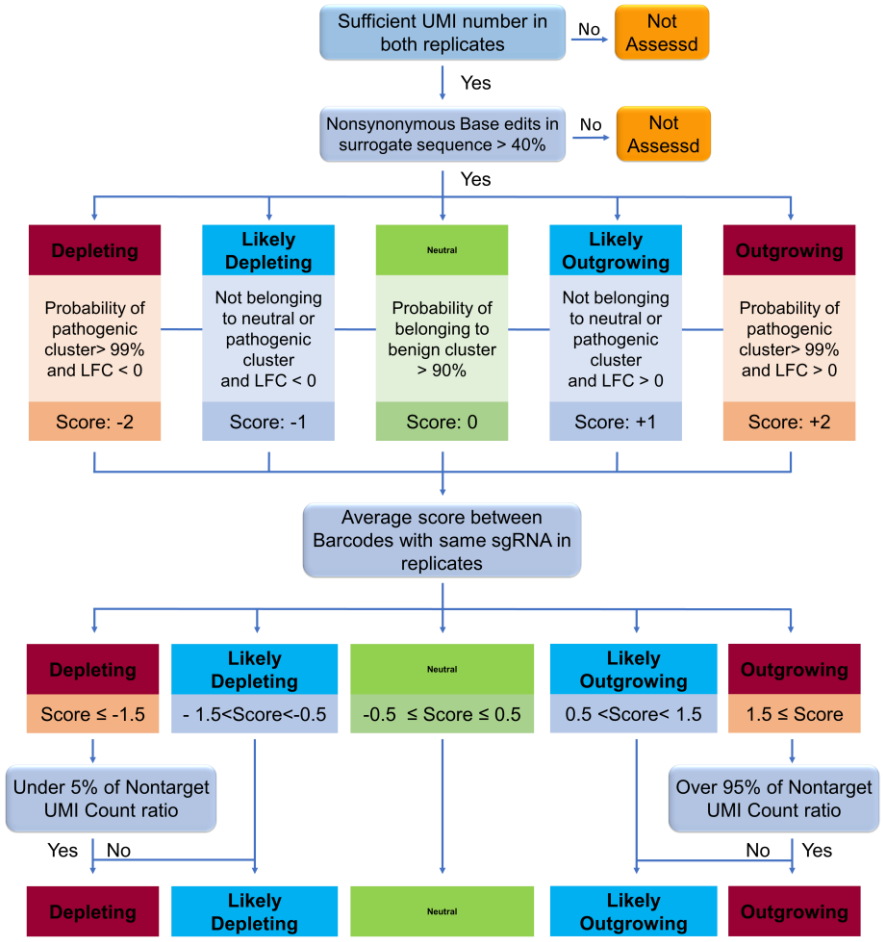


Figure 12. Flowchart of the functional scoring approach used in the focused libraries.

As a result of these steps, we classified each unique sgRNA into the defined clusters described above (**Figure 13**). As we expected, the majority of sgRNAs (63% to 77%) were placed into Neutral clusters, especially in focused library sets. We identified a small number of pathogenic sgRNAs that have robust effects on proliferation. Surprisingly, most mutations in 'Outgrowing' clusters represent well-known driver gene mutations that have been previously identified in multiple cancer types[55], affecting genes such as *TP53* and *KMT2C*; depleting mutations affected essential genes such as *CTCF*, *POLR1C*, and *SMARCB1* (**Figure 14A, 14B, 14C and 14D**). In particular, we identified sgRNAs targeting several TP53 variants (TP53_p.N239D and TP53_p.K120E) along with others targeting important residues such as p.T125, p.T155, p.Q192, and p.R280, although these sgRNAs caused expected amino acid changes that differed from those of previously identified mutations; mutations affecting these residues are among the 579 consensus driver mutations from the PanCancer dataset[5]. Additionally, we compared genes targeted by the depleting sgRNAs to the DepMap common essential gene set based on CRISPR knockout screening[56]. Among sgRNAs classified as Depleting, about 73% (24/35) in library C and 54% (7/13) in library A targeted DepMap common essential genes. In comparison, 17% of 'Likely Depleting' sgRNAs in library C and 4% in library A targeted common essential genes. These results suggest that our high-throughput platform effectively discriminates sgRNAs with either outgrowing or depleting functional consequences.

To determine whether the consequences of these genetic perturbations were consistent in different datasets, we compared correlations between LFCs associated with clustered mutations. LFCs between replicates showed high reproducibility in the Outgrowing and Depleting classes (R=0.84, 0.95, 0.89, and 0.93 in libraries C2, A1, C1, and A2, respectively) (**Figure 14E, 14F, 14G and 14H**). However, no correlations were seen between the LFCs of two

replicates in the Neutral cluster (**Figure 14I, 14J, 14K and 14L**), suggesting that the effects of genetic perturbations in the Neutral cluster were not sufficient to overcome the noise in the evaluation. However, outgrowing and depleting sgRNAs caused consistent changes in proliferation rates compared to neutral sgRNAs.

Next, we investigated whether the classification results obtained using the relatively low coverage screening library were consistent with those obtained using the high coverage focused libraries. First, we extracted mutations that were covered in both the screening and focused libraries. The majority of neutral sgRNAs in the screening library were annotated as belonging to the neutral cluster in the focused libraries (**Figure 15**). However, most outgrowing ("Outgrowing" and "Likely Outgrowing") and depleting ("Depleting" and "Likely Depleting") sgRNAs classified using the low coverage libraries were also annotated as belonging to the neutral cluster in the focused libraries, suggesting that low coverage screening was associated with high false positive rates, which is consistent with our previous findings. Therefore, we considered only neutral clusters in the screening libraries (library C or A) for further functional annotations of cancer-associated mutations.

