nature chemical biology

Article

https://doi.org/10.1038/s41589-024-01731-0

Functional profiling of serine, threonine and tyrosine sites

Received: 4 December 2023

Accepted: 14 August 2024

Published online: 23 September 2024

Check for updates

Yizhou Li[1,2,4], Tao Xu[1,4], Huazheng Ma[1,3,4], Di Yue[1], Qiezhong Lamao[1,2], Ying Liu[1,2], Zhuo Zhou[1,3] & Wensheng Wei[1,2] ✉

Systematic perturbation of amino acids at endogenous loci provides diverse insights into protein function. Here, we performed a genome-wide screen to globally assess the cell fitness dependency of serine, threonine and tyrosine residues. Using an adenine base editor, we designed a whole-genome library comprising 817,089 single guide RNAs to perturb 584,337 S, T and Y sites. We identified 3,467 functional substitutions affecting cell fitness and 677 of them involving phosphorylation, including numerous phosphorylation-mediated gain-of-function substitutions that regulate phosphorylation levels of itself or downstream factors. Furthermore, our findings highlight that specific substitution types, notably serine to proline, are crucial for maintaining domain structure broadly. Lastly, we demonstrate that 309 enriched hits capable of initiating cell overproliferation might be potential cancer driver mutations. This study represents an extensive functional profiling of S, T and Y residues and provides insights into the distinctive roles of these amino acids in biological mechanisms and tumor progression.

Understanding protein function at the level of individual amino acids is crucial for gaining mechanistic insights, as substitutions at this level can fundamentally alter the modification, expression, localization or interaction profiles of associated proteins[1–3]. Post-translational modification (PTM) has a critical regulatory role in various biological processes, with the majority occurring at the single amino acid level. Among these modifications, phosphorylation accounts for the highest proportion according to the dbPTM database, primarily occurring on serine, threonine and tyrosine residues (Fig. 1a). Phosphorylation influences protein function in various ways, such as the activation or inactivation, signal transduction and crosstalk with other PTMs[4,5]. Furthermore, the dysregulation of phosphorylation is frequently associated with cancer emergence and metastasis, drawing great attention to the development of kinase inhibitors to target these abnormalities. Therefore, it is crucial to globally identify phosphorylation-mediated critical functions.

Clustered regularly interspaced short palindromic repeats (CRISPR) base editors (BEs), which facilitate the conversion of C•G to T•A or A•T to G•C nucleotide transitions at designated loci, have been used to investigate human genetic variants. Unlike Cas9-induced double-strand breaks (DSBs), BEs introduce nicks at target sites, reducing genetic dependency bias in the DNA damage response[6,7]. Furthermore, the outcomes of BE-mediated editing are relatively predictable, producing specific substitutions within a defined window[8]. Recent studies benchmarked gene loss-of-function (LOF) screens using cytosine BEs (CBEs)[9–11] and evaluated the functions of variants in the DNA repair pathway. These endeavors highlight the value of extensively identification of novel mechanisms through amino acid perturbation using BEs.

In this study, we explored the capabilities of ABEmax[12], an optimized version of the adenine BE (ABE), to induce genome-wide missense mutations in S, T and Y codons and systematically examined the

[1] Biomedical Pioneering Innovation Center, Beijing Advanced Innovation Center for Genomics, Peking-Tsinghua Center for Life Sciences, Peking University Genome Editing Research Center, State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing, China. [2] Changping Laboratory, Beijing, China. [3] State Key Laboratory of Common Mechanism Research for Major Diseases, Suzhou Institute of Systems Medicine, Chinese Academy of Medical Sciences and Peking Union Medical College, Suzhou, China. [4] These authors contributed equally: Yizhou Li, Tao Xu, Huazheng Ma. ✉e-mail: wswei@pku.edu.cn

https://doi.org/10.1038/s41589-024-01731-0

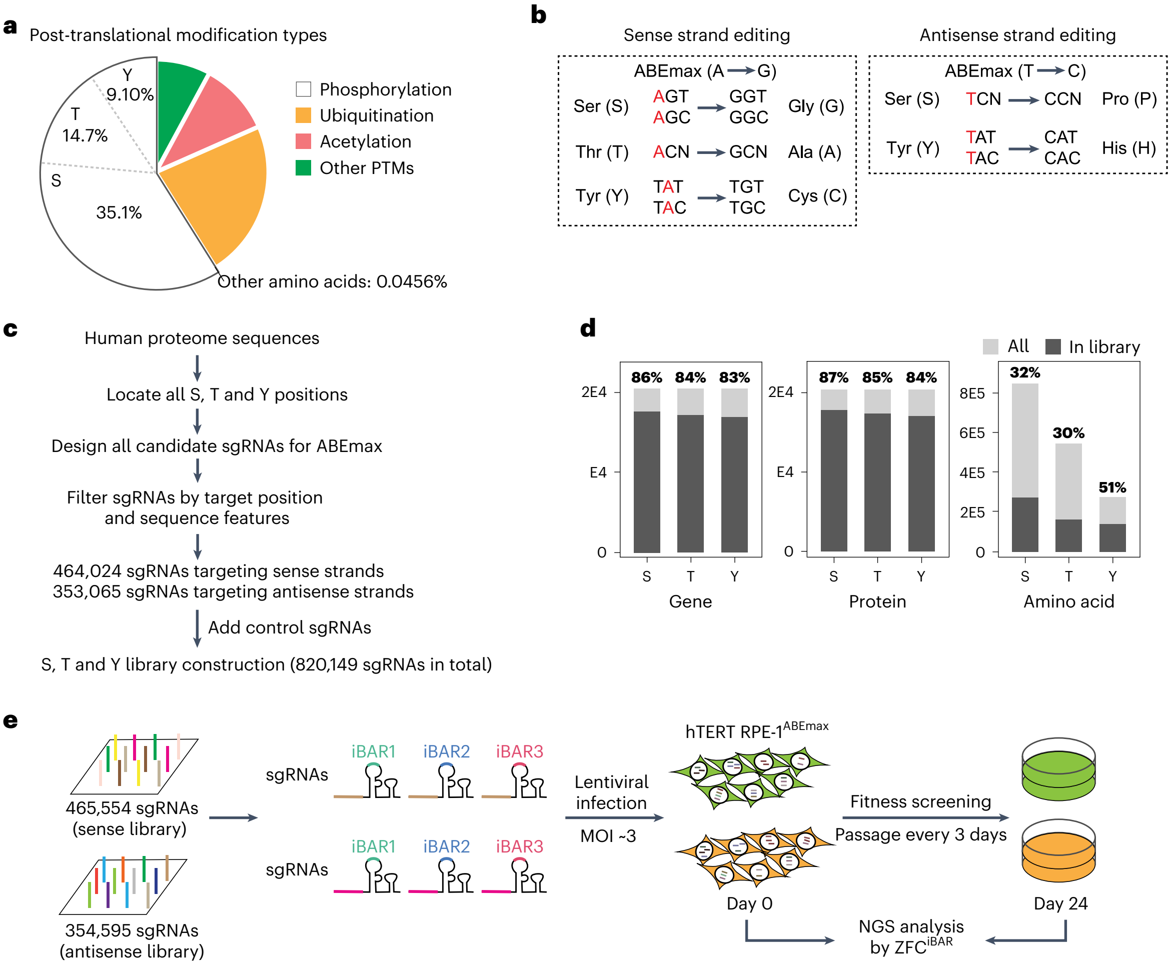

a
Post-translational modification types
Y 9.10%
T 14.7%
S 35.1%
Other amino acids: 0.0456%
Phosphorylation
Ubiquitination
Acetylation
Other PTMs
b
Sense strand editing
ABEmax (A→G)
Ser (S) AGT AGC → GGT GGC Gly (G)
Thr (T) ACN → GCN Ala (A)
Tyr (Y) TAT TAC → TGT TGC Cys (C)
Antisense strand editing
ABEmax (T→C)
Ser (S) TCN → CCN Pro (P)
Tyr (Y) TAT TAC → CAT CAC His (H)
c
Human proteome sequences
Locate all S, T and Y positions
Design all candidate sgRNAs for ABEmax
Filter sgRNAs by target position and sequence features
464,024 sgRNAs targeting sense strands
353,065 sgRNAs targeting antisense strands
Add control sgRNAs
S, T and Y library construction (820,149 sgRNAs in total)
d
All
In library
86% 84% 83%
87% 85% 84%
32% 30% 51%
2E4 E4 0
8E5 6E5 4E5 2E5 0
S T Y
Gene
Protein
Amino acid
e
465,554 sgRNAs (sense library)
354,595 sgRNAs (antisense library)
sgRNAs
iBAR1 iBAR2 iBAR3
Lentiviral infection
MOI ~3
hTERT RPE-1ABEmax
Fitness screening
Passage every 3 days
Day 0
Day 24
NGS analysis by ZFCiBAR

Fig. 1 | Performance of ABE screens targeting S, T and Y residues. a, The proportions of various PTMs recorded in the dbPTM database. Blank pie chart with red frame showing the percentage of phosphorylated amino acids in all types of PTMs. **b**, The outcomes of S, T and Y codon edits by ABE are displayed, depending on whether the sgRNA targets the sense strand or the antisense strand. **c**, The principles for selecting and filtering sgRNAs for the library design. **d**, Histograms illustrating the targeting coverage of annotated genes, proteins or S, T and Y codons in the datasets from UniProt Proteomes in the sgRNA library. **e**, The workflow of cell fitness screens in hTERT RPE1 cells.

impact of perturbing functional S, T and Y on cell fitness. To achieve an unbiased approach, we design a genome-wide iBARed[13] single guide RNA (sgRNA) library, which introduced three barcodes for each sgRNA serving as internal replicates to target all feasible S, T and Y residues in hTERT RPE1 cells. This library included not only phosphorylation sites but also other critical sites affecting fitness. By systematically targeting S, T and Y residues on a genome-wide scale, we aimed to gain functional insights into numerous S, T and Y sites and their roles in various cellular processes, with many representing essential phosphorylation-mediated functions.

Results

ABE-based genome-wide screens for functional S, T and Y sites

Building on our recent efforts to identify genome-wide functional K residues in the hTERT RPE1$^{\mathrm{ABEmax}}$ cell line[14], we developed an sgRNA library designed to target all feasible protein-coding regions of three amino acids (S, T and Y) within the defined six-nucleotide ABEmax window (13–18 nucleotides from the protospacer-adjacent motif (PAM)), resulting in nonsynonymous editing outcomes. By converting A•T to G•C on both sense and antisense strands, the codons for S, T and Y were transformed into G or P, A and C or H, respectively (Fig. 1b). The library comprised 818,619 sgRNAs, targeting 277,051 S, 165,599 T and 141,687 Y sites. Positive controls included 30 sgRNAs targeting splice sites of essential ribosomal genes, while negative controls consisted of 500 sgRNAs targeting the safe locus *AAVS1* and 1,000 nontargeting sgRNAs (Fig. 1c). Covering over 80% of human genes and proteins, the library encompassed 32% of S codons, 30% of T codons and 51% of Y codons (Fig. 1d). To better manage the screening effectively, we constructed two separate libraries targeting the sense strand (465,554 sgRNAs) and the antisense strand (354,595 sgRNAs), each with identical positive and negative controls. These two libraries were individually introduced into hTERT RPE1$^{\mathrm{ABEmax}}$ cells through lentiviral infection at a multiplicity of infection (MOI) of 3, ensuring over 1,500-fold coverage for each sgRNA. Genomic DNA (gDNA) was obtained after 24 days of screening and subjected to next-generation sequencing (NGS) analysis to assess the screening outcomes (Fig. 1e).

The fitness score (FS) for each sgRNA was computed by considering the log fold change (LFC) and the consistency of three associated barcodes using the ZFC$^{\mathrm{iBAR}}$ algorithm[11]. We then applied a robust rank aggregation (RRA) algorithm[15] to assess the significance of sgRNA enrichment in either the positive or the negative direction. With an RRA threshold of <0.001, a total of 3,467 functional residues were identified (Fig. 2a,b). Among these, sgRNAs targeting 309 S, T and Y sites in 161 genes were enriched (Supplementary Table 1 and Supplementary Table 2), while sgRNAs targeting 3,158 sites in 1,618 genes showed depletion (Supplementary Table 3 and Supplementary Table 4). Negative and positive control sgRNAs were used to assess the quality of our large-scale screenings, confirming the anticipated data quality for both libraries (Extended Data Fig. 1a,b). Moreover, we observed a high correlation of sgRNAs targeting the same S, T and Y sites, indicating the robustness of our screenings (Extended Data Fig. 1c,d). Notably, in the sense library, perturbations caused by sgRNAs on Y had more pronounced effects on cell growth compared to those on S and T, implying that a Y-to-C conversion likely exerts a stronger impact on the targeted protein function related to cell proliferation (Fig. 2c).

Our screens successfully identified functional sites in proteins known to impact cell fitness, including proteins encoded by established tumor suppressor genes such as *TP53*, *BRCA2* and *NF1* (ref. 16), as well as essential genes such as *RAF1*, *EIF4A3* and *MTOR* (ref. 11). For example, cells carrying the Y32C substitution in the essential protein KRAS experienced a significant depletion, while the Y71C mutant was enriched. These findings align with the understanding that dynamic phosphorylation on Y32 is essential for maintaining KRAS function in the canonical guanosine triphosphatase cycle and the alteration of Y71 increases the affinity with its downstream factor RAF, thereby activating the KRAS-mediated mitogen-activated protein kinase (MAPK) signaling pathway and conferring survival advantages[17,18]. Notably, we also observed a strong cell growth inhibition effect for the KRAS-Y40H mutant, indicating its potentially critical role in maintaining KRAS function through a different mechanism (Fig. 2d).

To validate the identified functional S, T and Y sites, we initially monitored cell proliferation in RPE1 cells infected with individual sgRNAs from the top-ranking sgRNAs in both libraries (Fig. 2e and Extended Data Fig. 2). We subsequently sequenced the targeted loci to confirm desired editing outcomes with high efficiency, considering potential bystander edits caused by the ABE editing window (Fig. 2f). In total, we validated 126 sgRNAs, including the top 15 and randomly sampled sgRNAs above threshold for both libraries. Of these, 122 exhibited significant and consistent phenotypes with our pool screens (Fig. 2g). In three of 25 depleted NGS cases, it was challenging to capture the amino acid conversion, likely because of the rapid depletion of cells carrying the designated mutations (Extended Data Figs. 3 and 4). However, for the majority of these cases, we confirmed the exact desired editing outcomes with high efficiency. Of note, some cases in our cell proliferation assay showed significant effects on cell fitness but with relatively low editing frequency in NGS. We suspected that such a discrepancy might be because of two main reasons. First, when the targeted base was 13 nucleotides away from the PAM, the overall editing efficiency was relatively low (we used 13–18 nucleotides as the editing window for ABEmax). Second, for the depleted hits, we could only extract the genome from living cells for NGS, causing a discrepancy when a mutation severely leads to cell death without an opportunity to capture the genome.

