nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-01172-3

Base editing sensor libraries for high-throughput engineering and functional analysis of cancer-associated single nucleotide variants

Francisco J. Sánchez-Rivera[1,10,11,14], Bianca J. Diaz[2,3,14], Edward R. Kastenhuber[1,2,14], Henri Schmidt[5], Alyna Katti[2,3], Margaret Kennedy[1,4], Vincent Tem[1], Yu-Jui Ho[1], Josef Leibold[1,12,13], Stella V. Paffenholz[1,4], Francisco M. Barriga[1], Kevan Chu[2,3], Sukanya Goswami[2], Alexandra N. Wuest[1], Janelle M. Simon[1], Kaloyan M. Tsanov[1], Debyani Chakravarty[7,8], Hongxin Zhang[7], Christina S. Leslie[5], Scott W. Lowe[1,6] and Lukas E. Dow[2,3,9] ✉

Base editing can be applied to characterize single nucleotide variants of unknown function, yet defining effective combinations of single guide RNAs (sgRNAs) and base editors remains challenging. Here, we describe modular base-editing-activity 'sensors' that link sgRNAs and cognate target sites in cis and use them to systematically measure the editing efficiency and precision of thousands of sgRNAs paired with functionally distinct base editors. By quantifying sensor editing across >200,000 editor-sgRNA combinations, we provide a comprehensive resource of sgRNAs for introducing and interrogating cancer-associated single nucleotide variants in multiple model systems. We demonstrate that sensor-validated tools streamline production of in vivo cancer models and that integrating sensor modules in pooled sgRNA libraries can aid interpretation of high-throughput base editing screens. Using this approach, we identify several previously uncharacterized mutant *TP53* alleles as drivers of cancer cell proliferation and in vivo tumor development. We anticipate that the framework described here will facilitate the functional interrogation of cancer variants in cell and animal models.

Genome sequencing studies have revealed a complex, heterogeneous mix of cancer-associated mutations, including both known and druggable oncogenic mutations (for example, BRAF-V600E), and a large collection of variants of uncertain significance (VUS). Understanding the impact of specific oncogenic mutations requires functional analysis. Even subtle changes in cancer-associated single nucleotide variants (SNVs) can have important functional consequences in tumorigenesis and drug sensitivity[1–6]. Thus, while DNA sequencing has enormous potential to support clinical decision making, it is limited by a lack of understanding of how specific variants contribute to disease.

Base editing (BE) can introduce SNVs with high specificity and in the absence of DNA double-strand breaks (DSBs) or exogenous DNA templates[7]. We and others have developed BE tools that enable efficient BE in cell lines, primary cells and in vivo[8–10]. However, unlike Cas9-mediated DNA targeting of DSBs, predicting the efficiency and precision of BE guides remains challenging.

To expand the capability and feasibility of studying VUS at scale, we set out to develop a framework for systematic engineering of thousands of cancer-associated genetic variants. To do this, we developed a modular 'BE sensor' platform that couples a sgRNA with its cognate genomic target in cis. Thus, in the presence of a base editor, sgRNAs drive editing of a physically linked surrogate, or 'sensor' target site. We find that sensor-based measurement of editing efficiency correlates closely with endogenous gene targeting and that sensor-validated sgRNAs can be used to streamline the engineering and characterization of cancer-associated SNVs in vivo. Further, integrated sensors support the interpretation of pooled BE library screens by providing a surrogate readout of sgRNA activity in parallel to sgRNA abundance or 'screen fitness'.

To aid the development of future mutation-focused sensor libraries, we developed a flexible computational pipeline (annotated mutation-informed nucleotide editing sgRNA search (AMINEsearch)) that generates BE sensor libraries from annotated genomic data, and a web application (BE-SCAN) that simplifies selection of effective BE tools for generating cancer-associated SNVs. We expect the resources described here will accelerate the functional interrogation of VUS.

Results

Development and validation of a BE sensor. To measure the activity of individual sgRNAs in a high-throughput manner, we designed

[1]Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Sandra and Edward Meyer Cancer Center, Department of Medicine, Weill Cornell Medicine, New York, NY, USA. [3]Weill Cornell Graduate School of Medical Sciences, Weill Cornell Medicine, New York, NY, USA. [4]Louis V. Gerstner Jr. Graduate School of Biomedical Sciences, New York, NY, USA. [5]Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [6]Howard Hughes Medical Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [7]Kravis Center for Molecular Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [8]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [9]Department of Medicine, Weill Cornell Medicine, New York, NY, USA. [10]Present address: David H. Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [11]Present address: Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [12]Present address: Department of Medical Oncology and Pneumology, University Hospital Tuebingen, Tuebingen, Germany. [13]Present address: iFIT Cluster of Excellence EXC 2180 'Image-Guided and Functionally Instructed Tumor Therapies', University of Tuebingen, Tuebingen, Germany. [14]These authors contributed equally: Francisco J. Sánchez-Rivera, Bianca J. Diaz, Edward R. Kastenhuber. ✉e-mail: lud2005@med.cornell.edu

a BE sensor, in which a sgRNA is linked to its cognate target site in a lentiviral vector, allowing for high-throughput measurement of editing efficiency by PCR amplification and sequencing of the sensor cassette (Fig. 1a). To test whether BE sensors could measure qualitative and quantitative features of BE across different targets and editors, we generated a library containing ten human and eight mouse sgRNAs, in which sensor target sites were modified to contain all 64 possible three-nucleotide protospacer adjacent motifs (PAMs) (ALL-PAM (AP) library; Fig. 1a and Supplementary Table 1a). We next generated MDA-MB-231 cells that stably expressed one of nine base editors that span a range of PAM specificities, editing window sizes and overall editing efficiency. These include FNLS[10], AncBE4max[9], FNLS-2X (F2X)[10], FNLS-HF1 (HF1)[10,11], FNLS-HiFi (HiFi)[12,13], FNLS-NG (NG)[12,14], FNLS-HiFi-NG (HiFi-NG)[12,14], FNLS-VQR (VQR)[11] and xFNLS[10,15] (Fig. 1a). We also generated Cas9 and Cas9-NG nuclease controls to assess the frequency of SNVs following DSBs. As expected, apart from Cas9 and Cas9-NG, each line showed efficient BE activity as measured by a fluorescent reporter[12] (Fig. 1b). We next transduced each base editor- or Cas9-expressing-line with the AP library in duplicate at >2,000× representation and cultured cells for 1 week to allow BE. We amplified and sequenced entire sgRNA-scaffold-target cassettes from each cell population and quantified insertion or deletion (indel) frequency and target cytosine editing (Supplementary Table 1b). All lines showed high correlation of C > T editing efficiency between replicates (Supplementary Fig. 1), and, as expected, Cas9 and Cas9-NG showed indel formation but little-to-no target C > T editing (Fig. 1c, Supplementary Figs. 2 and 3). The sensor assay accurately reported the relative efficiency and known PAM preferences of individual base editors (Fig. 1c). FNLS, AncBE4max, F2X, HF1, HiFi and xFNLS had maximum C > T editing at NGG PAMs and lower, but detectable, editing at NAG and NGA PAMs. Consistent with previous publications, VQR showed higher C > T editing at NGA PAMs[11], while FNLS-NG and FNLS-HiFi-NG showed broad editing capabilities at NGN PAMs[12,14] (Fig. 1c and Extended Data Fig. 1). In general, high-fidelity variants showed editing patterns identical to parental base editors, albeit with overall lower efficiency (Fig. 1d). Together, these data show that BE sensor libraries reliably report known features of well-characterized base editors.

AMINEsearch generates BE sensor libraries from genomics data. To establish a flexible pipeline to facilitate BE screens driven by clinical genomics data, we developed AMINEsearch (annotated mutation-informed nucleotide BE sgRNA search)—a BE sgRNA design algorithm that compiles ready-to-clone libraries of annotated sgRNAs to model user-defined mutations (Fig. 2a). We first implemented AMINEsearch to generate sensor libraries to model cancer-associated mutations derived from targeted sequencing data (MSK-IMPACT)[16], providing deep coverage of 462 cancer-relevant genes in >21,000 tumors at the time of library generation (Fig. 2b). BE is well suited to the creation of cancer-associated alterations, as such alterations are highly enriched for C to T to G to A transition mutations (Fig. 2c). Most BE-compatible SNVs were missense mutations, followed by nonsense and splice site alterations (Fig. 2d). We identified 2,608 SNVs as recurrent (at least four occurrences), with mutation frequency ranging from 0.02% to 5.1% (Fig. 2e).

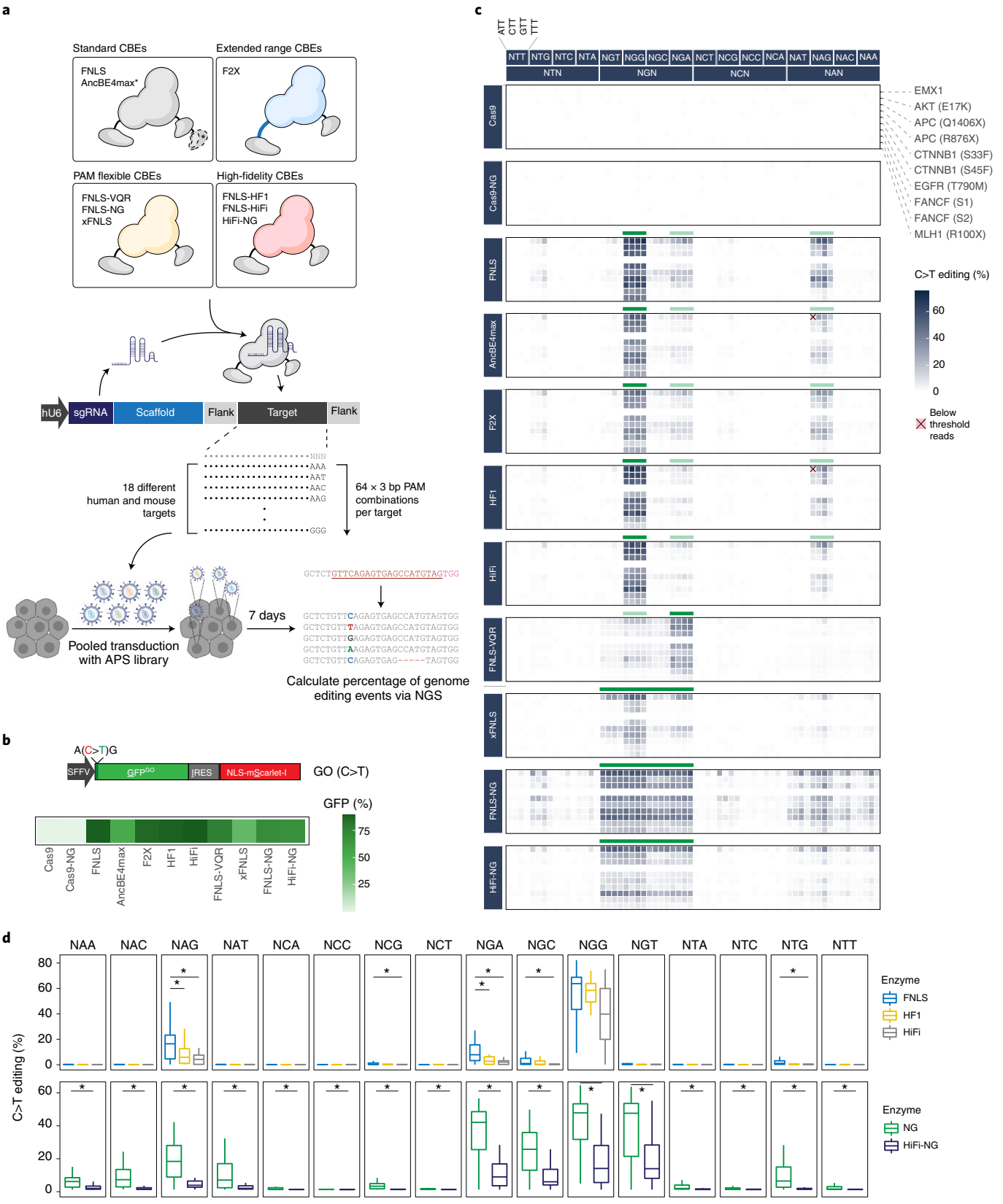
By inputting the parameters of well-characterized, efficient Cas modules (SpCas9, Cas9-NG, xCas9 and ScCas9) and a BE targeting window of 4–11 bp, we identified 5,855 sgRNAs covering 1,450 unique mutations. This human BE sensor (HBES) library represented ~56% of all recurrent mutations in the dataset (Supplementary Table 2c–d). While the MSK-IMPACT targeted sequencing assay is designed to focus specifically on known cancer-associated genes (Fig. 2f), cross-referencing with the OncoKB precision oncology knowledge base (https://sop.oncokb.org/)[17,18] showed that the plurality of SNVs targeted in HBES are VUS, with the most frequent subset of mutations being enriched in variant-level annotation (Fig. 2g and Extended Data Fig. 2).

To model cancer-associated mutations in the mouse genome, we included further steps in the AMINEsearch workflow to identify orthologous murine sites. For simplicity, unless otherwise stated, we refer to sgRNAs using the human mutation nomenclature. The mouse BE sensor (MBES) library contained 4,686 sgRNAs targeting 1,177 unique mutations (Supplementary Tables 2c–f). We noted modest attrition of BE sgRNAs when targeting the mouse genome due to lack of sequence conservation (Fig. 2d,f,h). The diversity of available BE tools allows for distinct tradeoffs. For instance, using base editors with an expanded editing window (F2X) or PAM flexibility (FNLS-NG) increases theoretical coverage (Fig. 2h) at the expected cost of reduced local specificity (Fig. 2i) or potentially increased global offtarget effects (Extended Data Fig. 3 and Supplementary Table 3), respectively.

BE sensor identifies optimal sgRNAs for engineering variants. To measure editing efficiency of sgRNAs in the HBES and MBES libraries, we transduced each in duplicate into MDA-MB-231 cells expressing one of three base editors: FNLS[10] (highest editing efficiency), F2X[10] (expanded editing range) or FNLS-NG[12,14] (PAM flexibility) (Fig. 3a and Supplementary Tables 4 and 5). Cas9-expressing cells served as a control to measure baseline C to T transitions in the absence of BE. Base editors were expressed to approximately equal amounts, while Cas9 showed higher protein abundance (Supplementary Fig. 4a). We observed excellent correlation of cytosine editing between replicates from each library and base editor combination (Supplementary Fig. 5). As observed in AP screens, editing efficiency was influenced strongly by PAM, target cytosine position and dinucleotide sequence context (TC > AC = CC > GC) (Fig. 3a,b and Supplementary Fig. 6). Consistent with previous studies, FNLS and FNLS-NG showed maximum editing efficiency at positions 3 to 9 of the protospacer, whereas F2X showed expanded editing at positions 3 to 13 (ref. [10]) (Extended Data Fig. 4). While F2X had an extended editing range, its average efficiency in the canonical editing window (3–9 bp) was lower than that of FNLS (Supplementary Fig. 7). As expected, FNLS and F2X had maximum editing efficiency at NGG PAMs, whereas FNLS-NG showed broad activity at NGN PAMs (Fig. 3a). Cas9 showed no detectable BE activity, with ≤0.1% C > T editing across all base editor-PAM combinations (Fig. 3a,b and Supplementary Fig. 6).

Fig. 1 | A high-throughput sensor assay to characterize BE outcomes at thousands of target sites. a, Schematic of the sensor assay. A sgRNA is paired with its cognate target site in cis such that editing outcomes can be assessed quantitatively in a massively parallel fashion using NGS. Here, we illustrate the design of an APS library that queries 18 target sites with all 64 possible PAM combinations upon lentiviral integration into cells expressing a range of base editors and cultured for 7 days followed by gDNA isolation and screen deconvolution via NGS. **b**, Schematic of the GO BE reporter (top) used to confirm the activity of each of the nine base editors used in APS screens by measuring C > T-dependent induction of GFP expression in mScarlet-infected cells (bottom). **c**, C > T editing efficiency at each AP library human target site across the full range of base editors ($n = 9$). Cas9 and Cas9-NG serve as nuclease controls. Rows denote target sites. Columns denote PAM subclass. See also Extended Data Fig. 1 and Supplementary Figs. 2 and 3. **d**, Head-to-head comparison of C > T editing efficiency at different PAMs by 'standard' CBEs (top row) and PAM flexible CBEs (bottom row). $^*P \leq 0.01$. *P* values were determined with two-sided Wilcoxon signed rank test. Boxplots show the median and interquartile range (IQR) and whiskers represent 1.5 × IQR.