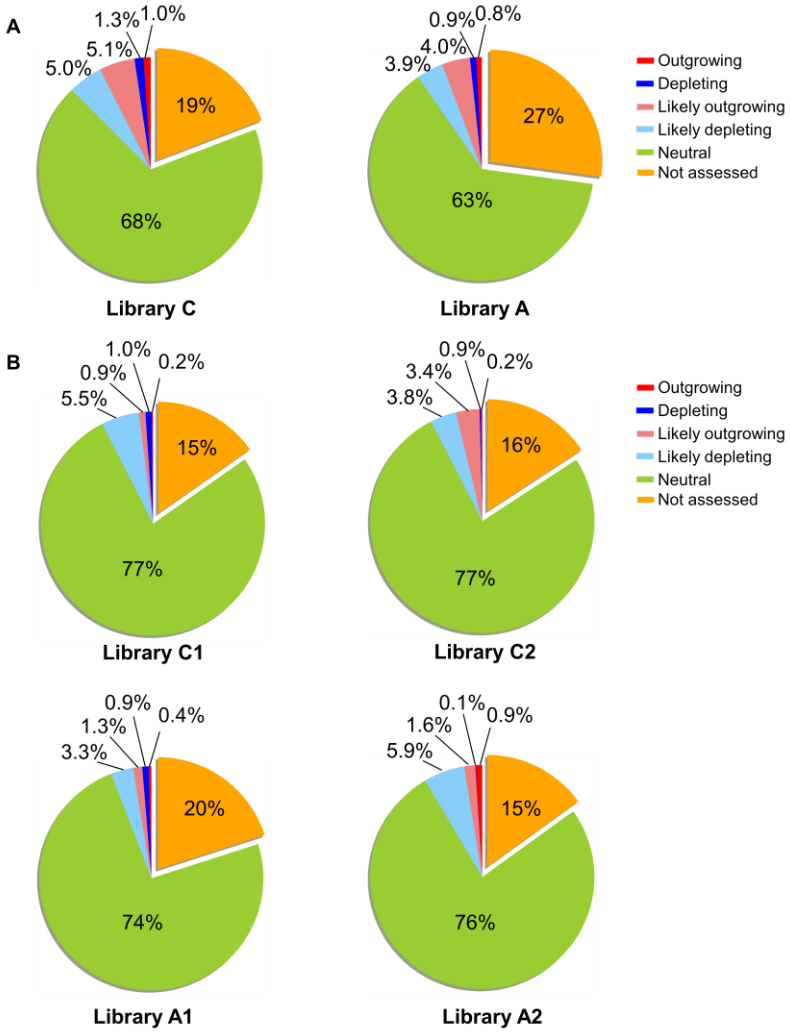


Figure 13. Final classification of sgRNAs in each library dataset. A representation of the functional classification of screened sgRNAs in the screening library datasets (A) and focused library datasets (B).

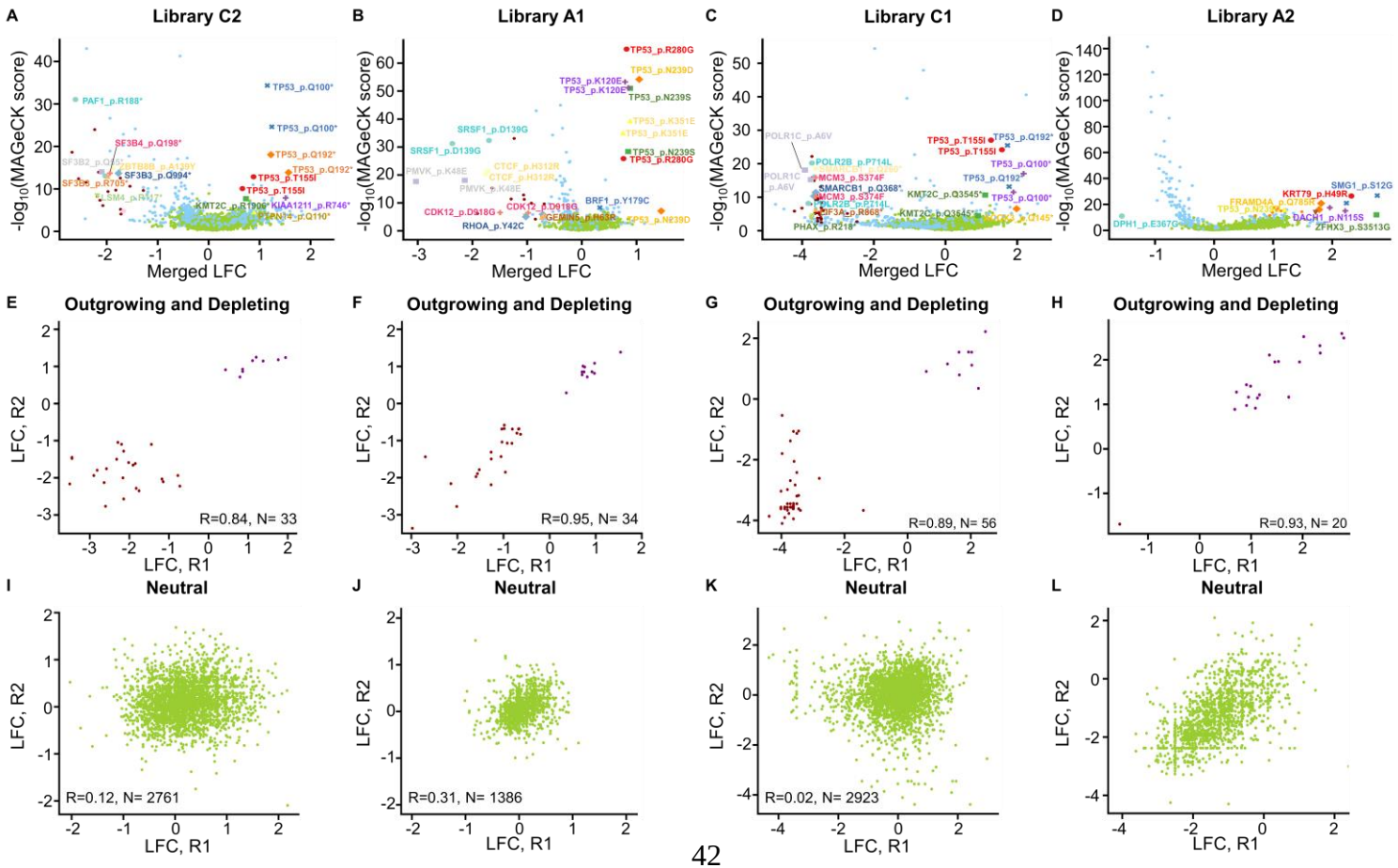


Figure 14. Classification of sgRNA in focused library identified candidate of mutations that affects proliferation. **(A-D)** Volcano plot of between LFC and -$\log_{10}$ MAGeCK score in library C2 (**A**), library A1 (**B**) library C1 (**C**) and library A2 (**D**). Functional clusters according to functional scoring result including 'Neutral' (green dots), 'likely depleting' or 'likely outgrowing' (light blue), 'Depleting' (red dots) were shown. All 'outgrowing' sgRNAs or representative sgRNAs in 'depleting' were shown separately. (**E-H**) The correlation of LFCs between two replicates in Outgrowing or Depleting sgRNAs in library C2 (**E**), library A1 (**F**), library C1 (**G**) and library A2 (**H**). (**I-L**) The correlation of LFCs between two replicates in Neutral sgRNAs in library C2 (**I**), library A1 (**J**), library C1 (**K**) and library A2 (**L**).

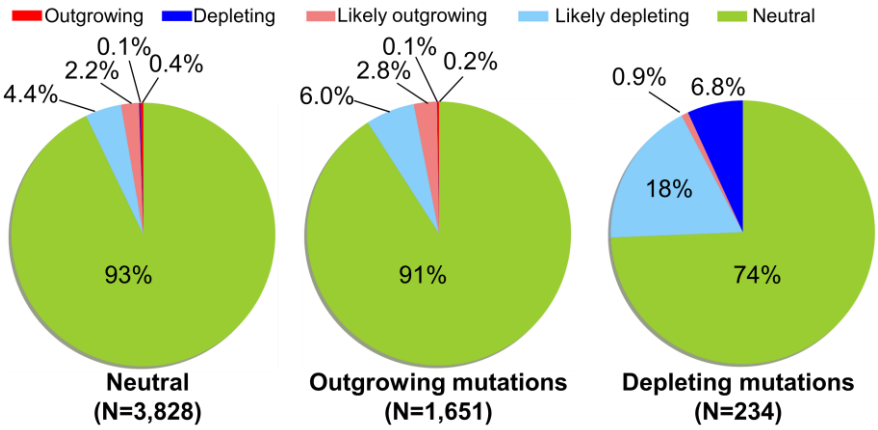


Figure 15. Validation of classifications made using the screening libraries. Overview of results in which each sgRNA that was clustered in the screening libraries was further classified in the focused libraries. sgRNAs from the CBE and ABE libraries were summed.

