nature biotechnology

Article

https://doi.org/10.1038/s41587-023-01949-8

Direct measurement of engineered cancer mutations and their transcriptional phenotypes in single cells

Received: 31 October 2022

Accepted: 15 August 2023

Published online: 11 September 2023

Check for updates

Heon Seok Kim[1,2,3], Susan M. Grimes[1], Tianqi Chen[1], Anuja Sathe[1], Billy T. Lau[1], Gue-Ho Hwang[4], Sangsu Bae[4,5] & Hanlee P. Ji[1,6]

Genome sequencing studies have identified numerous cancer mutations across a wide spectrum of tumor types, but determining the phenotypic consequence of these mutations remains a challenge. Here, we developed a high-throughput, multiplexed single-cell technology called TISCC-seq to engineer predesignated mutations in cells using CRISPR base editors, directly delineate their genotype among individual cells and determine each mutation's transcriptional phenotype. Long-read sequencing of the target gene's transcript identifies the engineered mutations, and the transcriptome profile from the same set of cells is simultaneously analyzed by short-read sequencing. Through integration, we determine the mutations' genotype and expression phenotype at single-cell resolution. Using cell lines, we engineer and evaluate the impact of >100 *TP53* mutations on gene expression. Based on the single-cell gene expression, we classify the mutations as having a functionally significant phenotype.

Ongoing genomic studies of cancer are cataloguing extensive numbers of somatic variants. For example, genome sequencing studies have identified numerous cancer mutations across a wide spectrum of tumor types. Many of these mutations result in amino acid substitutions. Given the sheer number of discovered mutations, determining the phenotype of cancer substitutions with functional characterization remains an enormous challenge. In silico functional predictions of cancer mutations are frequently used as a solution. However, these computational methods do not provide more discrete biological characterization. There remains a notable need for high-throughput approaches to functionally evaluate many mutations in an efficient manner. CRISPR base editors and single guide RNAs (sgRNAs) have been used for genetic screens, where they directly introduce specific variants into target genes at their native genomic loci among transduced cells[1–4]. Studies using this method examined the altered cellular fitness resulting from the introduced genetic variants, either by counting sgRNA or barcode sequences among the cell pool, but these approaches do not directly verify the presence of an engineered mutation, as the association with a genotype is imputed based on the sgRNA or the barcode sequence.

Base editors can introduce multiple variants into a target genomic sequence. Although a given sgRNA sequence is intended to generate a single variant, the actual base editing process introduces multiple different, unintended variants at the target genomic sequence. For example, when using the cytosine base editor (CBE), the conversion of either a C to T or a C to G produces different variants other than what was intended. CBEs exhibit cytosine editing in both the target and neighboring bystander cytosines in the editing window with the outcome being multiple different variants at the target sequence site. This variability points to the need to directly genotype the base editor target site as the best approach for verifying the intended mutation

[1]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Life Science, College of Natural Sciences, Hanyang University, Seoul, Republic of Korea. [3]Hanyang Institute of Bioscience and Biotechnology, Hanyang University, Seoul, Republic of Korea. [4]Medical Research Center of Genomic Medicine Institute, Seoul National University College of Medicine, Seoul, Republic of Korea. [5]Department of Biochemistry and Molecular Biology, Seoul National University College of Medicine, Seoul, Republic of Korea. [6]Department of Electrical Engineering, Stanford University, Stanford, CA, USA. e-mail: genomics_ji@stanford.edu

https://doi.org/10.1038/s41587-023-01949-8

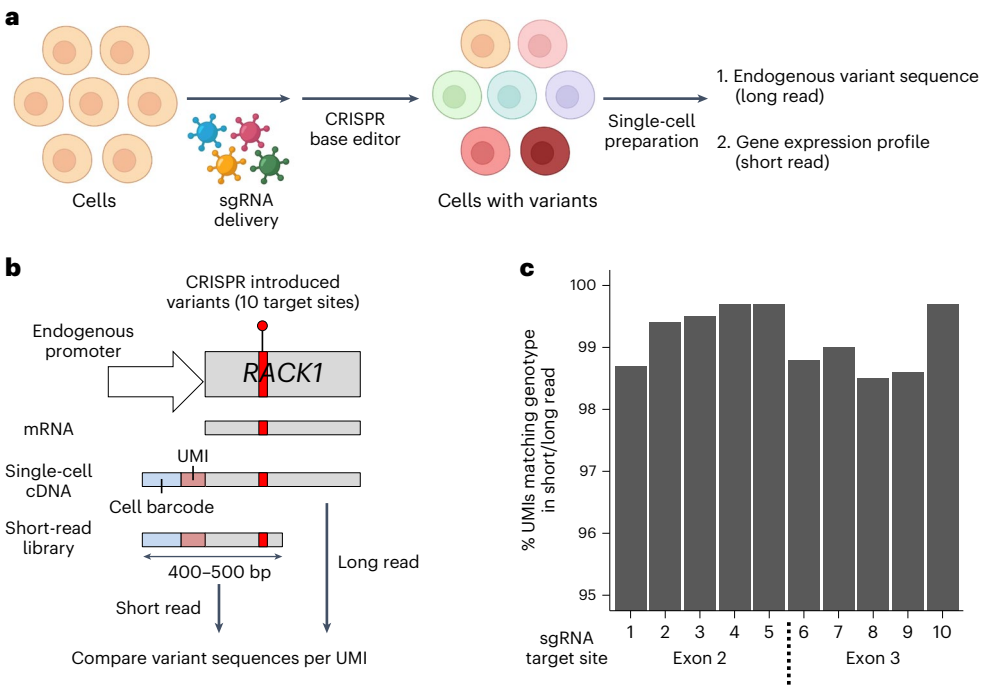


Fig. 1 | Schematic of TISCC-seq. a, Overview of direct detection and phenotyping of various *TP53* coding mutations. **b**, Schematic of the variant calling accuracy comparison between short- and long-read single-cell sequencing. **c**, Accuracy of the mutation calling of long-read sequencing. We compared mutation sequences of each sgRNA target site and calculated proportion of UMIs that have the same sequence in short- and long-read sequencing.

being present. Direct validation of an engineered mutation is a necessary step if one is to accurately determine the phenotype, and this requires examining individual cells.

Several studies have used a reporter system to infer the presence of engineered mutations[3,4], but this is an indirect approach and assumes the same genome edit has occurred in both the reporter and endogenous site. Also, these methods may not reflect the precise effects of mutations on gene expression. For example, the single-cell Perturb-seq method was adapted to exogenously express genes in the form of cDNAs containing a specific variant and then indirectly measure the mutated gene using a barcode sequence[5]. Although one can interrogate the resultant single-cell transcriptome changes induced by each variant, this approach has limitations. Specifically, the gene variant is expressed with an exogenous promoter that is not under canonical genetic regulation at the gene's native locus. Second, variants are delivered to cells with wild-type gene expression of the target gene, which can mask the effect of the variant on protein function. Third, only the barcode sequence is detected instead of the variant itself. Template switching in lentivirus packaging can induce swapping of the variant-barcode association[6], leading to artifacts in identification and transcriptional phenotyping.

We developed a method that addresses these challenges and resolves these issues. This method is referred to as transcript-informed single-cell CRISPR sequencing (TISCC-seq). This approach relies on CRISPR base editors to introduce multiple endogenous genetic variants into a given genomic target. Long-read sequencing identifies these mutations directly from a target's transcript sequence at single-cell resolution. Then, we integrate the short-read transcriptome profile from the same single cells (Fig. 1a). This integrative approach enables single-cell direct genotyping and phenotyping of various genetic variants introduced into the native gene locus. Single-cell characterization allows one to distinguish the base editor's intended versus unintended mutations among individual cells. We applied this approach to engineer a series of previously reported cancer mutations in *TP53*, the majority of which have never been functionally characterized.

Results

Identifying mutations with single-cell cDNA sequencing

We conducted an analysis comparing long- versus short-read single-cell cDNA sequencing. For this initial test, we designed an assay to introduce different genetic variants in exon2 and 3 of the *RACK1* gene (Fig. 1b). The length of *RACK1* cDNA up to exon3 is approximately 500 bp; this length interval can be fully covered with short reads. This gene is the most highly expressed in the HEK293T cell line as determined from single-cell short- and long-read gene expression data from our previous publication[7]. We designed 10 sgRNAs targeting exon2 and 3 of *RACK1* gene and transduced lentiviruses encoding those sgRNAs to HEK293T cells at 0.1 multiplicity of infection (Supplementary Table 2). Transduced cells were selected by puromycin. Then, we transfected a plasmid encoding an adenine base editor (ABE) into the cells. This step introduced multiple genetic variants at sgRNA target sites. After 6 days, we generated single-cell cDNAs and extracted genomic DNA from cells derived from the same suspension.

From the genomic DNA of transduced cells, we amplified exon2 or 3 of the *RACK1* gene and performed short-read sequencing to evaluate the frequency of genetic variants in *RACK1* genomic DNA. Based on the DNA sequencing, we identified genetic variants introduced by all ten sgRNAs. The frequency of ABE-induced genetic variants varied from 1.1% to 10.1% from the genomic DNA of pooled cells (Supplementary Fig. 1).

Next, we evaluated the presence of these variants at single-cell transcript level using single-cell cDNAs. These engineered variants were proximal to the 5′ end of the cDNA, allowing us to sequence them with short reads (that is, Illumina). Short-read sequences have a high base quality for variant calling and allowed us to compare the long- and short-read results. From the single-cell cDNA library, we prepared sequencing libraries for both short- and long-read sequencing to assess single-cell level genetic variants from the *RACK1* transcripts. For short-read sequencing, we amplified exon2 or 3 of *RACK1* from single-cell cDNA with cell barcodes and unique molecular identifier (UMI) sequences using the 5′ adaptor primer and exon-specific primers

(Fig. 1b). These libraries were sequenced on the Illumina MiSeq platform. In Illumina sequencing, each DNA fragment is sequenced from both ends, resulting in two reads per fragment. These two reads are referred to as read1 and read2. Similar to regular single-cell gene expression sequencing, we used 26 bp of read1 sequences for cell barcode and UMI extraction. The read2 sequences were used for the evaluation of the newly introduced *RACK1* genetic variants at target sites. Using the genetic coordinates of the sgRNA target window (that is, 3 bp to 8 bp), for a given read, we identified the corresponding cell barcode, UMI and the genetic variant.

For long-read sequencing, we amplified the entire *RACK1* cDNA using the 5′ adaptor and primers specific to the last 3′ exon from the same single-cell cDNA library (Fig. 1b). The intact cDNA amplicon was sequenced with an Oxford Nanopore instrument. Guppy was used for base calling, and minimap2 was used for alignment[8,9]. Each sequence read had the cell barcode, UMI and complete *RACK1* cDNA sequence. We extracted the cell barcodes and UMI as we previously described[7]. After genome alignment of the long-read data, the cell barcodes and UMI fell into soft-clipped sequence. Therefore, we extracted the soft-clipped portion of each read and compared that with the cell barcodes identified from gene expression library sequencing. Only reads with perfectly matching cell barcodes were used for further analysis. Using the aligned long-read data, we identified the *RACK1* genetic variants. Therefore, long-read information provided the genetic variants with accompanying cell barcode and UMI sequence. For additional quality control filtering, UMIs with less than three reads were filtered out. We generated consensus genetic variants for each UMI using multiple reads.

We compared the *RACK1* variant calls from short- and long-read single-cell data. We analyzed consensus *RACK1* genetic variants for each cell barcode and UMI combination. Across all target sites, we compared 479,509 UMIs: 99.2% of them had identical genetic variants in average (Fig. 1c). This result demonstrated the high accuracy of long-read identification of CRISPR-engineered genetic variants. Recent improvements in the accuracy of nanopore sequencing and UMI-based consensus generation enabled this analysis. We then compared the frequency of genetic variants from genomic DNA and aggregated single-cell cDNA for each of the 10 target sites introduced by base editors. The frequency of each variant between genomic DNA and single-cell cDNA had a high correlation (R^2 = 0.63; Supplementary Fig. 1).

Base editor guide RNA designs for *TP53* cancer mutations

We introduced a set of sgRNAs designed for multiple *TP53* mutations and used TISCC-seq to obtain the gene expression profile and *TP53* genotype from individual cells. First, we focused on the design of the genome engineering of *TP53* mutations (Fig. 2a). We identified *TP53* mutations which were reported more than nine times in the COSMIC database[10]. The majority of these frequent cancer mutations were within the *TP53* DNA-binding domain. The total number of coding mutations was 351. We designed base editor libraries targeting this mutation set. To cover as many mutations as possible, we used several base editor combinations: (1) CBE with NGG protospacer adjacent motif (PAM), (2) CBE with a NG PAM, (3) ABE with NGG PAM and (4) ABE with a NG PAM. Using the NGG PAM base editors, we designed 74 sgRNAs targeting 99 *TP53* variants. The NG PAM base editors have a more flexible PAM, so we were able to design an additional 88 sgRNAs targeting 159 variants (Supplementary Fig. 2). Most sgRNAs targeted the DNA-binding domain of p53 protein (Fig. 2b).