To determine whether sensor editing scores identified in one cell line could be extrapolated to other cell types, we repeated HBES and MBES screens in four more cell lines: human PC9 and murine *Kras*^{G12D};*Trp53*$^{-/-}$ mutant (KPT1) lung adenocarcinomas[19], as well as immortalized NIH3T3 and *Kras*$^{G12D/+}$;*Trp53*$^{WT/WT}$ pancreatic ductal epithelial cells (PDECs)[20]. In all, we measured editing across >200,000 base editor-sgRNA-cell line combinations. Each cell line showed high concordance between replicates and different base editors (Supplementary Fig. 5). Average editing efficiency varied by cell line; however, PAM specificity, editing range and relative



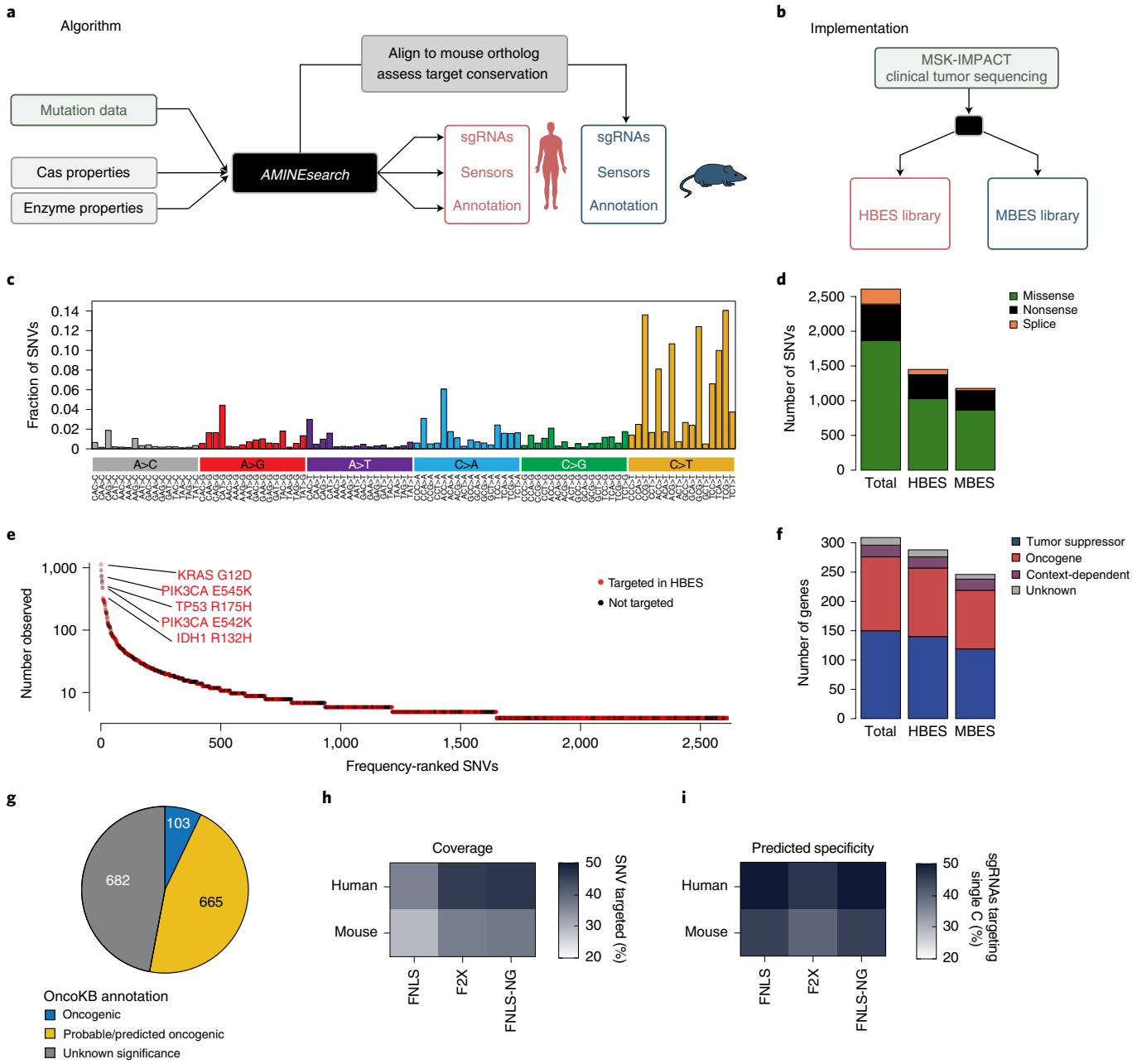


Fig. 2 | AMINEsearch: a versatile computational pipeline to identify cancer-associated mutations compatible with BE. a, AMINEsearch pipeline. Provided a *maf* mutation file, Cas protein properties (for example, PAM usage) and base editor editing ranges, AMINEsearch produces libraries of human and mouse sgRNAs and sensor constructs designed to engineer specific mutations using BE. **b**, Deploying AMINEsearch to analyze tumor mutation data from cancer patients profiled with the MSK-IMPACT clinical DNA sequencing platform. HBES and MBES libraries were developed to systematically interrogate thousands of cancer-associated mutations using BE. **c**, Mutation signatures of recurrent SNVs observed in the MSK-IMPACT dataset relative to their frequency. Note that this dataset is enriched for C•G to T•A mutations, which can be modeled using cytosine BE. **d**, Distribution of missense, nonsense and splice site variants among recurrent mutations in the MSK-IMPACT dataset, and variants targeted in the HBES and MBES libraries. Splice site mutations were less likely to be included in MBES libraries due to lower sequence conservation in noncoding regions (Fig. 2d) ($P = 0.0009$, two-tailed Fisher's exact test). **e**, Distribution observed frequency of cancer-associated mutations in the MSK-IMPACT dataset. Mutations were classified as targeted (red circle, compatible with BE) and not targeted (black circle, incompatible with BE). Well established oncogene and tumor suppressor gene mutant alleles are highlighted in red. **f**, Gene-level annotation of cancer-associated function of recurrently mutated genes in the MSK-IMPACT dataset and genes targeted in the HBES and MBES libraries. **g**, Pie chart denoting OncoKB annotations of variants targeted in the HBES library, split by level of evidence. **h**, Predicted coverage (fraction of sites with at least one predicted sgRNA) in HBES and MBES libraries relative to the base editor used. **i**, Predicted specificity (fraction of sgRNAs with no expected off-target editing of bystander nucleotides in the locus) in HBES and MBES libraries relative to the base editor used.

efficiency of individual sgRNAs remained highly correlated (Fig. 3c, Supplementary Fig. 4b and Extended Data Fig. 5). We observed a moderate, nonlinear relationship between Cas9-induced indels and BE across all cell lines, where sgRNAs with high BE scores were a subset of sgRNAs with efficient Cas9-mediated indel generation (Extended Data Fig. 6). These data suggest that the relative potency

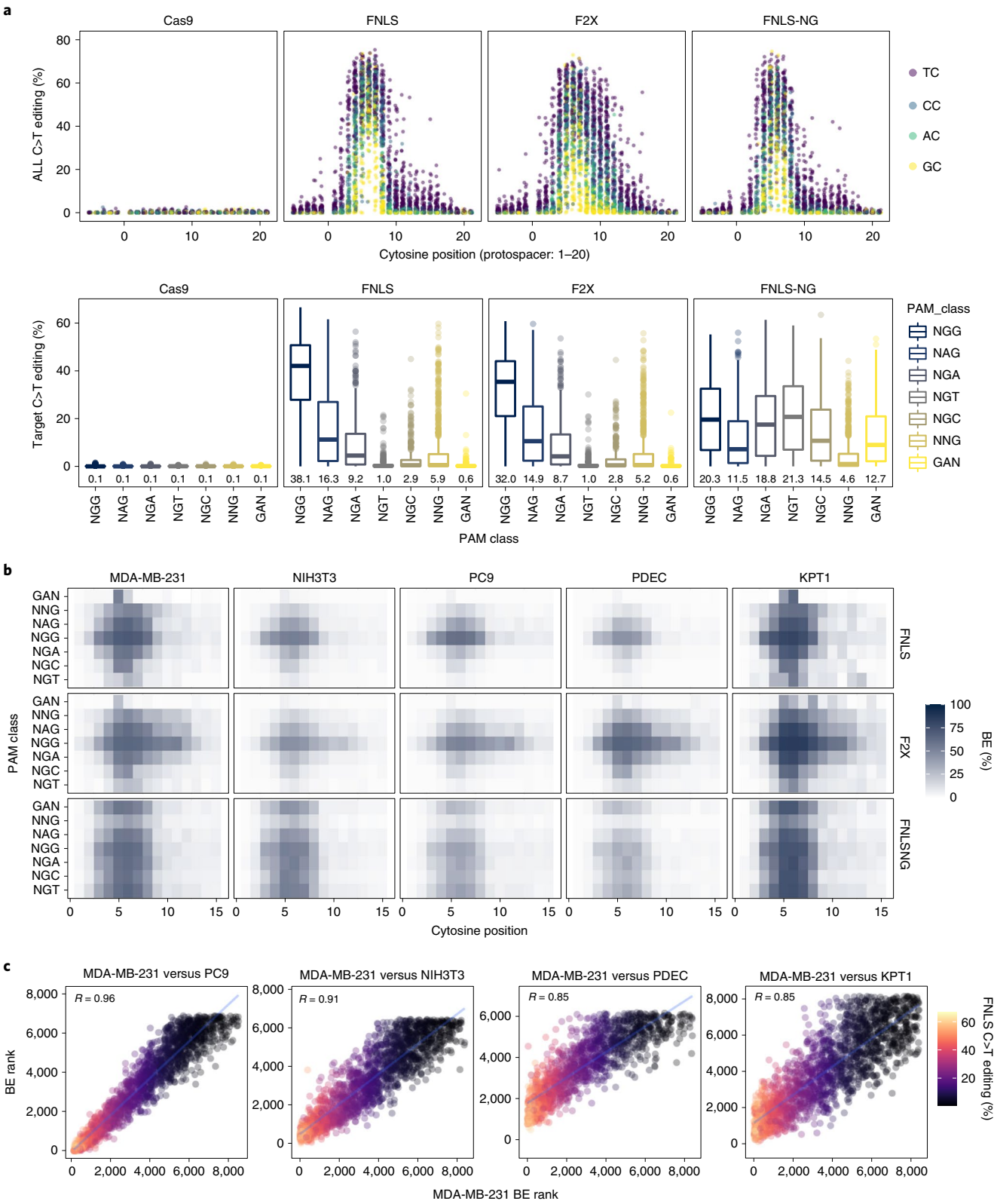


Fig. 3 | Massively parallel assessment of BE outcomes across thousands of cell line/editor/sgRNA combinations using the sensor assay. a, Top: Percentage of all C > T editing (*y* axis) across Cas9, FNLS, F2X and FNLS-NG at every cytosine among the −5 to 20 positions of the target site (*x* axis). Colored dots specify dinucleotide contexts. Bottom: Percentage of target C > T editing (*y* axis) across Cas9, FNLS, F2X and FNLS-NG relative to PAM class (*x* axis). Box plots show the median and interquartile range (IQR) and whiskers represent 1.5× IQR. Outliers are shown as dots. See also Supplementary Fig. 6. **b**, Heatmap of BE efficiency across MDA-MD-231, NIH3T3, PC9, PDEC and KPT1 cell lines (columns) at every cytosine among the −5 to 20 positions of the target site classified by cytosine base editor and PAM class (rows). See also Supplementary Fig. 6. **c**, Correlation of individual sgRNA efficiency across screen cell lines (PC9, NIH3T3, PDEC and KPT1) compared with MDA-MB-231. Only HBES sgRNAs that had >1% activity in the sensor were ranked. See also Extended Data Fig. 5.

of individual sgRNAs across different cell systems can be predicted en masse using the BE sensor assay.

Base editors can exhibit collateral (bystander) cytosine editing, whereby C > T mutations are induced in both target and neighboring cytosines in the editing window[7]. To investigate collateral editing, we calculated C > T editing 'purity' as the frequency of target C > T editing without further mutations. As expected, purity decreased with the presence of more cytosines in the target window, especially with immediately adjacent bystander cytosines (Supplementary Fig. 8). Collectively, these results demonstrate that the sensor platform can be used to assess ontarget and collateral cytosine editing across multiple base editors and thousands of target sites in a high-throughput manner.

To directly test how well the BE sensor scores predicted activity at endogenous targets, we measured editing at 12 independent genomic sites with a panel of 13 sgRNAs that showed high editing in the sensor assay. Using either FNLS or AncBE4max, endogenous editing aligned well with sensor-based estimates, with more than 50% of cases (7/13) within 10% of the sensor-reported efficiency (Fig. 4a and Supplementary Fig. 9a). To ask whether BE sensor scores could predict the relative efficiency of target editing given a range of possible options, we tiled the R213 site in *TP53* with a series of seven sgRNAs. In this case, we used the F2X base editor to allow editing across the wide range of target positions in this series (5–11 bp). Consistent with the data described above (Fig. 4a and Supplementary Fig. 9a), sensor estimates closely resembled editing at the endogenous locus (Fig. 4b).

Arbab et al. recently reported a machine learning tool (BE-Hive) for predicting BE outcomes[21]. We noted that BE-Hive predictions for the *TP53.R213* series did not accurately predict editing outcomes for non-NGG sgRNAs, probably because BE-Hive does not incorporate the PAM sequence as a prediction feature. To assess this more broadly, we determined the similarity of BE-Hive predictions with the sensor-measured editing activity for the HBES and MBES libraries. As expected, given the strong dependence on PAM for editing activity, comparison of all sgRNAs showed relatively low overall correlation between BE-Hive and BE sensor estimates (Fig. 4c, Supplementary Fig. 10 and Supplementary Table 6). Restricting our analyses to sgRNAs associated with NGG PAMs improved the correlation (Supplementary Fig. 11); however, much of it was driven by low-scoring guides, as focusing on sgRNAs with >5% BE sensor activity led to lower overall similarity (Fig. 4c and Supplementary Fig. 12). Together, these data show that BE sensor editing is well correlated with editing at endogenous sites, allowing reliable identification of sgRNAs with high editing efficiency across multiple biological contexts.

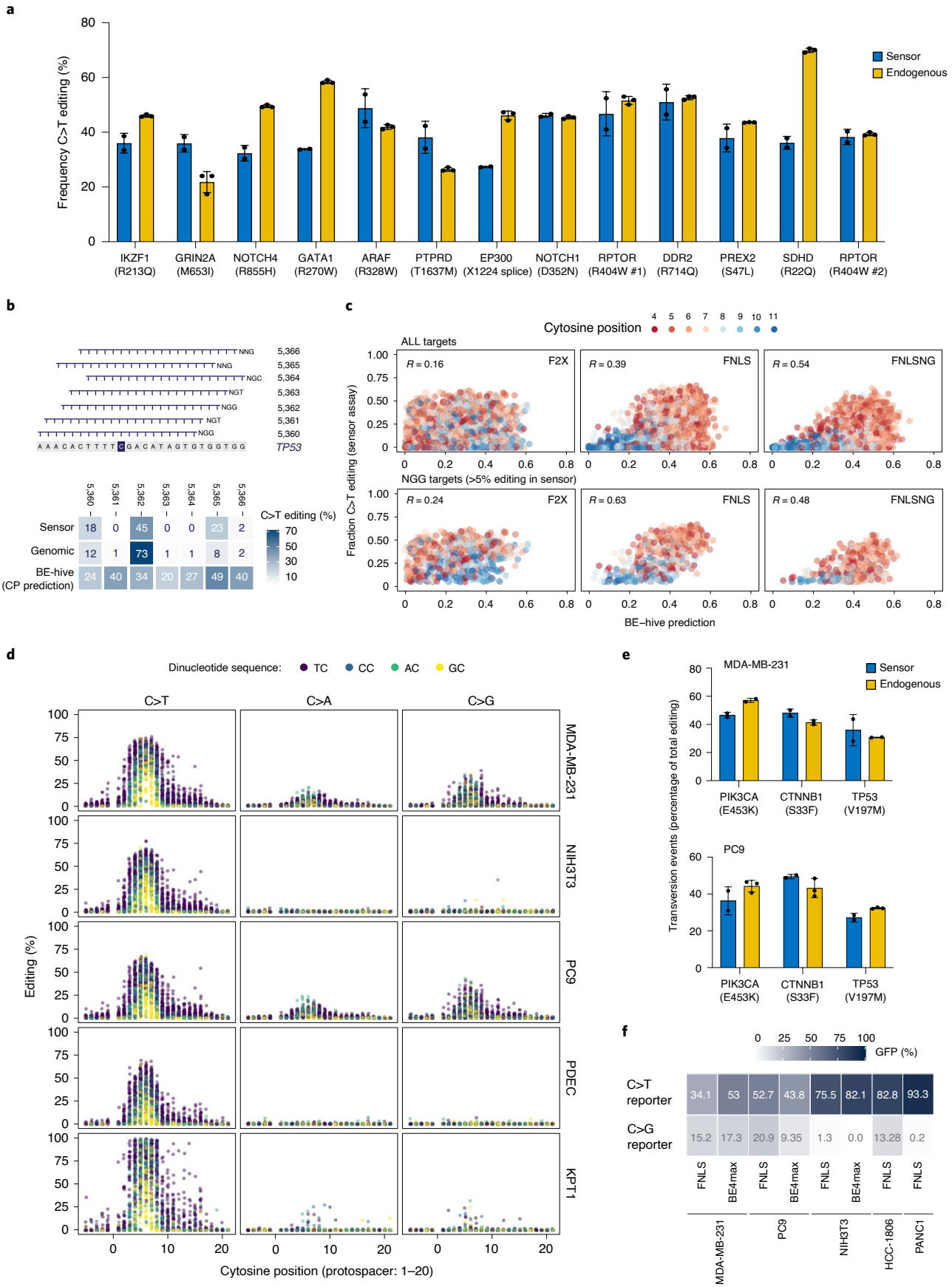
Noncanonical editing identified by BE sensor. APOBEC-driven mutation signatures in cancer include transitions (C > T; signature 2) and transversions (C > G; signature 13)[22]. Cytosine base editors (CBEs) containing rAPOBEC1 can induce C > G mutations in some contexts[12,21]. Such 'noncanonical' transversion editing could be leveraged to increase the breadth of mutations that can be modeled using BE (22% of MSK-IMPACT dataset) (Fig. 2c). Transversion editing was apparent in our sensor screen (Fig. 4d) and, for some targets, C > G editing occurred at levels greater than C > T editing (Extended Data Fig. 7a). Instances of transversion editing (C > R) closely resembled editing outcomes at three endogenous loci, chosen for their high (30–60%) predicted C > R editing rates (Fig. 4e). As expected, C > R editing by AncBE4max was slightly lower, probably due to the presence of a further uracil glycosylase inhibitor domain[23] (Supplementary Fig. 9b).

We next looked at sequence features affecting C > R editing outcomes and noted that transversion mutations were strongly disfavored at CC dinucleotides (average 4% of BE events) but were around threefold higher in AC and TC contexts (13% and 12%, respectively) (Fig. 4d and Extended Data Fig. 7b). Transversions were also disfavored when target cytosines were followed by another cytosine (NCC, particularly CCC and GCC) and enriched when followed by a thymine (NCT, particularly ACT and TCT) (Extended Data Fig. 7b). Comparison across lines revealed that not all cells induce transversions with equal efficiency. While MDA-MB-231 and PC9 cells showed frequent and high level transversion editing, NIH3T3, PDEC and KPT1 cells had a very low frequency of C > R alterations (Fig. 4d).

We developed a lentiviral BE reporter that drives GFP induction following target C > G editing (Extended Data Fig. 7c). This construct accurately reported cellular C > G editing bias, showing efficient C > G induction in MDA-MB-231, PC9 and HCC1806, but not NIH3T3 or PANC1, consistent with sensor measurements (Fig. 4f). Notably, C > G editing efficiency in the reporter was similar to average C > G editing seen at TCT motifs (Fig. 4d,f and Extended Data Fig. 7b), suggesting it is a useful tool for gauging C > R editing potential in different cells. Thus, transversion editing bias is not a universal feature of human cancer cells. Systematic studies employing this reporter could provide insight into mechanisms that dictate this activity.