We further noted that conserved amino acids between paralogs were identified by different sgRNAs, indicating reasonable functional similarity resulting from independent editing events. For example, sgRNA targeting *PRKCD*^{S359} and its paralogous site *PRKCE*^{S418}, located in the kinase domain, were both significantly enriched (Fig. 2h). It has been reported that alteration of phosphorylation at PRKCD-S359 is relevant to the structural determinant of PRKCD catalytic activity[19]. We hypothesized that PRKCE-S418 may perform a similar function by impacting the interaction of downstream factors such as signal transducer and activator of transcription 3, Rho A/C and protein kinase B (Akt), contributing to the behavior of the tumor suppressor[20]. Additionally, we identified 13 pairs of paralogous sites across four kinase groups, implying potential similar mechanisms among kinase proteins (Extended Data Fig. 5). Collectively, these findings demonstrate the high reliability and robustness of our screens and underscore that ABE-based screens have the potential to provide deeper insights into protein function.

Functional S, T and Y profiling depicted novel fitness dependency

Profiling functional S, T and Y enabled us to comprehensively identify a wide range of LOF and gain-of-function (GOF) mutations that affect cell fitness in a gene-specific context. We categorized gene knockout (KO) phenotypes using data from a previously reported CBE-mediated whole-genome gene KO screen called BARBEKO (using RRA < 0.05 to define gene fitness dependency)[11]. Functional S, T and Y substitutions were classified into six groups according to the fitness effects of gene KOs and amino acid substitutions: d^{E}, d^{N}, d^{D}, e^{E}, e^{N} and e^{D}. Here, ‘d’ indicates S, T and Y substitutions leading to cell death or growth inhibition, ‘e’ indicates substitutions promoting cell growth, ‘D’ represents gene KOs leading to cell death or cell growth inhibition, ‘E’ represents gene KOs promoting cell growth and ‘N’ represents gene KOs that do not affect cell growth (referred to as nonfitness genes) (Fig. 3a).

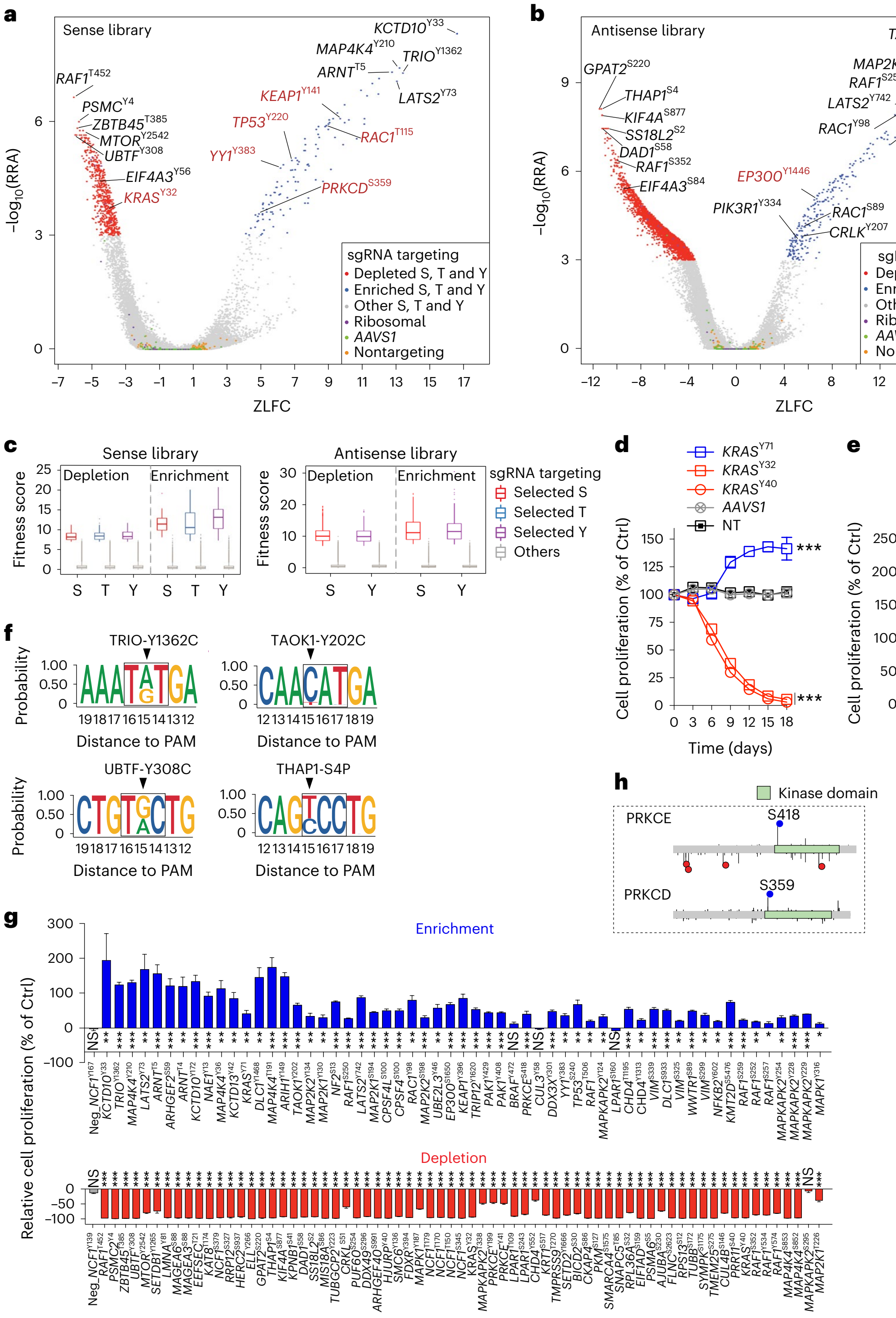
More than half of the mutants exhibited phenotypes resembling the respective gene KOs, indicating LOF because of the disruption of critical amino acids. However, some mutants showed the opposite effect, enhancing gene function. Notably, we identified 747 fitness-irrelevant genes with 1,164 mutations leading to cell growth inhibition (d^{N}) and 91 mutations leading to overproliferation (e^{N}),

Fig. 2 | S, T and Y targeting enables genome-wide screening of the mutations impacting cell growth and proliferation. a,**b**, Volcano plots depicting the phenotype of each sgRNA in the sense library (**a**) and antisense library (**b**) after fitness screens. S, T and Y substitutions above the threshold with depleted and enriched sgRNAs are plotted in red and blue, respectively. Nontargeting (NT) sgRNAs are plotted in orange, sgRNAs targeting the *AAVS1* locus are plotted in green and sgRNAs targeting splice sites of ribosomal genes are plotted in purple. Several top-ranked S, T and Y residues in both directions are labeled in black and several positive sites with known functions for proliferation are labeled in red individually. **c**, Box plots showing the distribution of FSs for sgRNAs targeting selected and nonselected sites in the sense library (top) (depletion, $n = 165$ for selected S, 80,928 for nonselected S, 231 for selected T, 125,782 for nonselected T, 218 for selected Y and 66,121 for nonselected Y; enrichment, $n = 21$ for selected S, 56,596 for nonselected S, 26 for selected T, 88,783 for nonselected T, 62 for selected Y and 45,092 for nonselected Y) and antisense library (bottom) (depletion, $n = 1,954$ for selected S, 135,196 for nonselected S, 719 for selected Y and 63,463 for nonselected Y; enrichment, $n = 134$ for selected S, 99,151 for nonselected S, 97 for selected Y and 52,349 for nonselected Y). The center lines represent the median and the whiskers represent the minimum to maximum values. The boundaries of the box indicate the first and third quantiles. **d**,**e**, Validation examples of selected S, T and Y substitutions by cell proliferation assay using individual sgRNAs in hTERT RPE1^{ABEmax} cells. Enriched sgRNAs are labeled in blue and depleted sgRNAs are labeled in red. The sgRNAs targeting *AAVS1* and NT sgRNAs are labeled in gray and black as negative controls. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to sgRNA targeting *AAVS*1 at the endpoint (day 18), calculated using a one-sided Student’s *t*-test without adjustment ($n = 3$ biologically independent experiments); $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$. Effects of sgRNAs targeting *KRAS*^{Y71}, *KRAS*^{Y32} and *KRAS*^{Y40}, with *KRAS*^{Y71} and *KRAS*^{Y32} shown as positive controls (**d**). Effects of top-ranking sgRNAs targeting *TRIO*^{Y1362} and *UBTF*^{Y308} in sense library and *TAOK1*^{Y202} and *THAP1*^{S4} in the antisense library (**e**). **f**, NGS results showing the editing outcomes of sgRNAs targeting S, T and Y residues in **e**. The targeted codons of S, T and Y are framed, with one additional nucleotide on each side beyond the editing window (13–18 nucleotides from PAM). The targeted nucleotides are indicated by arrows. Samples for NGS sequencing were obtained 3 days after infection (labeled as day 0 in the cell proliferation assay). **g**, Overview of 126 validated sgRNAs. Bar plots show the effect of the indicated sgRNAs on day 18. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to sgRNA targeting *AAVS*1 at the endpoint (day 18), calculated using a one-sided Student’s *t*-test without adjustment ($n = 3$ biologically independent experiments); $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$. NS, not significant. **h**, Schematics showing targeted S, T and Y substitutions in *PRKCE* and *PRKCD*. Negatively selected mutants are labeled in red and positively selected mutants are labeled in blue. Conserved residues with corresponding sgRNAs are indicated by same-colored circles with labeled names.

suggesting novel mechanisms of amino acid substitution that could be hardly observed in gene KO screens (Fig. 3b,c). Of note, these mutations in the N category are challenging to unveil in gene KO screens, implying that these novel sites could potentially serve as a molecular basis for therapeutic interventions in diseases. Similar comparisons were also made to genome-wide gene activation screening[21], revealing many functional sites whose genes have no impact on cell fitness when overexpressed (Extended Data Fig. 6), highlighting the necessity of probing protein function at the amino acid level.

We next investigated the functional S, T and Y sites with respect to pathways by comparing their fitness behaviors to gene KO outcomes (Fig. 3d and Extended Data Fig. 7a,b). Among the enriched pathways, defined by $P < 0.01$, the cell cycle was notably the most enriched in the d^D group, aligning with its well-known close relationship to cell growth and regulation by phosphorylation. Interestingly, for D, E and N groups, the depleted and enriched S, T and Y residue-affected pathways were largely complementary, suggesting that the observed opposite fitness effects arise from more complex regulatory mechanisms rather than simple upregulation or downregulation of specific pathways.

Our screening results also implicated many cancer-related pathways, highlighting their potential regulation at the level of single amino acid substitutions (Extended Data Fig. 7c). For example, members of the MAPK signaling pathway are frequently targeted in drug development because the dysregulation of kinase cascades is often associated with



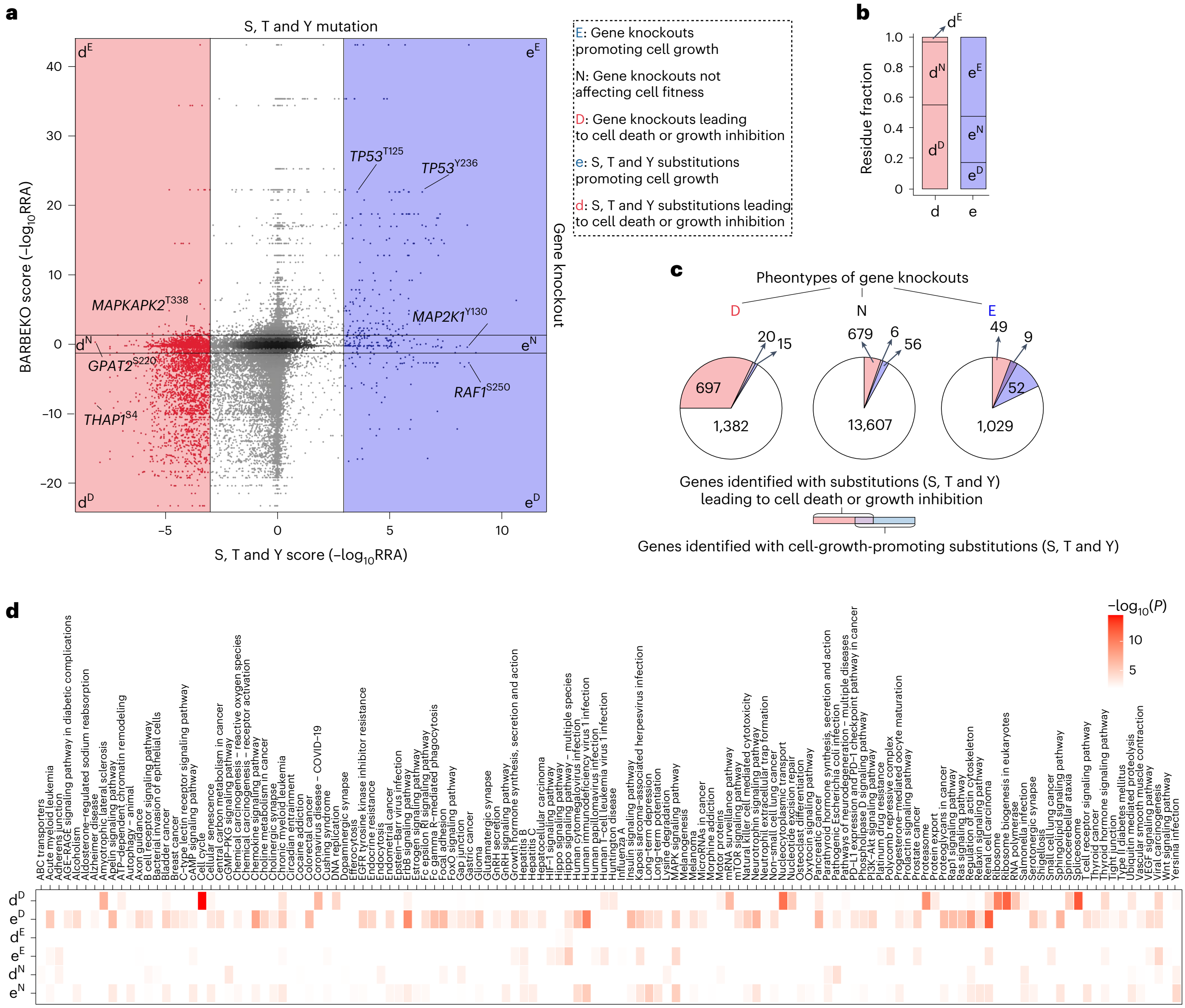


Fig. 3 | Functional S, T and Y profiling for cell fitness in gene context. **a**, Scatter plot showing the distribution of functional mutants RRA in S, T and Y mutagenesis screens and gene RRA in gene KO screening (BARBEKO). Red dots represent S, T and Y sites with depleted sgRNAs and blue dots represent S, T and Y sites with enriched sgRNAs. Amino acids with depleted and enriched sgRNAs are labeled as lowercase 'd' and 'e'. Genes with depleted, maintained and enriched sgRNAs are labeled as superscript 'D', 'N' and 'E'. **b**, Bar chart illustrating the fraction of selected amino acids in each cluster from **a**. **c**, Pie chart indicating the numbers and proportions of genes with positively (blue) or negatively (red) selected mutations in gene KO screens. **d**, Enrichment analysis for d^D, e^D, d^E, e^E, d^N and e^N S, T and Y substitutions involving KEGG pathways, performed using the clusterProfiler R package. Only statistically significant pathways ($P < 0.01$) are shown (Fisher's exact test, one-sided, without adjustment).

tumor development. Unlike traditional drugs that act as whole-protein inhibitors to regulate the activity of target kinase, our screen identified numerous amino acid substitutions with similar functional outcomes. These substitutions hold considerable value for potential drug development, especially for proteins with pockets that have been deemed undruggable so far (Extended Data Fig. 8).