Base editors can alter any target nucleotide in their target window (that is, 3 bp to 8 bp) which leads to different nucleotides at that position. TISCC-seq identified this variation among single cells. For example, the sgRNA introducing E258K mutation by C to T substitution induces the E258G mutation by C to G substitution (Supplementary Fig. 3). Similarly, the sgRNA introducing S127P mutation by A to G substitution at the third adenine induces the Y126H mutation by A to G substitution at the sixth adenine (Supplementary Fig. 3). Therefore, this result suggests that any given sgRNA can introduce multiple variants depending on the window sequence context. The entire number of amino acid changes that could be introduced by the NGG or NG PAM base editors and our sgRNA libraries was 920 and 1,999, respectively. For the final design, we targeted 251 known *TP53* mutations with the potential for introducing 2,892 possible amino acid changes (Supplementary Fig. 2).

CRISPR base editor engineering of *TP53* mutations

We used HCT116 and U2OS human cell lines for this study. Both cell lines have wild-type *TP53*, which we independently confirmed[11–13]. The p53 pathway is repressed by the negative regulator MDM2 in both cell lines[14]. The oncoprotein MDM2 is an E2 ubiquitin ligase[15]. MDM2 binds to and promotes the ubiquitin-dependent degradation of the p53 protein. The small molecule nutlin-3a can inhibit p53-MDM2 binding efficiently[16]. To activate the p53 pathway and select for *TP53* mutations with functional effects, we tested various concentrations of nutlin-3a, including 5 μM, 10 μM and 20 μM, based on previous reports[14]. Our results showed successful p53 pathway activation at 10 μM nutlin-3a, which we used for both cell lines.

We generated four sgRNA libraries for each base editor (NGG-CBE, NGG-ABE, NG-CBE and NG-ABE), and the combined libraries were designed to cover the preselected *TP53* mutations. We transduced those libraries using a lentivirus system to both the HCT116 and U2OS cell lines. The cells were transfected with each respective base editor plasmid. It had been reported that base editors can induce off-target RNA editing[17]. To minimize those effects, we chose transient transfection rather than stable expression of base editors. Typically, plasmid-based protein expression peaks after 24 h of transfection and diminishes after 5 or 6 days[18]. Six days after transfection, we used nutlin-3a to activate the p53 pathway.

TISCC-seq detection of *TP53* mutations

After 10 days of nutlin-3a treatment, we harvested the cells for suspension, prepared single-cell cDNA libraries and also extracted genomic DNA from a portion of the cell suspension (Methods). We amplified *TP53* transcripts from the single-cell cDNA library, sequenced their full-length transcript and determined the presence of the *TP53* mutation from the long-read data (Fig. 2a). As an important additional step, we extracted cell barcodes and UMI per each long-read as described earlier. To prevent the effect of sequencing error in UMI region, we filtered out any UMI with less than 10 long reads. As a quality control threshold, we used only the cell barcode and UMI combinations found in 10 or more reads. For generating a consensus, we also included UMIs with a low edit distance, assuming the differences were related to sequencing errors. For *TP53* variant calling, we extracted every nucleotide sequence in the sgRNA target window (for example, chr 17:7674940–7674945 for the sgRNA in Fig. 2d) and compared them with reference sequence (for example, CACTCG to CATTCG). Based on nucleotide changes of a given mutation, we determined the amino acid substitution at the target site (for example, V196M).

For independent validation, we used amplicon sequencing from the transduced cells' genomic DNA to independently assess the frequency of a subset of *TP53* mutations. This analysis compared the frequency of each *TP53* mutation introduced by 12 sgRNAs in genomic DNA versus the results from analyzing the single-cell cDNA from HCT116 cells. These *TP53* mutations were introduced efficiently with up to 12.1% for one variant and 27 variants were introduced with a frequency greater than 0.25%. The prevalence of each mutation from single-cell cDNA and genomic DNA was generally correlated (Fig. 2c and Supplementary Fig. 4; R^2 = 0.59). Some variants had higher frequency in genomic DNA and lower in cDNA (that is, W146Ter). This result means that for some mutations the corresponding transcripts were not expressed efficiently or were subjected to higher RNA degradation. The lower prevalence of cDNA mutations may reflect effects

https://doi.org/10.1038/s41587-023-01949-8

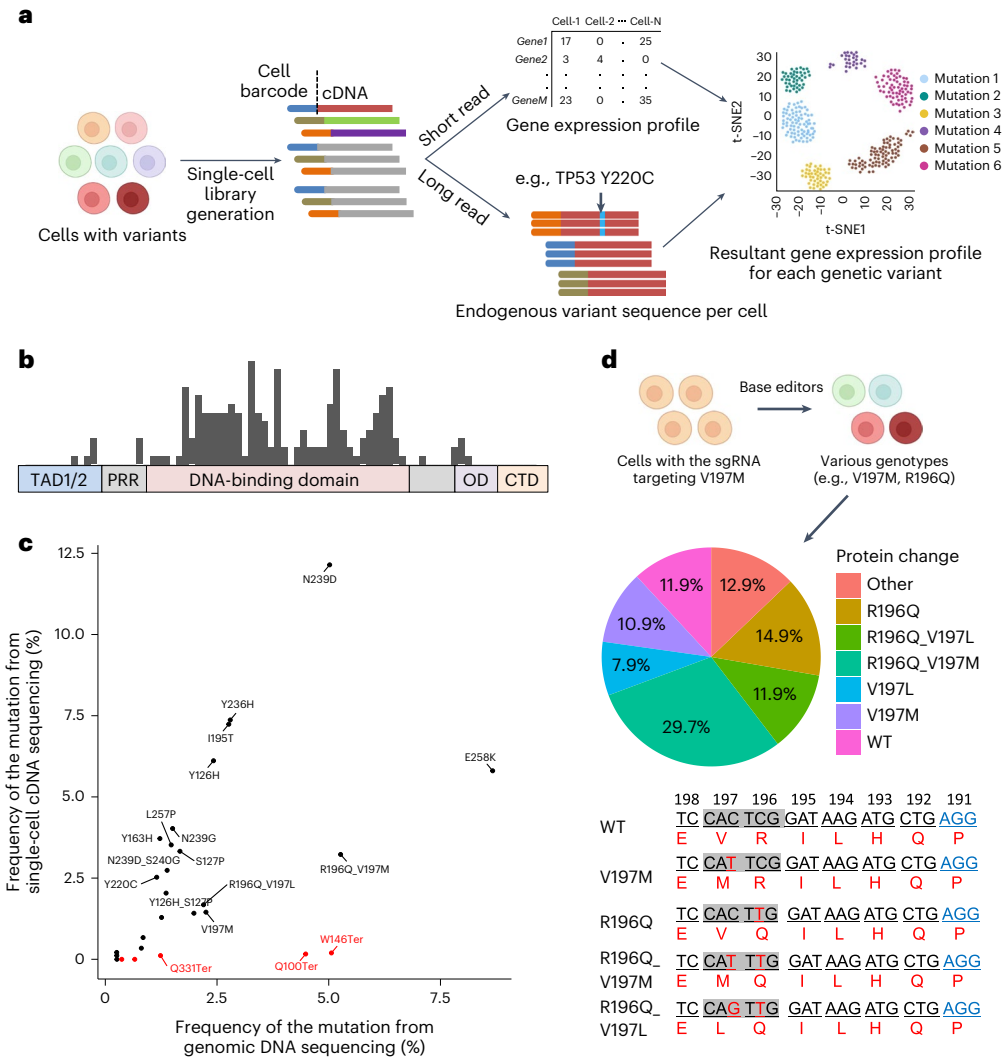


Fig. 2 | TISCC-seq identifies mutations directly. a, Overview of single-cell cDNA analysis pipeline. **b**, Structure of p53 protein and distribution of sgRNA target sites we used in this study. TAD, transactivation domain; PRR, proline-rich region; OD, oligomerization domain; CTD, C-terminal domain. **c**, Dot plot showing the proportion of each genetic variant detected from single-cell cDNA and genomic DNA. Red dots represent variants with a premature stop codon. **d**, Cells with same sgRNA can result in various genotypes. The pie chart shows the proportion of resultant amino acid changes from cells with sgRNA targeting V197M mutation. Proportions of mutations are calculated from the single-cell cDNA long-read sequencing. Underlines indicate each triplet codon and numbers indicate position of the codon. Red DNA sequences indicate substituted bases, and blue indicate PAM sequences. WT, wild type.

from nonsense-mediated decay (NMD). This process is a surveillance mechanism that eliminates mRNA transcripts containing premature stop codons. For example, although 5.1% of cells had a W146Ter mutation at the genomic DNA level, this mutation was not detected as frequently at the single cDNA level (0.2%) because the transcripts with the variant were degraded in cells by NMD (Fig. 2c).

As another type of validation, we also sequenced the sgRNA expressed in each cell from single-cell cDNA using a direct capture method previously described[7,19]. Most of the single-cell CRISPR screen studies have relied on an sgRNA sequencing (sgRNA-seq) method to infer the resultant genetic edits[20–23]. This method assumes that cells with the sgRNA have the targeted genomic edit. However, the efficiency of base editors is lower than that of Cas9 nuclease[24,25]. As described earlier, a base editor may introduce multiple genetic variants from the same sgRNA (Supplementary Fig. 3). Therefore, one cannot assume that cells transduced with base editors and a single sgRNA have the intended variant at the target position (Fig. 2d and Extended Data Fig. 1). Our results showed that this was the case. For example, we evaluated an sgRNA which was designed to introduce the *TP53* V197M mutation. The sgRNA's target site has three cytosines in its window. Among 101 cells expressing this specific sgRNA, 11 cells had V197M mutation, whereas 30 cells had both R196Q and V197M mutations (Fig. 2d). Therefore, the conventional single-cell CRISPR screening method using sgRNA-seq did not correctly identify the introduced variants among the various single cells. In contrast, with direct long-read sequencing of the full-length target transcripts from single cells, we bypassed this issue and directly identified the actual mutation introduced by the base editor from the cDNA.

TISCC-seq and analysis of HCT116 cells with *TP53* mutations

We performed gene expression analysis using the same single-cell cDNA library we used for long-read sequencing. As we described previously, we integrated the single-cell *TP53* mutation genotypes from long reads with the single-cell gene expression profile data from short reads[7]. We used cell barcode matching between the long-read data with a mutation genotype and the short-read data (Methods). This process allowed us to link those cells with *TP53* mutation to their individual gene expression profiles. To conduct a cluster analysis of the cells with different *TP53*

mutations, we used Uniform Manifold Approximation and Projection (UMAP) (Fig. 3). We investigated the effect of p53 pathway activation by nutlin-3a in HCT116 cells with *TP53* mutations using a subset of our sgRNA library (10 sgRNAs). When we compared the gene expression profiles between cells with wild-type or *TP53* mutations, there was a clearly delineated difference upon p53 pathway activation (Fig. 3a,b). When we visualized the expression of p53 pathway involved genes on a UMAP plot using a heatmap, we found that cells with deleterious *TP53* mutations displayed decreased p53 pathway involved gene expression compared to wild-type cells (Extended Data Fig. 2).

Next, we sequenced HCT116 cells transduced with our full *TP53* sgRNA library and activated by nutlin-3a. Among the 42,564 cells that were sequenced, we filtered out a set of high-quality long-read UMIs (UMI read count >9) covering *TP53* from 12,887 cells. This subset of high-quality reads were useful for confirming the mutation genotype. For each cell, we had an average of 898 *TP53* reads with a complexity of 4.5 UMIs for this subset. We filtered out cells which had a heterozygous mutation. Overall, we detected a total of 169 different mutations distributed among the various single cells.

We analyzed the single-cell gene expression for each mutation. To provide a robust measurement of single-cell expression, we filtered out those *TP53* mutations expressed in fewer than five cells. This step retained 74 mutations for further analysis. Via UMAP clustering, the cells with wild-type versus *TP53* mutations were separated among different clusters. Compared to the clustering observed in Fig. 3b, which included 11 mutations, this dataset encompasses 74 mutations with a wider range of impact. As a result, the separation between wild-type cells and other cells is less distinct in this dataset. Wild-type cells were predominantly clustered in clusters 5 and 9 (Fig. 3c and Supplementary Fig. 5). For each variant, we calculated its proportion within each cluster and performed hierarchical clustering of each variant based on the proportion (Fig. 3d). Cells with the following five mutations (R156C, V157I, V173A, R273C and A276V) clustered with the wild-type cells. This result was a preliminary indication that this set of mutations did not have a significant impact on the gene expression phenotype; we annotated them as wild-type-like and the others as functionally significant.