Sensor-validated sgRNAs streamline in vivo model development. A main advantage of sensor-based validation is the ability to identify active sgRNAs that generate specific missense mutations with little-to-no collateral editing. Such guides can be used to interrogate the impact of specific mutations in vitro and in vivo. As

Fig. 4 | Validation of canonical and noncanonical BE activity predicted by the sensor assay. a, Experimental validation of C > T editing activity observed in the sensor (blue) when targeting endogenous (yellow) loci in FNLS-PC9 cells. Each dot corresponds to a single replicate ($n = 2$ for sensor screening data; $n = 3$ for endogenous validation). Data are presented as mean values ± s.e.m. BE rates (efficiencies) across endogenous loci were determined via NGS of edited loci and analyzed using CRISPResso2 (ref. [48]). See also Extended Data Fig. 7 and Supplementary Fig. 9. No direct statistical comparisons were performed between sensor and endogenous C > T editing data because sensor screens were performed in duplicate. **b**, Top: schematic of the human *TP53-R213* locus. Horizontal bars denote sgRNAs, and numbers to the right denote sgRNA identifiers (based on HBES whitelists). Target cytosine is labeled in blue. Bottom: heatmap comparing C > T editing efficiency in an allelic series of *TP53-R213* sgRNAs between the sensor results, F2X-MDA-MB-231 cells targeting the endogenous locus and BE-Hive predictions (CP1028)[21]. **c**, Correlation of observed BE efficiency measured by the sensor in MDA-MB-231 cells versus efficiencies predicted by the BE-HIVE algorithm[21] classified by BE enzyme and cytosine position (fill). Here, the HBES library was stratified to include all targets (top) and all NGG targets showing >5% editing in the sensor (bottom). FNLS and FNLS-NG values were compared with BE4 prediction results and F2X values were compared with CP1028 prediction results. See also Supplementary Figs. 10–12. **d**, Canonical (C > T) and noncanonical (C > A and C > G) BE activity profiled across all screen cell lines (rows) at every cytosine in position −5 to −20 of HBES library targets. **e**, Validation of noncanonical C > R editing events at sensor target sites (blue) and endogenous targets (yellow). Each dot corresponds to a single replicate ($n = 2$ for sensor screening data; $n = 3$ for endogenous validation). Data are presented as mean values ± s.e.m. BE rates (efficiencies) across endogenous loci were determined via NGS of edited loci and analyzed using CRISPResso2 (ref. [48]). See also Extended Data Fig. 7 and Supplementary Fig. 9. **f**, A heatmap of a panel of mammalian cell lines expressing FNLS and/or AncBE4max transduced with either canonical C > T (top) or noncanonical C > G (bottom) GO reporters measuring GFP induction in mScarlet-infected base editor cells. See also Extended Data Fig. 7.


a
Frequency C>T editing (%)
Sensor
Endogenous
IKZF1 (R213Q)
GRIN2A (M653I)
NOTCH4 (R855H)
GATA1 (R270W)
ARAF (R328W)
PTPRD (T1637M)
EP300 (X1224 splice)
NOTCH1 (D352N)
RPTOR (R404W #1)
DDR2 (R714Q)
PREX2 (S47L)
SDHD (R22Q)
RPTOR (R404W #2)
b
TP53
Sensor
Genomic
BE-hive (CP prediction)
C>T editing (%)
c
Cytosine position
ALL targets
NGG targets (>5% editing in sensor)
Fraction C>T editing (sensor assay)
BE–hive prediction
F2X
FNLS
FNLSNG
d
Dinucleotide sequence: TC CC AC GC
C>T
C>A
C>G
MDA-MB-231
NIH3T3
PC9
PDEC
KPT1
Editing (%)
Cytosine position (protospacer: 1–20)
e
MDA-MB-231
PC9
Transversion events (percentage of total editing)
PIK3CA (E453K)
CTNNB1 (S33F)
TP53 (V197M)
f
GFP (%)
C>T reporter
C>G reporter
FNLS
BE4max
MDA-MB-231
PC9
NIH3T3
HCC-1806
PANC1

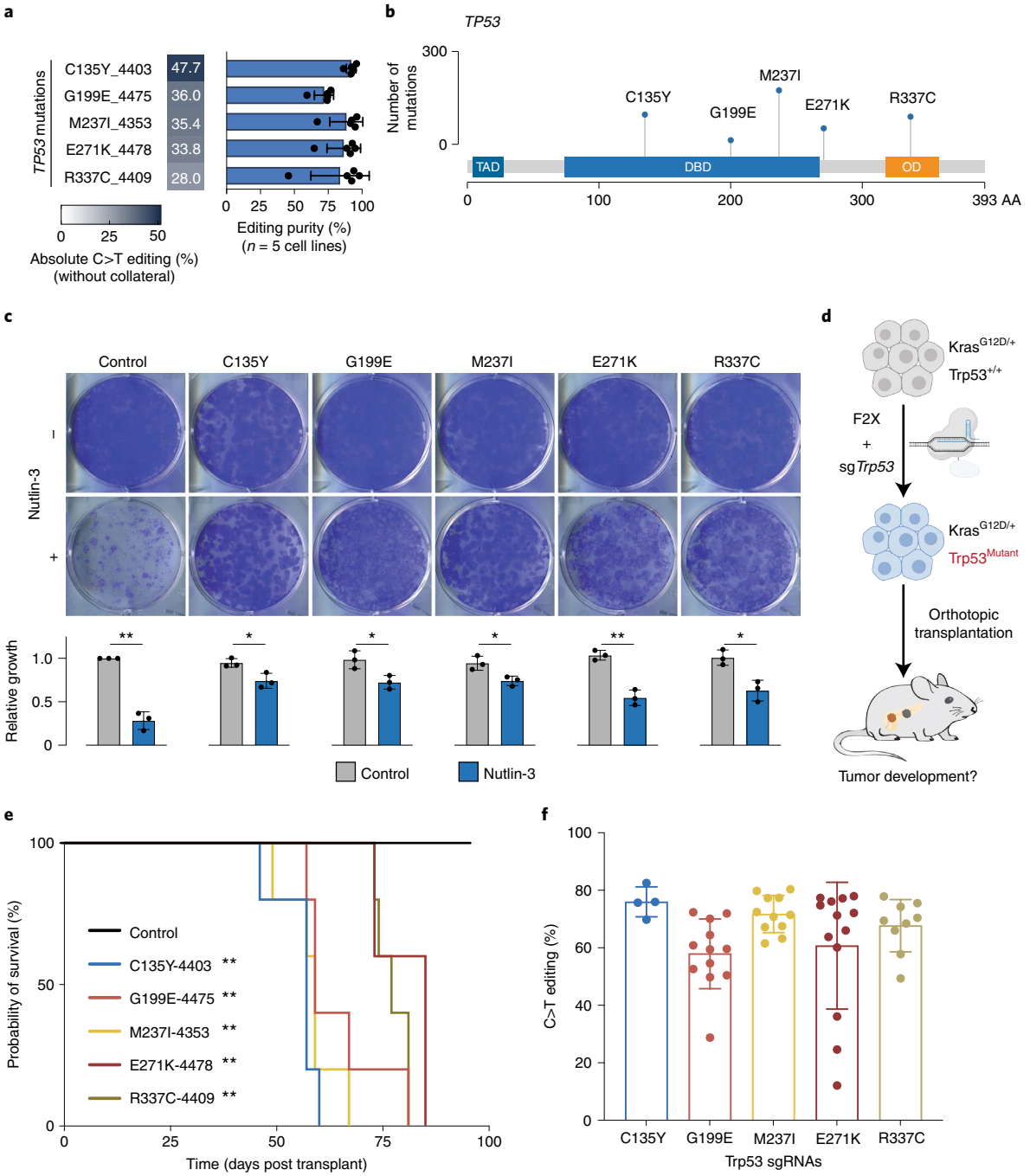


Fig. 5 | In vivo validation of cancer-associated single nucleotide *TP53* variants using BE. a, Candidate *TP53* variant-specific BE sgRNAs sorted by C > T efficiency scores obtained from FNLS-MDA-MB-231 MBES screening data. Only *TP53* sgRNAs with a percentage of C > T editing >25% and no collateral cytosine editing are shown. Data are presented as mean values ± s.e.m. **b**, Lollipop plot showing frequency of candidate *TP53* variants detected in the MSK-IMPACT cohort. TAD, transactivation domain; DBD, DNA binding domain; OD, oligomerization domain. **c**, FNLS-expressing *Kras*$^{G12D/+}$; *Trp53*$^{WT/WT}$ PDECs were transduced with sgRNAs designed to introduce defined mutations in the mouse *Trp53* gene followed by plating at low density (1,000 cells per well in six-well plates; three wells per variant) and treatment with DMSO or Nutlin-3 (10 μM). Upper panel: plates were stained with crystal violet to assess colony formation capacity. Bottom panel: quantification of crystal violet staining; $n = 3$ wells per variant (or control) per treatment arm (control or Nutlin-3). Data are representative of $n = 3$ independent experiments and are presented as mean values ± s.d. $^{*}P \leq 0.05$, $^{**}P \leq 0.01$. *P* values were calculated using unpaired, two-sided *t*-test. **d**, Schematic for in vivo validation of candidate *TP53* variants via orthotopic transplantation of F2X-expressing *Kras*$^{G12D/+}$; *Trp53*$^{WT/WT}$ PDECs transduced with sgRNAs designed to introduce defined mutations in the mouse *Trp53* gene. **e**, Survival analysis of mice transplanted with F2X-expressing PDECs transduced with specific *Trp53*-targeting BE sgRNAs; $n = 5$ mice per mutation. See also Extended Data Fig. 8 and Supplementary Figs. 13 and 14. $^{**}P \leq 0.01$. *P* values were calculated using the log-rank test. **f**, Frequency of target C > T editing in tumors from transplanted mice. Each dot corresponds to a single tumor or tumor fragment. Data are presented as mean values ± s.d. Target C > T editing was measured by NGS of amplified target loci and data were analyzed using CRISPResso2 (ref. [48]). See also Extended Data Fig. 8 and Supplementary Figs. 13 and 14.

proof-of-concept, we focused on the *TP53* tumor suppressor gene, which is the most frequently mutated gene in cancer and shows remarkable mutational heterogeneity[24]. Hundreds of *TP53* SNVs have been identified[16], most of which are missense variants that may have loss-of-function, gain-of-function, dominant negative or neomorphic behavior[24,25]. Our mouse sensor library contained 244 sgRNAs targeting 62 distinct and recurrent p53 mutations that were represented in the mouse library. To measure the tumorigenic potential of p53 variants, we used immortalized murine $Kras^{G12D/+}$;$Trp53^{WT/WT}$ PDECs, a genetically defined and physiologically relevant setting to model pancreatic cancer[20]. To test this concept, we cloned five sensor-validated sgRNAs to introduce specific missense mutations in *Trp53* with low collateral activity (high 'purity') (C135Y, M237I, G199E, E271K and R337C; human *TP53* gene nomenclature) (Fig. 5a,b). Introduction of each sgRNA into F2X-PDECs enabled low density growth in the presence of Nutlin-3 (ref. [26]) (Fig. 5c), suggesting these mutations compromise p53 function. To test whether these mutations impaired tumor suppression in vivo, we transplanted PDECs transduced with *Trp53* or control sgRNAs into the pancreas of recipient mice ($n = 5$ mice per sgRNA) (Fig. 5d). In cases where the sensor assay predicted multiple sgRNAs for a single mutation, we included all available sgRNAs to rule out off-target effects (Extended Data Fig. 8a). Orthotopic transplantation of control PDECs does not lead to pancreatic tumor development (up to 200 days), but all mice transplanted with PDECs carrying *Trp53* sgRNAs succumbed to pancreatic tumors (46–99 days) (Fig. 5e and Extended Data Fig. 8a). In each case, analysis of bulk tumor tissue showed high frequency of C > T mutations at their respective sites in the *Trp53* gene (Fig. 5f, Extended Data Fig. 8b,e and Supplementary Figs. 13 and 14). Identical results were obtained with FNLS-PDECs ($n = 5$ mice per mutation) (Extended Data Fig. 8c,d). Thus, BE-sensor validated sgRNAs can be used to synchronously engineer endogenous patient-derived mutations in experimental in vivo systems, facilitating systematic variant-to-phenotype studies in cancer and other diseases.

Pooled BE sensor screens to interrogate cancer variants. The experiments above demonstrated the robustness of BE sensor-validated sgRNAs for in vivo interrogation of cancer variants. Encouraged by these results, we set out to test whether BE sensor libraries could be coupled with high-throughput screening approaches for massively parallel functional interrogation of cancer-associated SNVs. An advantage of screening BE sensor libraries is that cells should harbor editing at both the sensor module and endogenous target site. Hence, variant-specific effects on cellular phenotype can be correlated with editing precision and efficiency at sensor target sites, minimizing false positives. In theory, this approach should also identify sgRNAs that edit their target but induce no phenotypic effect. To test this concept, we transduced $Kras^{G12D/+}$;$Trp53^{WT/WT}$ FNLS-PDECs with the MBES library at low multiplicity of infection and >1,000× representation (Fig. 6a). Six technical replicates of PDEC-MBES cells (Supplementary Fig. 15) were used for a multitime point in vitro proliferation screen performed for around 36 cumulative population doublings to quantitatively assess sgRNA activity and abundance in parallel.

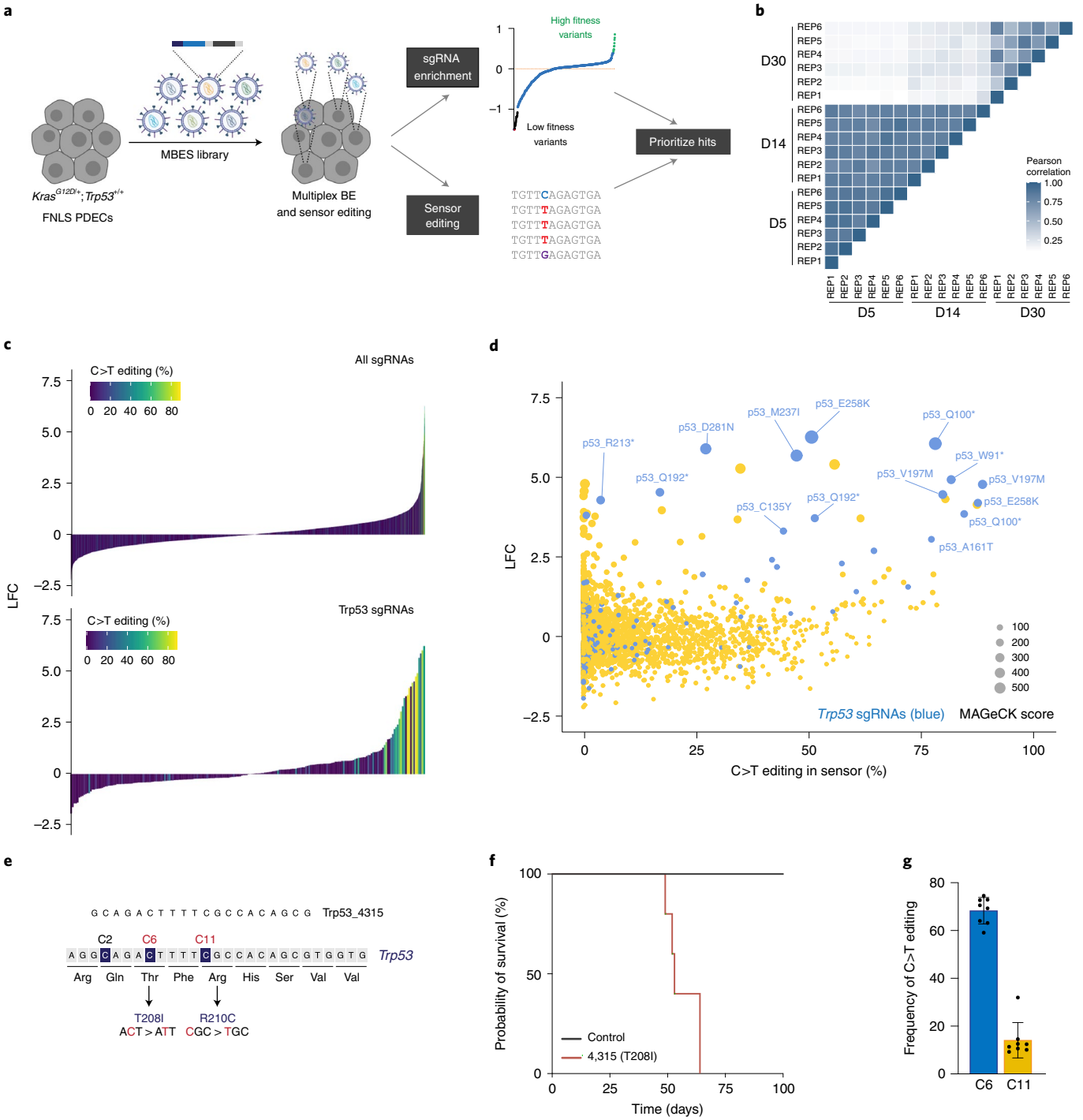
Pairwise correlation analyses at the first timepoint (day 5) demonstrated excellent technical screening performance and replicates diverged at later time points (Fig. 6b). To quantify sensor editing and sgRNA enrichment, we used our analytical pipeline to calculate target editing efficiency, followed by MAGeCK[27,28] to determine changes in sgRNA abundance (Supplementary Table 7). Focusing on day 30 versus day 5 comparisons, our analysis identified 150 sgRNAs that seemed to promote ($n = 125$; $\log_2$ fold change (LFC) ≥ 1.5) or inhibit ($n = 25$; LFC ≤ −1.5) PDEC proliferation (false discovery rate ≤ 0.01) (Fig. 6c,d and Supplementary Table 7b). Significantly enriched sgRNAs were predicted to install mutations in genes with known oncogenic activity, including *Jak3*, *Fgfr2* and *Egfr* (Supplementary Table 7d). Mutations in genes with known tumor suppressive function were also represented, including *Trp53*, *Apc*, *Fbxw7*, *Nf2* and *Chek2* (Supplementary Table 7d). In fact, after filtering for sgRNAs with more than 20% editing activity, 72% of enriched sgRNAs (26/36) targeted known or probable oncogenic mutations, compared with 38% in nonenriched sgRNAs ($P = 0.0003$; Fisher's exact test) (Extended Data Fig. 9a,b). Notably, more than half (19/36) of the enriched sgRNAs targeted *Trp53*, consistent with our proof-of-concept experiments (Fig. 6) and the role of p53 in suppressing mutant Kras-driven proliferation[29] (Fig. 6c,d). In fact, collapsing the data to 'gene-level' scores identified *Trp53* as the only significantly scoring gene in this screen (false discovery rate < 0.01) (Supplementary Table 7c).