6. Functional annotation of cancer mutations

Finally, we attempted to investigate the functional consequences of individual cancer-associated mutations. Each sgRNA used in base editing could induce multiple bystander edits due to editable Cs or As existing within or nearby the activity window[39,53]. To translate sgRNA-based screening results to individual mutation-based results, we calculated the proportion of each base-edited outcome from the surrogate sequences at the initial timepoint (day 10), and sorted them according to their final amino acid outcomes. Only mutations causing amino acid changes that were harbored in more than 5% of the total reads were considered to be major editing outcomes. About 62% and 49% of the sgRNAs only induced one or two major editing outcomes in libraries C and A, respectively; however, 38% and 51% of the sgRNAs induced more than three major editing outcomes.

We integrated information from functional clustering of sgRNAs and major outcomes from surrogate sequences to determine which mutations were effectively screened in our experimental settings (**Figure 16A;** see **Materials and Methods** for details). We only considered mutations that represented at least 5% of the editing outcomes as effectively assessed. First, we calculated a Coverage score, which is determined based on how many datasets were effectively screened with targeted mutations as major editing outcomes. Mutations that were not effectively screened due to low UMI counts, read counts, or sgRNA-induced nonsynonymous mutation rates or a low proportion of sgRNA-induced editing outcomes were considered to be “Not assessed”. For mutations that were only assessed in the screening library (Coverage score 1), we classified each mutation into neutral, intermediate, or pathogenic (i.e., outgrowing or depleting) clusters according to the functional score from the sgRNA. Only mutations that were induced by sgRNAs in Neutral clusters with a Coverage score of 1 were considered likely to be neutral; other mutations were annotated as “Uncertain”, considering that there were high

frequencies of neutral mutations in the Outgrowing and Depleting clusters in the the screening library (**Figure 15**). For mutations assessed in a focused library (with a Coverage score of 2 or more), we annotated each mutation according to the functional scores of the sgRNAs obtained from the focused libraries.

Finally, we selected mutations effectively screened from each library and annotated their functional consequences on proliferation and survival using sgRNA-based functional scores (**Figure 16B**). We also provide the assessed function of VUS although such mutation-based annotations have high rates of "Not assessed" compared to sgRNA-based annotations (**Figure 13**) due to the presence of mutations that are less likely to be induced by base editing. From the mutations identified in COSMIC, we could annotate 46,350 (55%) and 13,722 (59%) of them with functional and Coverage scores in the CBE and ABE libraries, respectively, excluding mutations defined as 'Not Assessed' and 'Uncertain'.

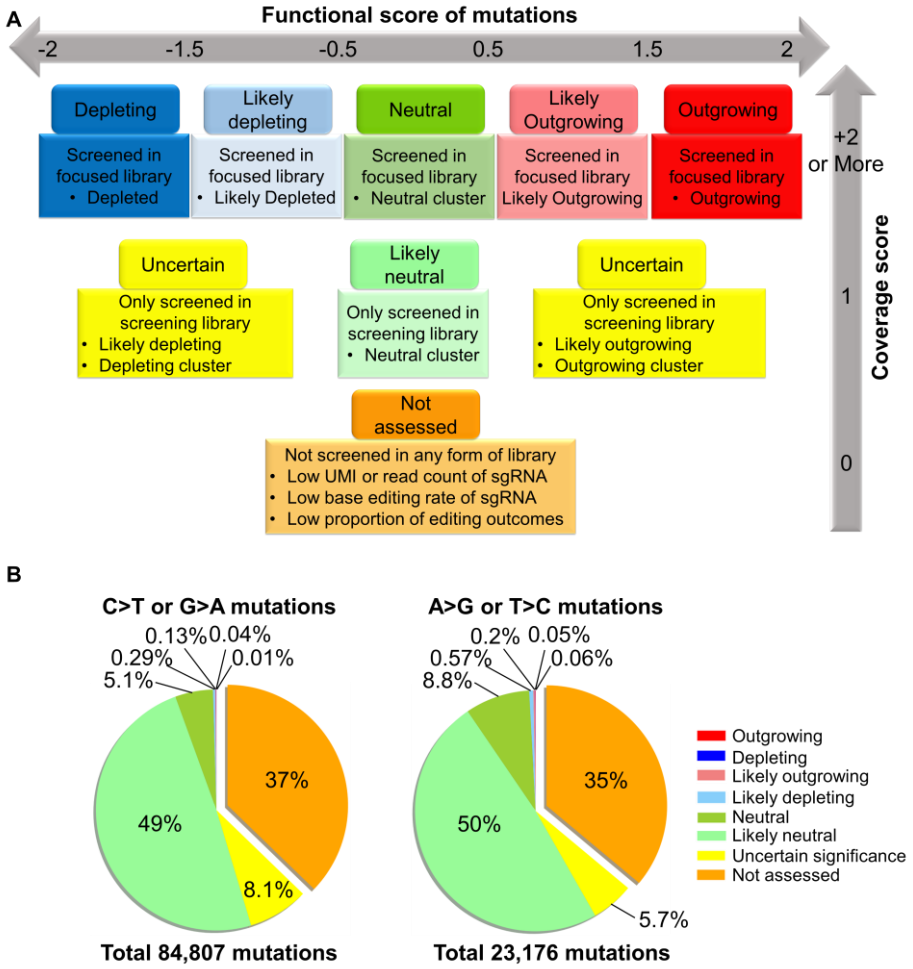


Figure 16. Functional annotations of cancer-associated mutations. **(A)** Characteristics of each classification category. (**B**) Final classification of cancer-associated mutations. Unlike sgRNA-based classifications, in this case mutations that were less likely to be induced by base editing were annotated 'Not assessed'.

7. Validation of library-based results by transduction of individual sgRNAs revealed amino acid changes that have substantial effects on proliferation.

Next, we sought to validate the effects of the expected individual amino acid changes induced by base editing. Generally, the effects of sgRNAs on proliferation were the sum of the effects of all induced amino acid changes. If the summed effects were neutral, all alleles induced by base editing could be considered to be neutral. In the case of Outgrowing or Depleting sgRNAs, specific alleles harboring pathogenic mutations would cause the cells to outgrow or be depleted relative to other cells in the population. To corroborate whether sgRNA-induced pathogenic alleles would cause outgrowing or depleting effects on proliferation, we tracked the proportion of sgRNA-induced mutations causing amino acid changes at endogenous loci by transducing individual sgRNAs into cells (**Figure 17A**).