Global insight of massive phosphorylation-mediated GOF

We investigated the correlation between functional S, T and Y sites and their phosphorylation by conducting a mass spectrometry (MS)-based quantitative phosphoproteomic analysis in RPE1 cells to identify phosphorylation sites. The results revealed 65 functional S, T and Y sites with detectable phosphorylation, including four novel phosphorylation sites (Fig. 4a and Supplementary Table 5). Of these, only nine phosphorylation sites had known regulatory functions according to the PhosphoSitePlus database. For instance, RAF1-S259 phosphorylation binds the protein 14-3-3 to regulate RAF1 activity[22] and MAPK1-Y187 phosphorylation persistently activates extracellular signal-regulated kinase (ERK) for cell growth promotion[23,24] (Fig. 4b). Larger-scale validation through a cell proliferation assay confirmed significant phenotypes in cell growth for all evaluated sgRNAs, including VIM^{S339}, VIM^{S325} and PKM^{S217} (Fig. 2g and Supplementary Table 6) and four novel sites ($SETD2^{Y1666}$, $BICD2^{S30}$, $CKAP4^{S86}$ and $DLC1^{S933}$) (Fig. 4c), indicating the causality of the phosphorylation for cell viability at these sites.

Given the highly dynamic nature of phosphorylation and the sensitivity of MS, we cross-referenced our screen results with all known phosphorylation sites in the PhosphoSitePlus database and a predicted functional phosphorylation sites repertoire from a previous study[25].

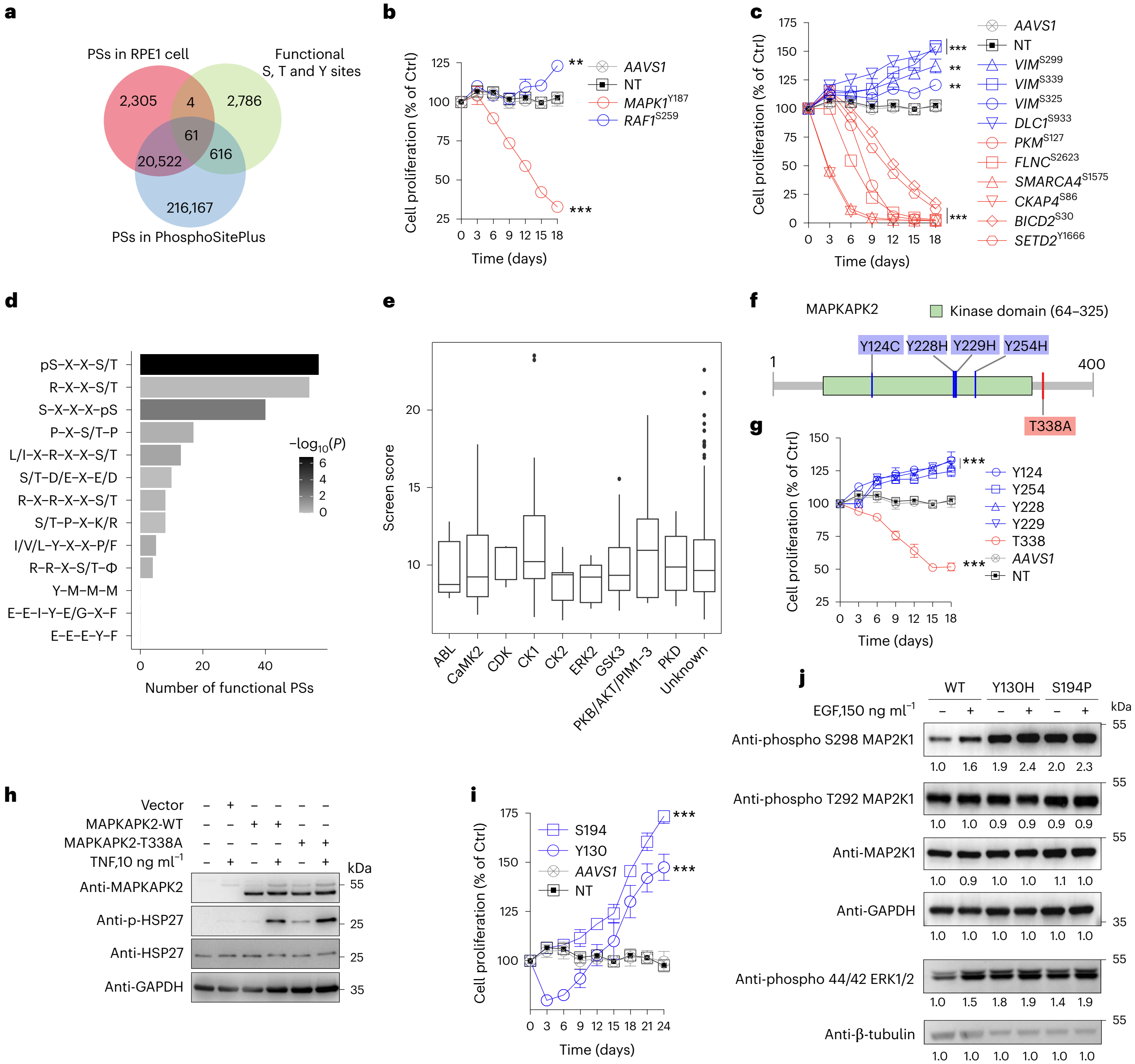


Fig. 4 | Insights into phosphorylation-mediated functional S, T and Y sites. a, Venn diagram showing the numbers of functional S, T and Y sites (green), phosphorylation sites (PSs) in RPE1 cells (pink) and PSs in the PhosphoSitePlus database (blue). **b**, Validation of MAPK1-Y187 and RAF1-S259, two known fitness-affecting PSs. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to sgRNA targeting *AAVS1* at the endpoint (day 18), calculated using a one-sided Student's *t*-test without adjustment (*n* = 3 biologically independent experiments); **P* < 0.05, ***P* < 0.01 and ****P* < 0.001. **c**, Validations of ten selected sgRNAs targeting functional PSs are shown up as examples, including four newly identified PSs in bold. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to sgRNA targeting *AAVS1* at the endpoint (day 18), calculated using a one-sided Student's *t*-test without adjustment (*n* = 3 biologically independent experiments); **P* < 0.05, ***P* < 0.01 and ****P* < 0.001. **d**, Bar plot showing the number and enrichment significance of functional PSs in different types of defined phosphorylation motifs (Fisher's exact test, one-sided, without adjustment). **e**, Box plot showing the distribution of FSs for sgRNAs targeting functional PSs grouped by their upstream kinase family (ABL, *n* = 5; CaMK2, *n* = 35; CDK, *n* = 4; CK1, *n* = 55; CK2, *n* = 10; ERK2, *n* = 11, GSK3, *n* = 39; Akt and PIM1–PIM3, *n* = 9; PKD, *n* = 16; unknown, *n* = 596). The center lines represent the median and the whiskers show the minimum to maximum values. The boundaries of the box indicate the first and third quantiles. **f**, Schematic showing the locations of selected S, T and Y substitutions in MAPKAPK2. **g**, Validation of selected S, T and Y substitutions in MAPKAPK2 by cell proliferation assay. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to sgRNA targeting *AAVS1* at the endpoint (day 18), calculated using a one-sided Student's *t*-test without adjustment (*n* = 3 biologically independent experiments); **P* < 0.05, ***P* < 0.01 and ****P* < 0.001. **h**, Western blot showing the effect of MAPKAPK2-T338A on MAPKAPK2 and Hsp27 expression through cDNA rescue under tumor necrosis factor treatment. The experiment was repeated independently three times with similar results. **i**, Validation of sgRNAs targeting *MAP2K1*^{Y130} and *MAP2K1*^{S194} by cell proliferation assay. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to sgRNA targeting *AAVS1* at the endpoint (day 18), calculated using a one-sided Student's *t*-test without adjustment (*n* = 3 biologically independent experiments); **P* < 0.05, ***P* < 0.01 and ****P* < 0.001. **j**, Western blot showing phosphorylation cascades in cell clones of MAP2K1 mutants (Y130H and S194P) in hTERT RPE1 cells upon EGF treatment. The experiment was repeated independently three times with similar results.

The 677 overlapping sites (Fig. 4a and Supplementary Table 7) indicated a potentially more widespread phosphorylation-dependent fitness effect worth further verification in various cell contexts and states. We also observed significantly higher functional scores from prediction for S, T and Y sites above our screening threshold, suggesting a close functional association between the prediction and our results (Extended Data Fig. 9a,b).

To globally understand the features of phosphorylation sites affecting cell fitness, we classified the 677 functional phosphorylation sites on the basis of their motif patterns in the surrounding sequence. Among well-defined consensus phosphorylation motifs[26], the CK1 motif (pS/pT-x-x-S/T) was the most significantly enriched (Fig. 4d), indicating a critical role for phosphorylation sites in such sequence contexts. We also classified them according to their kinase families and found that substrates of Akt and proviral integration site kinase (PIM kinases) families exhibited a more prominent fitness effect (Fig. 4e). Of note, as phosphorylation sites with a known upstream factor often receive more attention in traditional phosphoproteomic studies, yet only about 3% of detected phosphorylation sites have a known kinase so far[27], our results provide a more unbiased identification of functional phosphorylation sites, regardless of their upstream kinase. The numerous functional phosphorylation sites with unknown kinase should not be overlooked in further research.

Having identified a large number of mutations showing distinct phonotypes compared to gene KO (Fig. 3a), we aimed to clarify part of the GOF effect through phosphorylation modulation. We focused on T338 in MAPKAPK2, critical in many physiological contexts. While *MAPKAPK2* KO promotes cell proliferation[11], the T338A substitution caused striking cell growth inhibition (Fig. 4f,g), suggesting T338 as a GOF site compared to the gene itself. We identified that T338A enhances downstream factor Hsp27 phosphorylation through complementary DNA (cDNA) rescue in *MAPKAPK2* KO cells, indicating that T338A increases MAPKAPK2 kinase activity, potentially leading to cell growth suppression (Fig. 4h). Additionally, we investigated whether other mutants mimic gene KO by regulating phosphorylation. Y124C significantly reduced MAPKAPK2 protein abundance, while Y254H in the kinase domain decreased Hsp27 phosphorylation (Extended Data Fig. 9c,d).

We further suspected that some GOF sites influence phosphorylation levels of both downstream factors and selfproteins. Focusing on MAP2K1, an essential component of the MAPK signaling pathway, we found that although *MAP2K1* KO had no fitness effect, two positively selected substitutions (Y130H and S194P) validated by cell proliferation and NGS (Fig. 4i and Extended Data Fig. 9e) significantly improved the phosphorylation level of MAP2K1 on S298, even without epidermal growth factor (EGF) stimulation (Fig. 4j). As MAP2K1 activates ERK1 and ERK2, we also evaluated ERK1 and ERK2 phosphorylation, which was dramatically upregulated in the mutants compared to the wild type (Fig. 4j). The S194 and Y130 substitutions, located in the kinase domain, likely altered kinase activity. Together, our screen results revealed numerous phosphorylation-mediated functional substitutions, indicating that perturbation of these sites could critically control cell life-and-death decisions.

Functional S, T and Y sites have a critical role for protein structure

Our comprehensive genome-wide screening targeting S, T and Y sites also allowed us to explore the phosphorylation-independent functions of these three amino acids. Amino acid substitution can alter protein structure, stability, cellular localization or interactions. Because structure changes are the most universal and direct effect of amino acid substitution, we initially mapped our 3,467 functional sites to annotated structural regions in the proteome sequence (Fig. 5a). The results showed that about one third of these sites were located in domain regions, potentially disrupting domain structures. The substantial number of functional sites outside protein function domains indicated the existence of critical sites in other regions with ambiguous function, such as linkers, coiled coils and tandem repeats, possibly leading to abnormal protein folding. Interestingly, 32 functional sites were identified in protein intrinsically disordered regions (IDRs). Given that some studies have reported that IDRs regulate proteins by mediating phase separation[28], we suspected that such processes might be disrupted by specific amino acid substitutions.