Using *Trp53* as a case study, we next compared fitness scores with sensor editing data from the same screen. Most sgRNAs enriched in the proliferation screen showed high editing activity, including two potent sgRNAs we previously validated in vivo (C135Y and M237I) (Fig. 6e and Extended Data Fig. 8c). We identified several *Trp53* missense and nonsense mutations that were enriched exclusively in vitro or in vivo (E271K, R337C and G199E), highlighting the importance of the context in measuring p53 variant fitness advantage. Most enriched sgRNAs demonstrated relatively high sensor editing but, notably, *Trp53-R213*

Fig. 6 | Massively parallel interrogation of cancer-associated single nucleotide variants via pooled BE screening. a, Schematic of BE proliferation screen. Briefly, FNLS-expressing PDECs were transduced with the MBES library at 1,000× representation followed by selection and culture for a total of around 36 cumulative population doublings. A total of $n = 6$ independent transduction replicates were established and cultured separately. Cells were sampled at multiple time points over the course of the screen until reaching the final time point at day 30 post-transduction. Screens were deconvoluted using NGS (see Methods for more details). **b**, Heatmap for correlation coefficients between samples. See also Supplementary Fig. 15. **c**, Waterfall plots comparing sgRNA log fold changes between days 5 and 30 post-transduction. Colors represent mean C>T editing percentage of each target cytosine at day 30. Top plot denotes all sgRNAs (and corresponding sensor target sites). Bottom plot denotes *Trp53* sgRNAs (and corresponding sensor target sites). **d**, Bubble plot comparing sgRNA log fold changes with mean frequency of C > T editing in the sensor target site between days 5 and 30 post-transduction. Blue bubbles denote *Trp53* sgRNAs (and corresponding sensor target sites). Yellow bubbles denote all other sgRNAs (and corresponding sensor target sites). Inset denotes MaGeCK[27,28] score (Supplementary Table 7). Note the use of human gene-based nomenclature of protein residues (for example, *p53_Q100* corresponds to *Q97* in mouse *Trp53*). **e**, Schematic of the mouse *Trp53-R210* locus (*TP53-R213* in humans). The number to the right of the sgRNA is the sgRNA identifier (Trp53_4315; based on MBES whitelists). Target cytosines are labeled in red. As denoted by the black arrows in the diagram, C > T BE of C6 and C11 is predicted to produce the T208I and R210C mutations, respectively. See also Supplementary Fig. 13. **f**, In vivo validation of T208I mutation via orthotopic transplantation of F2X-expressing PDECs transduced with the Trp53_4315 sgRNA. Median survival for mice harboring tumors initiated by the Trp53_4315 sgRNA was 53 days; $n = 5$ mice per condition. Data are presented as mean values ± s.d. $^*P \leq 0.01$, Log-rank test. See also Supplementary Fig. 13. **g**, Frequency of target C > T editing in tumors from mice transplanted with F2X-expressing PDECs transduced with the Trp53_4315 sgRNA. Each dot corresponds to a single tumor or tumor fragment (total $n = 8$). Data are presented as mean values ± s.d. Target C > T editing was measured by NGS of amplified target loci and data were analyzed using CRISPResso2 (ref. [48]). See also Supplementary Fig. 13.

sgRNA showed more than tenfold enrichment in the screen, despite less than 3% sensor target editing. Inspection of the cognate sensor cassette revealed that this sgRNA showed a high level of editing at an adjacent cytosine, creating a T211I mutation (Fig. 6e)—a variant also observed in human cancers[16,30]. Mice transplanted with *Trp53*^{T208I} cells (corresponding to human *TP53*^{T211I}) succumbed to a fully penetrant disease (median survival of 53 days) (Fig. 6f). Sequencing analysis of bulk tumor tissue gDNA confirmed the C>T (*Trp53*^{T208I}) mutation, with <15% C>T editing at the cytosine in R210 (corresponding to human *TP53*^{R213}) (Fig. 6g), implying that T208I is the oncogenic driver in this case. These data identify multiple *TP53* missense mutations as drivers of proliferation in nontransformed pancreatic epithelium and establish *Trp53*^{T208I}/*TP53*^{T211I} as a bona fide driver mutation in this mouse model of pancreatic cancer.

A flexible platform of BE sensor predictions. Mutation databases are expanding, and new Cas and base editor variants are being identified at a rapid pace. Motivated by this reality, we expanded the capabilities of AMINEsearch (Methods) and applied it to a more recent release of MSK-IMPACT that contains sequencing of 47,550 tumors (Extended Data Fig. 10 and Supplementary Table 8g–j). The characteristics of an expanded set of Cas variants and base editors (including adenine base editors/ABEs[8]) were included as input and can be leveraged to select tools tailored to experiments that require maximum coverage or specificity (Supplementary Fig. 16).



Discussion

BE is an efficient strategy with which to engineer and study SNVs, yet the identification of effective sgRNA-base editor combinations remains challenging. Here, we describe a versatile sensor-based BE platform that enables identification of efficient sgRNAs from large, pooled libraries across multiple species, cell lines and base editor configurations. We show that sensor predictions can accelerate the characterization of cancer-associated SNVs in vivo and demonstrate that integrating a BE sensor can support the interpretation of BE genetic screens.

All-in-one library strategies have been described for measuring Cas9 and BE outcomes[21,31–34]. Such libraries have been used to develop machine learning tools to predict activity and purity of BE tools[21,32,33,35]. These tools are useful for their scope of prediction, but predictions can diverge significantly from experimentally observed editing at endogenous sites (Fig. 4 and Supplementary Figs. 6 and 10–12). Our work shows that sensor-based activity estimates closely reflect editing outcomes at endogenous loci. Moreover, although it is possible that cell-specific differences in DNA accessibility or expression levels of base editors could impact editing efficiency, our results show relative consistency across cell lines, suggesting that sequence context is an important determinant of editing.

In addition to expected target C > T editing, we observed frequent 'noncanonical' (C > R) transversion editing. Transversion frequencies were influenced by local sequence features surrounding target cytosines (Fig. 4 and Extended Data Fig. 7b) and cell line. Analysis of multiple cell lines revealed that transversion editing is not universal and cannot be predicted easily by association with Signature 13 (APOBEC-driven C > G mutations)[22]. Our data showed that C > R editing occurs most frequently at DCT motifs (ACT > TCT > GCT) and is strongly disfavored when the target C is flanked by another cytosine. This observation is similar, but not identical, to that reported by Arbab et al.[21], who reported RCT motifs as the most prone to transversion editing. This distinction may reflect different cell types used or enrichment of specific genomic sites from cancer-associated mutations in our dataset compared with rationally designed sequences used in their study. Recently, multiple groups have described new base editors (CGBEs) that enable efficient transversion editing[9,36–38]. The ability to engineer transversions will significantly expand the mutation repertoire that can be engineered using BE. It remains unclear whether CGBE editors can overcome cell-line-dependent effects that limit transversion editing.

Gene function is complex, reflected by the diversity of phenotypes, including therapeutic responses, that can be driven by distinct variant alleles[39]. Building defined genetic models of specific oncogenic alterations is critical to define their direct impact and to reveal new treatment strategies. Similarly, BE screens offer a new approach to interrogate gene variant function en masse[40,41]. Unlike traditional CRISPR screens that provide 'gene-level' information, BE screens reveal 'amino acid level' information and, as such, cannot always rely on the activity of multiple sgRNAs to define 'scoring' hits. We argue that incorporation of a BE sensor cassette in BE libraries will enhance the interpretation of BE screens by providing preliminary validation of sgRNA activity, flagging possible false positives and improving the classification of phenotype-neutral mutational events. Indeed, our proof-of-concept fitness screen in nontransformed pancreatic epithelial cells identified multiple candidate oncogenic variants spread across a collection of genes. These included *Trp53* mutations that drive increased proliferation, but also those that induce target mutations without driving increased cellular fitness (at least in vitro) or that potentially exhibit context-specific phenotypes. Future iterations of this approach could employ unique molecular identifiers embedded in sensor backbones or sgRNAs[42] to account for clonal phenotypes or explain variant-specific transcriptional effects using single cell RNA sequencing[43–45]. Furthermore, the sensor framework should be compatible with emerging genome editing technologies like Prime Editing[46,47].

For those who wish to use individual validated BE sgRNAs or design alternate BE sensor libraries, we developed a web application, BE-SCAN (BE sensor-validated cancer-associated mutations; https://dowlab.shinyapps.io/BEscan/) that allows browsing and selection of guides by species, gene, target mutation and/or base editor. Further, the expanded AMINEsearch-defined (nonsensor validated) collection of somatic cancer mutations is also available as an interactive portal in BE-SCAN. To enable the creation of sensor-based BE libraries beyond those described in this study, the full AMINEsearch pipeline is available (https://github.com/Kastenhuber/AMINEsearch) and can be run on any set of mutations and base editor characteristics. Beyond cancer-associated somatic mutations, we envision this approach could be employed to functionally annotate genome-wide association study variants and mutations associated with heritable genetic disease. Whereas we performed proliferation screens in immortalized cells, screening of genetic variants could just as easily be conducted using any sortable cellular feature or biosensor. We expect that the compendium of experimentally vetted BE tools described here will accelerate the development of next-generation allele-focused in vitro and in vivo cancer models.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01172-3.

Received: 2 July 2021; Accepted: 29 November 2021;
Published online: 14 February 2022

References

1. Gorelick, A. N. et al. Phase and context shape the function of composite oncogenic mutations. *Nature* **582**, 100–103 (2020).
2. Hyman, D. M. et al. AKT inhibition in solid tumors with AKT1 mutations. *J. Clin. Oncol.* **35**, 2251–2259 (2017).
3. Vasan, N. et al. Double PIK3CA mutations in cis increase oncogenicity and sensitivity to PI3Kalpha inhibitors. *Science* **366**, 714–723 (2019).
4. Zafra, M. P. et al. An in vivo *Kras* allelic series reveals distinct phenotypes of common oncogenic variants. *Cancer Discov.* **10**, 1654–1671 (2020).
5. Findlay, G. M. et al. Accurate classification of BRCA1 variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
6. Vivanco, I. et al. Differential sensitivity of glioma- versus lung cancer-specific EGFR mutations to EGFR kinase inhibitors. *Cancer Disco.* **2**, 458–471 (2012).
7. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
8. Gaudelli, N. M. et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
9. Koblan, L. W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
10. Zafra, M. P. et al. Optimized base editors enable efficient editing in cells, organoids and mice. *Nat. Biotechnol.* **36**, 888–893 (2018).
11. Kleinstiver, B. P. et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
12. Katti, A. et al. GO: a functional reporter system to identify and enrich base editing activity. *Nucleic Acids Res.* **48**, 2841–2852 (2020).
13. Vakulskas, C. A. et al. A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. *Nat. Med.* **24**, 1216–1224 (2018).
14. Nishimasu, H. et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science* **361**, 1259–1262 (2018).
15. Hu, J. H. et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).
16. Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).

17. Chakravarty, D. et al. A precision oncology knowledge base. *JCO Precis. Oncol.* **2017**, PO.17.00011 (2017).
18. Chakravarty, D. & Solit, D. B. Clinical cancer genomic profiling. *Nat. Rev. Genet.* **22**, 483–501 (2021).
19. Dimitrova, N. et al. Stromal expression of miR-143/145 promotes neoangiogenesis in lung cancer development. *Cancer Discov.* **6**, 188–201 (2016).
20. Lee, K. E. & Bar-Sagi, D. Oncogenic KRas suppresses inflammation-associated senescence of pancreatic ductal cells. *Cancer Cell* **18**, 448–458 (2010).
21. Arbab, M. et al. Determinants of base editing outcomes from target library analysis and machine learning. *Cell* **182**, 463–480.e430 (2020).
22. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
23. Komor, A. C. et al. Improved base excision repair inhibition and bacteriophage Mu Gam protein yields C:G-to-T:A base editors with higher efficiency and product purity. *Sci. Adv.* **3**, eaao4774 (2017).
24. Kastenhuber, E. R. & Lowe, S. W. Putting p53 in context. *Cell* **170**, 1062–1078 (2017).
25. Muller, P. A. & Vousden, K. H. Mutant p53 in cancer: new functions and therapeutic opportunities. *Cancer Cell* **25**, 304–317 (2014).
26. Vassilev, L. T. et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303**, 844–848 (2004).
27. Li, W. et al. Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol.* **16**, 281 (2015).
28. Li, W. et al. MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).
29. Morris, J. P. T. et al. α-Ketoglutarate links p53 to cell fate during tumour suppression. *Nature* **573**, 595–599 (2019).
30. Kanda, M. et al. Mutant TP53 in duodenal samples of pancreatic juice from patients with pancreatic cancer or high-grade dysplasia. *Clin. Gastroenterol. Hepatol.* **11**, 719–730 e715 (2013).
31. Koblan, L. W. et al.Efficient C*G-to-G*C base editors developed using CRISPRi screens, target-library analysis, and machine learning. *Nat. Biotechnol.* **39**, 1414–1425 (2021).
32. Shen, M. W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
33. Song, M. et al. Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. *Nat. Biotechnol.* **38**, 1037–1043 (2020).
34. Tycko, J. et al. Pairwise library screen systematically interrogates *Staphylococcus aureus* Cas9 specificity in human cells. *Nat. Commun.* **9**, 2962 (2018).
35. Marquart, K. F. et al. Predicting base editing outcomes with an attention-based deep learning algorithm trained on high-throughput target library screens. *Nat. Commun.* **12**, 5114 (2021).
36. Chen, L. et al. Programmable C:G to G:C genome editing with CRISPR-Cas9-directed base excision repair proteins. *Nat. Commun.* **12**, 1384 (2021).
37. Kurt, I. C. et al. CRISPR C-to-G base editors for inducing targeted DNA transversions in human cells. *Nat. Biotechnol.* **39**, 41–46 (2021).
38. Zhao, D. et al. Glycosylase base editors enable C-to-A and C-to-G base changes. *Nat. Biotechnol.* **39**, 35–40 (2021).
39. Hyman, D. M., Taylor, B. S. & Baselga, J. Implementing genome-driven oncology. *Cell* **168**, 584–599 (2017).
40. Cuella-Martin, R. et al. Functional interrogation of DNA damage response variants with base editing screens. *Cell* **184**, 1081–1097 e1019 (2021).
41. Hanna, R. E. et al. Massively parallel assessment of human variants with base editor screens. *Cell* **184**, 1064–1080 e1020 (2021).
42. Xu, P. Genome-wide interrogation of gene functions through base editor screens empowered by barcoded sgRNAs. *Nat. Biotechnol.* **39**, 1403–1413 (2021).
43. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882.e1821 (2016).
44. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
45. Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866.e1817 (2016).
46. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
47. Kim, H. K. et al. Predicting the efficiency of prime editing guide RNAs in human cells. *Nat. Biotechnol.* **39**, 198–206 (2021).
48. Clement, K. et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).

Methods

Plasmids and sgRNA cloning. *Base editor plasmids.* The following lentiviral BE plasmids were used in this manuscript: FNLS (Addgene, catalog no. 110841), AncBE4max (this manuscript), FNLS-2X (F2X) (Addgene, catalog no. 110840), FNLS-HF1 (HF1) (Addgene, catalog no. 110866), FNLS-HiFi (HiFi) (Addgene, catalog no. 136902), FNLS-NG (NG) (Addgene, catalog no. 136900), FNLS-HiFi-NG (HiFi-NG) (Addgene, catalog no. 136903), FNLS-VQR (VQR) (this manuscript) and xFNLS (Addgene, catalog no. 110872). All new plasmids and libraries will be available from Addgene.

CRISPR nuclease plasmids. The following CRISPR nuclease plasmids were used in this manuscript: lentiCas9-Blast (Addgene, catalog no. 52962), Cas9-NG (Addgene, catalog no. 117919) and Cas9-Puro (Addgene, catalog no. 110837).

sgRNA plasmids. The following sgRNA plasmids were used in this manuscript: LRT2B (Addgene, catalog no. 110854) (ref. [10]), pUSEPR (U6-sgRNA-EFS-Puro-P 2A-TurboRFP)[49] and pUSEBR (pUSE-Blast-P2A-TurboRFP) (this manuscript). We cloned *Esp*3I/*Bsm*BI-compatible annealed and phosphorylated oligos encoding sgRNAs into *Esp*3I/*Bsm*BI-linearized pLRT2B, pUSEPR or pUSEBR using high-concentration T4 DNA ligase (NEB). A 5′ G (to boost U6 transcriptional initiation) was added to sgRNAs that lacked it either by appending it to the 5′ or by substituting the first nucleotide in the 5′ position for a G. All sgRNA sequences used are listed in Supplementary Table 9.

Other plasmids. The gene ontology (GO) (C > G) reporter was cloned by modifying the GO reporter system as described in ref. [12]. Briefly a custom GFP(ATC) gBlock cassette was inserted to *Eco*RI- and *Bsrg*I-digested mUGIS^{GO} by standard InFusion assembly protocol. To insert GO3 sgRNA (C > G targeting guide), mU6-GO3-scaffold was amplified. Both inserts were digested with *Xho*I and *Nsi*I and ligated using T4 DNA ligase.

Cell culture. HEK293T (ATCC CRL-3216), A549 (CCL-185), MDA-MB-231 (ATCC HTB-26) and KPT1 cells were cultured in DMEM supplemented with 10% fetal bovine serum (FBS) and 100 IU ml^{-1} penicillin/streptomycin. KP cells were a kind gift from T. Jacks (MIT). PC9 cells were a kind gift from H. Varmus (Weill Cornell) and cultured in RPMI supplemented with 10% FBS and 100 IU ml^{-1} penicillin/streptomycin. NIH3T3 cells (ATCC CRL-1658) were cultured in DMEM supplemented with 10% fetal calf serum (FCS) and 100 IU ml^{-1} penicillin/streptomycin. PDECs[20] were a kind gift from D. Bar-Sagi (New York University) and cultured in collagen-coated plates (100 µg ml^{-1} PureCol 5005, Advanced Biomatrix) with Advanced DMEM/F12 supplemented with 10% FBS (Gibco), 100 IU ml^{-1} penicillin/streptomycin (Gibco), 100 mM Glutamax (Gibco), ITS Supplement (Sigma), 0.1 mg ml^{-1} soy trypsin-inhibitor (Gibco), bovine pituitary extract (Gibco), 5 nM T3 (Sigma), 100 µg ml^{-1} cholera toxin (Sigma), 4 µg ml^{-1} Dexamethasone (Sigma) and 10 ng ml^{-1} human EGF (Preprotech).

Virus production. Lentiviruses were produced by cotransfection of HEK293T cells with the relevant lentiviral transfer vector and packaging vectors psPax2 (Addgene, catalog no. 12260) and pMD2.G (Addgene, catalog no. 12259) using Lipofectamine 2000 (Invitrogen). Viral supernatants were collected at 48 and 72 h post-transfection and stored at −80°C.

Drug treatments. Nutlin-3 (Selleck Chemicals, S1061) was dissolved in DMSO at a stock concentration of 10 mM and used at a final concentration of 10 µM.

Flow cytometric analyses. GO validation experiments were measured in either a Thermo Fisher 2018 Attune NxT flow cytometer or a Guava Easycyte (Millipore). Fluorescence-assisted cell sorting was performed in either BD FACS Aria II or Sony MA900 cell sorters.