We examined the expression of 343 genes known to be involved in the p53 pathway from previous report using single-cell data analysis (Supplementary Table 6)[26]. Cells that were wild type or with mutations that were wild-type-like had higher expression of p53 pathway involved genes (Fig. 3e). Wild-type cells had higher p53 pathway gene expression scores compared to the majority of cells expressing functionally significant *TP53* mutations (Fig. 3f; $P < 0.03$; Supplementary Fig. 6). Additionally, we analyzed the expression of the *CDKN1A* gene, which encodes a p21 protein. p21 protein is a regulator of cell cycle progression and arrest. Wild-type cells had higher *CDKN1A* expression compared to the cells with functionally significant *TP53* mutations (Extended Data Fig. 3). Next, we performed pathway analysis between wild-type cells and cells with wild-type-like versus functionally significant variants. Cells with functionally significant mutations had lower p53 pathway activity and higher G2M checkpoint gene expression than the wild-type cells (Fig. 3g; $P = 1.66 \times 10^{-11}$ and 1.66×10^{-11}). In addition, cells with wild-type-like variants expressing R156C, V157I, V173A, R273C or A276V did not have differences in these two pathways compared to cells with wild-type *TP53* (Fig. 3g; $P = 0.95$ and 0.44). These results are evidence that this subset of the mutations had features similar to wild type and thus had less functional impact. In summary, wild-type cells had higher p53 pathway activity and related gene expression than cells with functionally significant *TP53* variants. These results validated the TISCC-seq method for high-throughput functional classification of these mutations.

TISCC-seq analysis of *TP53* mutations in U2OS cell line

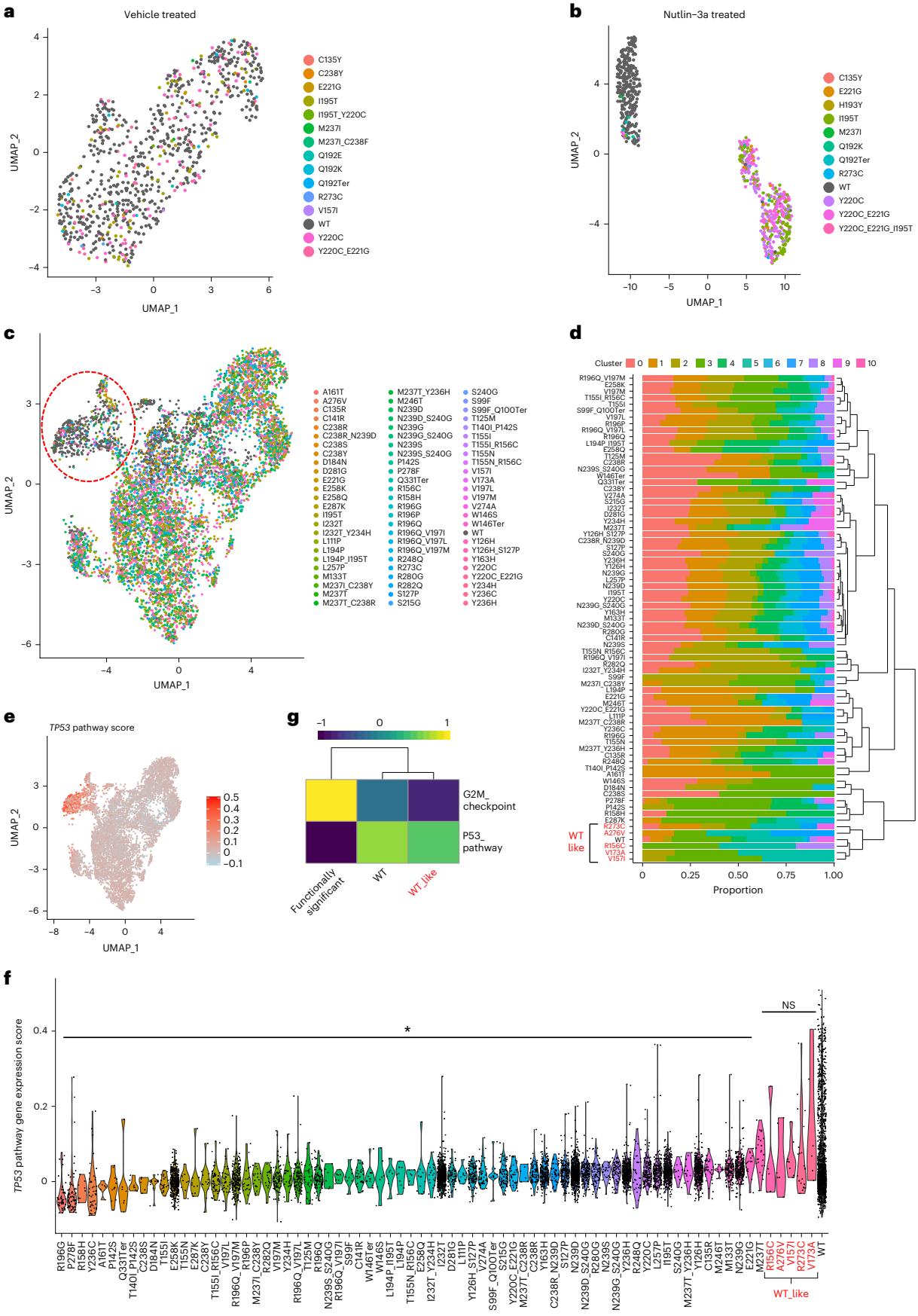
As an additional verification of our results, we performed a similar analysis with the U2OS cell line using the same sgRNAs for the *TP53* mutations. Among 38,451 cells that we sequenced, we were able to acquire high-quality long-read sequences from 12,155 cells. On average per each cell, we filtered out the high-quality *TP53* reads of which there were 890 with a complexity of 4.6 UMIs. As described, we applied a filtering strategy to eliminate heterozygous mutations. For the U2OS line, we characterized 161 mutations with TISCC-seq. For gene expression analysis, we used the 62 variants which were detected in more than five cells. From the UMAP analysis, wild-type cells and cells with *TP53* mutations separate into distinct clusters (Supplementary Fig. 7 and Extended Data Fig. 4). Wild-type cells were primarily associated with cluster 1. For each mutation, we calculated its proportion within each cluster and performed hierarchical clustering based on this cluster proportion (Extended Data Fig. 4b). From the hierarchical clustering results, we identified four mutations, T140I, R156C, T221I and R273C, that were associated with wild-type *TP53*. The R156C and R273C mutations had a similar association with the wild-type cells for both the HCT116 and U2OS cell lines. The wild-type U2OS cells had higher expression of *CDKN1A* and other p53 pathway involved genes compared to the majority of cells expressing functionally significant *TP53* mutations (Extended Data Figs. 4 and 5). The analysis of pathway activity showed that cells with functionally significant mutations had significantly lower p53 pathway activity and higher G2M checkpoint gene expression (Extended Data Fig. 4e; $P = 1.62 \times 10^{-12}$ and 1.62×10^{-12}). Conversely, cells with wild-type-like mutations were not statistically significant to the same extreme degree as the functionally significant mutations (Extended Data Fig. 4e; $P = 0.52$ and 0.001).

Confirmation of TISCC-seq using clonal cell lines

Our prior experiments were highly multiplexed in engineering different mutations. Providing additional confirmation of the single-cell results, we conducted simplex experiments of individual mutations using the HCT116 cell line. Using the ABE, we generated homozygous clonal cell lines with either the *TP53* I195T or Y220C mutation which were functionally significant and had enough cells from single-cell assay. To obtain clones, we used limiting dilution after ABE transfection. These two mutations have been reported to have a deleterious effect on function[10,27] and the multiplexed TISCC-seq results also demonstrated that they had a functional effect (Fig. 3d). We performed bulk RNA-seq from nutlin-3a treated wild-type cells and those clonal cells. We compared the result with single-cell results from HCT116 cell lines (Fig. 4 and Extended Data Fig. 6).

Fig. 3 | TISCC-seq on HCT116 cells. a–**c**, UMAP plots showing single-cell gene expression profile per each genetic variant. HCT116 cells are treated with vehicle (**a**) or nutlin-3a (**b**) after the introduction of variants using a subset of the sgRNA library. **c**, HCT116 cells are treated with nutlin-3a after the introduction of variants using the full sgRNA library. **d**, Proportion of UMAP cluster from cells with each genetic variant. Hierarchical clustering was performed based on the proportion to categorize genetic variants. Red indicates wild-type-like variants. **e**, UMAP embedding of cells colored by p53 pathway gene scores. **f**, Violin plots showing p53 pathway gene score per cells with each genetic variant. *$P < 0.03$; NS, not significant; two-sided *t*-test. $P = 1.7 \times 10^{-33}$, 3.7×10^{-29}, 1.3×10^{-6}, 2.1×10^{-14}, 1.5×10^{-6}, 3.8×10^{-7}, 2.1×10^{-2}, 9.5×10^{-9}, 7.8×10^{-5}, 2.6×10^{-7}, 7.2×10^{-27}, 6.9×10^{-4}, 3.9×10^{-7}, 1.5×10^{-88}, 2.8×10^{-6}, 2.0×10^{-30}, 8.7×10^{-23}, 5.0×10^{-67}, 1.4×10^{-9}, 4.4×10^{-14}, 5.7×10^{-14}, 3.3×10^{-37}, 3.0×10^{-13}, 5.8×10^{-38}, 1.5×10^{-10}, 1.5×10^{-43}, 7.5×10^{-4}, 8.6×10^{-9}, 5.5×10^{-5}, 4.3×10^{-23}, 3.1×10^{-7}, 9.2×10^{-3}, 1.2×10^{-3}, 1.4×10^{-5}, 1.3×10^{-5}, 6.3×10^{-4}, 2.3×10^{-12}, 8.6×10^{-65}, 7.2×10^{-41}, 1.1×10^{-10}, 1.8×10^{-49}, 2.1×10^{-25}, 3.8×10^{-4}, 7.2×10^{-35}, 4.2×10^{-20}, 2.0×10^{-4}, 5.0×10^{-35}, 2.0×10^{-50}, 8.0×10^{-23}, 8.9×10^{-43}, 1.4×10^{-52}, 8.2×10^{-42}, 4.2×10^{-29}, 3.8×10^{-21}, 1.8×10^{-31}, 1.7×10^{-47}, 7.3×10^{-8}, 2.2×10^{-34}, 8.7×10^{-31}, 2.2×10^{-45}, 6.1×10^{-8}, 8.2×10^{-6}, 7.6×10^{-40}, 7.0×10^{-14}, 5.7×10^{-10}, 2.1×10^{-25}, 8.6×10^{-32}, 5.3×10^{-5}, 5.3×10^{-1}, 5.7×10^{-1}, 4.6×10^{-1}, 2.6×10^{-1}, 3.7×10^{-1}, 3.8×10^{-1}. **g**, Heatmap showing average Gene Set Variation Analysis (GSVA) enrichment score of selected Hallmark pathways per each category of genetic variant.

https://doi.org/10.1038/s41587-023-01949-8



https://doi.org/10.1038/s41587-023-01949-8

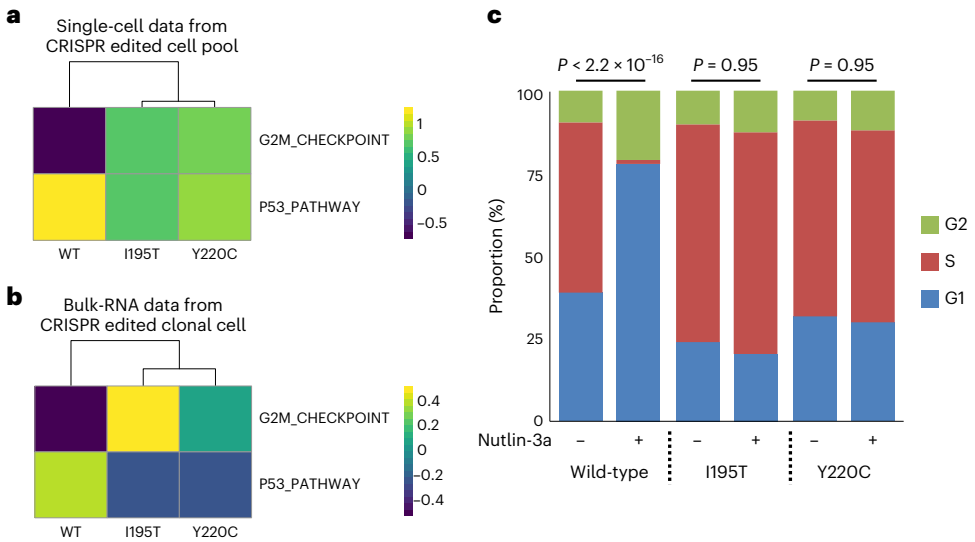


Fig. 4 | Confirmation of TISCC-seq. a,**b**, Heatmaps showing the average GSVA enrichment score of selected Hallmark pathways. **a**, Scores are calculated from single-cell analysis of heterogenous *TP53* genetic variants pool. **b**, Scores are calculated from bulk RNA-seq from clonal cells with indicated *TP53* genetic variants. **c**, Cell cycle analysis using DNA content staining using clonal cells. Genetic variant per cells and nutlin-3a treatments are indicated. $n = 2$ biologically independent cells. $P < 2.2 \times 10^{-16}$, $P = 0.95, 0.95$. *P* values are calculated by chi-squared test, two sided.