We selected the sgRNAs associated with the most substantial effects from the “Outgrowing” or “Depleting” categories in the high-throughput evaluation, and validated their functional consequences. First, we tracked allele frequencies by target-specific deep sequencing of the endogenous genomic region after individually delivering sgRNAs into P-C or P-A cells via lentiviral transduction. Cells were selected with puromycin treatment for 48 hours and maintained with doxycycline for an additional 7 days to induce base editor expression. At 10 days post-infection, the cells were seeded in duplicate and cultured for an additional 2 weeks. The cells were harvested at 6, 10, 17, and 24 days post-infection, and subjected to deep sequencing to track the frequency of individual alleles longitudinally (**Figures 17B-D, Figures 18-19**). Representative Outgrowing sgRNAs induced high proportions of mutated alleles, along with multiple patterns of amino acid changes; the proportion of wild-type alleles gradually decreased over time (**Figure 17B**).

In contrast, wild type alleles in cells transduced with Depleting sgRNAs gradually increased after removal of doxycycline at 10 days post-infection(**Figure 17C, 17D and Figure 19**), suggesting edited alleles showed robust depleting phenotype. For comparison, we picked two sgRNAs that were targeting common essential genes but annotated as Neutral (ABESg.POLE_p.Y1889C and ABESg.ACTL6A_p.T405A). In case of CBESg. POLR1C_p.A6V and CBESg.POLR2B_p.P714L, the expected alleles were not observed due to presence of bystander C in active target window. In particular, CBESg.POLR1C_p.A6V induced stop codons by converting glutamine at position 5. In addition, we found another sgRNA (CBESg.POLG_p.Q1029*) that induces stop codons in *POLG* gene, which were not common essential gene in current DepMap dataset, suggesting *POLG* gene is essential for survival and proliferation at least in our experimental setting.

We assumed that the fold change in the frequency of each allele could be explained by the sum of the effect of the allele frequency change induced by the activity of the base editor and the effect of competition between each allele, considering possible background activity of tetracycline dependent regulatory system without doxycycline[57]. In the case of depleting mutations, the effect of base editing would be to reduce the frequency of wild-type alleles, but cells containing wild-type alleles would overcome the base editing effect and proliferate compared to cells containing mutated alleles if the mutated alleles showed sufficient depleting effects on proliferation. However, in the case of outgrowing mutations, both base editing effects and allele-specific effects on proliferation would act in the direction of reducing the proportion of wild-type alleles. Therefore, we also performed competitive proliferative assays[58] to compare the proliferation of sgRNA-transduced cells and non-transduced cells. P-A or P-C cells were infected by lentivirus harboring an empty sgRNA cassette and the puromycin resistance-p2A-GFP

gene fusion and cultured independently with cells harboring the sgRNA-of-interest (**Figure 17A**). We calculated the relative enrichment of the sgRNA-transduced cells by comparing the ratio of GFP^+ cells to GFP^- cells at each timepoint to the ratio at the "background" timepoint (3 days after cells were mixed) (see **Materials and Methods** for details). Longitudinal flow cytometric assays revealed that several *TP53*-targeting sgRNAs (CBEsg.TP53_p.T155I, ABEsg.TP53_p.K120E, and ABEsg.TP53_p.N239D) along with sgRNAs that introduced nonsense mutations in *TP53* (CBEsg.TP53_p.Q100* and CBEsg.TP53_p.Q192*) caused robust proliferative changes compared to sgRNAs targeting *GRP52* (P-A cells) and *OTX2* (P-C cells), which were curated as nonessential genes[54] (**Figure 17E**), suggesting that our high-throughput evaluation enables us to discriminate SNVs that affect the rate of cell proliferation. Also, representative Depleting sgRNAs (CBESg.POLR2B_p.P714L, CBESg.POLR1C_p.A6V and ABESg.SRSF1_p.D139G) caused robust depletion relative to the control GFP+ cells, suggesting our high-throughput evaluation successfully identified sgRNAs with definite effects on cellular proliferation.

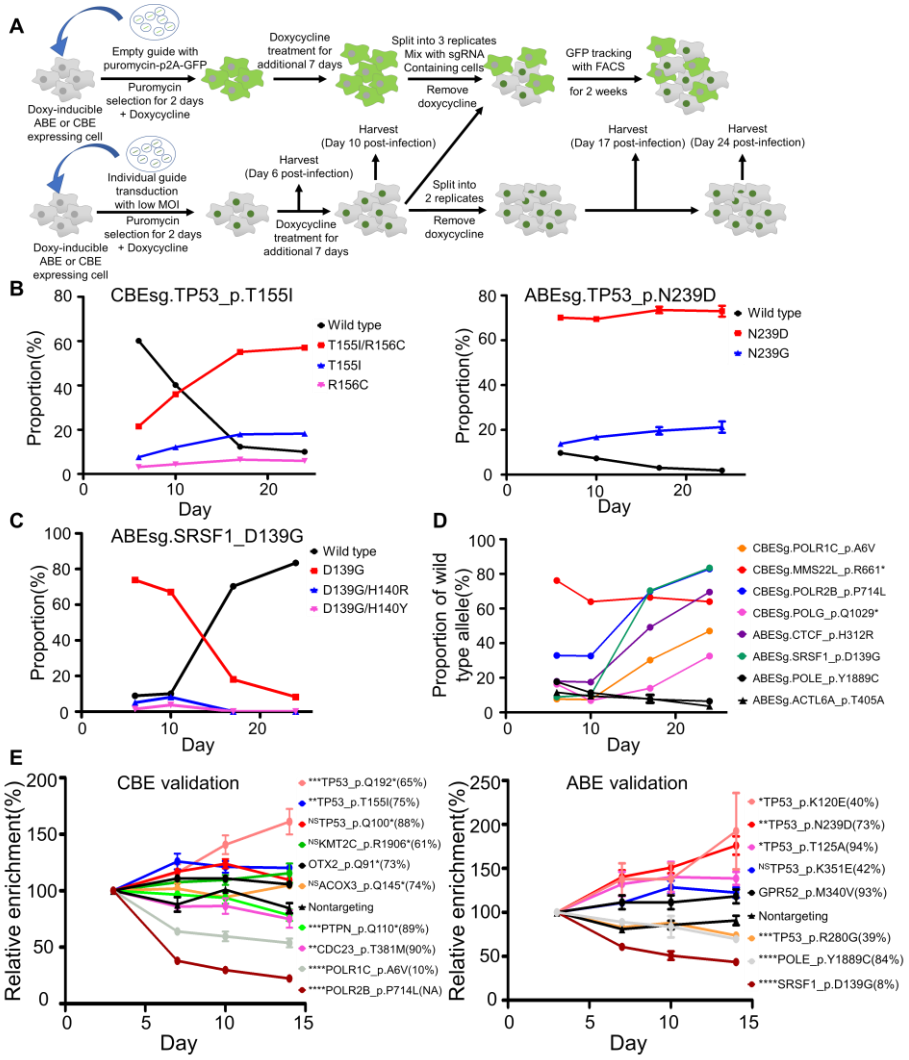


Figure 17. Validation of library-based results for sgRNAs associated with the most substantial effects on proliferation. (**A**) Schematic of validation experiments, including a competitive proliferation assay (upper panel) and allele frequency tracking (lower panel), for individual sgRNAs. (**B-C**) The relative proportions of various patterns of amino acid changes caused by representative Outgrowing (**B**) and Depleting (**C**) sgRNAs were analyzed and tracked longitudinally by targeted deep sequencing. The experiments were