Protein analysis. 231, PC9 and 3T3 screen pellets were resuspended with 500 µl radioimmunoprecipitation assay buffer then centrifuged at 4°C at 13,000 rpm to collect protein lysates. Antibodies used for western blot analyses were: Cas9 (CellSignaling, catalog no. 19526S) and Actin (Abcam, catalog no. ab49900).

Animal work. *Animals.* All mouse experiments were approved by the Memorial Sloan-Kettering Cancer Center (MSKCC) Internal Animal Care and Use Committee under MSKCC IACUC protocol 11-06-018. Mice were maintained under specific pathogen-free conditions, and food and water were provided ad libitum. Foxn1^{nu} (Swiss nude) mice were purchased from Envigo. All mice used were 6- to 8 week-old females.

Pancreatic orthotopic transplants. For transplantation of PDEC cells into the pancreas of adult mice, animals were anesthetized and a survival surgery was performed to expose the pancreas. Independent of genotype, a total of 1×10^5 PDEC cells resuspended in 25 µl of growth factor reduced Matrigel (354230; Corning) diluted 1:1 with cold OptiMEM (Gibco) were injected into the tail region of the pancreas of each mouse. Mice were monitored for tumor development over time by abdominal palpation and were euthanized upon developing overt disease and becoming moribund following disease monitoring guidelines of IACUC and the MSKCC Animal Facility.

Genomic DNA isolation. *Isolation of genomic DNA from cells.* Genomic DNA (gDNA) was extracted from cells using the DNeasy Blood and Tissue Kit (Qiagen) following the manufacturer's instructions. Cell pellets were processed in parallel and the resulting gDNA was resuspended in 100–200 µl of 10 mM Tris-Cl; 0.5 mM EDTA; pH 9.0. Samples from corresponding replicates from MBES and HBES screens were pooled at the gDNA level, measured using a NanoDrop 2000 (ThermoFisher) and normalized before performing sequencing deconvolution.

Isolation of gDNA from tumor tissues. gDNA was extracted from tissues using the DNeasy Blood and Tissue Kit (Qiagen) following manufacturer's instructions. Multiple tumor fragments or nodules were microdissected and either processed immediately by finely mincing the tissue and incubating overnight in a lysis buffer containing proteinase K and following the manufacturer protocol or snap-frozen in liquid nitrogen and stored at −80°C until day of processing. Resulting gDNA was resuspended in 100–200 µl of 10 mM Tris-Cl; 0.5 mM EDTA; pH 9.0, measured using a NanoDrop 2000 (ThermoFisher), and normalized before assessing genome editing at the relevant locus of interest using deep sequencing.

AMINEsearch bioinformatic pipeline. We developed a genome editing design tool, AMINEsearch, implemented in R, to comprehensively build libraries of annotated gene editing reagents to model a user-defined set of mutations. This algorithm can be applied to any sequencing dataset that uses standard maf format files. For further description of the algorithm and analysis, including the process of library design, offtarget analysis and conservation of variant protein sequence between human and mouse, see Supplementary Note 1.

HBES and MBES library design. MSK-IMPACT sequencing data ($n = 21{,}694$ tumors) was used to design sgRNAs and sensors compatible with commonly used BE configurations, incorporating Cas variants (SpCas9, Cas9-NG, xCas9 and ScCas9) combined in the FNLS or F2X (expanded window) BE vector variants (Supplementary Table 2a,b). IMPACT-derived outputs of AMINEsearch (Supplementary Table 2c,e) were used to compile unique sensor constructs to construct HBES (Supplementary Table 2d) and MBES libraries (Supplementary Table 2f) that target the human and mouse genome, respectively. These libraries served as the basis for experimental validation and screening of BE sensors, which are available under the 'Sensor validated' tab of the BE-SCAN web portal (https://dowlab.shinyapps.io/BEscan/).

AMINEsearch v.2. Modifications to the algorithm were made to increase functionality of the AMINEsearch algorithm. Specifically, modifications were made to accommodate BE variants that edit outside the region complementary to the sgRNA (CDA-BE4). As the demands of running larger dataset grew, we incorporated the capacity to run parallel execution on multiple cores or processors. The option to reverse the effects of mutations, rather than model them, given a list of pathogenic mutations as input, was added. A known issue was addressed to handle multiple genotypes that converge on the same protein sequence substitution as independent mutations. Version 2 includes the ability to track expected variant protein sequence when modeling human mutations in the mouse genome (Supplementary Note 1). Versions 1 and 2 of the algorithm can be accessed at https://github.com/Kastenhuber/AMINEsearch/tree/AMINEsearch_v1.0 and https://github.com/Kastenhuber/AMINEsearch/tree/AMINEsearch_v2.0, respectively, and are generalizable to analyze new mutation datasets and/or new base editor configurations.

Exploratory set of BE sensor predictions. We applied the algorithm to the recently updated mutation dataset from the MSK-IMPACT platform, containing 341,736 total somatic cancer SNVs derived from targeted sequencing of 47,550 tumor samples (Supplementary Fig. 17). This targeted sequencing panel captures the coding region of up to 580 genes. Candidates for BE included 5,542 unique SNVs, classified as missense, nonsense, splice site or nonstop mutations, which were observed six or more times (>0.01% frequency). We considered all combinations of 13 Cas9 orthologs and 11 deaminases, yielding 143 possible base editor configurations (Supplementary Table 8). This includes configurations that have been characterized extensively as well as combinations of Cas9 orthologs and deaminases that have not yet been assembled and used experimentally. Collectively, this exploratory set of sgRNA predictions provides a broad set of options to generate mutations in human and mouse (Supplementary Table 8). A searchable, filterable interface for the exploratory predicted set of sgRNAs are available under the 'Sensor Design—Human' and 'Sensor Design—Mouse' tabs in the shiny app web portal alongside sensor validated sgRNAs in BE-SCAN (https://dowlab.shinyapps.io/BEscan/).

Design and construction of MBES and HBES libraries. *BE sensor module design.* Each sensor module is composed of the following parts: (1) a 22 nt long 5′ adapter/priming site with an *Esp*3I restriction site; (2) a 20 nt long 5′ G-containing sgRNA; (3) a 93 nt long improved SpCas9 sgRNA scaffold partially based on Chen et al.[50];

(4) an 11 nt long sequence corresponding to the 5′ flanking sequence of the endogenous target site; (5) the 23 nt cognate target site; (6) a 7 nt long sequence corresponding to the 3′ flanking sequence of the endogenous target site and (7) a 28 nt long 3′ adapter/priming site with a *Eco*RI restriction site. Thus, oligos encoding individual sensor modules are 204 nt long.

Cloning of MBES and HBES libraries. Due to longer-than-average oligonucleotide length, early attempts at design and construction of sensor libraries showed unacceptable synthesis and assembly error rates where, in some instances, over half of the sensors before or after assembly into the backbone were found to harbor insertions, deletions, single nucleotide mutations and incorrect chimeric sgRNA-target site molecules (data not shown). Through extensive trial and error, we found that assembling sensor libraries using Agilent's High Fidelity oligonucleotide synthesis platform significantly mitigated these issues.

All-PAM sensor (APS) (1,152 oligos white-listed), MBES (4,686 oligos) and HBES (5,855 oligos) libraries were cloned into the pLRT2B backbone[10] as follows (all library oligos are listed in Supplementary Table 2). Briefly, each oligonucleotide pool was amplified using forward and reverse primers that append *Esp*3I and *Eco*RI sites to the 5′ and 3′ ends of the sensor insert, purified using the QIAquick PCR Purification Kit (Qiagen) and ligated into *Esp*3I-digested and dephosphorylated pLRT2B vector using high-concentration T4 DNA ligase (NEB) (all cloning and sequencing oligos are listed in Supplementary Table 9). To ensure maximum library recovery, we set up $n = 24$ parallel PCR reactions per pool. A minimum of 2.4 µg of ligated pLRT2B plasmid DNA per pool (corresponding to $n = 8$ ligations) was electroporated into Endura electrocompetent cells (Lucigen), recovered for 1 h at 37°C, plated across four 15-cm Luria-Bertani (LB)-carbenicillin plates (Teknova) and incubated at 37°C for 16 h. The total number of bacterial colonies per pool was quantified using serial dilution plates to ensure a library representation of >10,000×. The next morning, bacterial colonies were scraped and briefly expanded for 4 h at 37°C in 500 ml of LB-carbenicillin. Plasmid DNA was isolated using the Plasmid Plus Maxi Kit (Qiagen). To assess sensor distribution and fidelity of assembly per pool, we amplified the sensor region using primers that append Illumina sequencing adapters on the 5′ and 3′ ends of the amplicon, as well as a random nucleotide stagger and unique demultiplexing barcode on the 5′ end (Supplementary Table 9). Library amplicons were size-selected on a 2.5% agarose gel, purified using the QIAquick Gel Extraction Kit (Qiagen) and sequenced on an Illumina MiSeq instrument.

Analysis of BE activity using the GO reporter system. Base-editor-expressing cells were plated at a density of 5,000 cells per well in 12-well plates and transduced 24 h later with a defined amount of GO reporter to achieve 20–50% transduction efficiency. Virus-containing medium was replaced with complete medium 24 h post-transduction, and cells were harvested for flow cytometry at 96 h post-transduction. We used an Attune NxT flow cytometer (Thermo Fisher). Cells were trypsinized with a 100 µl of 0.25% Trypsin+EDTA and resuspended in 300 µl of complete medium in a 96-well U-bottom plate. Data were acquired at a flow rate of 500 µl min^{-1} and at least 10,000 events from the single cell population gating were recorded.

Screening and deconvolution of MBES and HBES libraries. *Screening of MBES and HBES libraries.* We first screened the APS library in MDA-MB-231 cells expressing one of nine different base editors, as well as either the Cas9 or Cas9-NG nucleases as cutting controls. APS screens were performed essentially as described below in detail. We then screened a total of five mouse and human base-editor-expressing cell lines with either MBES or HBES libraries using the following approach. Human cell lines (MDA-MD-231 and PC9) were screened with MBES to minimize fitness differences between sensor modules due to endogenous targeting of genes that suppress cellular proliferation. Following the same rationale, mouse cell lines (KPT1, NIH3T3 and PDECs) were screened with HBES. Each screen (including the APS set) was performed as follows. To ensure that most cells harbor a single sgRNA integration event, we determined the volume of viral supernatant that would achieve an MOI between around 0.3 and 0.5 upon standard transduction of a population of base editor-expressing cells. All screens were performed in technical duplicate and each step of the screen—from infection to sequencing—was optimized to achieve a minimum representation of 1,000×. For instance, to ensure a representation of >1,000× for HBES libraries at the transduction step, we spinfected a total of 24 million cells across two 12-well plates per technical replicate using the volume of viral supernatant that would achieve a 30% infection rate (around 7.2 million transduced cells per technical replicate). At 24 h after infection, cells from each corresponding replicate were pooled into a minimum of 2 × 150 mm tissue culture dishes (Corning) and selected with Blasticidin S (Gibco) at an empirically determined final concentration ranging from 5 µg ml^{-1} to 30 µg ml^{-1} depending on the cell line. Cells were cultured and kept under Blasticidin selection for 7 days post-transduction. When needed, cells were trypsinized and replated at a minimum of 6 million cells per replicate to ensure a minimum representation of 1,000×. For PDEC screens, cell representation per replicate was maintained at greater than 600× at all points. Subsequently, at least 6 million cells were pelleted and stored at −20°C. gDNA from cells was isolated using the DNeasy Blood and Tissue Kit (Qiagen) following the manufacturer's guidelines. Genomic DNA was harvested from all timepoints and both sensor BE activity and sgRNA abundance were assessed via next-generation sequencing (NGS).

Deconvolution of MBES and HBES screens. We assumed that each cell contains approximately 6.6 pg of gDNA. Therefore, screen deconvolution at 1,000× required sampling around 6 million × 6.6 pg of gDNA, or around 39.6 µg. We employed a modified two-step PCR version of the protocol published by Doench et al.[51] adapted to our unique library design. Briefly, we performed an initial PCR, whereby the integrated sensor cassettes were amplified from gDNA, followed by a second PCR to append Illumina sequencing adapters on the 5′ and 3′ ends of the amplicon, as well as a random hexamer and unique demultiplexing barcode on the 5′ end. Each 'PCR1' reaction contained either 25 µl of Q5 High-Fidelity 2× Master Mix (NEB), 2.5 µl of Sensor_v6_Fwd Primer (10 µM), 2.5 µl of Sensor_v6_Rev Primer (10 µM), and 5 µg of gDNA in 20 µl of water (for a total volume of 50 µl per reaction) or 10 µl of Herculase II 5× Master Mix (Agilent), 0.5 µl dNTPs, 2.5 µl of Sensor_v6_Fwd Primer (10 µM), 2.5 µl of Sensor_v6_Rev Primer (10 µM), 1 µl of Herculase II polymerase, and 5 µg of gDNA in 33.5 µl of water (for a total volume of 50 µl per reaction). The number of PCR1 reactions was scaled accordingly; therefore, we performed eight PCR1 reactions per technical replicate and timepoint for all screens. PCR1 amplicons were purified using the QIAquick PCR Purification Kit (Qiagen) and used as template for 'PCR2' reactions. Each PCR2 reaction contained either 25 µl of NEBNext 2× Master Mix (NEB), 2.5 µl of a uniquely barcoded PCR2_Fwd Primer (10 µM), 2.5 µl of a common PCR2_Rev Primer (10 µM), and 300 ng of PCR1 product in 20 µl of water (for a total volume of 50 µl per reaction). We performed two PCR2 reactions per PCR1 product. Library amplicons were size-selected either on a 2.5% agarose gel and purified using the QIAquick Gel Extraction Kit (Qiagen) or using AMPure XP beads (Beckman Coulter) followed by normalization, pooling and sequencing on an Illumina NextSeq 500 instrument (150 nt paired-end reads). All primer sequences are available in Supplementary Table 9. PCR program for PCR1 using Q5 High-Fidelity 2× Master Mix (NEB) was: (1) 98°C for 30 s; (2) 98°C for 10 s; (3) 55°C for 30 s; (4) 72°C for 30 s; (5) Go to step 2 for 24 cycles; (6) 72°C for 2 min; (7) 4°C forever. When using Herculase II, denaturation steps were done at 95°C and the initial denaturation lasted for 2 min. PCR program for PCR2 using NEBNext 2X Master Mix (NEB) was: (1) 98°C for 30 s; (2) 98°C for 10 s; (3) 65°C for 30 s; (4) 72°C for 30 s; (5) Go to step 2 for 17 cycles; (6) 72°C for 2 min; (7) 4°C forever.

Analysis of MBES and HBES screening data. To quantify BE outcomes, raw paired-end FASTQ reads were paired using Pandaseq and merged FASTQ files were used as input for downstream analysis. We first removed reads with mutated sgRNAs or scaffolds, or reads with nonmatching sgRNA and target sequences (due to template switching during PCR amplification). Next, the 5′ scaffold and linker were used to associate each sgRNA with the read; sgRNAs that did not match the whitelist were also discarded. All remaining reads were aligned to their cognate target found in the whitelist, and aligned reads with no indels were considered for BE analysis. Sensors that deviated from the expected length were flagged as indels and their actual frequency calculated as a specific insertion or deletion. Editing events were classified over all cytosines within a −5 to 20 position window of the target, where 1 is defined as the first position of the protospacer. Target cytosine editing (tCTN, tCGN, tCAN) quantified the frequency of editing at the target cytosine regardless of editing at other adjacent cytosines. Target cytosine editing without collateral editing (tCT) was measured as specific C > T editing without associated mutations at adjacent sites. Purity values were calculated as the ratio of tCT/tCTN. Custom code to perform the BE analysis is available at: https://github.com/schmidt73/base-editing-analysis. All data for APS, HBES and MBES screens can be found in Supplementary Tables 1, 4 and 5, respectively.

Comparison of BE sensor screening outcomes with BE-Hive. To consider the performance of the BE-Hive BE outcome prediction model in relation to our data, we pulled the fully trained network from the Github repository linked by Arbab et al.[21]. This model is split into two parts: a component that predicts the probability of any edit occurring, and another that predicts the probability of a specific BE outcome, conditioned on any editing occurring. These are referred to as the editing efficiency and bystander model, respectively. Per instructions, we fed in our input spacer and its 50-mer context to both models and computed the posterior probability of each observed outcome using the chain rule. One caveat is that the editing efficiency model directly predicts an untransformed score, not a probability. To convert this to a probability, we sigmoid transformed it into the unit interval [0, 1]. Following the recommendations provided in their README file, we first rescaled the score linearly using the mean and s.d. of reads before sigmoid transforming it into a probability to account for variance in base editor expression by experimental condition and cell type. We note that this is a monotonic operation; therefore, it should not affect any SpearmanR correlations used to analyze performance.

BE-Hive does not consider PAM sequence as a feature in their prediction mode. Given that the PAM is an important determinant of the activity of standard base editors, we considered two different sets of sgRNAs in our comparisons. First, we considered all sgRNAs used in our screen. Second, we considered only the sgRNAs with canonical NGG PAMs.

Phenotypic screens using BE sensor libraries. Stable base editor-expressing *Kras*$^{G12D/+}$; *Trp53*$^{WT/WT}$ PDECs[20] were generated by lentiviral transduction with FNLS (Addgene, catalog no. 110841) and validated using GO[12] (Addgene, catalog no. 136896). Phenotypic MBES screens in FNLS-PDECs were set up essentially as described above for HBES/MBES sensor screens with a few modifications. MBES FNLS-PDEC screens were performed across six independent transduction replicates in parallel. Each replicate was maintained at a minimum 500× representation at every step of the screen by replating 3 million cells per timepoint and pelleting the rest of the cells for gDNA isolation and screen deconvolution. Screens ran for approximately 36 cumulative population doublings across 34 days, after which we isolated gDNA and proceeded to perform screen deconvolution essentially as described above for MBES/HBES screens.

Proliferation screen analysis. Paired-end reads were joined using Pandaseq. Merged reads were processed as described for BE analysis above. Total read counts for each replicate were used as input for MAGeCK[27,28] analysis. Any sgRNA with read counts <100 were removed from analysis. Comparisons of T0 (day 5 post-transduction) versus T1 (day 14) and T2 (day 30 post-transduction) for each replicate were performed using MAGeCK to determine log fold changes. BE outcomes at the sensor target site were measured concurrently using the sensor screen pipeline described above.