From the single-cell results, both mutations demonstrated lower p53 pathway activity and higher G2M checkpoint gene expression than wild-type cells (Fig. 4a; I195T: $P = 2.2 \times 10^{-11}$ and 1.7×10^{-3}, Y220C: 2.2×10^{-11} and 9.4×10^{-2}). From the conventional, bulk-based RNA-seq results, we observed the same effect on the same pathways (Fig. 4b; I195T: $P = 3.4 \times 10^{-6}$ and 2.4×10^{-7}, Y220C: 1.0×10^{-4} and 2.6×10^{-7}). Next, we performed differential gene expression (DGE) analysis between wild-type and mutation-bearing cells. We compared the DGE results from single-cell RNA-seq and standard RNA-seq. For the I195T or the Y220C mutation, we identified the top 100 genes determined from single-cell RNA-seq data. For the I195T mutation, 94 out of 100 were confirmed as showing differential expression per the conventional RNA-seq. Likewise for the Y220C mutation, 80 out of 100 genes were confirmed as showing differential expression per the conventional RNA-seq (Supplementary Tables 7 and 8; $P < 1.0 \times 10^{-5}$).

Overall, the I195T and Y220C cell lines had higher G2M checkpoint gene expression as an indicator of more active cell division compared to the cells with wild-type *TP53*. To validate this result, we evaluated cell division and cell cycling from wild-type and *TP53*-mutated HCT116 cells using 5-ethynyl-2′-deoxyuridine (EdU) and a propidium iodide (PI) flow cytometry assay. The PI assay detects total DNA amounts for G1- and G2-phase comparison. The EdU assay labels newly synthesized DNA to detect S-phase. The cell cycle of wild-type HCT116 cells was arrested by nutlin-3a treatment (Fig. 4c and Extended Data Fig. 7; $P < 2.2 \times 10^{-16}$). In contrast, the cell cycle of HCT116 cells with either the I195T or the Y220C mutation did not undergo arrest with nutlin-3a treatment (Fig. 4c and Extended Data Fig. 7; $P = 0.95$ and 0.95).

We expanded our analysis by generating five additional clones with *TP53* mutations and conducted RNA-seq analysis (Extended Data Fig. 6e,f). The V157I mutation was categorized as wild-type-like, whereas the remaining mutations were deemed functionally significant based on the TISCC-seq analysis. Our results revealed that HCT116 cells with the V157I mutation exhibited a gene expression profile that was similar to that of wild-type cells, whereas cells with functionally significant mutations showed distinct differences in gene expression. To further investigate the impact of *TP53* mutations on cell growth, we conducted growth assays using HCT116 cells with 10 different *TP53* mutations that were categorized as functionally significant (Extended Data Fig. 8). Our data demonstrated that cells with these mutations exhibited a growth advantage over wild-type cells when treated with nutlin-3a, further supporting the notion that these mutations confer a growth advantage. This result established that this single-cell approach accurately identified the phenotypes of these mutations.

Discussion

In this study, we report a multiplexed method that uses base editors to introduce specific cancer mutations and single-cell sequencing to identify the genotype and phenotypes of the induced cancer mutations. Referred to as TISCC-seq, this approach overcomes issues with short-read based single-cell or bulk CRISPR screens, neither of which verify endogenous DNA variants that are engineered into the genomes of cells. This approach integrated single-cell long-read and short-read sequencing for CRISPR base editor screens. As a result, endogenous genetic variants introduced by the CRISPR base editor are directly confirmed from the target gene transcript. At single-cell resolution, the genetic variant and its resultant transcriptome changes become evident. Therefore, we can determine the functional consequences of *TP53* mutations across different cell lines. Some mutations had a greater functional impact on the cells' gene expression while a smaller subset had a wild-type-like phenotype. Our results corroborated some in silico predictions (Supplementary Table 9). For example, the R156C mutation is predicted to have neutral effect on p53 pathway[10,28]. This was confirmed experimentally among our results. In both cell lines used in this study, this mutation had a wild-type phenotype. Overall, this approach has the potential for enabling highly multiplexed functional evaluation of cancer mutations and germline variants. Following functional assays using cell lines with desired genetic variants will help deepen understanding of the phenotype of each variant as shown in Fig. 4.

Although we used four base editors for this study, there were some mutations that we were unable to target (Supplementary Fig. 2). We anticipate that modification of base editor properties such as their enzymatic activity[29,30], window[31] and PAM restriction[32] will broaden the types of mutations and other variants which can be engineered into genomes. The prime editor which can introduce any genetic variant at the target site will even enable saturation mutagenesis of the target gene[33].

Mutually exclusive *TP53* mutations were observed in HCT116 and U2OS cell lines through TISCC-seq analysis (Supplementary Fig. 8). Our analysis suggests that differences in CRISPR base editing efficiencies between the two cell lines may account for these mutations. For instance, the C135Y mutation, which was only detected in U2OS cells and deemed functionally significant, exhibited low editing efficiency (~1%) when we attempted to introduce it into HCT116 cells using a guide RNA with a CRISPR base editor. Consequently, the mutation was not observed in the HCT116 cell TISCC-seq data. Nevertheless, our findings revealed that the C135Y mutation conferred a growth advantage in HCT116 cells (Extended Data Fig. 8). We also investigated four functionally significant *TP53* mutations (I195T, Y220C, Y236H and L257P) in non-cancer MMNK1 cells. We treated these cells with nutlin-3a and found no evidence of a growth advantage in cells carrying these *TP53* mutations (Extended Data Fig. 9). This observation is consistent with the known role of the p53 pathway, which frequently triggers cell cycle arrest or apoptosis in response to various stresses that are more prevalent in developed cancer cells than in non-cancer cells. Our results underscore the potential utility of TISCC-seq in revealing the functional consequences of mutations across diverse cellular contexts, including primary cells and developed cancer cells.

We further demonstrated that TISCC-seq can be applied to longer genes by targeting *SF3B1*, which has a transcript longer than 6 kb, and introducing multiple mutations using CRISPR base editors in K562 cells. Our analysis using TISCC-seq successfully genotyped these mutations at the single-cell level (Extended Data Fig. 10). These results illustrate the versatility of TISCC-seq and its potential to enable the assessment of genetic variants across a broad range of genomic contexts, including longer genes.

The complexities of high-throughput CRISPR engineering, single-cell sequencing and its higher cost limit the scalability of single-cell CRISPR screens compared to conventional genetic screens done with conventional bulk assays. TISCC-seq provides some potential benefits that may be useful for standard CRISPR screens. For example, one can use a bulk-based cellular genetic screen for hundreds of thousands of sgRNAs generating variants and then narrow down the sgRNAs to the hundreds with remarkable impact on cell survival or drug response. Then, TISCC-seq can be used for a deeper analysis of sgRNAs by detecting genuine endogenous mutations and their resultant phenotype at single-cell level resolution. This combination may enable more accurate evaluation of CRISPR-based screens in the future.

The sensitivity of single-cell RNA-seq is limited. Therefore, we can only detect a limited number of transcripts for each gene. It is challenging to detect any transcripts from low-expressed genes in individual cells. This sparsity in single-cell RNA-seq data restricts the application of TISCC-seq to genes with extremely low expression levels. However, advancements in single-cell reverse transcription and transcript enrichment technology can greatly enhance the efficiency of TISCC-seq.

Online content

Any methods, additional references, Nature Portfolio 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/s41587-023-01949-8.

References

1. Cuella-Martin, R. et al. Functional interrogation of DNA damage response variants with base editing screens. *Cell* **184**, 1081–1097 (2021).
2. Hanna, R. E. et al. Massively parallel assessment of human variants with base editor screens. *Cell* **184**, 1064–1080 (2021).
3. Kim, Y. et al. High-throughput functional evaluation of human cancer-associated mutations using base editors. *Nat. Biotechnol.* **40**, 874–884 (2022).
4. Sanchez-Rivera, F. J. et al. Base editing sensor libraries for high-throughput engineering and functional analysis of cancer-associated single nucleotide variants. *Nat. Biotechnol.* **40**, 862–873 (2022).
5. Ursu, O. et al. Massively parallel phenotyping of coding variants in cancer with Perturb-seq. *Nat. Biotechnol.* **40**, 896–905 (2022).
6. Hill, A. J. et al. On the design of CRISPR-based single-cell molecular screens. *Nat. Methods* **15**, 271–274 (2018).
7. Kim, H. S., Grimes, S. M., Hooker, A. C., Lau, B. T. & Ji, H. P. Single-cell characterization of CRISPR-modified transcript isoforms with nanopore sequencing. *Genome Biol.* **22**, 331 (2021).
8. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
9. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
10. Tate, J. G. et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
11. Berglind, H., Pawitan, Y., Kato, S., Ishioka, C. & Soussi, T. Analysis of p53 mutation status in human cancer cell lines: a paradigm for cell line cross-contamination. *Cancer Biol. Ther.* **7**, 699–708 (2008).
12. de Andrade, K. C. et al. The TP53 Database: transition from the International Agency for Research on Cancer to the US National Cancer Institute. *Cell Death Differ.* **29**, 1071–1073 (2022).
13. Leroy, B. et al. Analysis of TP53 mutation status in human cancer cell lines: a reassessment. *Hum. Mutat.* **35**, 756–765 (2014).
14. Tovar, C. et al. Small-molecule MDM2 antagonists reveal aberrant p53 signaling in cancer: implications for therapy. *Proc. Natl Acad. Sci. USA* **103**, 1888–1893 (2006).
15. Honda, R., Tanaka, H. & Yasuda, H. Oncoprotein MDM2 is a ubiquitin ligase E3 for tumor suppressor p53. *FEBS Lett.* **420**, 25–27 (1997).
16. Vassilev, L. T. et al. In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303**, 844–848 (2004).
17. Grunewald, J. et al. Transcriptome-wide off-target RNA editing induced by CRISPR-guided DNA base editors. *Nature* **569**, 433–437 (2019).
18. Kim, S., Kim, D., Cho, S. W., Kim, J. & Kim, J. S. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res.* **24**, 1012–1019 (2014).
19. Replogle, J. M. et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat. Biotechnol.* **38**, 954–961 (2020).
20. Adamson, B. et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* **167**, 1867–1882 (2016).
21. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
22. Jaitin, D. A. et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-seq. *Cell* **167**, 1883–1896 (2016).
23. Rubin, A. J. et al. Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. *Cell* **176**, 361–376 (2019).
24. Kim, H. K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci. Adv.* **5**, eaax9249 (2019).
25. Song, M. et al. Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. *Nat. Biotechnol.* **38**, 1037–1043 (2020).
26. Fischer, M. Census and evaluation of p53 target genes. *Oncogene* **36**, 3943–3956 (2017).
27. Landrum, M. J. et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res.* **44**, D862–D868 (2016).

28. Kakudo, Y., Shibata, H., Otsuka, K., Kato, S. & Ishioka, C. Lack of correlation between p53-dependent transcriptional activity and the ability to induce apoptosis among 179 mutant p53s. *Cancer Res.* **65**, 2108–2114 (2005).
29. Richter, M. F. et al. Phage-assisted evolution of an adenine base editor with improved Cas domain compatibility and activity. *Nat. Biotechnol.* **38**, 883–891 (2020).
30. Thuronyi, B. W. et al. Continuous evolution of base editors with expanded target compatibility and improved activity. *Nat. Biotechnol.* **37**, 1070–1079 (2019).
31. Huang, T. P. et al. Circularly permuted and PAM-modified Cas9 variants broaden the targeting scope of base editors. *Nat. Biotechnol.* **37**, 626–631 (2019).
32. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290–296 (2020).
33. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).