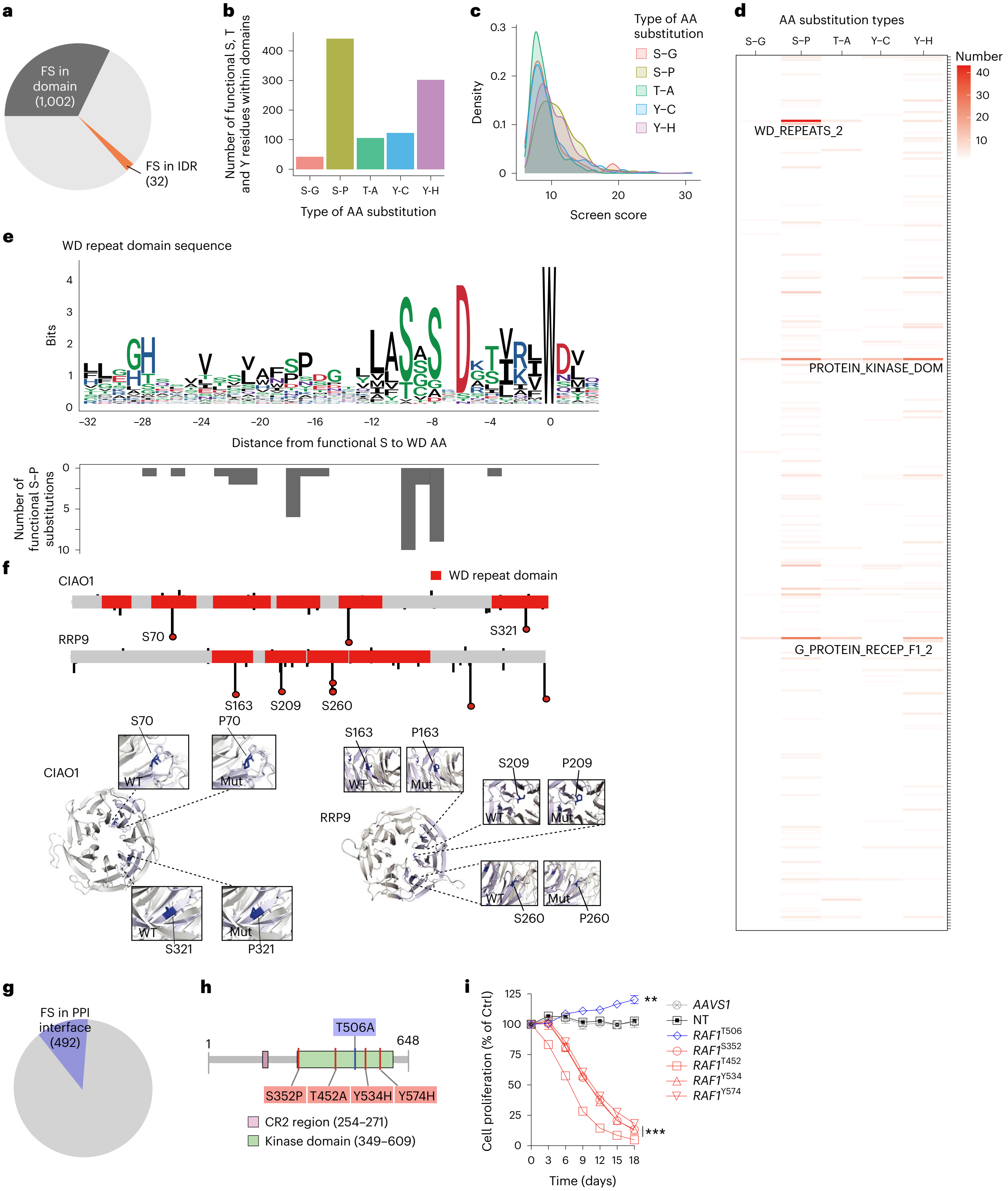
We then focused on functional substitutions within domain regions, as these are the most well-studied and closely related to protein function. Using ABE editing, S, T and Y residues could be replaced by specific amino acids with various physicochemical properties. For example, serine-to-proline mutants have a high likelihood to cause structural defects, tyrosine-to-cysteine mutants increase the probability of oxidation and disulfide bond formation and tyrosine-to-histidine mutants, while preserving aromaticity, introduce a partial positive charge. We classified the 1,002 within-domain functional sites by their substitution types and found that serine-to-proline substitutions accounted for the largest proportion and strongest fitness effect, followed by tyrosine-to-histidine substitution (Fig. 5b,c). To identify which domains were most impacted, we calculated the number of functional S, T and Y residues within all known domains across various substitution types and performed statistical tests to assess the enrichment of these functional S, T and Y residues relative to those designed in the library (Fig. 5d and Extended Data Fig. 9f). The analysis indicated that 1,034 functional S, T and Y residues were located in 301 kinds of domains according to the PROSITE database, with serine-to-proline substitutions being the most widespread, suggesting a global disruption of domain structures.

Among the involved domains, the kinase domain, WD repeat domain and the G-protein-coupled receptor domain contained more functional S, T and Y residues. The results were expected for the kinase domain and G-protein-coupled receptor domain, as the kinase domain is often regulated by selfphosphorylation sites[27] and the G-protein-coupled receptor domain is well known for phosphorylation signaling transduction[29]. We focused on the WD repeat domain, which had the largest number of functional serine-to-proline substitutions, and found that two subtypes of the WD repeat domain were significantly enriched (Extended Data Fig. 9f).

There are 349 known WD repeat-containing proteins that regulate important processes such as signal transduction, apoptosis and RNA

Fig. 5 | Role of functional S, T and Y sites in maintaining protein structure. **a**, The numbers of functional S, T and Y residues located within protein domain regions and IDRs. **b**, Histogram showing the number of functional S, T and Y residues within domain regions, grouped by substitution types. **c**, Distributions of screen scores across different types of functional amino acid substitutions. **d**, Heatmap showing the number of functional amino acid substitutions in various domain types from the PROSITE database. **e**, Sequence logo showing amino acid conservation preference of WD repeat domains with functional serine-to-proline residues (top) and the relative positions of functional serine-to-proline residues to WD in the C terminus (bottom). **f**, Targeted and selected sites in CIAO1 and RRP9. Black bars represent locations with a designed sgRNA and bars with red dots represent selected functional serine-to-proline substitutions in WD repeat domains. Examples of CIAO1-S70P, CIAO-S321P, RRP9-S163P and RRP9-S260P in protein 3D structure are shown (Protein Data Bank (PDB) 3FM0 for CIAO1 and PDB 4J0W for RRP9). WT, wild type. **g**, The numbers of functional S, T and Y residues located within PPI interface regions according to the Interactome INSIDER database. **h**, Schematic showing the locations of selected S, T and Y substitutions in the RAF1 kinase domain. **i**, Validation of sgRNAs targeting *RAF1*^{T506} and other depleted sites in the RAF1 kinase domain by cell proliferation assay. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to sgRNA targeting *AAVS1* at the endpoint (day 18), calculated using a one-sided Student's *t*-test without adjustment (*n* = 3 biologically independent experiments); **P* < 0.05, ***P* < 0.01 and ****P* < 0.001.

processing, acting as adaptors for protein–protein or protein–DNA interactions[30]. On a one-dimensional sequence, the WD repeat domain features two amino acids 'WD' at the C terminus, with some degree of conservation in other positions, appearing as tandem repeats. Despite low conservation in sequence, the WD repeat domain possesses high conservation in three-dimensional (3D) structure, with each unit forming a β-sheet arranged as the blades of a propeller around a central cavity, indicating structure-dependent function. Previous studies identified several residues critical for maintaining the WD repeat structure (V14, V17, F19, L26, A27, I35 and V37)[30]. We scanned all 43 functional serine-to-proline substitutions and found that all led to cell growth inhibition, consistent with most WD repeat-containing proteins being



essential (Fig. 5e). These substitutions were predominantly enriched at the eighth and tenth amino acids upstream from the WD motif, which showed high conservation of serine and were located at the top of the β-sheet, surrounding the central cavity in the 3D structure (Fig. 5f). These observations suggested that S28 and S30 might be novel critical sites for maintaining the WD repeat domain structure, highlighting that perturbation at the amino acid level provides subtle insights into domain function.

Single amino acid substitution can affect selfprotein function in various ways. Identifying functional residues for protein–protein interaction (PPI) interfaces is crucial for achieving PPI regulation. We mapped our functional S, T and Y residues to the Interactome INSIDER database, which includes experimental and predicted PPI interface regions. To our surprise, 492 of the 3,467 functional S, T and Y residues were located in PPI interface regions with statistical significance (Fisher's exact test, P value = 3.43×10^{-28}) (Fig. 5g), which would not be detected by gene KO screening. We further suspected that enhancing binding affinity with interacting proteins might contribute to some of the GOF effects. For example, in the essential kinase RAF1, we noticed that, while most selected sites in its kinase domain behaved destructively to kinase activity, T506A had an opposite effect on cell growth (Fig. 5h,i). The function of RAF1-T506 was unclear, except for one study suggesting that it might be a putative phosphorylation site based on its high evolutionary conservation but showing no evidence of regulating RAF1 activity[31]. We found that T506, located at the interface of RAF1 interacting with MAP2K1 and MAP2K2, might function by altering PPI affinity with these proteins, thus regulating cell proliferation. Together, we demonstrated the enrichment of functional sites on protein interfaces, indicating their crucial role in regulating PPI for cell survival.

Enriched mutations behaved as potential cancer drivers

After screening the normal RPE1 cell line, we aimed to investigate the link between substitutions that cause cell overproliferation and known cancer drivers. To do this, we cross-referenced clinical sequence data from cancer patients in the International Cancer Genome Consortium (ICGC) with all 309 positively selected mutants identified in our screens. Among these mutants, 71 sites exhibited missense mutations detected in ICGC across 34 tumor types, with seven of them having known associations with tumorigenesis (Fig. 6a). Our findings implicated several cancer driver proteins, notably the critical tumor suppressor protein p53, where seven substitutions were located in its DNA-binding domain and demonstrated dominant effects (Fig. 6b). For instance, TP53-Y220C was validated as a key missense mutation leading to a functionally inactive conformation, directly contributing to cancer formation[32]. Additionally, Y126 was identified as a phosphorylation site that positively affects p53 ISGylation to control its structure stability[33]. Concurrently, S240, representing the other five mutants, was successfully validated by individual sgRNA (Fig. 6c).

To further explore whether these positively selected mutants could be potential cancer driver mutations, we validated additional cases using individual sgRNAs, all of which significantly promoted cell growth (Fig. 6c). We then focused on MAP4K4-Y210, which exhibited enhanced cell growth in individual sgRNA perturbation and ranked among the top three in the sense library. We established two cell clones of *MAP4K4*$^{\mathrm{Y210C}}$ in RPE1 cells for further characterization, confirming the absence of bystander mutations in the editing window (Fig. 6d). As expected, the clones showed a progressive advantage in cell growth compared to wild-type cells (Fig. 6e). Additionally, a clonogenic assay revealed that the Y210C substitution had a stronger capability for clone formation than the wild type (Fig. 6f).

We then evaluated the role of MAP4K4-Y210 in tumor development using a xenograft model with hTERT RPE1$^{\mathrm{ABEmax}}$ cells. Injection of MAP4K4-Y210C mutant cells led to tumor formation within 10 days, with steady growth until day 24 in six mice, after which three were killed on day 28 because of tumor shrinkage. Another three tumors dissipated in the experimental group after 35 days, while all *AAVS1*-targeting RPE1 cells failed to result in tumor within the detection period (eight mice for control group and eight mice for experimental group; Extended Data Fig. 9g,h). Although MAP4K4-Y210C was not sufficient for typical malignant tumors, the initial tumor development in vivo indicated that the substitution was required for tumor development. Lastly, as the *MAP4K4*$^{\mathrm{Y210}}$ mutation has been identified in melanoma patients, we tested the growth advantage of this mutation in human melanoma A375 cells. The pooled sgRNA abundance dramatically increased, suggesting that the Y210C substitution is a potential cancer driver for melanoma (Fig. 6g).

In summary, we identified 309 S, T and Y substitutions in 161 genes with the potential to initiate tumor formation, including 25 annotated cancer driver genes in the Cancer Gene Census[16] and 136 novel genes with potential driver mutations (Fig. 6h), providing a valuable resource for the study of tumor initiation.

Functional effects of selected residues in other cell lines

Considering the size of our library, we performed our screens in a single cell line at this stage. We acknowledge that functional dependencies observed in the RPE1 cell line may not be fully generalizable to other cell lines. To address this, we validated the top hits (14 depleted mutations and 12 enriched mutations) in both A375 and SW620 cell lines (Extended Data Fig. 10). In total, 12 of 14 depleted mutations and 2 of 12 enriched mutations showed consistent cell fitness phenotypes across cell lines, indicating the reliability of our results and their potential extrapolation. The cell-specific functional mutations may represent distinct regulatory mechanisms in different cell types, which is reasonable because of the heterogeneity of cell lines in context of the genome, expression and PTMs.

Discussion

In our investigation, we used ABEmax to systematically screen S, T and Y sites across the entire genome, aiming to provide an unbiased functional map of these three phosphorylable amino acids. Given the crucial role of protein phosphorylation in signal transduction, it is unsurprising that protein kinases and their phosphorylated substrates have been extensively studied. While phosphoproteomics studies have identified numerous phosphorylation sites, some potential sites may be difficult to detect because of their low abundance or transient nature. Our high-throughput screening, targeting genome-wide S, T and Y amino acids closely associated with phosphorylation, enabled us to characterize thousands of subtle regulatory units influencing cell growth at an amino acid resolution, some of which represent phosphorylation-mediated critical functions. Although we identified only 677 functional S, T and Y sites as phosphorylation sites, we propose that many other functional sites, while not all, may influence cell fitness through phosphorylation.

We acknowledge that other effects caused by editing might interfere with the interpretation of the causality between functional S, T and Y sites and phosphorylation. For instance, some substitutions with potential severe structural change, especially serine to proline, could be involved because of the nature of ABE editing. However, considering that most serine-to-proline substitutions in our library had no effect on cell fitness and phosphorylation sites are often located in flexible protein regions with relatively high tolerance for structural changes[34], we believe that this may not inherently disrupt for large-scale screens but could be problematic for some individual cases. To validate the direct impact by phosphorylation, we suggest using Prime Editor (PE) to substitute functional phosphorylation sites to Ala, which is known to have a minimal impact on structure.