performed in duplicate and an average proportion was calculated for each pattern. Error bars represent the standard deviation of two replicates. (**D**) The relative proportion of wild type alleles caused by representative Depleting sgRNAs were analyzed and tracked. Representative neutral sgRNAs were marked as black dots. The experiments were performed in duplicate and an average proportion was calculated for each pattern. Error bars represent the standard deviation of two replicates. (**E**) The relative enrichment of transduced cells tracked by flow cytometry. Day 3 after sgRNA-expressing cells were mixed with GFP-expressing cells was assigned as the reference timepoint. The intended sgRNA-induced mutation patterns are indicated together with their proportions. Error bars represent the standard deviation of three replicates. Student's t test was performed under the null hypothesis that the relative enrichment of sgRNAs would be the same as that of nonessential sgRNAs (OTX2_Q91* for CBE and GPR52_M340V for ABE). The enrichment value of nontargeting control sgRNAs were obtained using the mean of the relative enrichment of two nontargeting sgRNAs. (CBESg.Nontargeting_87 and CBESg.Nontargeting_116) (Left) ***P = 0.001 (TP53_Q192*), **P = 0.002 (TP53_T155I) ^{NS}P = 0.104 (TP53_Q100*), ^{NS}P =0.125 (KMT2C_R1906*), ^{NS}P =0.729 (ACOX3_Q145*), ***P=0.001 (PTPN14_Q110*), **P=0.003 (CDC23_T381M), ****P=$3.6X10^{-5}$ (POLR1C_A6V) and ****P=$2.2X10^{-6}$ (POLR2B_P714L). (Right) **P = 0.002 (TP53_N239D), *P = 0.043 (TP53_K120E), *P = 0.031 (TP53_T125A), ^{NS}P = 0.492 (TP53_K351E), ***P=0.001 (TP53_R280G), ****P=$4.4X10^{-4}$ (POLE_Y1889C) and ****P=$8.4X10^{-5}$ (SRSF1_D139G).

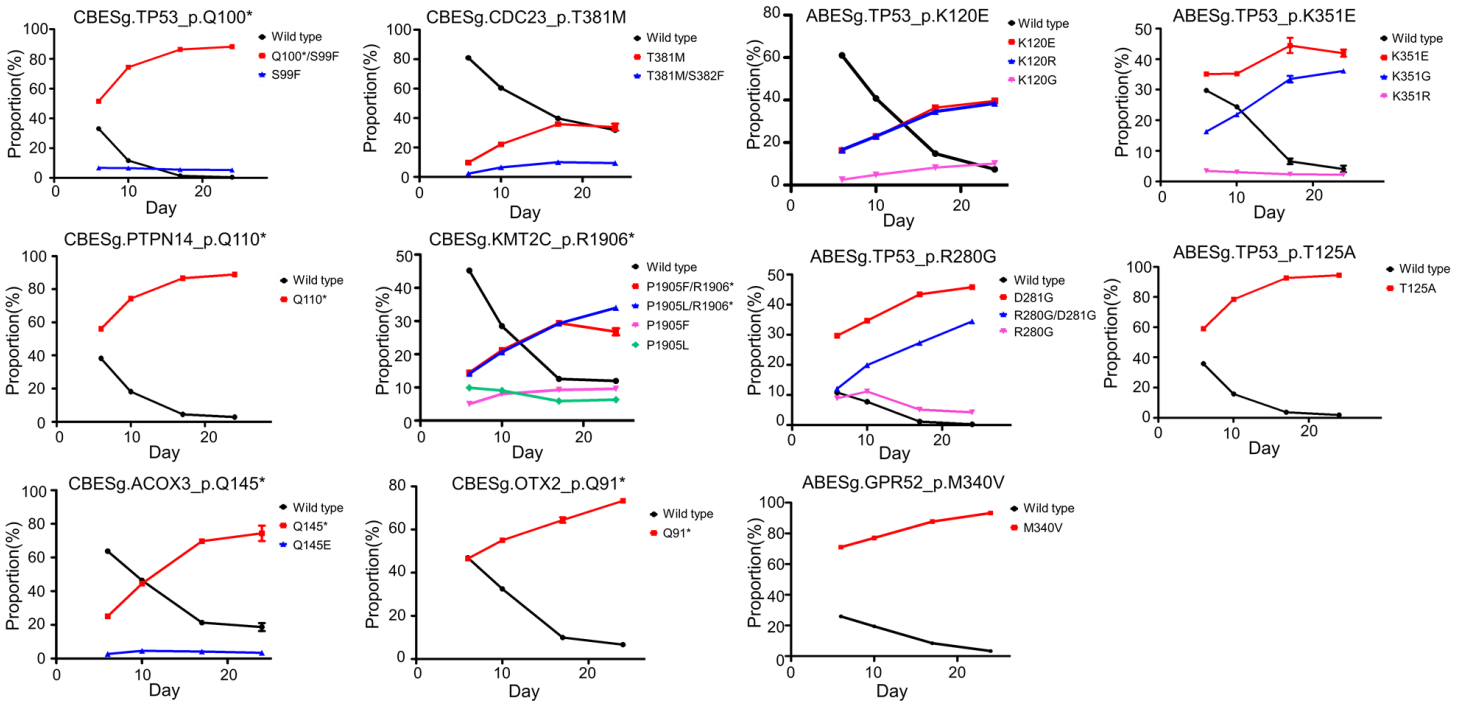


Figure 18. Allele frequency tracking for Outgrowing sgRNAs. Longitudinal changes in the allele frequencies induced by representative sgRNAs classified as “Outgrowing” and nonessential control sgRNAs (CBESg.OTX2_p.Q91* and ABESg.p.GPR52_M340V). The experiments were performed in duplicate and an average proportion was calculated for each pattern. Error bars represent the standard deviation of two replicates.