Methods

Cell culture conditions

HEK293T (ATCC CRL-11268) and MMNK-1 (JCRB1554) cells were maintained in Dulbecco's modified Eagle's medium (DMEM) with 10% fetal bovine serum (FBS). HCT116 (ATCC CCL-247) cells and U2OS (ATCC HTB-96) were maintained in McCoy's 5 A modified medium supplemented with 10% FBS. We stimulated p53 pathway of cells with 10 μM Nutlin-3a. K562 (ATCC CCL-243) cells were maintained in RPMI 1640 with 10% FBS. Cells were authenticated by STR profiling. All cell lines were confirmed by PCR to be free of mycoplasma contamination.

Lentiviral gRNA library production

The oligonucleotides for sgRNA library generation were ordered using IDT oPools Oligo Pools. Amplified gRNA cassettes were cloned using NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs) into lentiGuide-Puro (Addgene plasmid #52963). Purified plasmids were electroporated to ElectroMAX Stbl4 Competent Cells (New England Biolabs) and amplified.

Lentivirus production

Approximately 2.0×10^6 HEK293T cells were plated 24 h before transfection. Cells were transfected with pMD2.G (500 ng, Addgene plasmid #12259), psPAX2 (1.500 ng, Addgene plasmid #12260) and lentiviral sgRNA library (2,000 ng) using Lipofectamine 2000 (Invitrogen) as per the manufacturer's protocol. The viral supernatant was collected after 48 h of transfection. The supernatants were filtered through a 0.45 μm filter and transduced to cells.

Lentivirus transduction

HCT116 and U2OS cells were diluted to 1.4×10^5 and 0.7×10^5 cells ml^{-1} and plated a day before the transduction. Lentiviral supernatant and polybrene (8 μg ml^{-1}, Sigma-Aldrich) were added to the cells. After 24 h, transduced cells were selected by puromycin (Life Technologies) at concentration of 0.4 μg ml^{-1} and 1.0 μg ml^{-1}.

Transfection and electroporation conditions

We used 1.2×10^6 HEK293T cells to transfect the base editor plasmids (2,000 ng) using Lipofectamine 2000 (Invitrogen) as per the manufacturer's protocol. We used 1.0×10^6 HCT116, U2OS and K562 cells to transfect the base editor plasmids (2,600 ng) using SE or SF solution and 4D-nucleofector (Lonza) as per the manufacturer's protocol. We used SE solution and DN-100 program for MMNK-1 cells. Base editor plasmids pCMV_AncBE4max_P2A_GFP and pCMV_ABEmax_P2A_GFP were gifts from D. Liu (Addgene plasmid # 112100 and 112101)[34]. Base editor constructs pCAG-CBE4max-SpG-P2A-EGFP (RTW4552) and pCMV-T7-ABEmax(7.10)-SpG-P2A-EGFP (RTW4562) were gifts from Benjamin Kleinstiver (Addgene plasmid #139998 and #140002)[32]. After 6 days of electroporation, cells were subjected to chemical treatment or single-cell library preparation. For *TP53* variant clone generation, base editor plasmids (2,250 ng) and sgRNA plasmid (750 ng) were electroporated to cells. We conducted single-cell subcloning with limiting dilution and confirmed the genotype of the target with PCR amplification and sequencing.

Single-cell library preparation

Single-cell cDNA and gene expression libraries are generated using Chromium Next GEM Single Cell 5′ Library & Gel Bead Kit v2 (10x Genomics) according to the manufacturer's protocol. The cDNA and gene expression libraries are amplified with 16 and 14 cycles of PCR respectively. The quality of gene expression libraries is confirmed using 2% E-Gel (ThermoFisher Scientific). We quantified the sequencing libraries using Qubit (Invitrogen) and sequenced on Illumina sequencers (Illumina).

Single-cell sgRNA capture and sequencing

The sgRNA direct capture was performed as previously described[7,19]. Briefly, 6 pmol sgRNA scaffold binding primer was added to RT master mix. After cDNA amplification, the sgRNA fractions were purified using SPRIselect bead (Beckman Coulter Life Sciences). The library was amplified and sequenced with gene expression library.

Long-read sequencing

Ten ng of the single-cell full-length cDNA were used to amplify transcripts. Primer sequences are shown in Supplementary Table 5. We used KAPA HiFi HotStart ReadyMix (Roche) for amplification. Libraries were prepared with 900 fmol of each amplicon for Promethion flow cell FLO-PRO002 (Oxford Nanopore Technologies) using Native Barcoding Expansion and Ligation Sequencing Kit (Oxford Nanopore Technologies) according to the manufacturer's protocol. Libraries were sequenced on a Promethion over 72 h.

Single-cell transcript analysis

Short-read transcripts. Base calling for 5′ gene expression libraries was performed using cellranger 6.0 (10x Genomics). In preparation for integrated analysis, the transcript count matrices generated by cellranger were processed by Seurat 3.0.2 (ref. 35). QC filtering removed cells with fewer than 100 or more than 8,000 genes, cells with more than 30% mitochondrial genes and cells predicted to be doublets by DoubletFinder[36]. Additionally, any genes present in three or fewer cells were removed. Batch effects between each single-cell cDNA generation reaction and base editors were corrected by Harmony[37]. Cell cycle phase were also corrected by Harmony.

Long-read variant calling. Base calling was performed using guppy 5 with super accuracy mode and alignment to the GRCh38 reference genome using minimap2 (refs. 8,9). Cell barcodes and UMIs are extracted as previously described[7]. For validating *TP53* mutation genotyping, we filtered out UMIs less than 10 reads and consolidated UMIs with high similarity (edit distance less than 3). A custom python script utilizing the pysam module was used to identify reads spanning the sgRNA target windows and extracted the base calls at each position within the window. Base calls were used predict amino acid changes per each cell. Cells with heterozygous amino acid changes were excluded for the gene expression analysis. Output from this script was summarized to provide expected amino acid change per cell barcode.

Integration of long and short reads. The variant per cell barcode table were added to the Seurat object metadata as a new column. Cells without high-quality long-read data were filtered. For gene expression analysis, we filtered variants which were detected in less than 5 cells. A hierarchical clustering was done in R using hclust, cutree and dendextend. Biological pathway analysis was performed with the GSVA tool[38].

Cell cycle analysis

We used Click-iT Plus EdU Alexa Fluor 488 Flow Cytometry Assay Kit (Life Technologies) according to manufacturer's protocol. Briefly, we plated cells a day before nutlin-3a or vehicle treatment. After 24 h of chemical treatment, cells in S-phase were labeled with 10 mM EdU solution for 2 h. FxCycle PI/RNase Staining Solution (Life Technologies) was used for PI staining. After the staining, cells were analyzed by NovoCyte Quanteon Flow Cytometer Systems (Agilent).

RNA-seq

We used KAPA mRNA HyperPrep Kit (Roche) for mRNA-seq library preparation according to manufacturer's protocol. For each cell type, we used triplicate library preparations with 1 μg total RNA as an input. Libraries were sequenced by NextSeq (Illumina) by 75 bp paired-end sequencing. The reads were aligned to the reference genome GRCh38

by a two-pass method with STAR and gene expression level was measured using HT-Seq[39,40]. We used DESeq2 for DE analysis[41]. Biological pathway analysis was performed with the GSVA tool[38].

Reporting summary

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

Data availability

High-throughput DNA sequencing files are available from the NCBI SRA under BioProject PRJNA880341.

Code availability

Scripts for analysis are available on Zenodo (https://zenodo.org/badge/latestdoi/555044610) under the MIT license terms.

References

34. Koblan, L. W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
35. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
36. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329–337 (2019).
37. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
38. Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14**, 7 (2013).
39. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
40. Anders, S., Pyl, P. T. & Huber, W. HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
41. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Acknowledgements

This work was fully supported by the US National Institutes of Health grants R33 CA278469-01 (H.P.J.), R33 CA247700 (H.P.J., H.S.K. and B.T.L.), R01HG006137 (H.P.J. and H.S.K.) and R35HG011292-01 (B.T.L. and H.S.K.). H.S.K. received additional support for conventional gene expression studies through the National Research Foundation of Korea grant from the Korean Ministry of Science and ICT (RS-2023-00243993). H.P.J. and H.S.K. also received support from the Clayville Foundation.

Author contributions

H.S.K. and H.P.J. were involved in conception and design of the study, development of methodology, acquisition of data, analysis and interpretation of data and writing of the manuscript. H.S.K., S.M.G., T.C., A.S., B.T.L., G.-H.W., S.B. and H.P.J. were involved in analysis and interpretation of data. All authors contributed to the writing of the manuscript. H.P.J. oversaw all aspects of the study. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

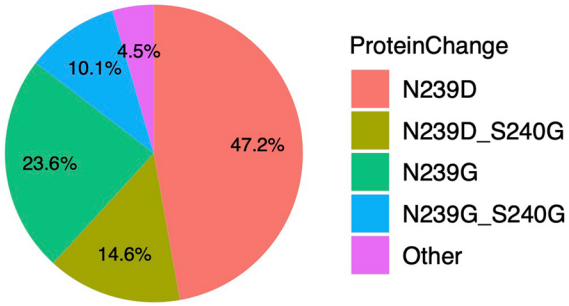
Extended data is available for this paper at https://doi.org/10.1038/s41587-023-01949-8.

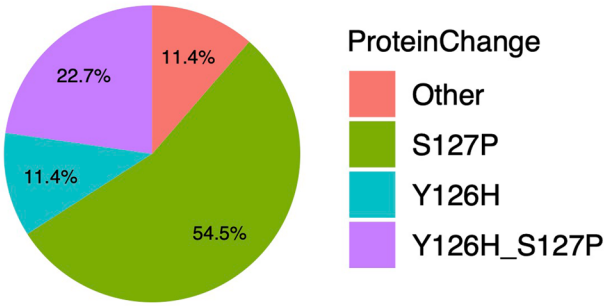
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01949-8.

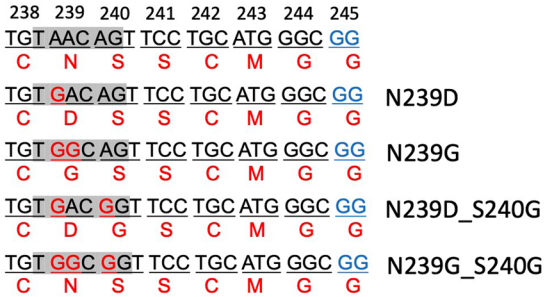
Correspondence and requests for materials should be addressed to Hanlee P. Ji.