Another limitation of our phosphorylation-associated strategy is its inability to investigate proteins shaped by multisite phosphorylation, such as Elk-1 (ref. 35), because we only assessed the effect

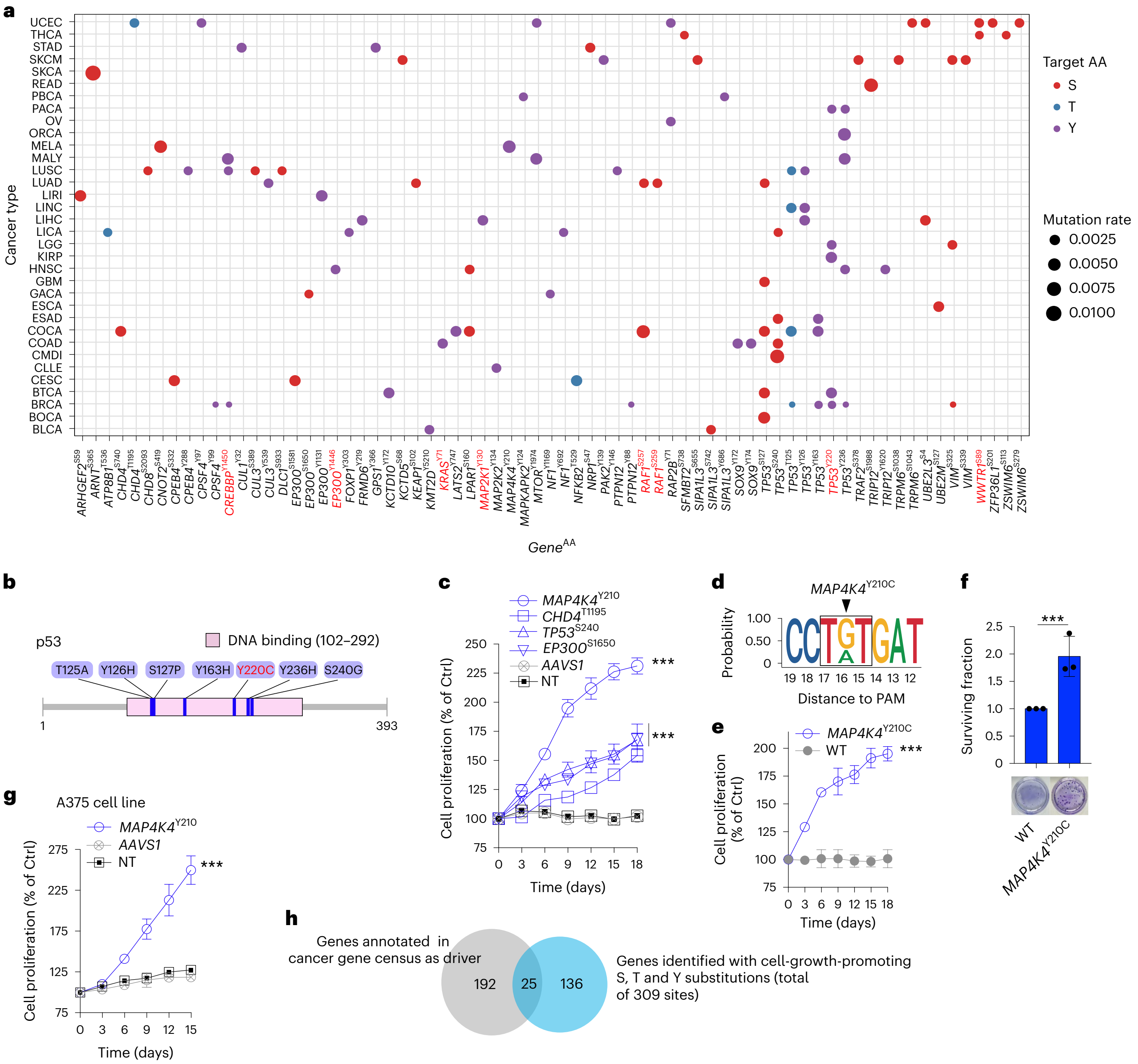


Fig. 6 | Positively selected S, T and Y substitutions showed strong clinical relevance and acted as potential cancer drivers. a, Schematic showing all 309 positively selected sites with identified mutations in the ICGC database. The *y* axis denotes different cancer types defined in ICGC and the red labels on the *x* axis denote amino acids with known oncogenic substitutions. The dot size represents the detected missense mutation rate for each amino acid. Sites in red represent known cancer driver mutations. **b**, Schematic showing p53 with selected S, T and Y substitutions located in the DNA-binding region. Substitutions with enriched phenotypes are indicated in blue. Y220, a known mutant contributing to cell growth, is listed in red. **c**, Validation of MAP4K4-Y210, CHD4-T1195, TP53-S240 and EP300-S1650 from **a** by cell proliferation assay. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to sgRNA targeting *AAVS1* at the endpoint (day 18), calculated using a one-sided Student's *t*-test without adjustment (*n* = 3 biologically independent experiments); **P* < 0.05, ***P* < 0.01 and ****P* < 0.001. **d**, NGS results showing the editing outcomes of sgRNA targeting MAP4K4-Y210 in hTERT RPE1^{ABEmax} cells. **e**, Cell proliferation assay of MAP4K4-Y210C cell clone in hTERT RPE1^{ABEmax}. The WT clone is labeled in gray and the MAP4K4-Y210C clone is labeled in blue. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to the WT clone at the endpoint (day 18), calculated using a one-sided Student's *t*-test without adjustment (*n* = 3 biologically independent experiments); **P* < 0.05, ***P* < 0.01 and ****P* < 0.001. **f**, Bar chart displaying the survival fraction of MAP4K4-Y210C cell clone compared to the WT. Fractions are presented as the mean ± s.d. of three independent experiments; **P* < 0.05, ***P* < 0.01 and ****P* < 0.001. **g**, Validation of sgRNA targeting MAP4K4-Y210 in A375 ^{ABEmax} cells. Data are presented as the mean ± s.d. of three independent experiments. *P* values represent comparisons to sgRNA targeting *AAVS1* at the endpoint (day 18), calculated using a one-sided Student's *t*-test without adjustment (*n* = 3 biologically independent experiments); **P* < 0.05, ***P* < 0.01 and ****P* < 0.001. **h**, Venn diagram showing the numbers of known cancer driver genes and genes with cell-growth-promoting S, T and Y substitutions in our screens.

of single substitutions in the screenings. To address this, tools such as paired sgRNA or PE could be used to achieve multisite editing or even large fragment replacement for multisite phosphorylation disruption.

Given the genome-wide scale of our strategy, we also identified a notable number of S, T and Y sites functioning in a phosphorylation-independent manner. To gain insights into this, we introduced structural considerations to interpret our results. While some substitutions, such as serine to glycine or threonine to alanine, are structurally innocuous, others, specifically serine to proline and tyrosine to histidine, greatly disrupt domain structure. If these effects were confirmed more universally, for example, by substituting other amino acids such as leucine to proline and verifying notable protein destruction through traditional gene KO strategies, it could present a broadly applicable alternative approach. This is relevant because more amino acids can be edited to proline than accessible splice sites.

Remarkably, most functional sites identified exhibited a GOF effect on cell fitness compared to gene context, a challenging outcome to achieve in gene KO screens. Our screens also revealed numerous critical amino acids contributing to the protein's intrinsic function, indicative of LOF. The correspondence between novel amino acids and phenotypes suggests potential strategies for precise cancer treatments. For instance, our data could support designing small-molecule inhibitors targeting substituted S, T and Y residues in 'undruggable' oncogenes.

In conclusion, the mutagenesis of endogenous S, T and Y sites enabled the study of amino acid function at endogenous loci. We anticipate that this work will provide a comprehensive resource on a mechanistic basis, expanding our understanding of tumor progression and guiding the development of novel therapies.

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/s41589-024-01731-0.

References

1. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
2. Woodsmith, J. & Stelzl, U. Studying post-translational modifications with protein interaction networks. *Curr. Opin. Struct. Biol.* **24**, 34–44 (2014).
3. Iqbal, S. et al. Comprehensive characterization of amino acid positions in protein structures reveals molecular effect of missense variants. *Proc. Natl Acad. Sci. USA* **117**, 28201–28211 (2020).
4. Ardito, F., Giuliani, M., Perrone, D., Troiano, G. & Lo Muzio, L. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (review). *Int. J. Mol. Med.* **40**, 271–280 (2017).
5. Viéitez, C. et al. High-throughput functional characterization of protein phosphorylation sites in yeast. *Nat. Biotechnol.* **40**, 382–390 (2022).
6. Bowden, A. R. et al. Parallel CRISPR–Cas9 screens clarify impacts of p53 on screen performance. *eLife* **9**, e55325 (2020).
7. Haapaniemi, E., Botla, S., Persson, J., Schmierer, B. & Taipale, J. CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. *Nat. Med.* **24**, 927–930 (2018).
8. 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).
9. Cuella-Martin, R. et al. Functional interrogation of DNA damage response variants with base editing screens. *Cell* **184**, 1081–1097 (2021).
10. Hanna, R. E. et al. Massively parallel assessment of human variants with base editor screens. *Cell* **184**, 1064–1080 (2021).
11. Xu, P. et al. Genome-wide interrogation of gene functions through base editor screens empowered by barcoded sgRNAs. *Nat. Biotechnol.* **39**, 1403–1413 (2021).
12. Koblan, L. W. et al. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat. Biotechnol.* **36**, 843–846 (2018).
13. Zhu, S. et al. Guide RNAs with embedded barcodes boost CRISPR-pooled screens. *Genome Biol.* **20**, 20 (2019).
14. Bao, Y. et al. Unbiased interrogation of functional lysine residues in human proteome. *Mol. Cell* **83**, 4614–4632 (2023).
15. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
16. Martinez-Jimenez, F. et al. A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* **20**, 555–572 (2020).
17. Kano, Y. et al. Tyrosyl phosphorylation of KRAS stalls GTPase cycle via alteration of switch I and II conformation. *Nat. Commun.* **10**, 224 (2019).
18. Cirstea, I. C. et al. Diverging gain-of-function mechanisms of two novel *KRAS* mutations associated with Noonan and cardio-facio-cutaneous syndromes. *Hum. Mol. Genet.* **22**, 262–270 (2013).
19. Gong, J. et al. The C2 domain and altered ATP-binding loop phosphorylation at Ser^{359} mediate the redox-dependent increase in protein kinase C-δ activity. *Mol. Cell. Biol.* **35**, 1727–1740 (2015).
20. Newton, A. C. & Brognard, J. Reversing the paradigm: protein kinase C as a tumor suppressor. *Trends Pharmacol. Sci.* **38**, 438–447 (2017).
21. Gilbert, L. A. et al. Genome-scale CRISPR-mediated control of gene repression and activation. *Cell* **159**, 647–661 (2014).
22. Dumaz, N. & Marais, R. Protein kinase A blocks Raf-1 activity by stimulating 14-3-3 binding and blocking Raf-1 interaction with Ras. *J. Biol. Chem.* **278**, 29819–29823 (2003).
23. Grethe, S. & Porn-Ares, M. I. p38 MAPK regulates phosphorylation of Bad via PP2A-dependent suppression of the MEK1/2–ERK1/2 survival pathway in TNF-α induced endothelial apoptosis. *Cell Signal* **18**, 531–540 (2006).
24. Lavoie, H., Gagnon, J. & Therrien, M. ERK signalling: a master regulator of cell behaviour, life and fate. *Nat. Rev. Mol. Cell Biol.* **21**, 607–632 (2020).
25. Ochoa, D. et al. The functional landscape of the human phosphoproteome. *Nat. Biotechnol.* **38**, 365–373 (2020).
26. Ubersax, J. A. & Ferrell, J. E. Jr. Mechanisms of specificity in protein phosphorylation. *Nat. Rev. Mol. Cell Biol.* **8**, 530–541 (2007).
27. Needham, E. J., Parker, B. L., Burykin, T., James, D. E. & Humphrey, S. J. Illuminating the dark phosphoproteome. *Sci. Signal* **12**, eaau8645 (2019).
28. Martin, E. W. & Holehouse, A. S. Intrinsically disordered protein regions and phase separation: sequence determinants of assembly or lack thereof. *Emerg. Top. Life Sci.* **4**, 307–329 (2020).
29. Tobin, A. B. G-protein-coupled receptor phosphorylation: where, when and by whom. *Br. J. Pharmacol.* **153**, S167–S176 (2008).
30. Xu, C. & Min, J. Structure and function of WD40 domain proteins. *Protein Cell* **2**, 202–214 (2011).
31. Chong, H., Lee, J. & Guan, K. L. Positive and negative regulation of Raf kinase activity and function by phosphorylation. *EMBO J.* **20**, 3716–3727 (2001).

32. Sundar, D. et al. Wild type p53 function in p53^{Y220C} mutant harboring cells by treatment with Ashwagandha derived anticancer withanolides: bioinformatics and experimental evidence. *J. Exp. Clin. Cancer Res.* **38**, 103 (2019).
33. Huang, Y. F. & Bulavin, D. V. Oncogene-mediated regulation of p53 ISGylation and functions. *Oncotarget* **5**, 5808–5818 (2014).
34. Liu, N., Guo, Y., Ning, S. & Duan, M. Phosphorylation regulates the binding of intrinsically disordered proteins via a flexible conformation selection mechanism. *Commun. Chem.* **3**, 123 (2020).
35. Mylona, A. et al. Opposing effects of Elk-1 multisite phosphorylation shape its response to ERK activation. *Science* **354**, 233–237 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Cell culture

HEK293T cells were from C. Zhang's laboratory (Peking University) and A375 cells were purchased from the American Type Culture Collection (CRL-1619). Both cell lines were cultured in DMEM (Corning). hTERT RPE1 cells were from Y. Sun's laboratory (Peking University) and were cultured in DMEM/F12 medium (Gibco). All mentioned media were supplemented with 10% FBS (Biological Industries) and 1% penicillin–streptomycin. All cells were cultured with 5% CO_2 at 37 °C with the confirmation of the absence of *Mycoplasma* contamination.

sgRNA design

The reference data were obtained from UniProt Proteomes (UP000005640.fasta released on November 17, 2018) and the Illumina iGenome (hg38) website. First, we captured all positions of S, T and Y residues from the UniProt protein sequence file using a Python script. Second, all positions of the three types of amino acids were mapped to genomic locations. We selected nonspliced S, T and Y sites (where the triplet genetic codon locates in one exon) because editing spliced S, T and Y sites (where the triplet genetic codon is separated into two exons) may cause gene KO instead of a single amino acid change. Third, sgRNAs for the S, T and Y sites were generated from the S, T and Y genome position with NGG (GG preceded by any nucleotides) PAM nearby. According to our experiment results and the conclusion from a previous study[36], with single targeting adenine at the 14th, 15th and 16th nucleotides to PAM, the sgRNA length should be 19 nt to ensure a high base-editing efficiency. With single targeting adenine at the 13th or 17th nucleotides, the sgRNA length should be 20 nt. For single targeting adenine at the 18th nucleotide, the sgRNA length should be extended to 21 nt. The whole-genome-wide S, T and Y library was separated into a sense library, as targeted by adenine in the coding strand leading to A to G in codons, and an antisense library, as targeted by adenine in the noncoding strand leading to T to C in codons. Then, we removed sgRNAs with more than ten hits across the whole genome using Bowtie, an ultrafast memory-efficient short read aligner[37], and also filtered sgRNAs with 'TTTT' or extreme G+C content (less than 0.2 or higher than 0.8). Last, redundant sgRNAs for single S, T and Y sites were removed. The library also contained 1,000 nontargeting sgRNAs and 500 sgRNAs targeting the *AAVS1* safe harbor locus (chr19: 55113873–55117983 in human hg38 assembly) as negative controls and 30 sgRNAs targeting splice sites of ten essential ribosomal genes causing gene KO as positive controls.

Plasmid construction

The pLenti-ABEmax-GFP plasmid was constructed by amplifying ABEmax from pCMV-ABEmax-P2A-GFP (Addgene, 112101). Other pLenti-cDNA-3×Flag-SV40-mCherry and pLenti-CMV-cDNA-3×Flag-IRES-GFP plasmids were constructed through Gibson Assembly (New England Biolabs, E2611L) based on cDNA sequences in the National Center for Biotechnology Information (NCBI) database. The cDNA-expressing plasmids were constructed by inserting each cDNA sequence into the multiple cloning sites before the Flag tag of the pLenti-SV40-mCherry vector following the standard cloning protocol.