Validation experiments. For validation of individual targets, sgRNAs were cloned into the lentiviral guide expression vector LRT2B (Addgene, catalog no. 110854) and lentiviral particles were produced as described above. Base-editor-expressing cells were plated at a density of 25,000 cells per well in 12-well plates and were infected 24-h later with enough virus to achieve 50% transduction efficiency. Virus-containing medium was replaced with complete medium 24 h post-transduction and cells were plated into selection medium containing 3 μg ml^{-1} Blasticidin S (Gibco). Experimental cells remained in selection medium until the final collection time point at 7 days post-transduction. Final LRT2B infection efficiency was determined by measuring the levels of tdTomato in 10% of the cells remaining at day 7 using flow cytometry. Genomic DNA was isolated using the protocol found on dowlab.org/protocols, and targets were amplified using a 100 μl reaction following the standard NEB *Taq* 2× MM protocol with primers found in Supplementary Table 9. Each PCR was performed three times per target and pooled. Amplicons were confirmed on a 2% agarose gel and PCR purified using Qiagen QIAquick PCR purification kit. DNA concentration was measured using a Nanodrop and samples were normalized to 20 ng μl^{-1} and sequenced using EZ-amplicon sequencing (MiSeq; 2 × 250 bp) by GENEWIZ, Inc.

Analysis of deep sequencing data from validation experiments. CRISPResso2 (ref. [48]) was used to process sequencing reads from the validation experiments and the corresponding sensor sequencing results for each individual target. The data was analyzed on default CRISPResso2 base editor mode with exceptions to the following parameters for endogenous locus results: -quantification_window_center -15 and sensor results: --quantification_window_size 10 --quantification_window_size 10 --base_editor_output --quantification_window_center -15 --exclude_bp_from_right 1 --plot_window_size 18. To calculate target C > T editing and noncanonical editing, we used the 'Alleles_frequency_table_around_sgRNA.txt' file to get the read counts for a specific allele.

Statistical analyses and data visualization. *Analysis and data visualization in R.* Heatmaps, dotplots and correlation analyses (including correlation graphs) were performed in R v.3.6.3 and plots were produced using the ggplot2 and ggpubr package. Statistical considerations are reported in each figure legend.

Analysis and data visualization in GraphPad PRISM. More bar plots, survival curves and associated statistical analyses were generated using Prism 8 (GraphPad) and are indicated in figure legends. Error bars represent s.d., unless otherwise noted. We used Student's *t*-test (unpaired, two-tailed) to assess significance between experimental and control groups, and to calculate *P* values. $P < 0.05$ was considered statistically significant. Schematics were created using BioRender.com.

Source data availability. All source data (including *P* values) are available in Supplementary Table 10. Processed screening data is available in Supplementary Tables 1,4,5 and primary data has been deposited in the SRA repository under accession PRJNA746395.

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

Data availability

All source data (including *P* values) are available in Supplementary Table 10. Processed screening data is available in Supplementary Tables 1,4,5 and primary sequencing data is available at the Sequence Read Archive (SRA) under accession PRJNA746395.

Code availability

Code for analysis and data visualization is available at: https://github.com/schmidt73/base-editing-analysis, https://github.com/Kastenhuber/AMINEsearch and https://github.com/lukedow/BEsensor

References

49. Soto-Feliciano, Y. M. et al. A molecular switch between mammalian MLL complexes dictates response to Menin-MLL inhibition. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.22.465184 (2021).
50. Chen, B. et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).
51. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).

Acknowledgements

We thank D. Solit, Ni. Schultz, M. Berger and B. Gross for access to MSK-IMPACT data, T. Jacks for sharing KP cells, D. Alonso-Curbelo and Dafna Bar-Sagi for sharing PDEC cells, M. Paz Zafra for sharing primers to assess tumor purity, T.M. Norman for conceptual advice and L. Cantley for support and mentorship. We gratefully acknowledge the members of the Molecular Diagnostics Service in the Department of Pathology, the Integrated Genomics Operation and Bioinformatics Core (P30 CA008748) and the Marie-Josée and Henry R. Kravis Center for Molecular Oncology. This work was supported by a project grant from the NIH/NCI (R01CA229773-01A1), P01 CA087497 (SWL), a MSKCC Functional Genomics Initiative (FGI) grant (SWL) and an Agilent Technologies Thought Leader Award (SWL). F.J.S.-R. was supported by the MSKCC TROT program (5T32CA160001), a GMTEC Postdoctoral Researcher Innovation Grant, and is an HHMI Hanna Gray Fellow. B.J.D. was supported by an F31 Ruth L. Kirschstein Predoctoral Individual National Research Service Award (F31-CA261061-01). E.R.K. was supported by an F31 Ruth L. Kirschstein Predoctoral Individual National Research Service Award (F31-CA192835) and is currently supported by NCI R35CA197588, awarded to L. Cantley. A.K. was supported by an F31 Ruth L. Kirschstein Predoctoral Individual National Research Service Award (F31-CA247351-02). J.L. was supported by the German Research Foundation (DFG) and the Shulamit Katzman Endowed Postdoctoral Research Fellowship. S.V.P. was supported by the German Academic Scholarship Foundation. F.M.B. was supported by a GMTEC Postdoctoral Fellowship, an MSKCC's Translational Research Oncology Training Fellowship (5T32CA160001-08), and a Young Investigator Award from the Edward P. Evans Foundation. K.M.T. is supported by the Jane Coffin Childs Memorial Fund for Medical Research. D.C. and H.Z. acknowledge funding from the MSKCC Marie-Josée and Henry R. Kravis Center for Molecular Oncology for supporting OncoKB. S.W.L. is the Geoffrey Beene Chair of Cancer Biology and an Investigator of the Howard Hughes Medical Institute. L.E.D. is the Burt Gwirtzman Research Scholar in Lung Cancer at Weill Cornell Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author contributions

F.J.S.-R, B.J.D., E.R.K. and L.E.D. conceived the project. F.J.S.-R. and B.J.D. performed experiments, analyzed data, and wrote the paper. E.R.K. wrote code related to AMINEsearch and associated computational analyses and wrote the paper. H.S. and Y.-J.H. performed computational analyses and wrote code for sgRNA predictions and analysis. A.K., M.K., V.T., J.L., S.V.P., F.M.B., K.C., S.G., A.N.W., J.M.S. and KMT performed experiments and analyzed data. D.C. and H.Z. assisted with OncoKB analyses. C.S.L. and S.W.L. supervised experimental and computational work. L.E.D. performed and supervised experiments, analyzed data, and wrote the paper.

Competing interests

L.E.D. is a scientific advisor and holds equity in Mirimus Inc. and is a consultant for Volastra Therapeutic and Fraizer Healthcare S.W.L. is an advisor for and has equity in the following biotechnology companies: ORIC Pharmaceuticals, Faeth Therapeutics, Blueprint Medicines, Geras Bio, Mirimus Inc. and PMV Pharmaceuticals. S.W.L. acknowledges receiving funding and research support from Agilent Technologies for the purposes of massively parallel oligo synthesis. The remaining authors declare no competing interests.

Additional information

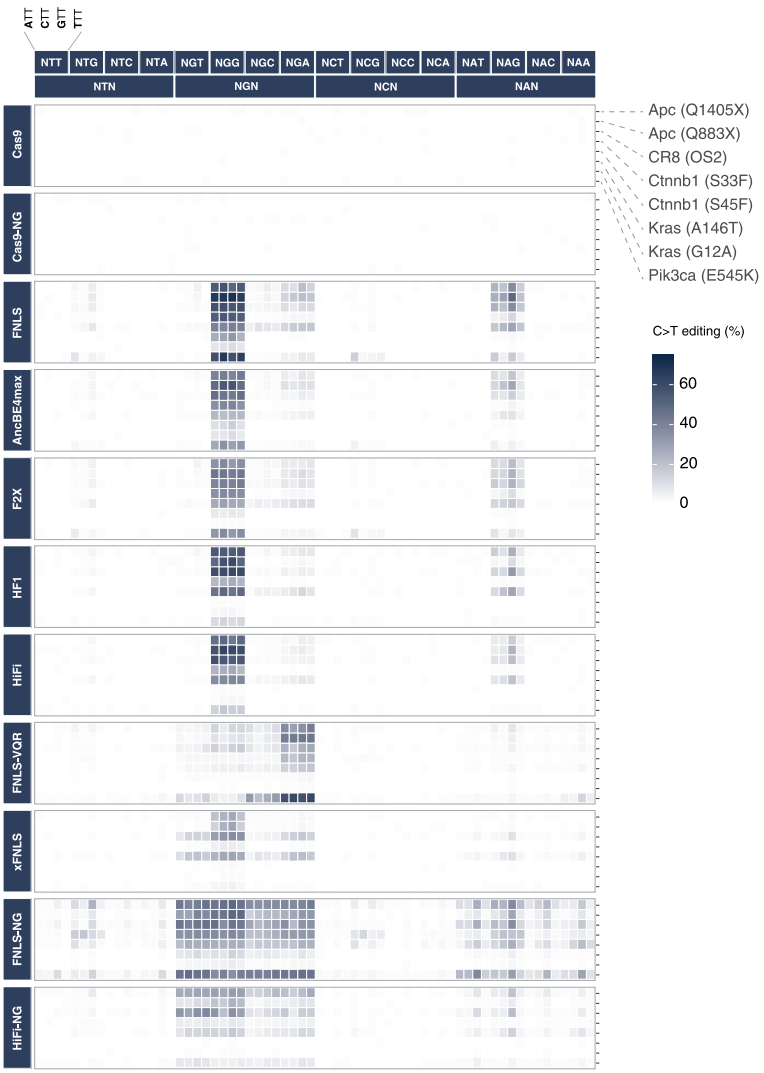
Extended data is available for this paper at https://doi.org/10.1038/s41587-021-01172-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01172-3.

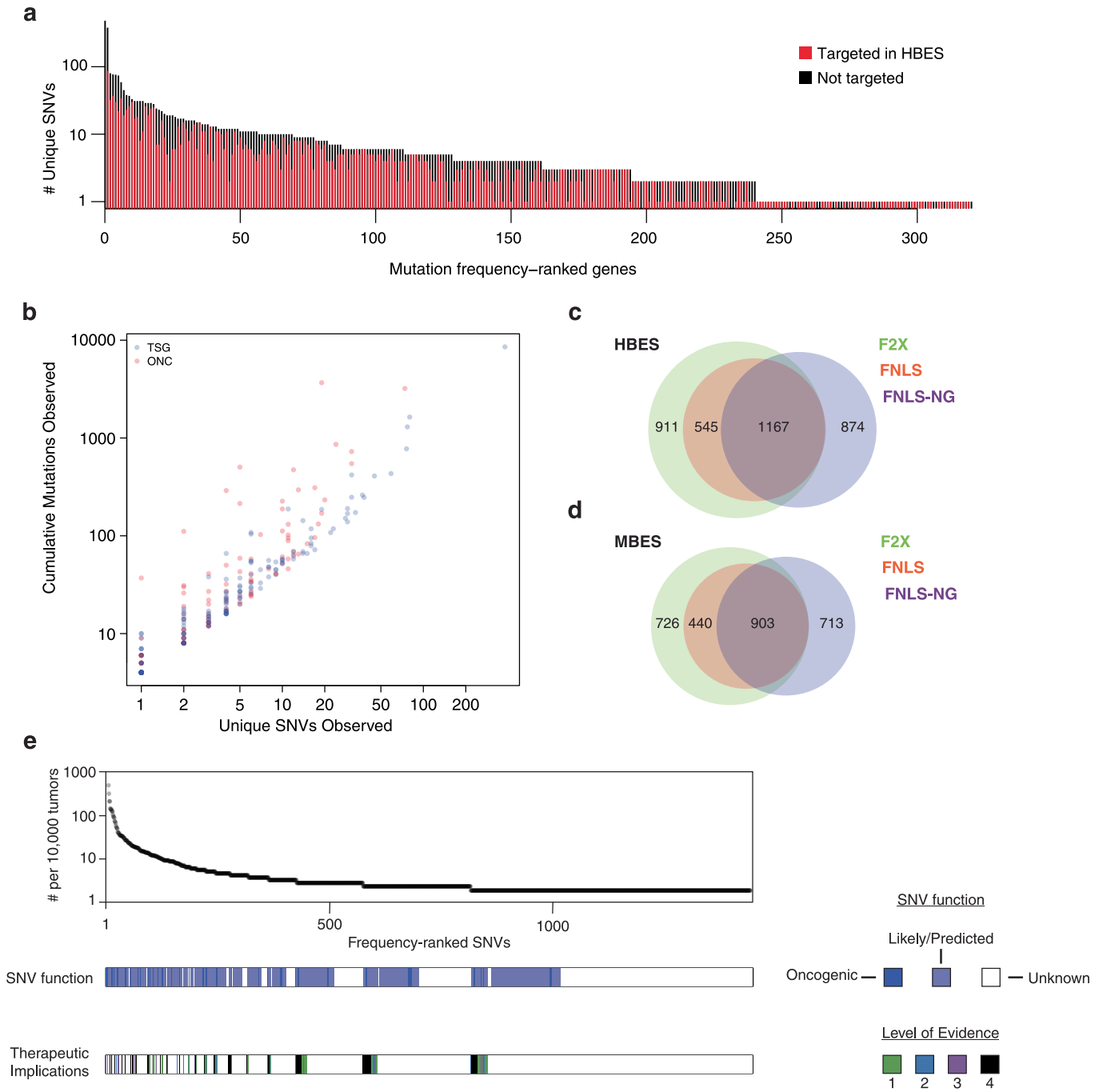
Correspondence and requests for materials should be addressed to Lukas E. Dow.

Peer review information *Nature Biotechnology* thanks the anonymous reviewers for their contribution to the peer review of this work.

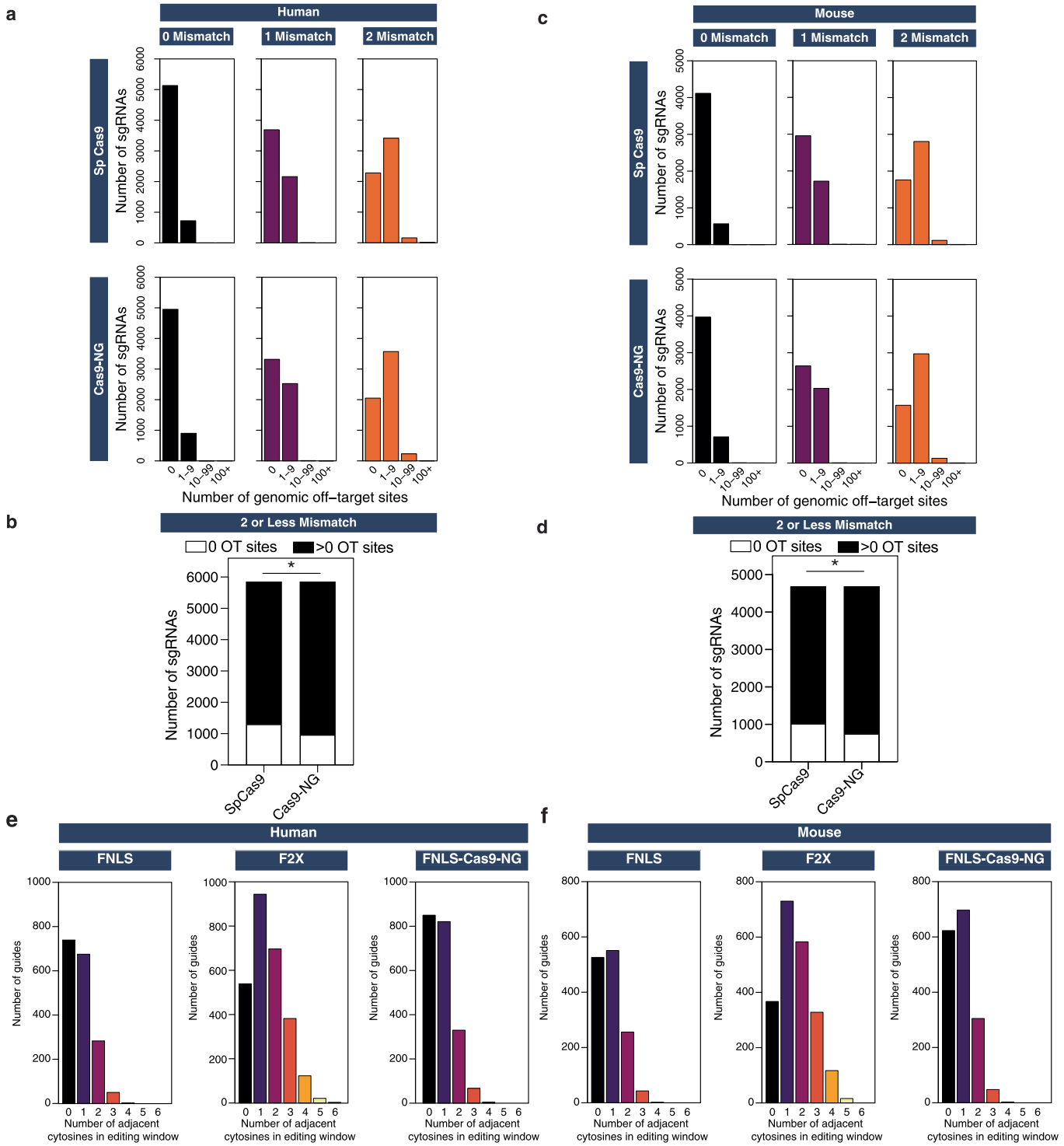
Reprints and permissions information is available at www.nature.com/reprints.