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

https://doi.org/10.1038/s41587-023-01949-8

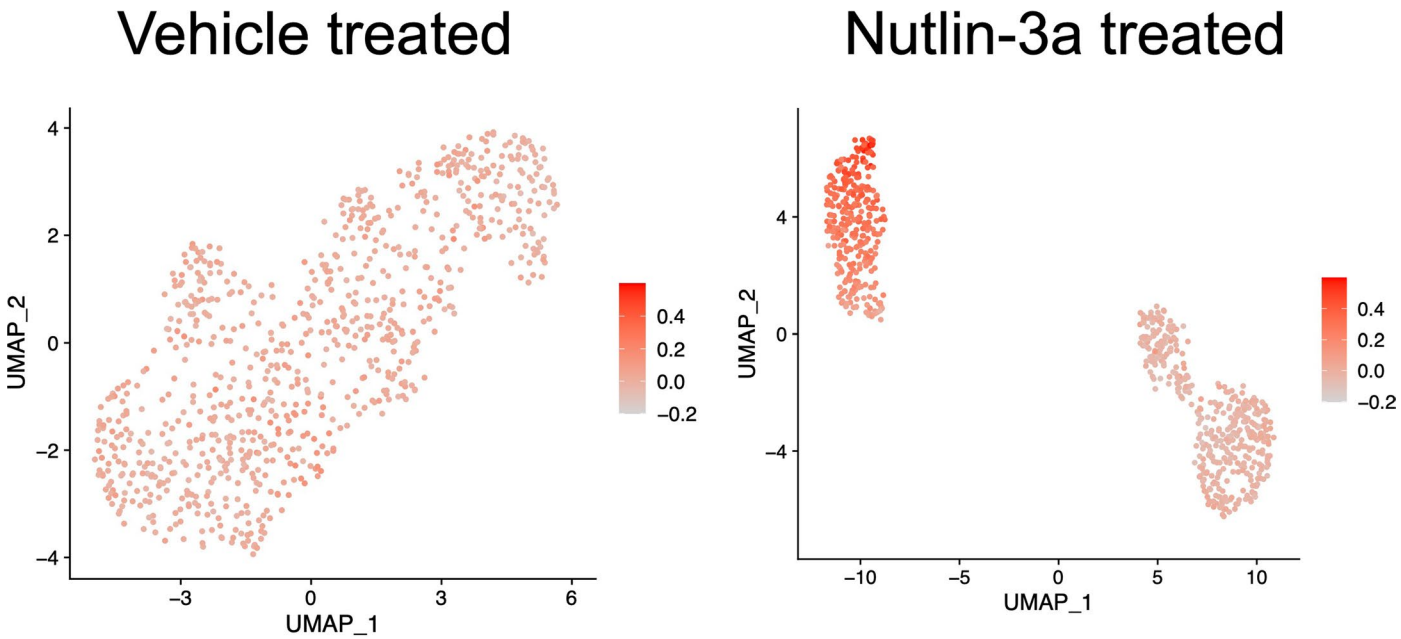






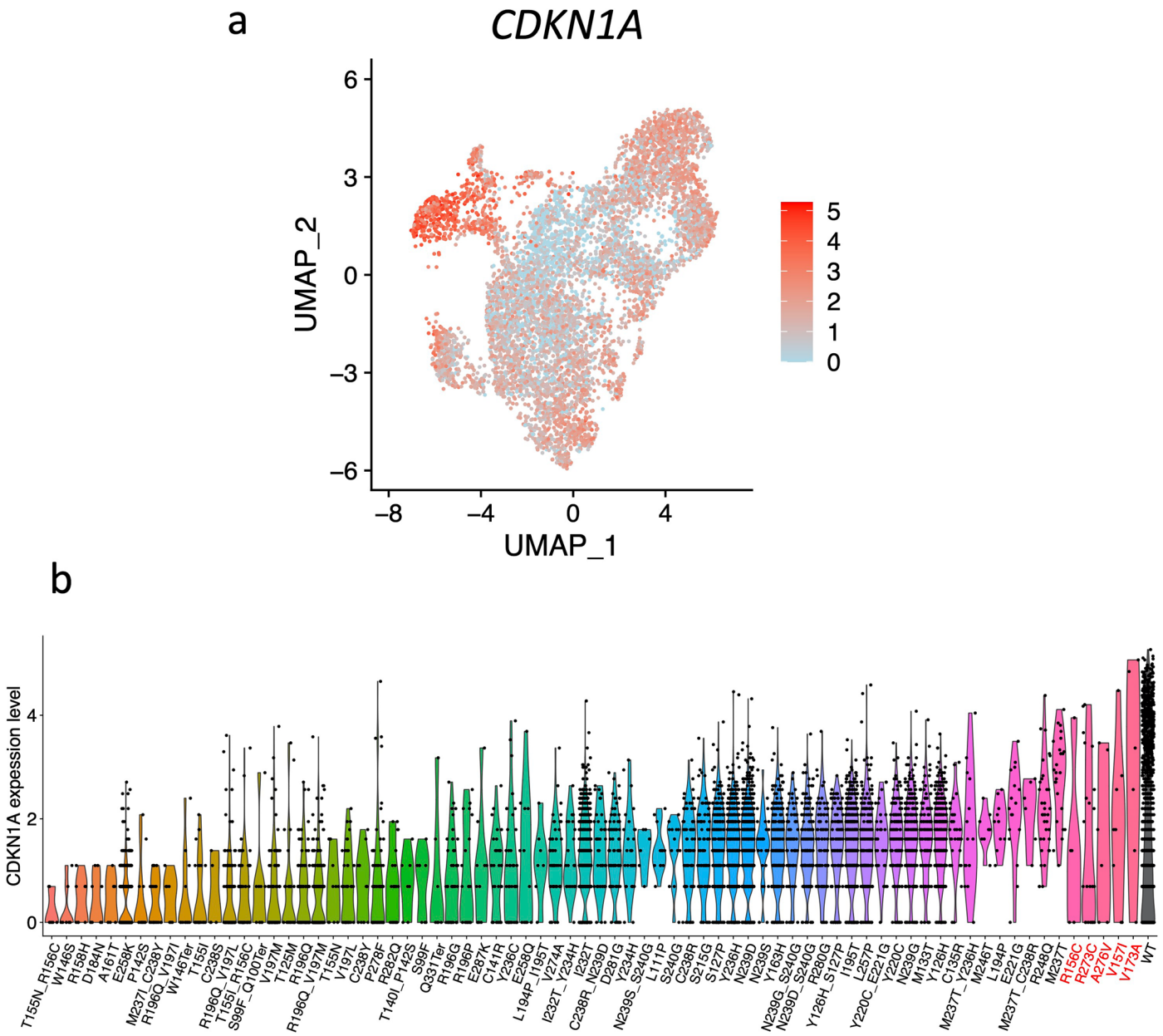
Extended Data Fig. 1 | Pie charts showing the proportion of resultant amino acid changes from cells with sgRNA targeting N239D or S127P mutations. Proportions of mutations are calculated from the single-cell cDNA long-read sequencing. Underlines indicate each triplet codon and number indicate position of the codon. Red DNA sequences indicate substituted bases and blues indicate PAM sequences.

https://doi.org/10.1038/s41587-023-01949-8



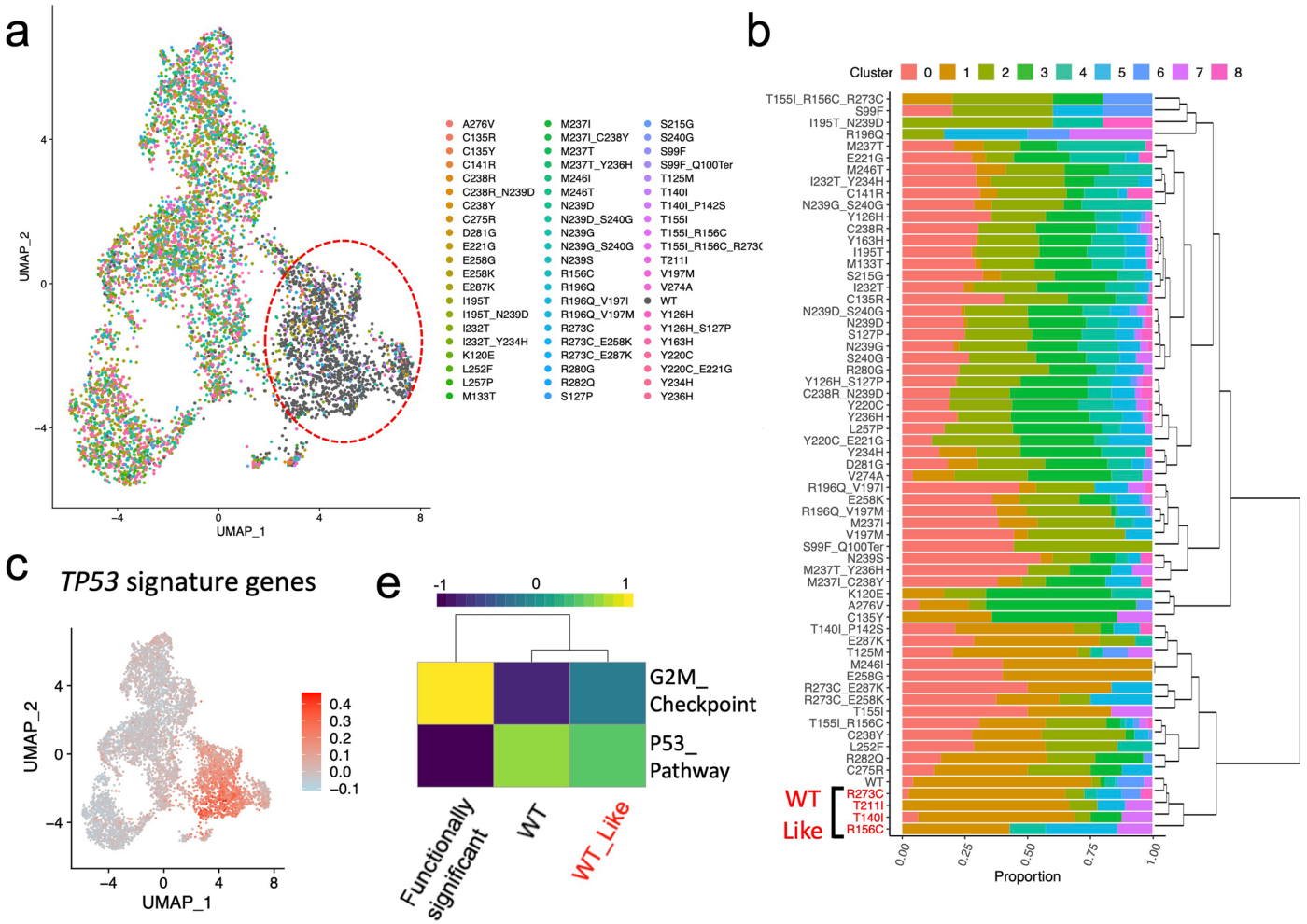
Extended Data Fig. 2 | UMAP embedding of cells colored by P53 pathway gene scores. *TP53* variants were introduced to HCT116 cells using a subset of our sgRNA library and analyzed.

https://doi.org/10.1038/s41587-023-01949-8



Extended Data Fig. 3 | *CDKN1A* expression level from HCT116 cells with various *TP53* genetic variants. (a) UMAP embedding of cells colored by *CDKN1A* gene expression. **(b)** Violin plot showing *CDKN1A* gene expression level per cells with each genetic variant. Reds indicate wild type like variants.