Construction of the ABEmax sgRNA^{iBAR} plasmid library

The sgRNA oligonucleotides were synthesized in array by Synbio Technologies and amplified by PCR with primers that included the BsmBI recognition site at the 5′ end from chip (F primer, CGCTCCGTGAACAGTAATTAGGTG; R primer, CAGGAATCGTGATACCATGCGTTG). After purification with a DNA Clean and Concentrator-25 kit (Zymo Research Corporation, D4034), the purified PCR products were respectively inserted into the three sgRNA^{iBAR}-expressing backbones constructed above through the Golden Gate cloning strategy (iBAR1, CTCGCT; iBAR2, GATGGT; iBAR3, GCACTG). The ligation mixture of each group was separately purified with a DNA Clean & Concentrator-5 kit (Zymo Research Corporation, D4014) and was electrotransformed into *Escherichia coli* HST08 Premium Electro-Cells (Takara, 9028) according to the manufacturer's protocol using a Gene Pulser Xcell (Bio-Rad). The plasmid of each sgRNA^{iBAR} library was extracted using an EndoFree Plasmid Maxi kit (Qiagen, 12362) and further mixed in a 1:1:1 molar ratio.

Primers of PCR amplification

The gDNA as templates could be PCR-amplified by KAPA HiFi HotStart ReadyMixPCR Kit (KAPABIOSYSTEMS, KK2602) with 26 cycles of reactions using five pairs of primers (Supplementary Table 8). Up to 6 µg of gDNA coud be used in one 100-µl PCR reaction and the number of PCR reactions was determined by the amount of gDNA extracted from library cells.

Production and infection of lentivirus

HEK293T cells were seeded 24 h before lentivirus packaging. The lentivirus library was produced by cotransfection of library plasmids with pVSVG and pR8.74 (Addgene, 12259) into HEK293T cells using the X-tremeGENE HP DNA transfection reagent (Roche) according to the manufacturer's instructions. The cell supernatant containing lentivirus was collected 48 h after transfection. The virus for cDNA overexpression was concentrated using the Lenti-X Concentrator (Clontech, 631232). For cell infection, the virus was added into culturing medium with polybrene. After 24 h, the medium was changed to conventional medium.

High-throughput functional S, T and Y screenings in RPE1 cells

The hTERT RPE1^{ABEmax} cells were seeded 24 h after lentiviral infection and were further infected with the library lentivirus at an MOI of 3 with high coverage (at least 1,500-fold) for each sgRNA. We used a high MOI because of the large size of our library; moreover, it made our library more operable. Benefiting from the iBAR strategy, we could obtain high-quality results with a minor impact of the potential free-rider effect. Then, 1 day after lentiviral infection, the library cells were subjected to puromycin selection for 48 h (20 µg ml^{-1}). After puromycin treatment, the library cells (labeled as day 0) were collected as the reference sample and were continuously cultured and passaged every 3 days. Experimental groups were isolated on day 24 (labeled as day 24). The sense library and antisense library were processed separately following the abovementioned steps.

Genome preparation and sequencing

The gDNA was extracted from the reference and experiment cells (Qiagen, 69506). The sgRNAs with a barcode were amplified by primers through 26 cycles of PCR (NEBNext Ultra II Q5 Master Mix). The target region of individual sgRNAs was amplified by primers for detecting the sequence. PCR products were purified using the DNA Clean & Concentrator-5 kit (Zymo Research Corporation, D4014) and indexed with different adaptors (New England Biolabs, 7335 and 7500) for NGS analysis.

Computational analysis algorithm for screening

To analyze the NGS data of screens, we used the ZFC^{iBAR} algorithm[11] to evaluate the change in sgRNA^{iBAR} abundance between the reference group and the experimental group. On the basis of the ZFC^{iBAR} workflow, we calculated the zLFC for each sgRNA^{iBAR} and the zLFC of sgRNA was calculated as the mean of the zLFC of the corresponding sgRNA^{iBAR}. RRA was also used to calculate the ranking significance for a certain sgRNA by ranking the sgRNA^{iBAR} in the whole library. For our bidirectional fitness screens, RRA was calculated twice according to a ranking of enrichment or depletion. The final sgRNA screen scores were based on the sgRNA zLFC and RRA as follows:

$$\text{Screen Score}_{\text{sgRNA}} = |\text{zLFC}_{\text{sgRNA}}| + (-\log_{10}\text{RRA}_{\text{sgRNA}})$$

Validation of S, T and Y fitness screening

The sgRNAs for validation were selected from the library. For all validation candidates, the sgRNAs were cloned into the lentiviral sgRNA-expressing vector individually with a cytomegalovirus (CMV) promotor-driven enhanced green fluorescent protein (EGFP) marker. The lentivirus was transduced into hTERT RPE1 cells at an MOI of less than 1. The percentage of mCherry-positive cells was measured every 3 days by fluorescence-activated cell sorting, indicating the fraction of sgRNA-infected cells. The first cell sorting analysis started 3 days after infection (labeled as day 0) and then the pooled cells were passaged on day 0.

The validation steps were as follows:

1. The pCMV-ABEmax-P2A-GFP was obtained from Addgene. The coding sequence of ABEmax was cloned into the pLenti-P2A-EGFP vector through restriction enzyme double digestion (New England Biolabs) and T4 DNA ligase ligation. The pLenti-ABEmax-EGFP plasmid was used for packing the lentivirus expressing ABEmax.
2. The sgRNA plasmids used for individual validation were generated by cloning the spacer sequences into the pCG-2.0-CMV-mCherry vector through Golden Gate assembly.
3. The cell library was established in two sublibraries because the size of the sgRNA library was huge. For the sense library, 4 × 10^6 cells were seeded into 15-cm dishes and 120 dishes were infected (~4.7 × 10^5 sgRNA); for the antisense library, 100 dishes were infected (~3.6 × 10^5 sgRNA).
4. The validation assay was based on competitive growth. ABEmax-expressing RPE1 cells were infected with lentiviruses carrying individual sgRNAs at an infection efficiency of 40–60%. sg*AAVS1* served as a negative control. The lentiviral plasmids expressing sgRNA contained an mCherry marker gene. The percentage of mCherry-positive cells was assessed through flow cytometry (LSRFortessa, Becton Dickinson). The flow cytometry analysis started on the third day after infection, designated as day 0, serving as the baseline for data normalization. Then, the percentage of mCherry-positive cells was analyzed every 3 days, extending to day 15 or day 18 (Supplementary Fig. 1–7).

Western blotting

Cells cultured in a six-well plate were treated with EGF (150 ng ml^{-1}) for 10 min before cell lysis for the experimental group and without EGF stimulation for control group. The concentrations of proteins in lysates were quantified using the Pierce BCA protein assay kit (Thermo Fisher Scientific). Cell lysates with SDS were denatured by boiling and subjected to SDS–PAGE following standard procedures. The proteins were transferred to a PVDF membrane (Bio-Rad) and blocked with PBS-T (PBS with 0.1% Tween 20) containing 5% milk at room temperature for 1 h. The membranes were incubated with primary antibodies at 4 °C overnight and then with secondary antibody at room temperature for 2 h. The horseradish peroxidase (HRP) signals were detected with Clarity Western enhanced chemiluminescence substrate kit (Bio-Rad, 1705060) and imaged with the Chemidoc Imaging system (Bio-Rad, 1708370).

The antibodies used in immunoblotting were anti-GAPDH (Abcam, ab9485; 1:10,000 dilution), anti-MAPKAPK2 (Proteintech, 13949-1-AP; 1:1,000), anti-phospho-Hsp27 (S15) (Cell Signaling Technology, 2404; 1:1,000 dilution), anti-MEK1 (Abcam, ab32091; 1:1,000 dilution), anti-MEK1 (phosphorylated on S298) (Abcam, ab96379; 1:1,000 dilution), anti-phospho-p44/42 MAPK (T202 and Y204) (Cell Signaling Technology, 8544; 1:1,000 dilution), goat antimouse IgG–HRP secondary antibody (Jackson ImmunoResearch, 115035003; 1:10,000 dilution), goat antirabbit IgG–HRP secondary antibody(Jackson ImmunoResearch, 111035003; 1:10,000 dilution).

MS and data analysis[38]

Cell pellets stimulated with EGF (experimental group) or not (control group) were lysed, digested and labeled with tandem mass tags. Equal quantities of peptides from each sample were pooled and fractionated by basic pH reverse-phase liquid chromatography (LC) with a gradient of 8–32% acetonitrile in 10 mM ammonium bicarbonate pH 9 over 60 min into 60 fractions. Then, the peptides were combined into eight fractions and dried by vacuum centrifuging. Phosphopeptides were enriched[39] before LC–MS/MS analysis. An Orbitrap mass analyzer was used at a resolution of 60,000 to acquire full MS with an m/z range of 350–1,400. The 20 most intense multiply-charged precursors were selected for higher-energy collision dissociation fragmentation with a normalized collisional energy of 28. The MS/MS fragments were measured at an Orbitrap resolution of 30,000 using an automatic gain control target of 8.3 × 10^4 ions per s and maximum fill times of 50 ms. MaxQuant-MS/MS data alignment (version 1.5.2.8 http://www.maxquant.org/) was used for analysis. The stimulation conditions were as follows: 150 ng ml^{-1} and 30-min incubation.

Clonogenic assay

At five passages of hTERT RPE1$^{\text{ABEmax}}$ cells and MAP4K4-Y210 cell clones, 30, 60 and 100 cells were seeded into each 10-cm plate and cultured for 20 days. Colonies were fixed with 4% paraformaldehyde for 10 min and then stained with 0.5% crystal violet for 10 min after PBS washing twice. Plating efficiency = number of colonies counted/number of cells plated. Surviving fraction = (number of colonies counted/number of cells plated)/plating efficiency (wild type)[40].

Xenograft tumor formation

hTERT RPE1$^{\text{ABEmax}}$ cells (3 × 10^6) and MAP4K4-Y210 cell clones (3 × 10^6) at passage 3 were trypsinized and mixed with Matrigel (BD Biosciences) in a 100-µl volume and implanted subcutaneously into 7-week-old BALB/C nude mice (Laboratory Animal Center of Peking University). Mice were housed in a 12-h light–dark cycle with food and water available ad libitum under specific pathogen-free conditions in the Laboratory Animal Center of Peking University. The temperature was maintained at ~18–23 °C, with 60% ± 10% relative humidity. Eight mice were prepared as the control group (hTERT RPE1$^{\text{ABEmax}}$ cells) and eight mice were prepared as the experimental group (MAP4K4-Y210 cell clones). Mice were monitored every other day until tumors appeared and tumor volumes were measured with a caliper twice per week (volume = (length × width × width)/2) until mice were killed when tumors reached 1 cm^3 or on day 28 or day 50. The animal experiments were approved by the Peking University Laboratory Animal Center. All experiment protocols were approved by the respective laboratory animal care and use committees of Peking University and undertaken in accordance with the National Institutes of Health Guide for Care and Use of Laboratory Animals.

Statistics and reproducibility

R (version 3.6.0) scripts (ggplot2 (version 3.4.4), dplyr (version 1.1.3) and clusterProfiler (version 4.8.2)) and Python (version 3.7.1) scripts (biopython (version 1.81), pandas (version 2.1.0), numpy (version 1.25.2) and scipy (version 1.11.3)) were used for statistical analysis and data visualization. Statistical methods were included in the related figure legends. The exact P values for comparisons in the cell proliferation assay are included in the Source Data files of the related figures. Customized bash scripts (with AWK version 4.0.2) were used to generate raw counts from FASTQ files. Bowtie (version 1.3.1) software was used to calculate sgRNA off-target sites. ZFC (version 0.1.6) software (https://github.com/wolfsonliu/zfc) was used to analyze cell fitness screens with iBARs. Metascape (http://metascape.org/gp/index.html#) was used for gene ontology enrichment analysis. PyMOL 3.0 was used for protein structure visualization. FlowJo 10 and GraphPad Prism 7 were used for basic statistical analysis and graph production.

Reporting summary

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

Data availability

FASTQ files of raw sgRNA reads obtained through NGS are available from the NCBI under BioProject PRJNA874493. Raw screening NGS data are available from the NCBI under BioProject PRJNA874493. Processed screen data and MS data are provided in the Supplementary Information. Reference human proteome data used in this study were UniProt Proteomes (UP000005640.fasta; https://www.uniprot.org/proteomes/). The human reference genome used was hg38 from UCSC (https://genome.ucsc.edu/). Databases involved in this study included dbPTM (https://awi.cuhk.edu.cn/dbPTM/), KEGG (https://www.genome.jp/kegg/), PhosphoSitePlus version 6.7.4 (https://www.phosphosite.org/homeAction.action), PROSITE (https://prosite.expasy.org/), DisProt (https://www.disprot.org/), Interactome INSIDER (http://interactomeinsider.yulab.org/) and ICGC (https://dcc.icgc.org/). Source data are provided with this paper.

Code availability

Custom code used in the analysis is available on GitHub (https://github.com/yzjoyceli/functional-STY/).

References

36. Ryu, S. M. et al. Adenine base editing in mouse embryos and an adult mouse model of Duchenne muscular dystrophy. *Nat. Biotechnol.* **36**, 536–539 (2018).
37. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
38. Tan, H. et al. Integrative proteomics and phosphoproteomics profiling reveals dynamic signaling networks and bioenergetics pathways underlying T cell activation. *Immunity* **46**, 488–503 (2017).
39. Tan, H. et al. Refined phosphopeptide enrichment by phosphate additive and the analysis of human brain phosphoproteome. *Proteomics* **15**, 500–507 (2015).
40. Rafehi, H. et al. Clonogenic assay: adherent cells. *J. Vis. Exp.* **49**, 2573 (2011).

Acknowledgements

We extend our gratitude to the National Center for Protein Sciences (Beijing) at Peking University for their support through the PKU-EdiGene High-Throughput Screening Core and the flow cytometry platform, MS platform at the School of Life Sciences and high-performance computing platform at Peking University for their assistance with NGS data analysis. This study was also supported by Changping Laboratory. Special appreciation goes to Y. Yu at Peking University for her contribution to the preparation of the NGS library and to Y. Fan for his essential role in establishing the xenograft mouse model. We are grateful for the collaboration with PTM BIO for their expertise in phosphoproteomic detection and analysis. T.X. received partial support from the Postdoctoral Fellowship of Peking-Tsinghua Center for Life Sciences. W.W. received funding from the National Science Foundation of China (82341207 and 31930016), the Beijing Advanced Innovation Center for Genomics at Peking University, the Peking-Tsinghua Center for Life Sciences and Changping Laboratory.