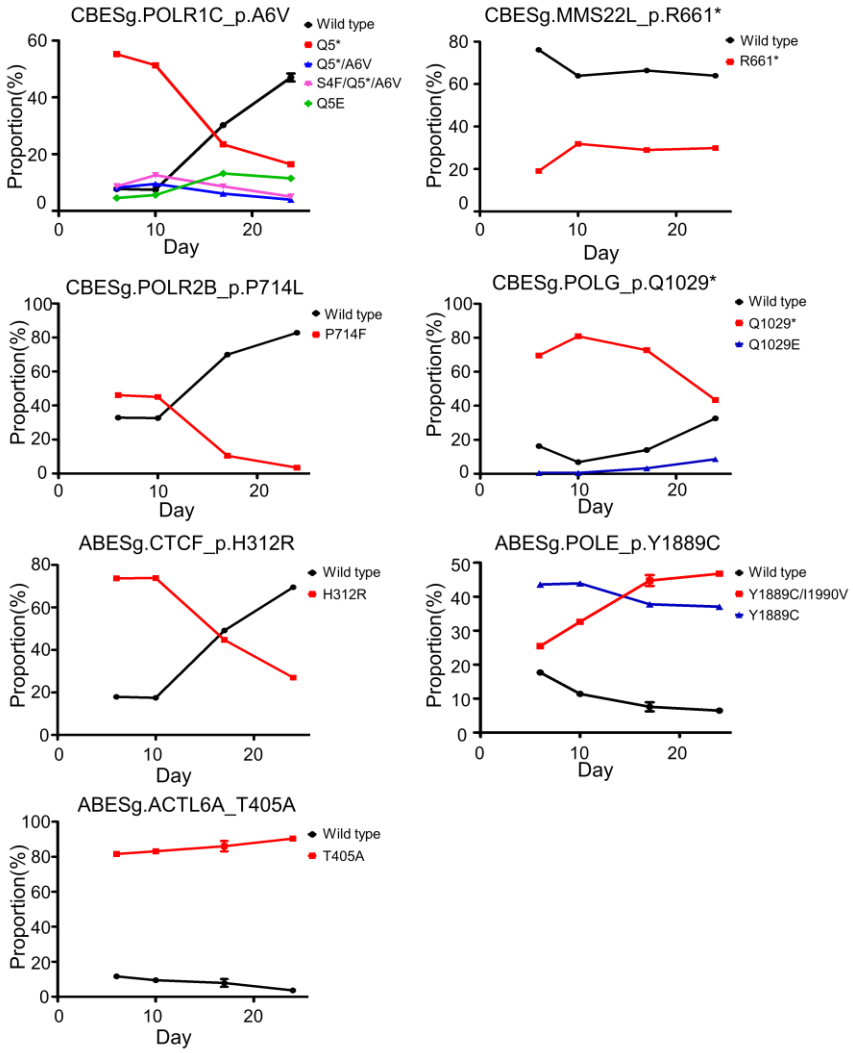


Figure 19. Allele frequency tracking for Depleting sgRNAs. Longitudinal changes in the allele frequencies induced by representative sgRNAs classified as “Depleting” and Neutral control sgRNAs (ABESg.POLE_p.Y1889C and ABESg.p.ACTL6A_T405A). The experiments were performed in duplicate and an average proportion was calculated for each pattern. Error bars represent the standard deviation of two replicates.

For a number of sgRNAs with substantial effects on proliferation in the competitive assay, we identified several base editing-induced amino acid variant outcomes using targeted deep sequencing of the endogenous locus. For example, CBEsg.TP53_p.T155I, ABEsg.TP53_p.K120E, and ABEsg.TP53_p.N239D caused various patterns of amino acid changes that were present in different proportions. In the case of neutral mutations, we assumed that the outcome of base editing in the surrogate target sequences would mainly be affected by the activity of the base editor itself and would be well correlated with that in the endogenous locus. However, in the case of Outgrowing or Depleting mutations, the proportions of different base editor-induced alleles might be biased due to competition between alleles. To determine which amino acid changes have a significant effect on proliferation, we normalized the fold change of each allele in its endogenous genomic site with that of the surrogate sequence to estimate the effect of proliferation on each allele while excluding the influence of the base editor to calculate an adjusted LFC (aLFC; see **Methods** for details). We compared LFCs and aLFCs for all amino acid changes (**Figure 20**), and found that aLFCs successfully discriminated the key changes that affect proliferation. For example, TP53_p.K120E, TP53_p.K120R, and TP53_p.K120G were similarly enriched in endogenous alleles, but TP53_p.K120G showed less pronounced effects on proliferation after adjusting for base editor effects. Additionally, whereas both TP53_p.N239D and TP53_p.239G variants became enriched, only TP53_p.N239D was found to play a key role in the enriching phenotype after adjusting for base editor effects.

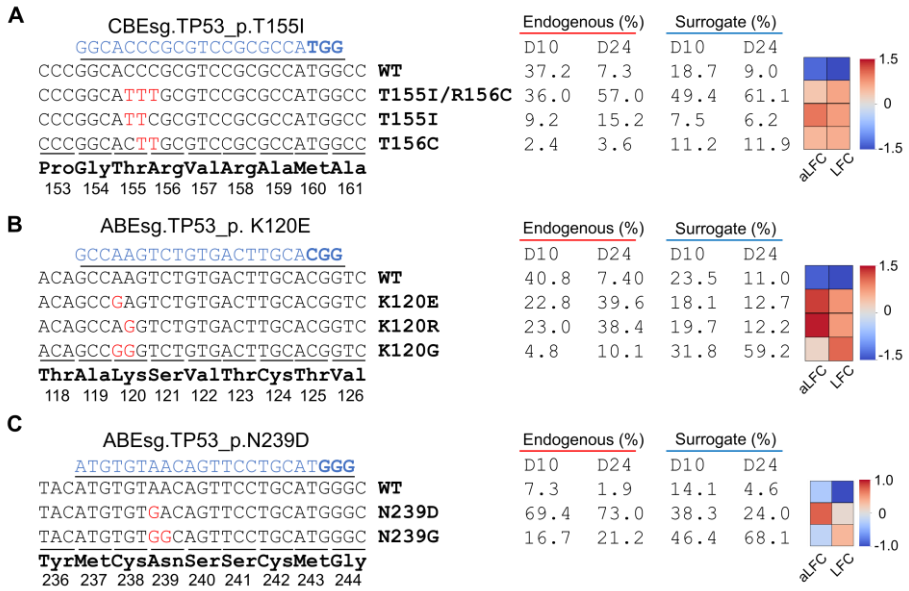


Figure 20. Fold change normalization of the allele frequency in the endogenous sequence relative to that in surrogate sequences to reveal phenotypes caused by individual mutations.(**Left panel**) Genomic sequences and frequencies of amino acid changes induced by CBEsg.TP53_p.T155I (**A**), ABEsg.TP53_p.K120E (**B**), and ABEsg.TP53_p.N239D (**C**). The positions of the sgRNA (blue) and PAM (blue/bold) are shown at the top of the sequence alignments. Only alleles that were present at a frequency of at least 1% of the total reads are shown. (**Right panel**) The proportion of alleles encoding each amino acid change in the endogenous genomic sequence and integrated target sequences in focused library screening induced by CBEsg.TP53_p.T155I (**A**), ABEsg.TP53_p.K120E (**B**), and ABEsg.TP53_p.N239D (**C**). aLFCs and LFCs of each allele are shown in the heatmaps.

IV. DISCUSSION

We have presented the most comprehensive methods for assessment of somatic cancer mutations using base editors. Direct functional assessments of human somatic cancer mutations revealed several mutations that affect cell proliferation and a large number of mutations that have a neutral effect in the cell context in which our experiments were performed. Many of them were *TP53*-related mutations; our approach might have been particularly sensitive to such mutations because *TP53* is repressed in the cell line we used here. In addition, we identified Depleting mutations, which generally were loss-of-function variants of common essential genes.

In contrast to CRISPR knockout screening, base editors can generate multiple alleles, which might affect proliferation differently; as a result, phenotypic effects might be weaker than those seen with a knockout screening system. In addition, to determine whether individual mutations affect proliferation, we must differentiate between mutations that are “effectively edited with low effects on proliferation” from mutations that are “not effectively edited with low effects on proliferation”; if the proportion of an intended allele among all alleles is too low, its effects on proliferation may be masked by the other alleles. Therefore, we introduced surrogate reporter sequences to estimate the proportion of each intended mutation; analysis with sgRNAs designed to induce stop codons in common essential genes revealed that effects on cellular proliferation are induced only when a substantial proportion of sequences have undergone base editing (**Figure 10A**). In addition, we classified mutations present at low frequency (<5% of total reads) in surrogate sequences as “Not assessed”.