https://doi.org/10.1038/s41587-023-01949-8

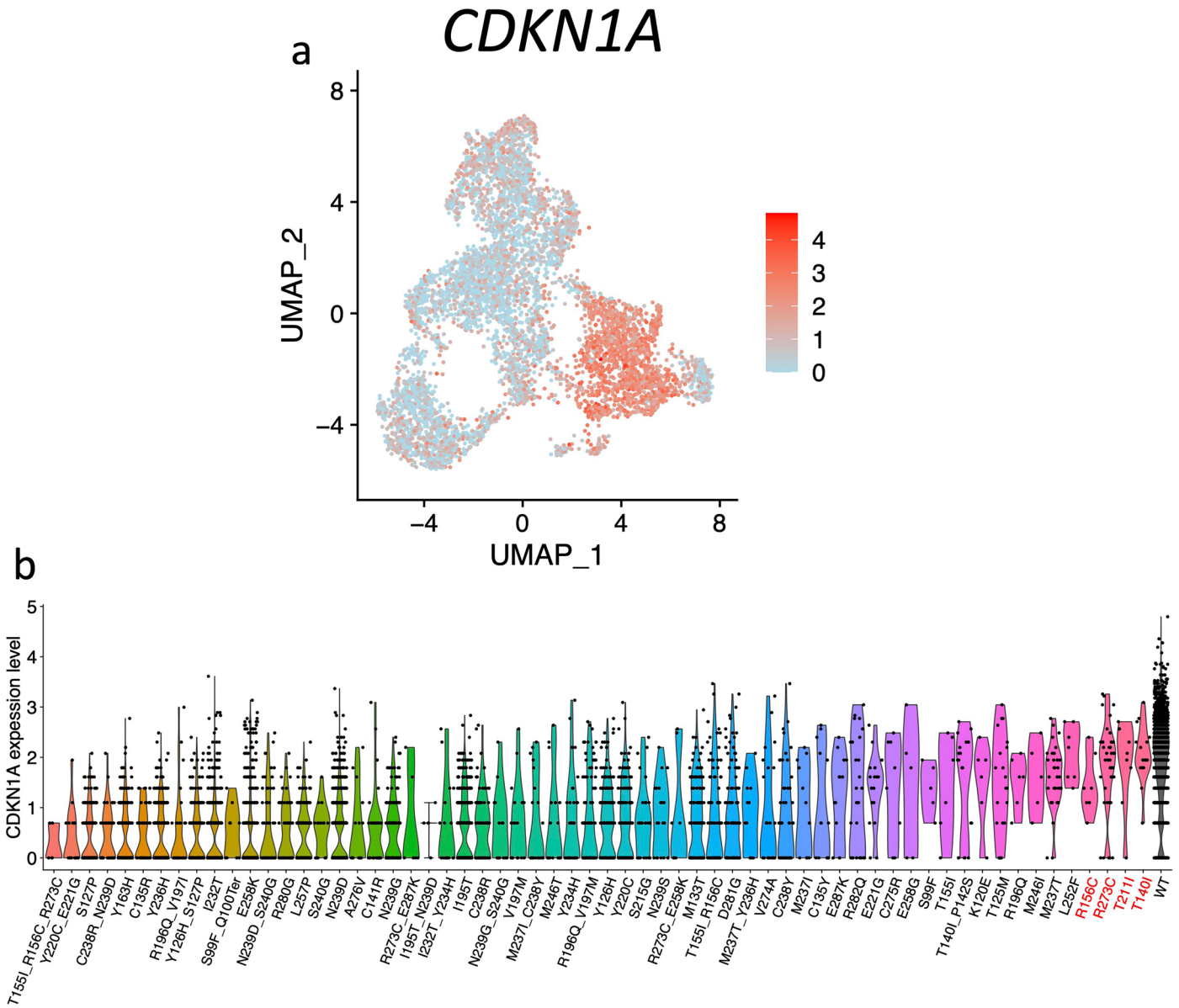


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41587-023-01949-8

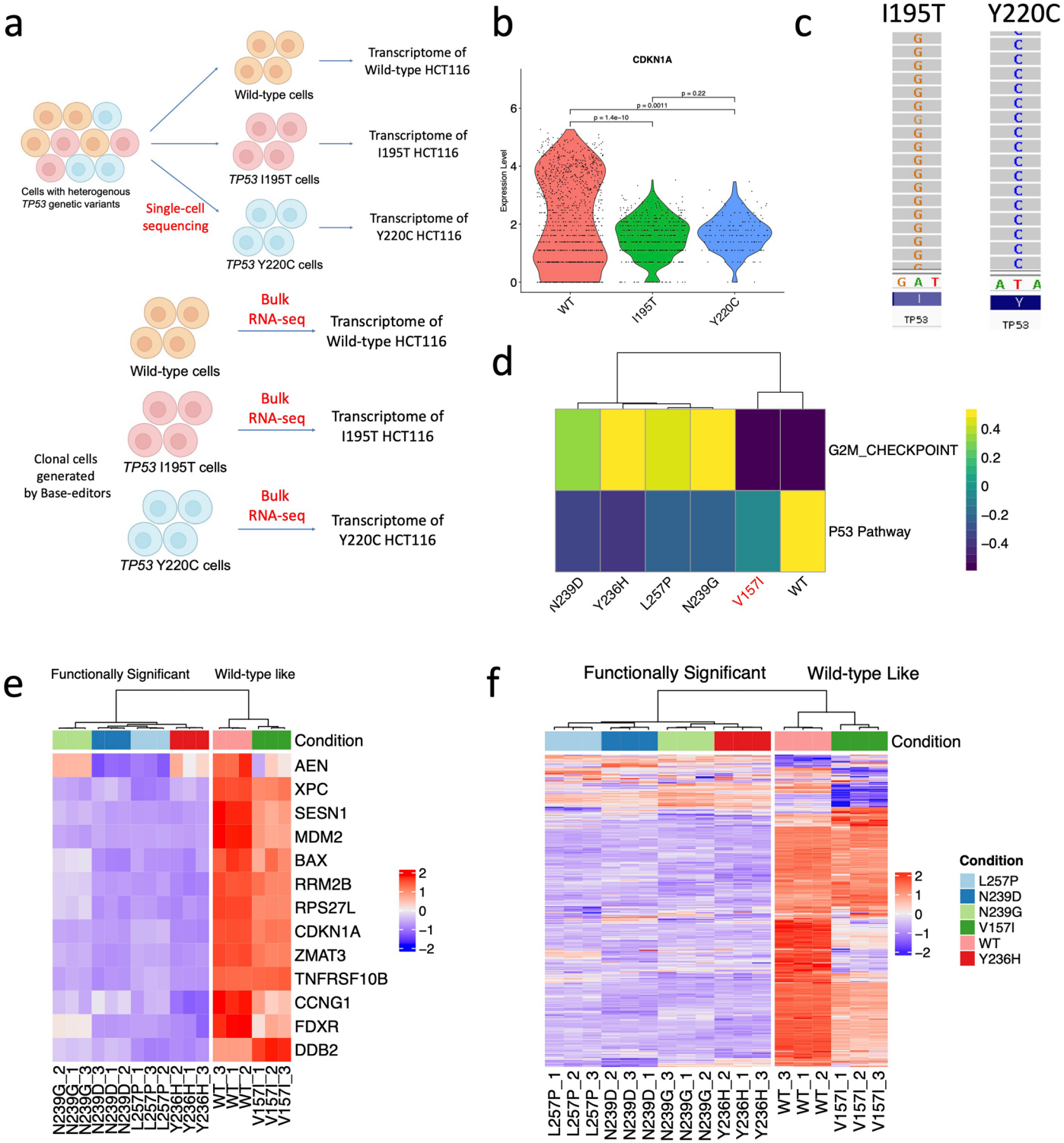
Extended Data Fig. 4 | *TP53* variants analysis in U2OS cells. (a) UMAP plot showing single-cell gene expression profile per each genetic variant. U2OS cells are treated with Nutlin-3a after introduction of variants using full sgRNA library. **(b)** Proportion of UMAP cluster from cells with each genetic variant. Hierarchical clustering was performed based on the proportion to categorize genetic variants. Reds indicate wild type-like variants. **(c)** UMAP embedding of cells colored by p53 pathway gene scores. **(d)** Violin plot showing the p53 pathway gene score per cells with each genetic variant. *: P < 0.03, n.s: Not significant; two-sided t-test. P = 8.0e-08, 2.0e-11, 6.5e-65, 6.8e-16, 0.0e + 00, 4.9e-12, 1.6e-259, 4.2e-65, 1.6e-20, 4.2e-04, 0.0e + 00, 5.5e-195, 9.2e-300, 3.8e-311, 4.9e-63, 1.8e-15, 1.0e-19, 2.0e-224, 1.6e-10, 1.6e-58, 0.0e + 00, 0.0e + 00, 9.6e-05, 7.4e-09, 3.4e-13, 5.9e-173, 1.4e-183, 1.0e-52, 6.4e-191, 5.3e-120, 5.8e-14, 1.3e-06, 9.6e-80, 1.5e-13, 1.3e-52, 1.4e-03, 1.2e-08, 5.6e-13, 2.3e-07, 3.9e-05, 5.6e-10, 1.9e-28, 7.8e-21, 1.4e-05, 7.7e-07, 5.8e-05, 2.2e-05, 4.3e-12, 1.1e-02, 4.9e-04, 3.5e-05, 1.5e-01, 3.5e-03, 2.7e-02, 1.1e-11, 3.0e-03, 1.0e-01, 3.6e-01, 4.8e-02, 4.0e-02, 5.0e-01, 7.2e-01. **(e)** Heatmap showing average GSVA enrichment score of selected Hallmark pathways per each category of genetic variant.

https://doi.org/10.1038/s41587-023-01949-8



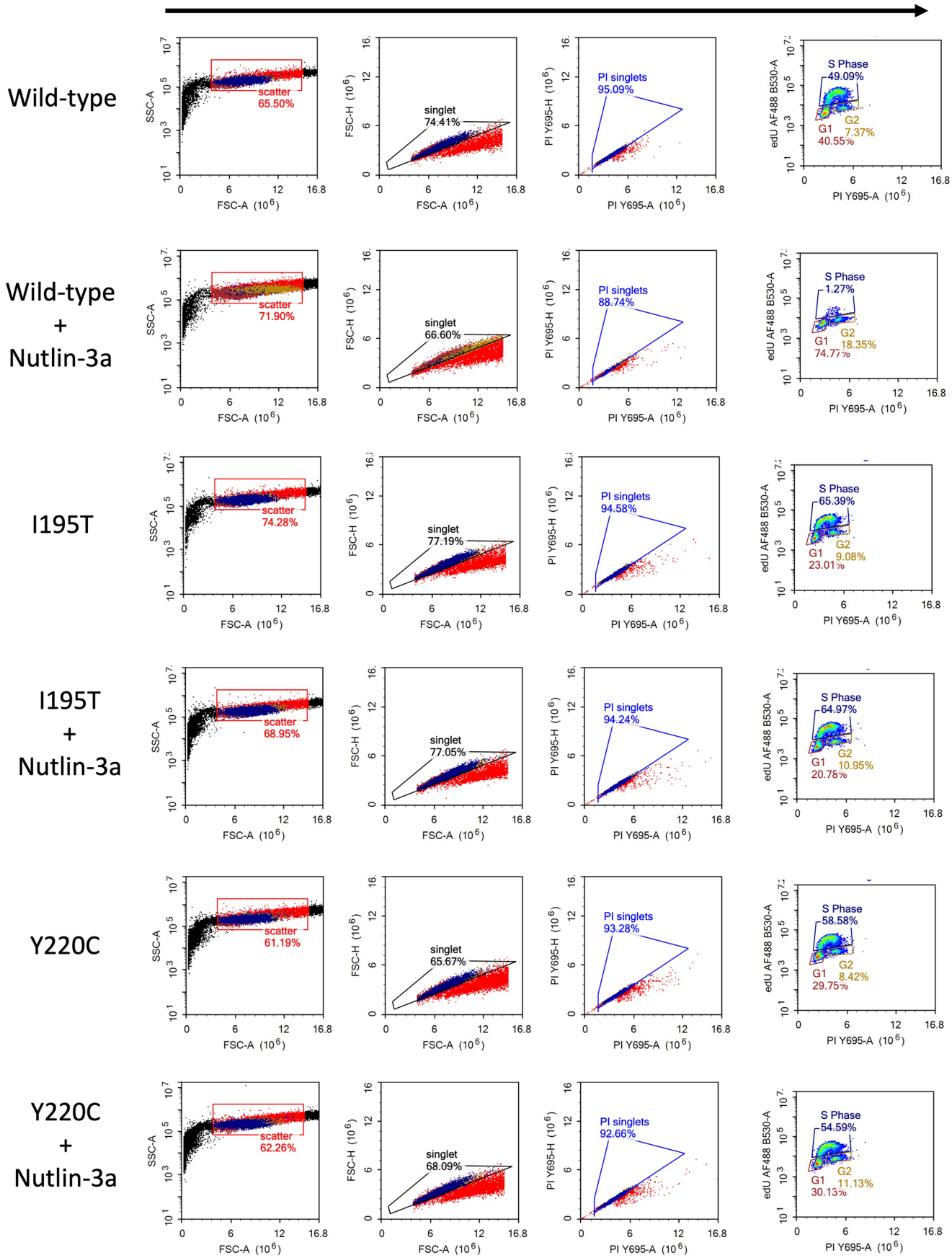
Extended Data Fig. 5 | *CDKN1A* expression level from U2OS cells with various *TP53* genetic variants. (a) UMAP embedding of cells colored by *CDKN1A* gene expression. **(b)** Violin plot showing *CDKN1A* gene expression level per cells with each genetic variant. Reds indicate wild type like variants.

https://doi.org/10.1038/s41587-023-01949-8



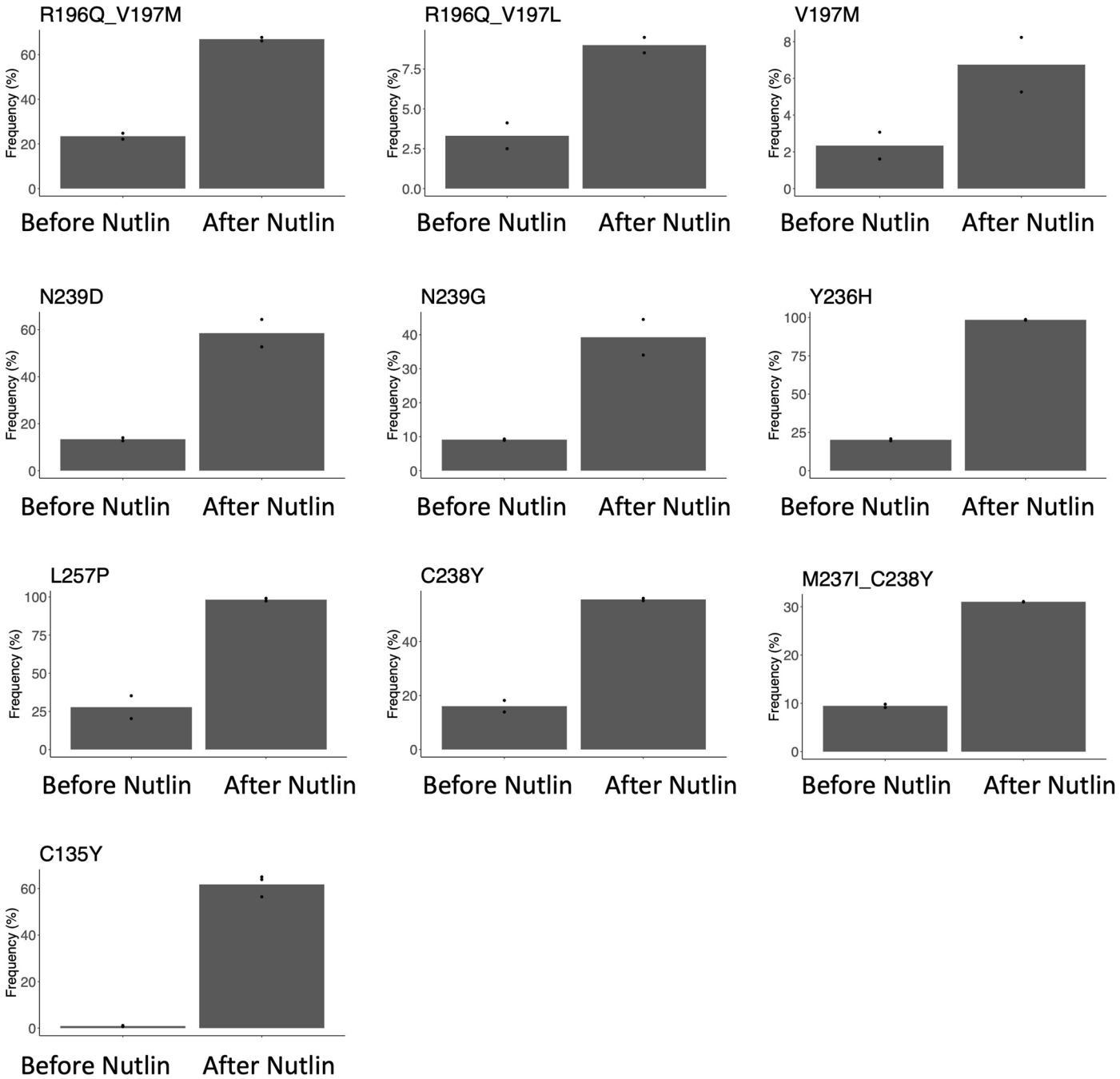
Extended Data Fig. 6 | Confirmation of single-cell variants screen using clonal cell-lines. (a) Overview of the confirmation experiments. **(b)** Violin plot showing the expression of *CDKN1A* from single-cell sequencing result. P-values are from a two-sided t-test. **(c)** Identified genetic variants from bulk RNA-seq from isolated clonal cells. **(d)** Heatmap showing the average GSVA enrichment score of selected pathways on various *TP53* mutant HCT116 cell-lines. The V157I mutation was considered as wild-type like from TISCC-seq. **(e)** Heatmap displays the expression of the 13 genes involved in the p53 pathway. **(f)** Heatmap shows the expression of 343 genes associated with the P53 pathway.