Author contributions

W.W. conceptualized and supervised the project. The experimental design was collaboratively developed by W.W., Y.Li, T.X., H.M. and Z.Z. The experiments, including base-editing screening and cell proliferation assays, were carried out by T.X., H.M., D.Y. and Q.L., with support from Y.Liu. T.X. and H.M. conducted the western blot assays. Y.Li handled the NGS analysis and data interpretation. The manuscript was written by Y.Li, T.X. and W.W., with contributions from all authors.

Competing interests

W.W. is the founder and scientific advisor of EdiGene and Therorna. The authors declare no other competing interests.

Additional information

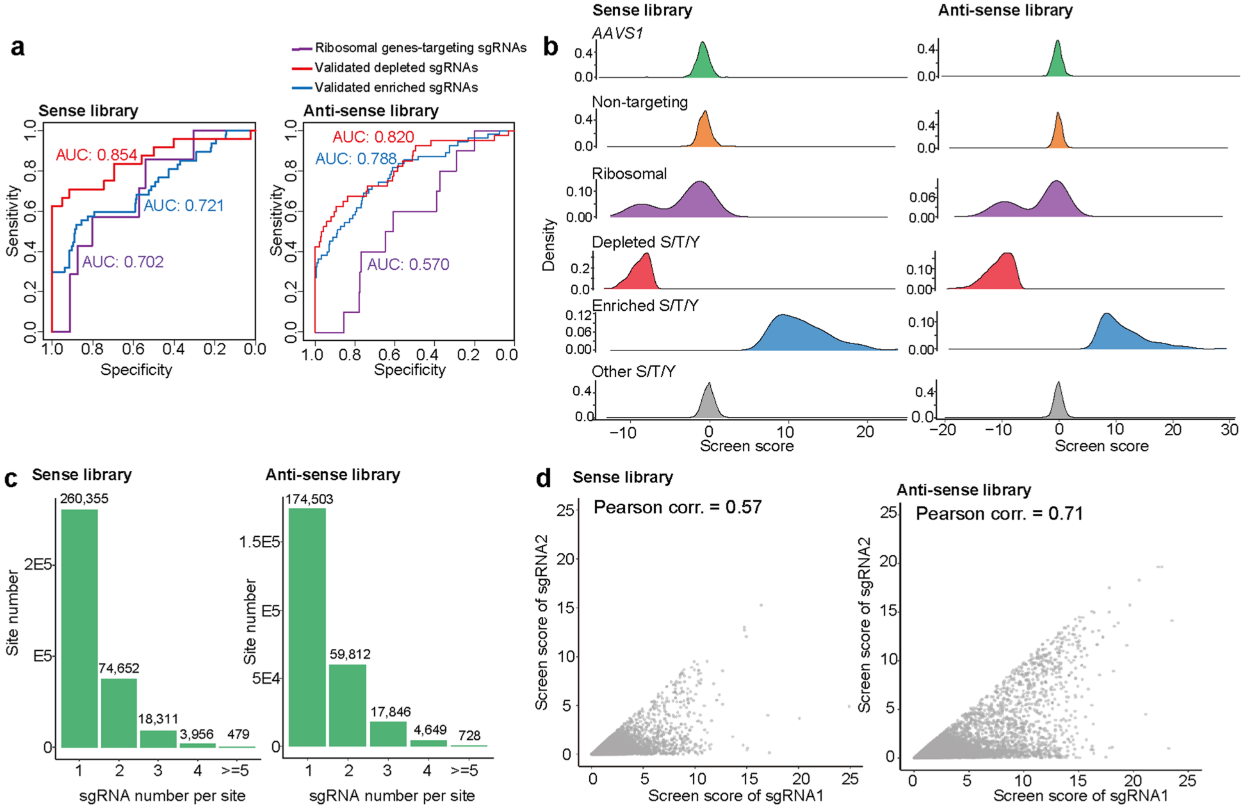
Extended data is available for this paper at https://doi.org/10.1038/s41589-024-01731-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41589-024-01731-0.

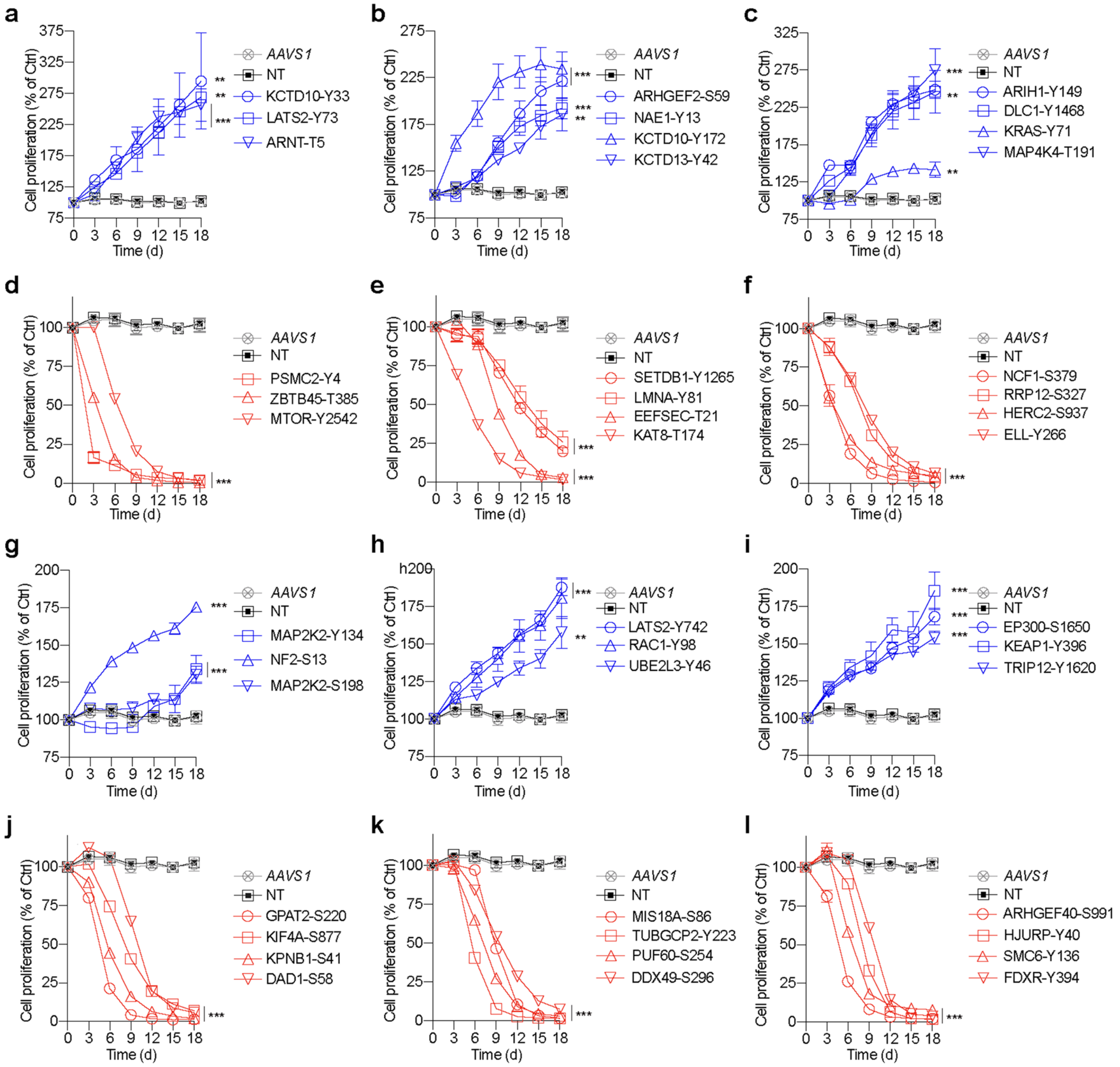
Correspondence and requests for materials should be addressed to Wensheng Wei.

Peer review information *Nature Chemical Biology* thanks Pedro Cutillas and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

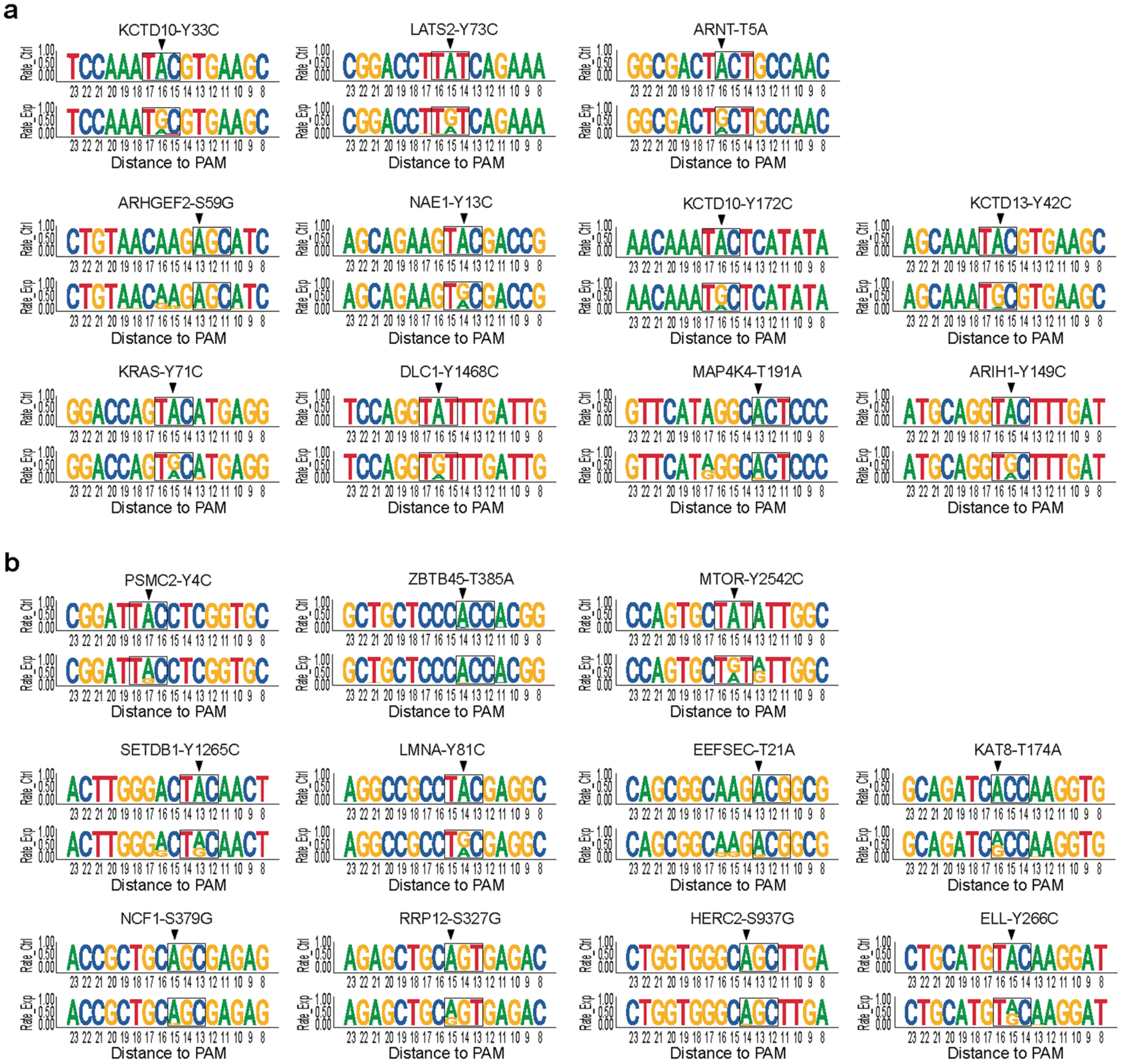
Reprints and permissions information is available at www.nature.com/reprints.



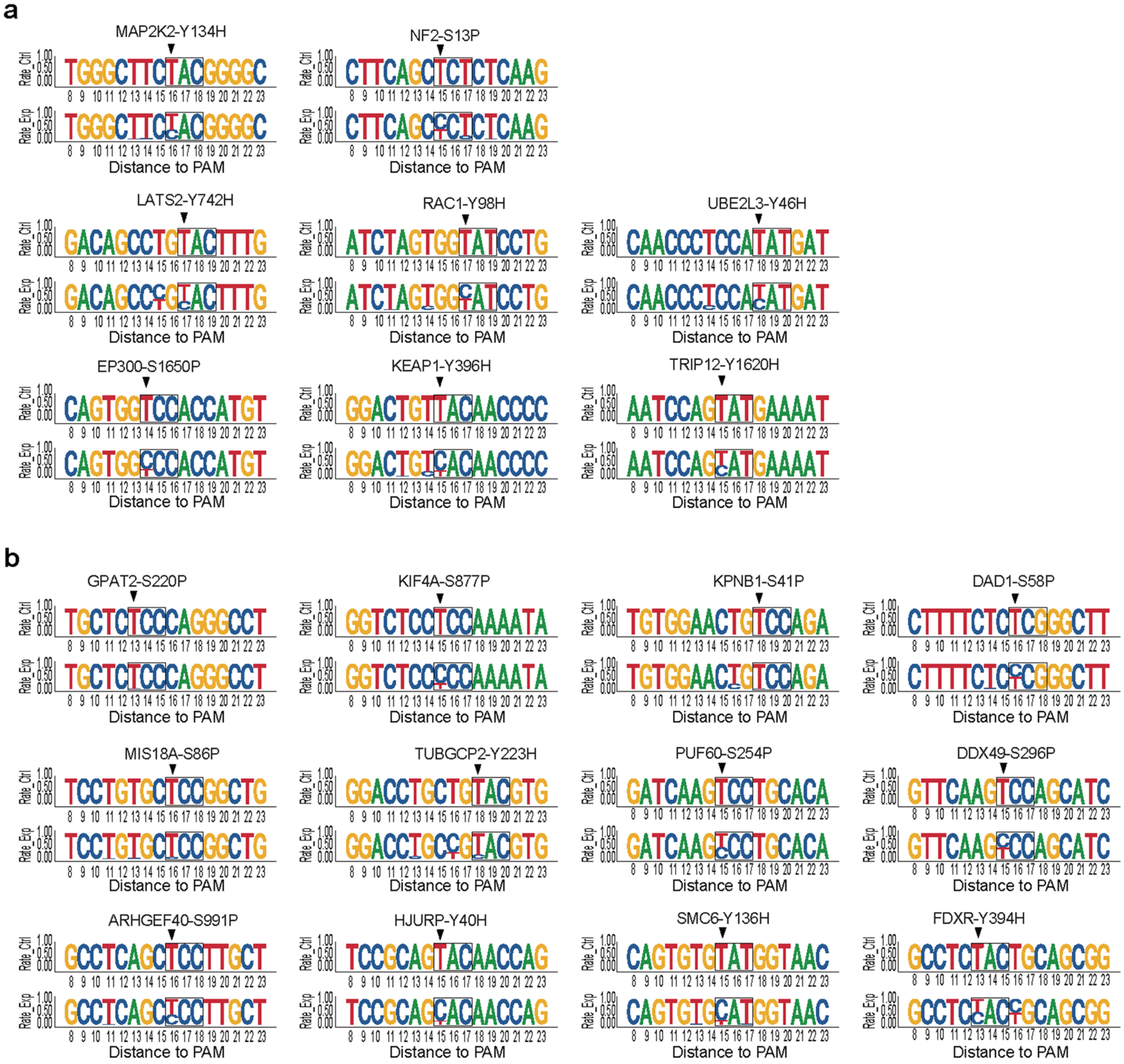
Extended Data Fig. 1 | Quality assessment of whole-genome S/T/Y screenings. **a**, AUC analysis of sense library (left) and anti-sense library (right) based on sgRNAs targeting negative controls and validated actual positives. The red and blue curve represent the ROC curve based on validated depleted sgRNAs and enriched sgRNAs, respectively; the purple curve represents ROC curve based on the 30 sgRNAs targeting ribosomal genes. **b**, Density plot of screen score for sgRNAs in different categories in sense library (left) and anti-sense library (right). **c**, Bar plots showing the number of sites targeting by single or multiple sgRNAs in sense (left) and anti-sense (right) library. **d**, Scatter plots showing the screen score correlation of sgRNAs targeting the same sites in sense (left) and anti-sense (right) library. The method used for assessing correlation is Pearson's r (n = 97,021 for sense library and 82,707 for anti-sense library).