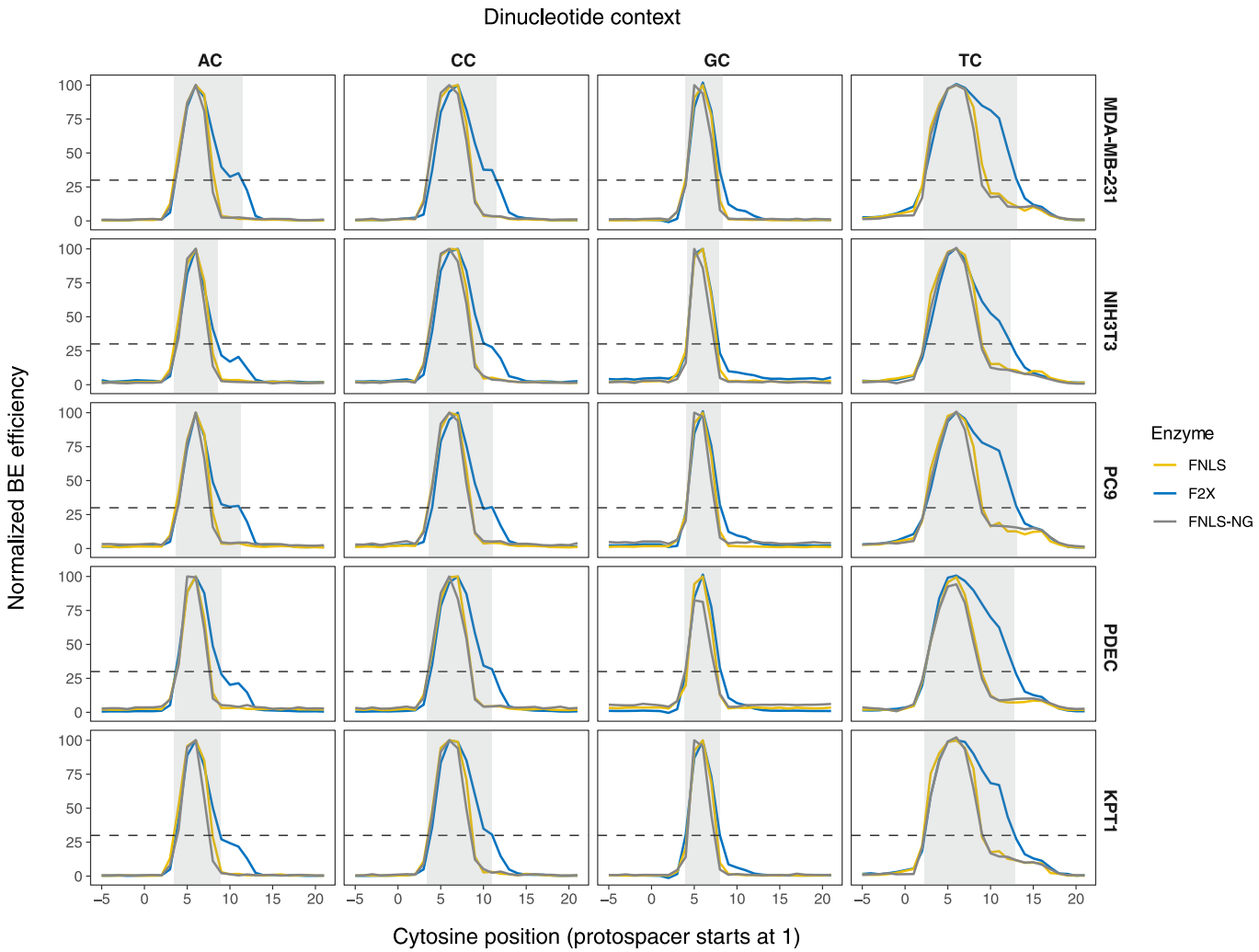
Extended Data Fig. 1 | BE efficiency for mouse sgRNAs in the APS library. C > T editing efficiency (%) at each APS library mouse target site across base editor enzymes, as indicated. Cas9 and Cas9-NG serve as nuclease controls. Rows denote sgRNAs; columns denote PAM subclass.



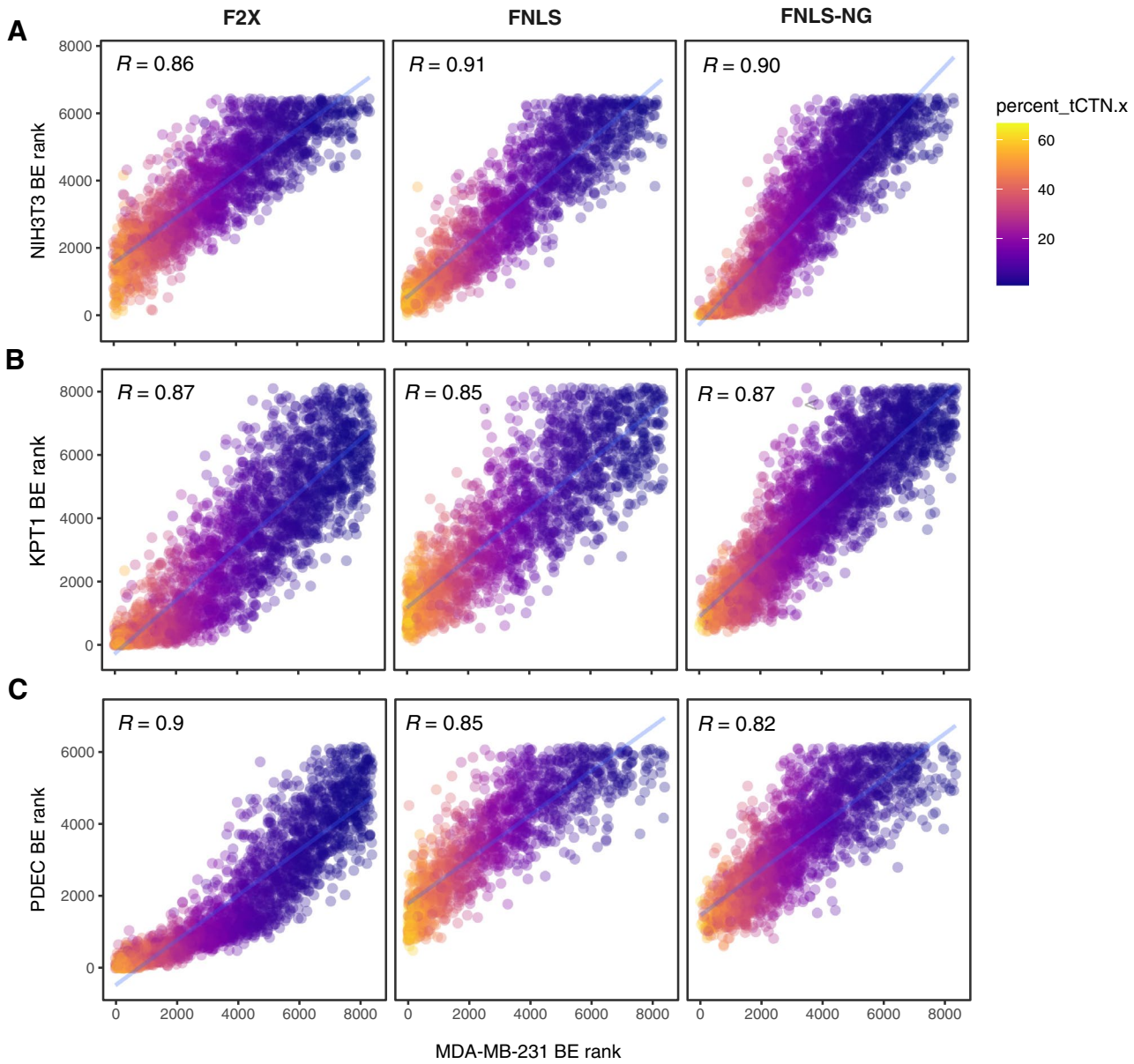
Extended Data Fig. 2 | Cancer somatic mutation-derived base editing sensor libraries. (a) Number of unique recurrent SNVs per gene, ordered by mutation frequency of gene. Bars are split to indicate proportion of SNVs targeted (red) or not (black) in the HBES library. **(b)** Focality of mutations by cancer gene classification. Number of cumulative mutations observed in recurrent sites with respect to the number of unique SNVs observed per gene. Oncogenes are indicated by red dots and tumor suppressor genes are indicated by blue dots. Mutations in oncogenes tend to be more focal on distinct hotspot sites, with greater number of recurrent mutations per unique SNV allele (11.1 vs 6.2 mutations per unique recurrent SNV, $p = 0.011$, two-tailed t-test). **(c)** Venn diagram of sgRNAs in HBES library compatible with each base editor configuration. **(d)** Venn diagram of sgRNAs in MBES library compatible with each base editor configuration. **(e)** SNV-level annotation with each color bar sorted in order of observed mutation frequency (top). NV characteristics are indicated, including oncogenic function (OncoKB assertion of oncogenic/Likely oncogenic/VUS) and therapeutic implications (OncoKB highest level of evidence for drug sensitivity or resistance) {Chakravarty, 2017 #76;Chakravarty, 2021 #105}.



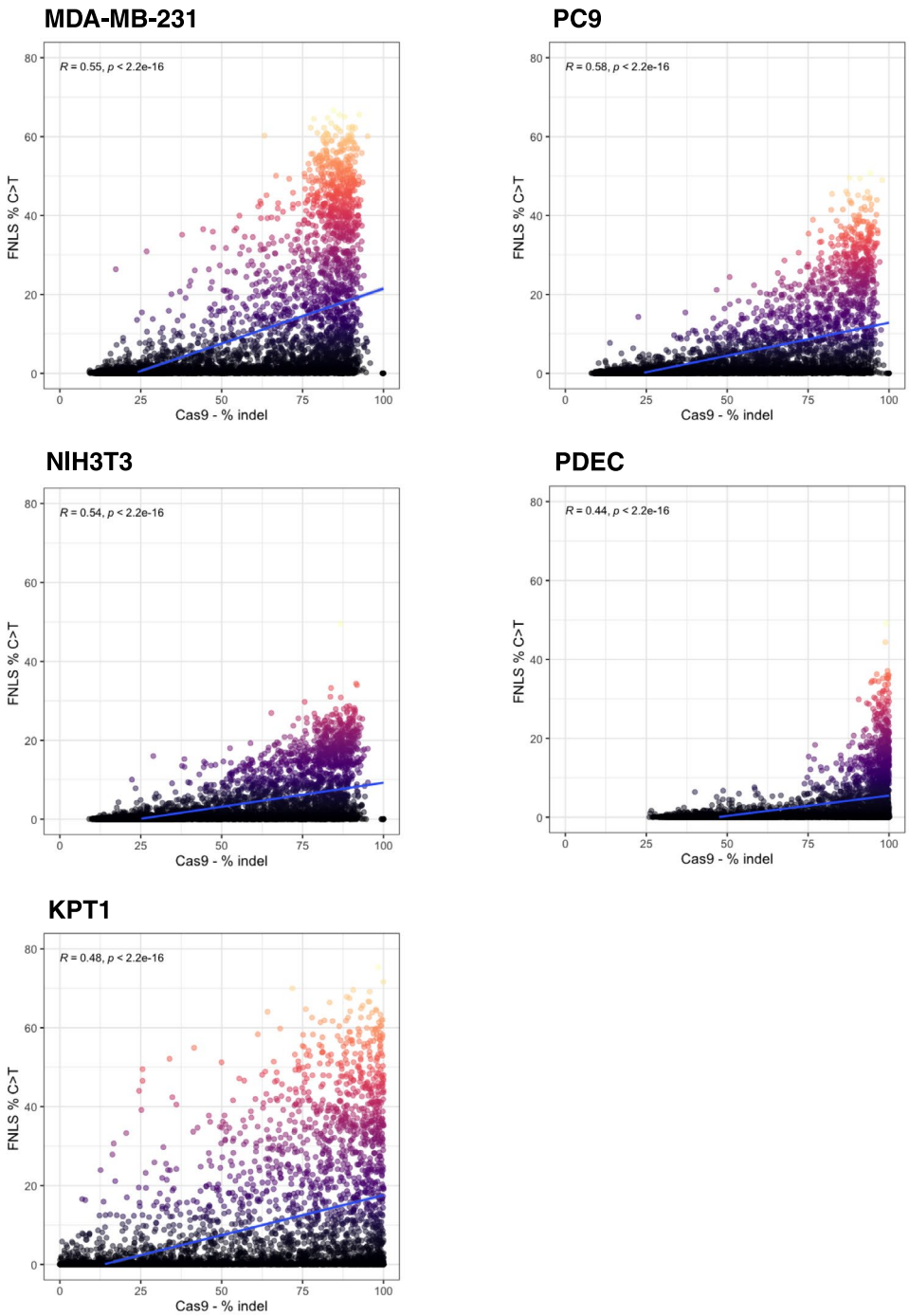
Extended Data Fig. 3 | Off-target editing predictions for base editing sensor libraries. (a) For sgRNAs in HBES library, distribution of potential off-target (OT) sites identified by PAM specificity and extent of mismatch. **(b)** Number of sgRNAs in HBES library targeting the human genome with 0 (white) and 1 or more (black) predicted OT sites depending on SPCas9 or Cas9-NG PAM specificity. A greater number of sgRNAs have no predicted OT sites used in conjunction with SpCas9 than with Cas9-NG. $p < 2.2e\text{-}16$, 2-sided Fisher's exact test. **(c)** For sgRNAs in HBES library, distribution of potential OT sites identified by PAM specificity and extent of mismatch. **(d)** Number of sgRNAs in MBES targeting mouse genome with 0 (white) and 1 or more (black) predicted OT sites depending on SPCas9 or Cas9-NG PAM specificity. A greater number of sgRNAs have no predicted OT sites used in conjunction with SpCas9 than with Cas9-NG. $p < 2.2e\text{-}16$, 2-sided Fisher's exact test. (e) Distribution of not-target editable bases (C for CBE) within the editing window for HBES library targeting human genome. **(f)** Distribution of not-target editable bases (C for CBE) within the editing window for MBES library targeting mouse genome.



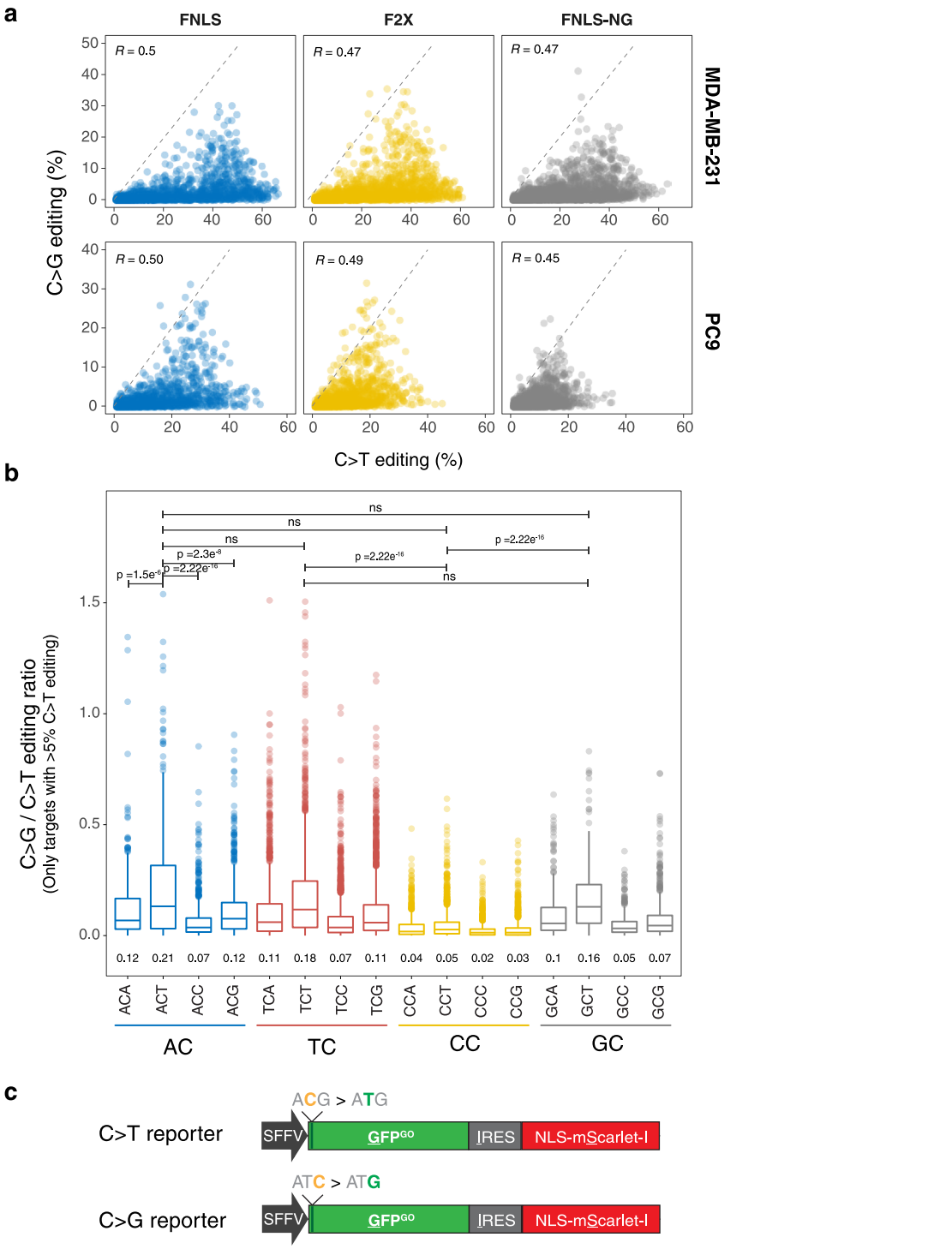
Extended Data Fig. 4 | Comparison of editing range (editing window) across FNLS, F2X, and FNLS-NG base editors as a function of dinucleotide context. Plots represent the mean normalized BE editing efficiency for each base editor (FNLS = yellow, F2X = blue, FNLS-NG = gray) across 5 cell lines (rows) and 4 dinucleotide contexts (columns). Area shaded in grey denotes maximum editing range in each condition where normalized BE is above 30% (dotted line).