https://doi.org/10.1038/s41587-023-01949-8



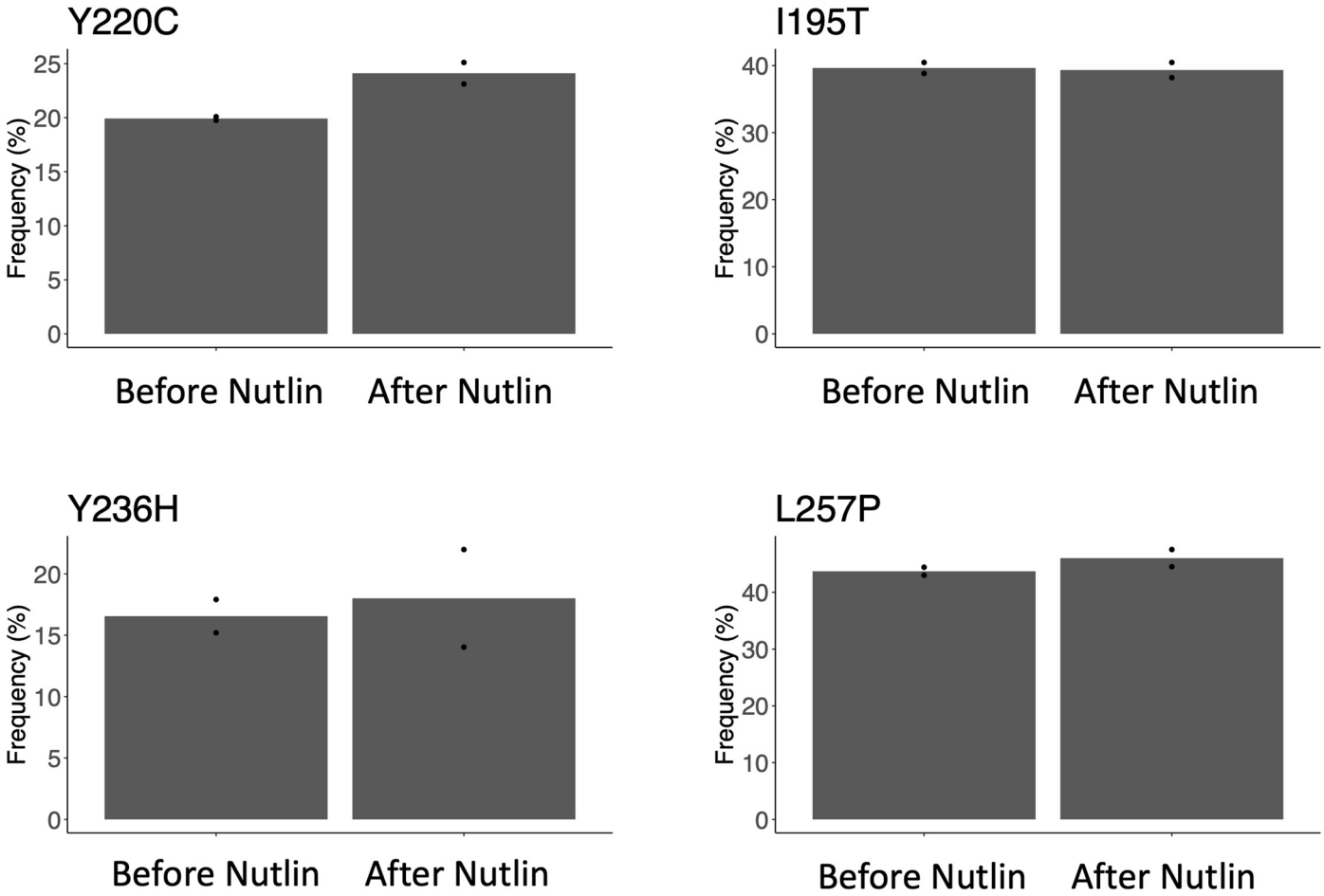
Extended Data Fig. 7 | FACS gating strategy. Gating strategy for cell cycle assay presented on Fig. 4c.

https://doi.org/10.1038/s41587-023-01949-8



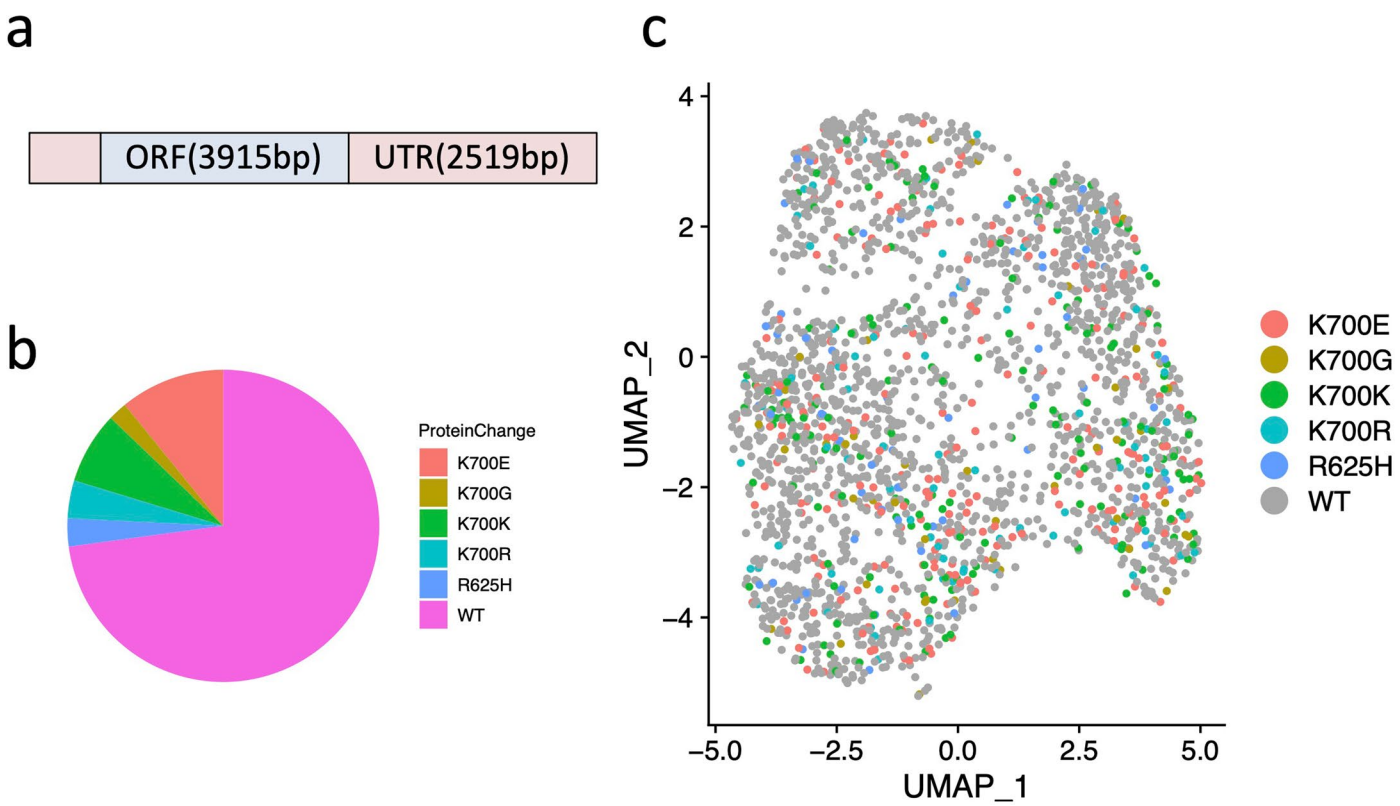
Extended Data Fig. 8 | Growth advantage of *TP53* mutations identified from TISCC-seq in HCT116 cells confirmed by CRISPR base editing and nutlin-3a treatment. We introduced various *TP53* mutations to HCT116 cells using CRISPR base editors and subsequently cultured the cells under media containing nutlin-3a. Analysis of the resulting population revealed an increasing frequency of the introduced mutations, indicating a growth advantage for cells with *TP53* mutations compared to wild-type cells. These results provide evidence for the selective advantage conferred by *TP53* mutations in HCT116 cells. N = 3 biologically independent cells for C135Y and N = 2 biologically independent cells for others. P = 0.00352, 0.0417, 0.161, 0.0786, 0.109, 0.00165, 0.064, 0.0287, 0.00885, 0.00189; two-sided t-test.

https://doi.org/10.1038/s41587-023-01949-8



Extended Data Fig. 9 | Analysis of the impact of *TP53* mutations in non-cancerous MMNK1 cells. We introduced various *TP53* mutations to MMNK1 cells using CRISPR base editors and subsequently cultured the cells under media containing nutlin-3a. Analysis of the resulting population revealed no growth advantage for cells with *TP53* mutations compared to wild-type cells. N = 2 biologically independent cells. P = 0.14, 0.85, 0.78, 0.34; two-sided t-test.

https://doi.org/10.1038/s41587-023-01949-8



Extended Data Fig. 10 | Single-cell genotyping of *SF3B1* gene using TISCC-seq. We used the CRISPR base editor to introduce various mutations in *SF3B1* in K562 cells and genotyped them using TISCC-seq. **(a)** Transcript structure of *SF3B1*. **(b)** Result of single-cell genotype of *SF3B1*. **(c)** Gene expression profile based on *SF3B1* mutations.

nature portfolio

Corresponding author(s): Hanlee P. Ji

Last updated by author(s): Aug 1, 2023

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	Default softwares for sequencing management were used for all sequencing equipment.
Data analysis	Every softwares are described in the manuscript (e.g., Cell ranger, Seurat, guppy, minimap, harmony, python, R). Codes are deposited at zenodo (https://zenodo.org/badge/latestdoi/365008149).

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

High-throughput DNA sequencing files are available from the NCBI SRA under BioProject PRJNA880341.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

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

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	No statistical methods were used to predetermine sample size for experiments.
Data exclusions	There are no excluded data.
Replication	Genetic variants replicated in more than 5 cells are used for statistical analysis.
Randomization	Samples were not randomized.
Blinding	The investigators were blinded to group allocation.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HEK293T, K562, HCT116 and U2OS cells are purchased from ATCC. MMNK1 cells are purchased from JCRB.
Authentication	Cells are authenticated by STR profiling
Mycoplasma contamination	All cell lines were confirmed by PCR to be free of mycoplasma contamination.

Commonly misidentified lines (See ICLAC register)	Not applicable.

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	HCT116 cells are detached by trypLE and filtered before FACS
Instrument	NovoCyte Quanteon
Software	NovoExpress 1.4.4
Cell population abundance	Cells are not sorted
Gating strategy	FSC/SSC was used for primary gating. Cells cycles are custom gated by PI and EdU staining.

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