Additionally, we overcome one limitation of conventional screening by using UMIs to analyze sgRNAs at the subclone level. As described previously[51], UMI-based analysis can differentiate outlier clones that disproportionally proliferate and dominate the total read space compared to other clones

harboring the same sgRNA. This strategy was especially useful for reducing false positives in our evaluation system in which only a few mutations result in a phenotype different from wild type, whereas most mutations are not associated with a distinct phenotype. In addition, subclonal analysis allowed us to track subclones in a bulk population of sgRNA-sorted reads, thereby enhancing the statistical power to detect pathogenic sgRNAs.

Although we designed our library to include most mutations that could be generated by base editors, we restricted possible SNVs to those located in the window spanning position 4 to 7 relative to the PAM, which is the canonical base editor target window[53] to maximize editing efficiency. This requirement for a nearby PAM restricted our library design to cover only 5.9% of C•G to T•A conversions for CBE and 6.2% of A•T to G•C conversions for ABE among the total missense and nonsense mutations in the COSMIC database, and to 26% of the C•G to T•A mutations and 28% of the A•T to G•C mutations of pathogenic SNVs reported in ClinVar[59]. Base editors constructed from other CRISPR-associated nucleases such as SaCas9[60], Cas12a[61], or engineered Cas9 variants with broadened PAM compatibility[62] together with high-activity base editors[63-65] would expand the targeting scope of our high-throughput evaluation.

Recently, functional assessment of more than ten thousand human genetic variants using base editors was attempted[40]. However, the experiments in this study involved a conventional CRISPR base editor scheme: one-sgRNA-per-one-variant, which resulted in enormous errors in differentiating negative controls and an insignificant false discovery rate (**Figure 7**) even in very high coverage (over 3000 cells per sgRNA) screening. In addition, these authors used DNA damaging reagents in their screen, which could induce biased results. Additionally, the screen was performed using a cancerous cell line, which could mask the effects of driver mutation candidates due to the presence of strong putative driver mutations already

found in cancer. Finally, only CBEs were used in this screen; our study, to the best of our knowledge, is the first large-scale high-throughput screen to use ABE.

V. CONCLUSION

Identifying causal mutations that drives tumor evolution is a central key to cancer genomics. Here, we showed that massive parallel screening of more than tens of thousands of cancer mutations revealed probable candidate of mutations that affect the proliferation of cells in single nucleotide resolution by base editors. Also, in contrast to conventional CRISPR screening that used multiple sgRNAs to reduce false positive rates and negative rates, base editor screening only used one sgRNA to induce desired nucleotide changes. These limitations inevitably cause many false positive and negative errors that makes difficult to distinguish between true hits and outlier clones. Our UMI mediated screening reduces errors by dividing one sgRNA transduced clones into many subclones, thereby greatly increases screening power. Further, we tried unbiased classification that could identify neutral mutations and pathogenic mutations, which could functional annotations of many variants that were previously unknown for their function in proliferation. Our screening platform using base editors should facilitate the cancer genomics study by identifying functional consequences of individual variants and provide tools on developing new candidate therapeutic options on cancer.

REFERENCES

1. McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, M. Mastrogianakis G, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061-8.
2. Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature 2010;464:993-8.
3. Campbell PJ, Getz G, Korbel JO, Stuart JM, Jennings JL, Stein LD, et al. Pan-cancer analysis of whole genomes. Nature 2020;578:82-93.
4. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature Reviews Cancer 2018;18:696-705.
5. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell 2018;173:371-85.e18.
6. Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. Nature 2020;578:102-11.
7. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009;458:719.
8. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature 2013;502:333-9.
9. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer Genome Landscapes. Science 2013;339:1546-58.

10. Niu B, Scott AD, Sengupta S, Bailey MH, Batra P, Ning J, et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat Genet 2016;48:827-37.
11. Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, et al. Identification of cancer driver genes based on nucleotide context. Nat Genet 2020;52:208-18.
12. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214-8.
13. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 2012;366:883-92.
14. Yachida S, Jones S, Bozic I, Antal T, Leary R, Fu B, et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. Nature 2010;467:1114-7.
15. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 2018;15:81-94.
16. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 2016;44:D862-8.
17. Porta-Pardo E, Kamburov A, Tamborero D, Pons T, Grases D, Valencia A, et al. Comparison of algorithms for the detection of cancer drivers at subgene resolution. Nat Methods 2017;14:782-8.
18. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nature Methods 2014;11:801-7.
19. Giacomelli AO, Yang X, Lintner RE, McFarland JM, Duby M, Kim J, et al. Mutational processes shape the landscape of TP53 mutations in human cancer. Nat Genet 2018;50:1381-7.

20. Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, et al. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. Mol Cell 2018;71:178-90.e8.
21. Majithia AR, Tsuda B, Agostini M, Gnanapradeepan K, Rice R, Peloso G, et al. Prospective functional classification of all possible missense variants in PPARG. Nat Genet 2016;48:1570-5.
22. Brenan L, Andreev A, Cohen O, Pantel S, Kamburov A, Cacchiarelli D, et al. Phenotypic Characterization of a Comprehensive Set of MAPK1/ERK2 Missense Mutants. Cell Rep 2016;17:1171-83.
23. Kim H, Kim JS. A guide to genome engineering with programmable nucleases. Nat Rev Genet 2014;15:321-34.
24. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature 2018;562:217-22.
25. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 2016;533:420-4.
26. Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, et al. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. Nature 2017;551:464-71.
27. Ramirez RD, Sheridan S, Girard L, Sato M, Kim Y, Pollack J, et al. Immortalization of human bronchial epithelial cells in the absence of viral oncoproteins. Cancer Res 2004;64:9027-34.
28. Kim HS, Mendiratta S, Kim J, Pecot CV, Larsen JE, Zubovych I, et al. Systematic identification of molecular subtype-selective vulnerabilities in non-small-cell lung cancer. Cell 2013;155:552-66.
29. Ellis BL, Potts PR, Porteus MH. Creating higher titer lentivirus with caffeine. Hum Gene Ther 2011;22:93-100.

30. Kim N, Kim HK, Lee S, Seo JH, Choi JW, Park J, et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. Nat Biotechnol 2020;38:1328-36.
31. Dang Y, Jia G, Choi J, Ma H, Anaya E, Ye C, et al. Optimizing sgRNA structure to improve CRISPR-Cas9 knockout efficiency. Genome Biol 2015;16:280.
32. Shalem O, Sanjana NE, Hartenian E, Shi X, Scott DA, Mikkelsen TS, et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. Science 2014;343:84-7.
33. Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res 2017;27:491-9.
34. Li W, Xu H, Xiao T, Cong L, Love MI, Zhang F, et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. Genome Biol 2014;15:554.
35. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, et al. High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities. Cell 2015;163:1515-26.
36. Reynolds DA. Gaussian Mixture Models. Encyclopedia of biometrics 2009;741.
37. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British journal of cancer 2004;91:355-8.
38. Kim HK, Kim Y, Lee S, Min S, Bae JY, Choi JW, et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. Sci Adv 2019;5:eaax9249.