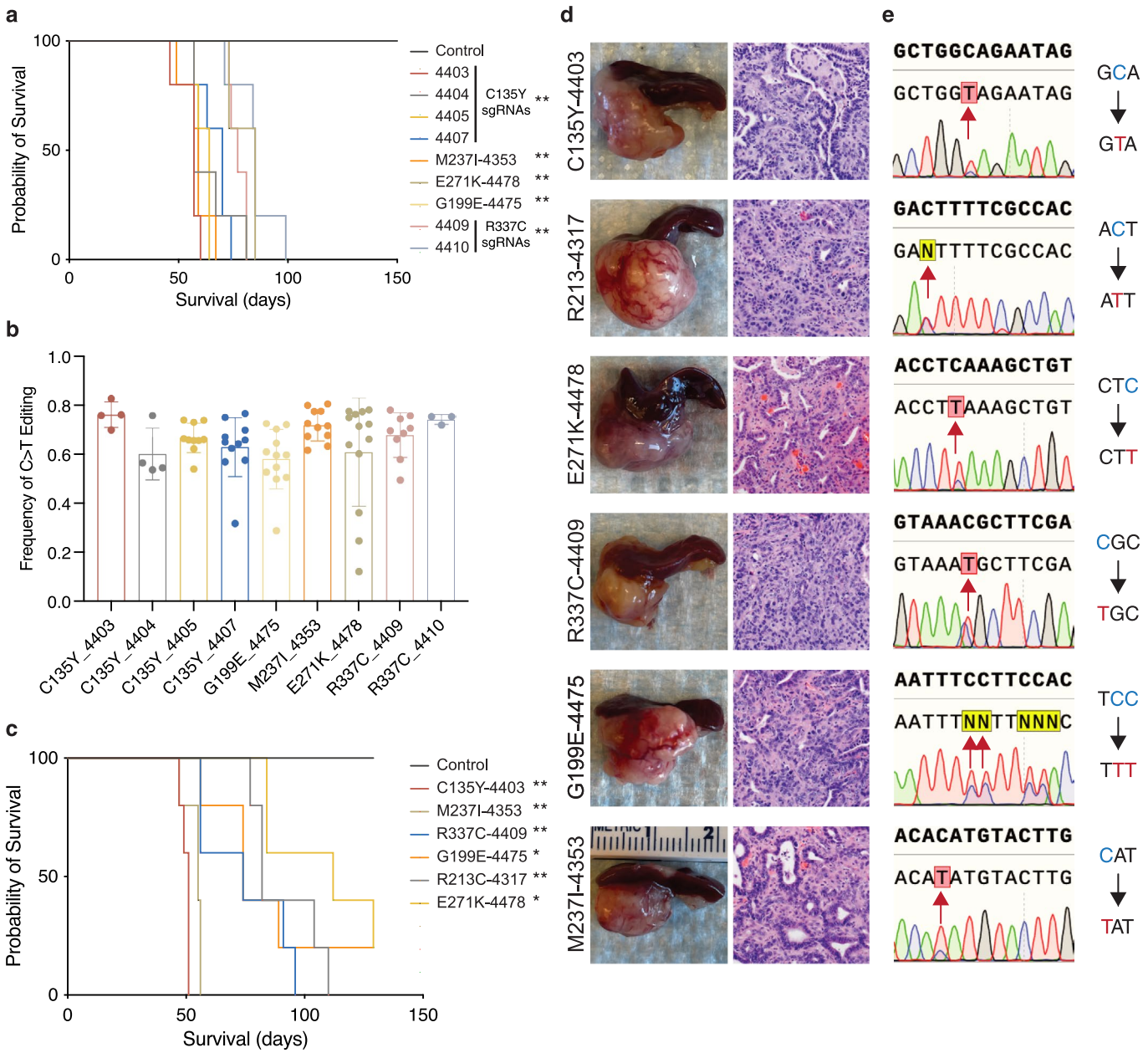
Extended Data Fig. 5 | Correlation of sgRNA efficiency ranking. Plots represent correlation of individual sgRNA efficiency rankings between MDA-MB-231 and NIH3T3, KPT1, and PDEC cells, as indicated. To reduce noise created by low efficiency sgRNAs, only HBES sgRNAs that had >1% activity in the sensor were included. Pearson correlation coefficients are shown; for all comparisons, $p < 2.22$ e-16.



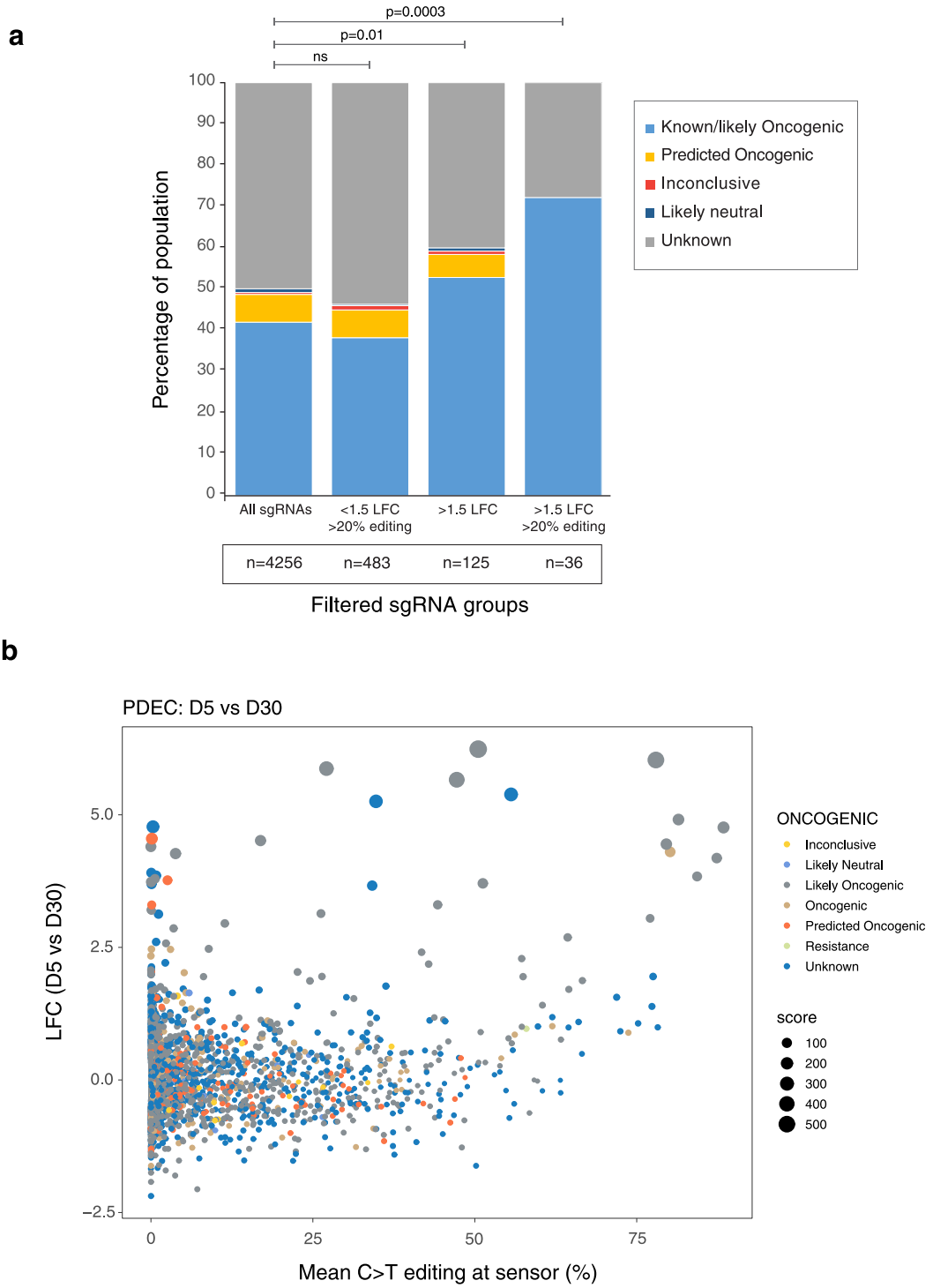
Extended Data Fig. 6 | Indel and BE correlation across cell lines. Correlation of indel and C > T editing frequencies for all sgRNAs in the HBES library across 5 screen cell lines. Pearson correlation coefficients were calculated using ggpubr(0.4.0) package in R, the p value represents the significance of two-sided t-test.



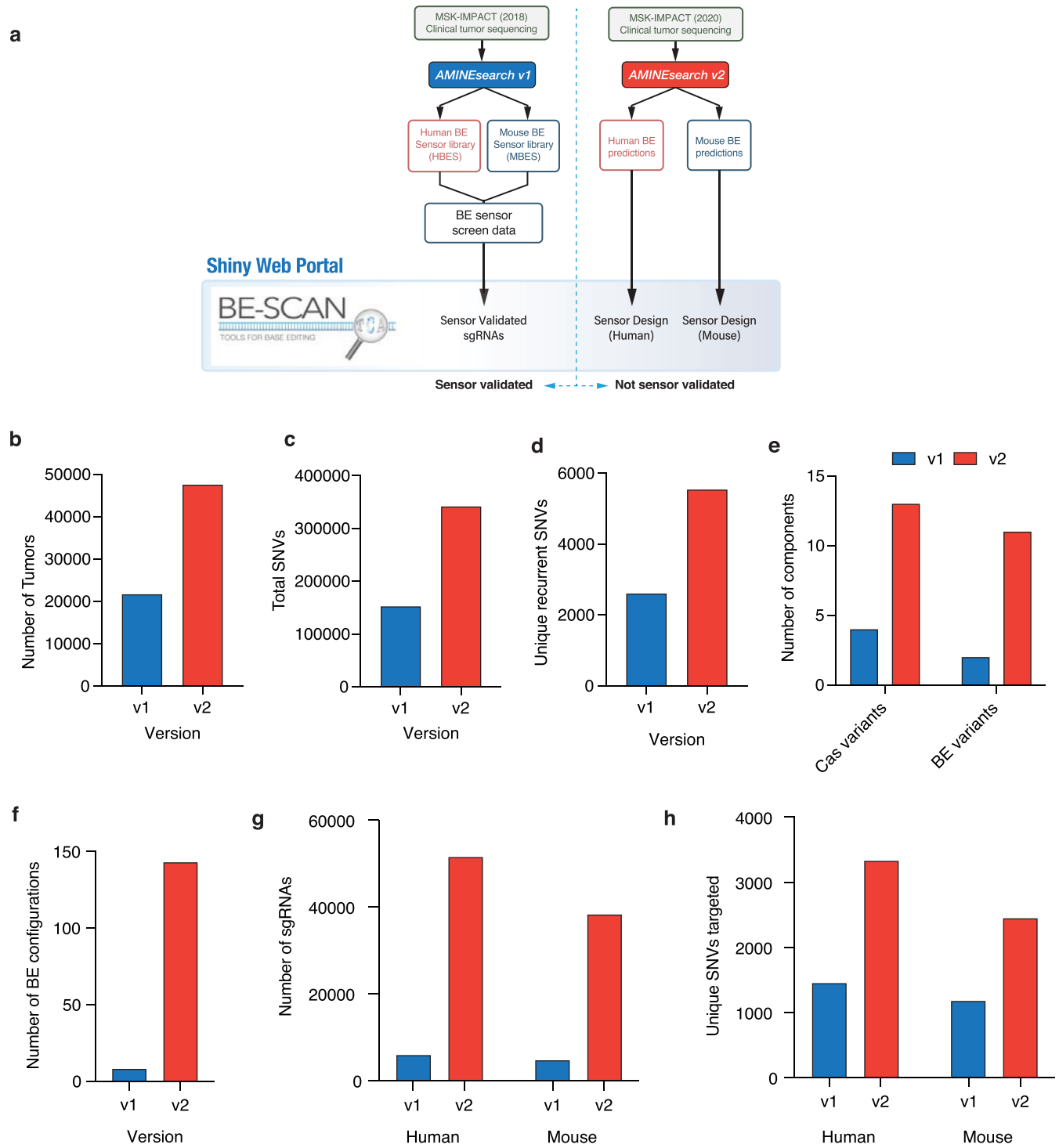
Extended Data Fig. 7 | Non-canonical cytosine editing identified by BE Sensor. (a) Dotplots show percent C>T and C>G editing for individual target cytosines in the HBES library across three BE enzymes (FNLS, F2X and FNLS-NG) and two cell lines (MDA-MB-231 and PC9). Scales on x and y axes are not the same; dotted lines indicate 1:1 ratio **(b)** Ratio of C > G/C > T editing in FNLS-MDA-MB-231 cells transduced with the HBES library classified by dinucleotide context (fill) and trinucleotide context (column). Data includes all base editors (FNLS, F2X and FNLS-NG) and is filtered for sgRNAs that show more than 5% C > T editing in the sensor assay. Boxplots show the median and interquartile range (IQR) and whiskers represent 1.5*IQR. Outliers are shown as individual points. ns indicate p > 0.05; p values were determined with two-sided Wilcoxon signed rank test. Complete list of all comparisons is available in Supplementary Table 10g. **(c)** Schematic of (C > G) reporter developed by modifying the GO (C > T) reporter.



Extended Data Fig. 8 | In vivo validation of cancer-associated TP53 missense mutations using BE. (a) Survival analysis of mice transplanted with F2X-expressing PDECs transduced with specific Trp53-targeting base editing sgRNAs. N = 5 mice per sgRNA per mutation. **(b)** Frequency of target C > T editing in tumors from transplanted mice. Each individual point represents a single isolated tumor (n = 3+ per sgRNA) Target C > T editing was measured by next generation sequencing of amplified target loci and data was analyzed using CRISPResso2. Data are presented as +/- SD. **(c)** In vivo validation of M237I and C135Y mutations via orthotopic transplantation of FNLS-expressing PDECs transduced with sgRNAs designed to introduce the corresponding mutations in the mouse Trp53 gene (M234I and C132Y, respectively). N = 5 mice per mutation. **(d)** Representative macroscopic (left) and microscopic (right; H&E) images of pancreatic tumors isolated from mice transplanted with FNLS-expressing PDEC cells transduced with specific Trp53-targeting base editing sgRNAs. **(e)** Representative Sanger sequencing traces from tumors in **(d)**. Red arrows denote target cytosines that, when mutated to thymine, give rise to the corresponding amino acid changes in the p53 protein. Nucleotide triplets on the right denote the precise mutational events that give rise to mutant p53 proteins. * $p \leq 0.05$, ** $p \leq 0.01$. P-values were calculated using the log-rank test.



Extended Data Fig. 9 | Classification of screen hits by OncoKB. (a) sgRNAs from the MBES proliferation screen were binned by categories: i) all sgRNAs; ii) sgRNAs depleted by <1.5 LFC and exhibiting 20% editing at the sensor; iii) sgRNAs enriched by >1.5 LFC; or iv) sgRNAs enriched >1.5 LFC and exhibiting 20% editing at the sensor followed by calculation of the percentage of each OncoKB classification. P-values indicate two-sided Fisher's exact test comparison of the frequency of known or likely oncogenic mutations in each subset. **(b)** Bubble plot comparing sgRNA log fold changes with mean frequency of C > T editing in the sensor target site between days 5 and 30 post-transduction. Bubbles were colored by their OncoKB classification. Size denotes MaGeCK score (see Supplementary Table 6d).



Extended Data Fig. 10 | Expanded base editing predictions. (a) We used the MSK-IMPACT clinical tumor sequencing dataset and the characteristics of commonly used base editors to inform the design of base editing sensor libraries used in the experiments in Figs. 3–6. These results are available in the Shiny web portal (https://dowlab.shinyapps.io/BEscan/). Using updated and expanded versions of MSK-IMPACT sequencing data, base editing configurations, and *AMINEsearch v2*, we generated an exploratory set of sgRNA and sensor predictions, which are also available in the Shiny web portal. The more recent version of MSK-IMPACT contains increased numbers of **(b)** tumors sequenced, **(c)** total SNVs observed, and **(d)** candidate unique recurrent SNVs. These factors in the input led to to an increase in the exploratory set (v2) compared to the HBES and MBES libraries (v1) in respect to **(e)** Cas variants (determining PAM recognition) and base editor variants (determining editing window), collectively making base editor configurations with distinct properties **(f)**. These factors in the input led to to an increase in the exploratory set (v2) compared to the HBES and MBES libraries (v1) in respect to (**g)** number of sgRNAs designed and **(h)** unique SNVs targeted by one or more sgRNAs in the database.

nature portfolio

Corresponding author(s): Lukas Dow

Last updated by author(s): November 22, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Attune NxT flow cytometer software version 3.1, Illumina MiSeq, Illumina NextSeq, BSgenome
Data analysis	Graphpad Prism 8.0, Geneious Prime 11.0.6 + 10, Adobe Illustrator 24.3, R 3.6.3, R studio 1.2.5033, CRISPREsso2, MAGeCK 0.5.9, Pandaseq 2.9, cutadapt 3.5, R packages: ggplot2 3.3.5, ggpubr 0.4.0. Attune NxT flow cytometer sofware. Custom java script to analyze merged FASTQ files is available at github repository: https://github.com/schmidt73/base-editing-analysis

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

Data

Policy information about availability of data

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

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Plasmids and sensor libraries will be made available at Addgene. Raw sequencing data is available at the sequence read archive (SRA) under submission: PRJNA746395. Processed data tables for BE-SCAN are available at GitHub (https://www.github.com/lukedow/data). All source data (including p-values) are available in Supplementary Table 10.

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	For experiments involving animals, 5 mice were used per sgRNA, per enzyme, unless otherwise indicated. This number allows detection of 50% difference between samples, with 80% power (assuming 20% variance). For sensor screens 5 cell lines in duplicate replicates was performed. We reasoned we would use both mouse and human systems in different tissues to represent a variety of cellular contexts. For validation experiments we chose 1 of 5 screening cell lines to individually validate sgRNAs and directly compare to sensor screen results as we expect the efficiency of sgRNAs to remain relatively similar across cell lines as shown in the results.
Data exclusions	Sequencing reads with mutations in sgRNA or sgRNA scaffold were excluded from analysis as this is due to synthesis errors during cloning of the library these reads don't represent actual experimental results.
Replication	Sensor screens were performed in duplicate in 5 different cell lines. Duplicate screens represented two separate transduction of the library. The functional sgRNA screen in PDEC cells was performed in 6 replicates transductions.
Randomization	No samples were randomized because it is not relevant to our study as we do not include new/exisiting drugs and or treatments on animals/ humans.
Blinding	Researchers were not blinded to experiments or analyses because it is not relevant to our study as no experimental treatments were provided.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used	Cas9 (Cell Signaling, #19526S) and Actin (Abcam, #ab49900)
Validation	Cas9 antibody was validated by western blot on cells overexpressing Cas9

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	MDA-MB-231, NIH3T3, and HEK293T cells were purchased from ATCC. PC9 cells were a gift from H. Varmus laboratory originally obtained from sigma aldrich(90071810-1VL) . KPT1 cells were a gift from T. Jacks laboratory derived from mouse tumors. PDEC cells were a gift from D. Bar-Sagi laboratory.
Authentication	MDA-MB-231 and NIH3T3 cells were purchased from ATCC as used at low passage. Other cell lines were not authenticated
Mycoplasma contamination	Cell kept in culture were routinely tested for mycoplasma and discarded if tested positive

Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Mice were maintained under specific pathogen-free conditions, and food and water were provided ad libitum. C57Bl/6N and Foxn1nu (Swiss nude) mice were purchased from Envigo. All mice used were 6 to 8 week-old females.
Wild animals	No wild animals were used.
Field-collected samples	No field collected samples were used.
Ethics oversight	Studies involving animals were approved by the MSKCC Institutional Animal Care and Use Committee under protocol 11-06-018

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cells are lifted with trypsin and re-suspended into 96 well plate in 200 ul of media.
Instrument	Thermo Fisher 2018 Attune NxT Flow cytometer and Guava Easycyte (Millipore)
Software	FlowJo v10.8.1
Cell population abundance	First cells expressing NLSmScarletI fluorescence were measured to determine infected cell population and GFP florescence was measured as cell population. GFP cell population varied with each cell line/base editor expression depending on efficiency.
Gating strategy	Live cells are detected by FSC-A and SSC-A and single cells are detected from live cell population by FSC-A and FSC-H. Single cells are then gated by YL2-A and SSC-A to determine NLSMScarlet population where any cell beyond 10^3 on the axis is considered positive. Population of GFP positive cells from infected cells was gated by BL1-A and FSC-A where any cell beyond 10^3 on the axis was considered GFP positive.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.