39. Song M, Kim HK, Lee S, Kim Y, Seo SY, Park J, et al. Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. Nat Biotechnol 2020;38:1037-43.
40. Hanna RE, Hegde M, Fagre CR, DeWeirdt PC, Sangree AK, Szegletes Z, et al. Massively parallel assessment of human variants with base editor screens. bioRxiv 2020; doi:10.1101/2020.05.17.100818.2020.05.17.100818.
41. Kuscu C, Parlak M, Tufan T, Yang J, Szlachta K, Wei X, et al. CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. Nat Methods 2017;14:710-2.
42. Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat Biotechnol 2014;32:1262-7.
43. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol 2016;34:184-91.
44. Koblan LW, Doman JL, Wilson C, Levy JM, Tay T, Newby GA, et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. Nat Biotechnol 2018;36:843-6.
45. Kim HK, Song M, Lee J, Menon AV, Jung S, Kang YM, et al. In vivo high-throughput profiling of CRISPR-Cpf1 activity. Nat Methods 2017;14:153-9.
46. Kim HK, Min S, Song M, Jung S, Choi JW, Kim Y, et al. Deep learning improves prediction of CRISPR-Cpf1 guide RNA activity. Nat Biotechnol 2018;36:239-41.

47. Kim HK, Lee S, Kim Y, Park J, Min S, Choi JW, 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 2020;4:111-24.
48. Kim HK, Yu G, Park J, Min S, Lee S, Yoon S, et al. Predicting the efficiency of prime editing guide RNAs in human cells. Nat Biotechnol 2020; doi:10.1038/s41587-020-0677-y.
49. Allen F, Crepaldi L, Alsinet C, Strong AJ, Kleshchevnikov V, De Angeli P, et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat Biotechnol 2018;37:64-72.
50. Arbab M, Shen MW, Mok B, Wilson C, Matuszek Z, Cassa CA, et al. Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning. Cell 2020;182:463-80 e30.
51. Michlits G, Hubmann M, Wu SH, Vainorius G, Budusan E, Zhuk S, et al. CRISPR-UMI: single-cell lineage tracing of pooled CRISPR-Cas9 screens. Nat Methods 2017;14:1191-7.
52. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet 2015;16:299-311.
53. Rees HA, Liu DR. Base editing: precision chemistry on the genome and transcriptome of living cells. Nat Rev Genet 2018;19:770-88.
54. Hart T, Brown KR, Sircoulomb F, Rottapel R, Moffat J. Measuring error rates in genomic perturbation screens: gold standards for human functional genomics. Mol Syst Biol 2014;10:733.
55. Martinez-Jimenez F, Muinos F, Sentis I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer 2020;20:555-72.
56. Institute DaB. DepMap: The Cancer Dependency Map Project at Broad Institute. Available at: https://depmap.org/portal/ [Accessed Oct 30, 2020

57. Markusic D, Oude-Elferink R, Das AT, Berkhout B, Seppen J. Comparison of single regulated lentiviral vectors with rtTA expression driven by an autoregulatory loop or a constitutive promoter. Nucleic Acids Research 2005;33:e63-e.
58. Eekels JJM, Pasternak AO, Schut AM, Geerts D, Jeeninga RE, Berkhout B. A competitive cell growth assay for the detection of subtle effects of gene transduction on cell proliferation. Gene Therapy 2012;19:1058-64.
59. Hu JH, Miller SM, Geurts MH, Tang W, Chen L, Sun N, et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature 2018;556:57-63.
60. Kim YB, Komor AC, Levy JM, Packer MS, Zhao KT, Liu DR. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. Nat Biotechnol 2017;35:371-6.
61. Li X, Wang Y, Liu Y, Yang B, Wang X, Wei J, et al. Base editing with a Cpf1-cytidine deaminase fusion. Nat Biotechnol 2018;36:324-7.
62. Nishimasu H, Shi X, Ishiguro S, Gao L, Hirano S, Okazaki S, et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science 2018;361:1259-62.
63. Thuronyi BW, Koblan LW, Levy JM, Yeh WH, Zheng C, Newby GA, et al. Continuous evolution of base editors with expanded target compatibility and improved activity. Nat Biotechnol 2019;37:1070-9.
64. Richter MF, Zhao KT, Eton E, Lapinaite A, Newby GA, Thuronyi BW, et al. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. Nature Biotechnology 2020;38:883-91.
65. Gaudelli NM, Lam DK, Rees HA, Sola-Esteves NM, Barrera LA, Born DA, et al. Directed evolution of adenine base editors with

increased activity and therapeutic application. Nat Biotechnol 2020;38:892-900.

ABSTRACT(IN KOREAN)

염기변환자를 이용한 인간 종양 단일염기변이의 대량산출 기능평가

<지도교수 김 형 범 >

연세대학교 대학원 의과학과

김 영 광

종양 발생에 직접적으로 연관된 체세포 변이 (발암 변이, driver mutation)를 찾아내는 것은 종양발생의 과정을 이해하거나 치료제 개발에서 중요한 과제임. 본 연구에서는 크리스퍼 염기변환자와 고유 분자 표지자 (Unique molecular identifier, UMI)를 이용한 가이드 RNA 라이브러리를 이용하여 인간 종양에서 발견되는 점돌연변이변이를 유도하여 기능을 평가할 수 있는 새로운 대량산출 스크리닝 방법을 개발하였음.

본 연구에서는 83,731개의 C>T 혹은 G>A 변이와 23,613개의 A>G 혹은 T>C 변이 (Single nucleotide variants)를 유도할 수 있는 guide RNA (sgRNA) 라이브러리를 통해 107,952개의 인간 종양 단일염기변이를 유도하였음. 이를 이용한 실험과 이후 높은 coverage를 가지는 소규모의 라이브러리 실험을 통해 세포의 증식과 생존에 유리한 혹은 불리한 효과를 가지는 변이를 발굴하였음. 본 연구에서는 고유 분자 표지자를 이용한 본 방법이 기존의 변이 하나당 하나의 guide RNA를 사용하는 방식보다 변이를 발굴하는데 더 우수함을 보였음. 본 연구에서 제시한 스크리닝 방법은 단일 염기 수준의 종양 세포 변이의 기능변화를 구함으로써 종양 유전학에서 유용한 도구로 사용될 수 있으며, 이를 통해 종양에서의 새로운 치료 방법을 발굴할 수 있을 것으로 기대됨.

핵심되는 말 : 크리스퍼/카스9, 염기변환자, 유전체 스크리닝, 체세포 변이, 불확실형 변이형

PUBLICATION LIST

1. Kim HK*, **Kim Y***, Lee S, Min S, Bae JY, Choi JW, et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. Sci Adv 2019;5:eaax9249. (**co-first author**)
2. Song M, Kim HK, Lee S, **Kim Y**, Seo S-Y, Park J, et al. Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. Nature Biotechnology 2020.