Extended Data Fig. 2 | Validation of the selected S/T/Y mutations in sense library and anti-sense library. a to l, Effects of indicated sgRNAs in sense library (**a** to **f**) and anti-sense library (**g** to **l**) on cell proliferation in hTERT RPE1$^{\mathrm{ABEmax}}$ cells. Cell proliferation assay and data analysis are same as Fig. 2d. Data are presented as the mean ± SD of three independent experiments. *p*-values represent comparisons with sgRNA targeting *AAVS1* at the end point (day 18), calculated using one-sided Student's t-test without adjustment (n = 3, biologically independent experiments), **p*-value < 0.05, ***p*-value < 0.01, ****p*-value < 0.001. NT, non-targeting.

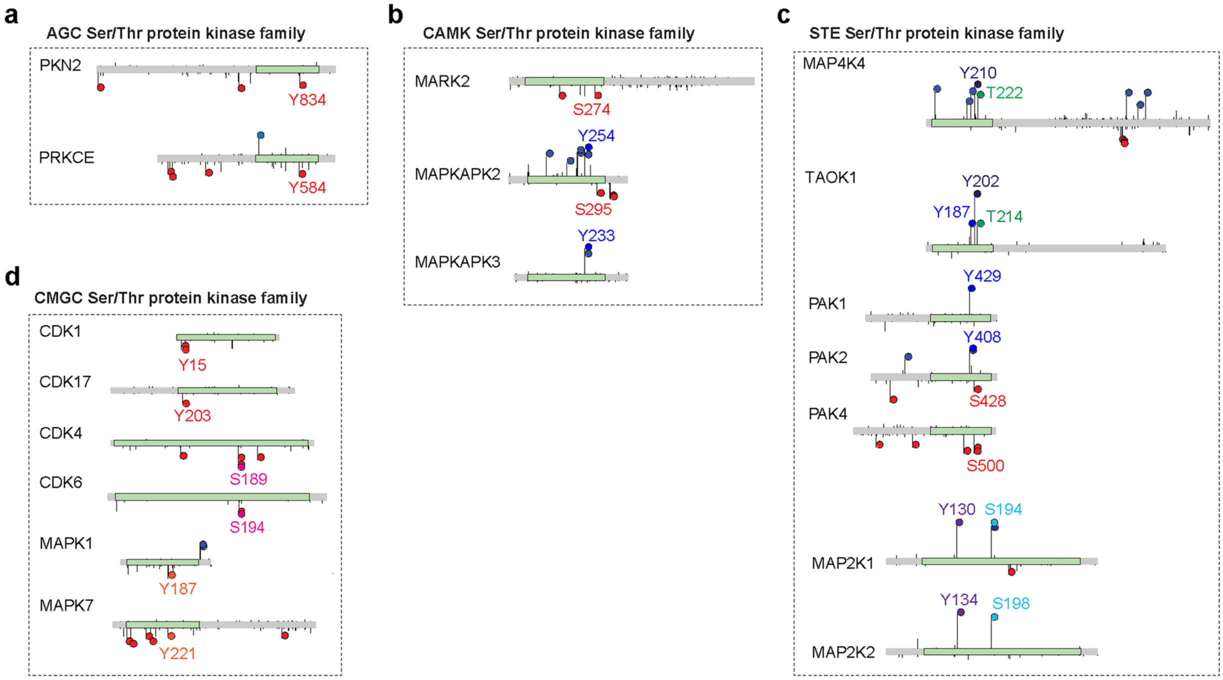


Extended Data Fig. 3 | The NGS results showing the editing outcomes of sgRNAs targeting corresponding S/T/Y in sense library. a, Sequences present the residues with enriched sgRNAs related to Extended Data Fig. 2a-c. **b**, Sequences present the residues with depleted sgRNAs related to Extended Data Fig. 2d-f.

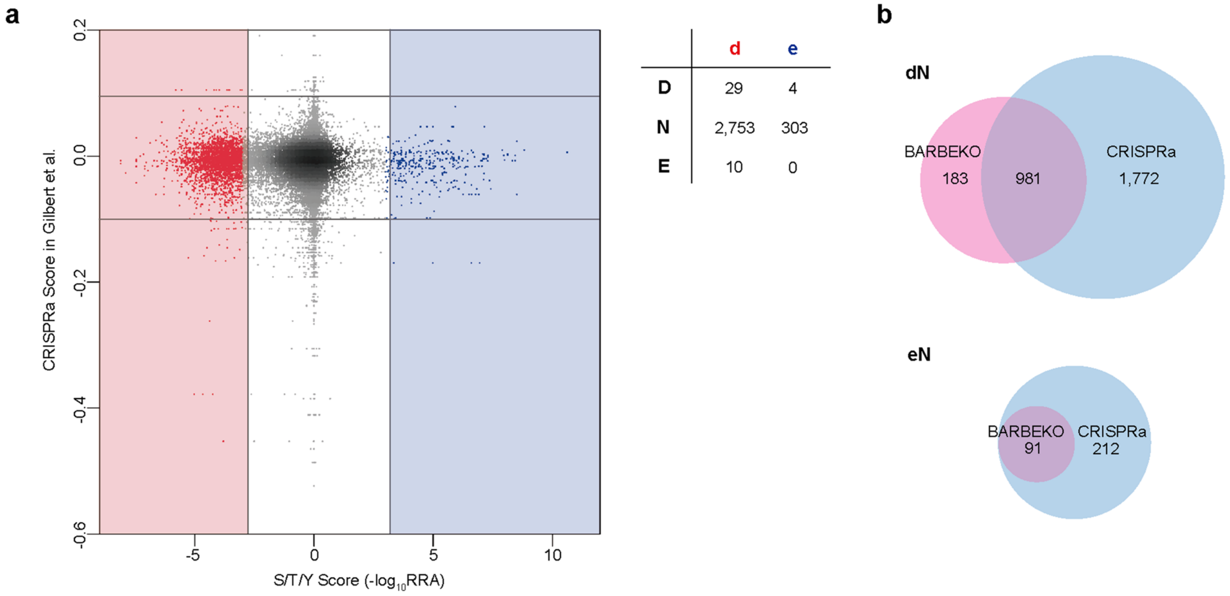


Extended Data Fig. 4 | The NGS results showing the editing outcomes of sgRNAs targeting corresponding S/T/Y in anti-sense library. **a**, Sequences present the residues with enriched sgRNAs related to Extended Data Fig. 2g-l. **b**, Sequences present the residues with depleted sgRNAs related to Extended Data Fig. 2j-l.

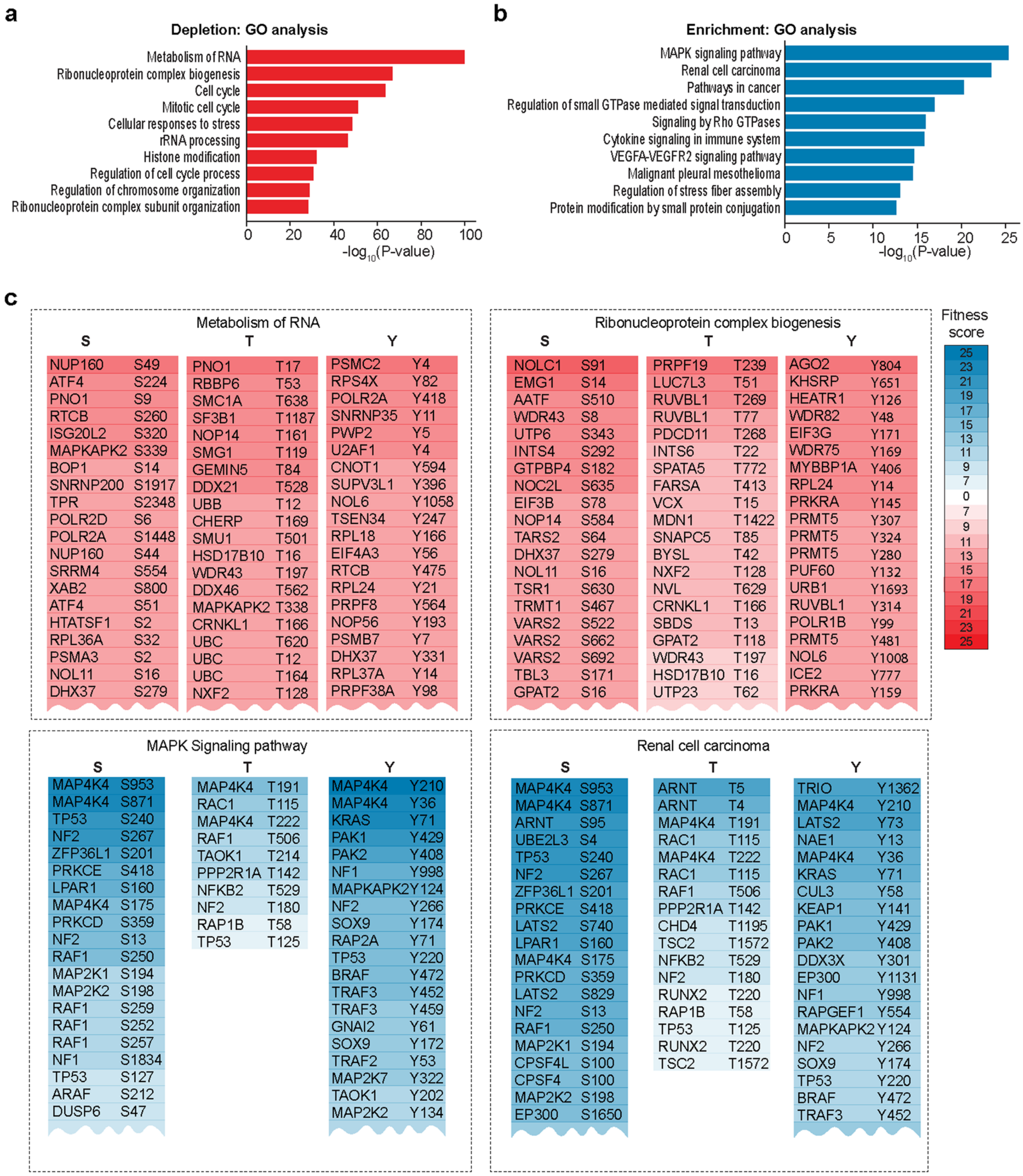
https://doi.org/10.1038/s41589-024-01731-0



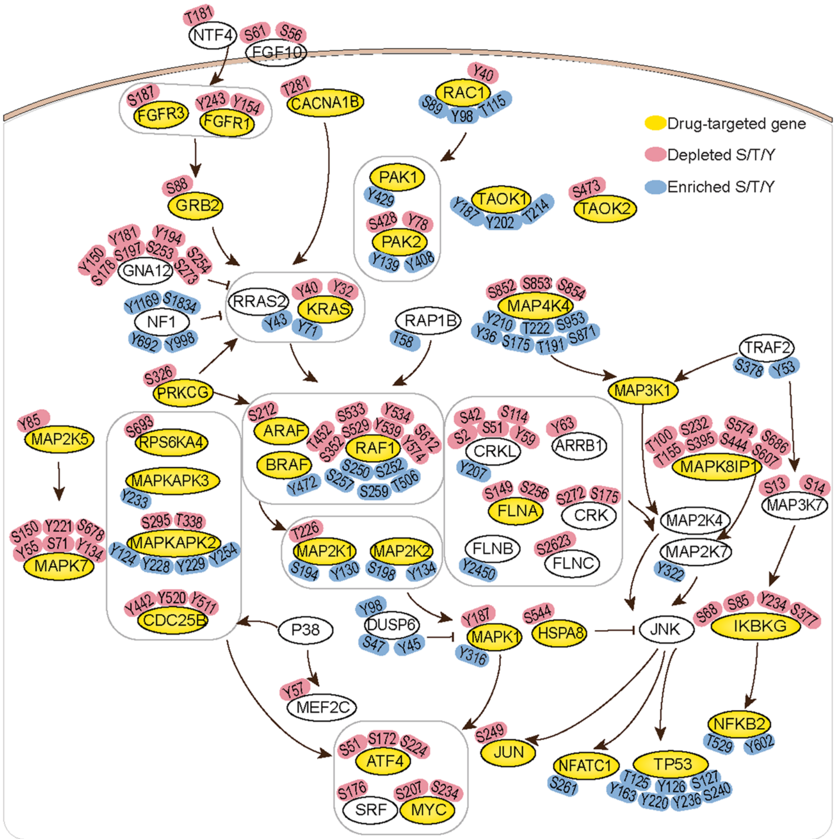
Extended Data Fig. 5 | Conserved functional S/T/Y sites in kinase domain. **a** to **d**, Schematics showing 12 pairs of highly conserved sites in four kinase subfamilies. Conserved residues with corresponding sgRNAs are indicated by the same color circles with labeled name. The color of circle stands for conserved amino acids and each pair of homolog sites shared same color. The direct of lollipop represents the guides were either depleted or enriched and the height correspond to the extent of enrichment or depletion.



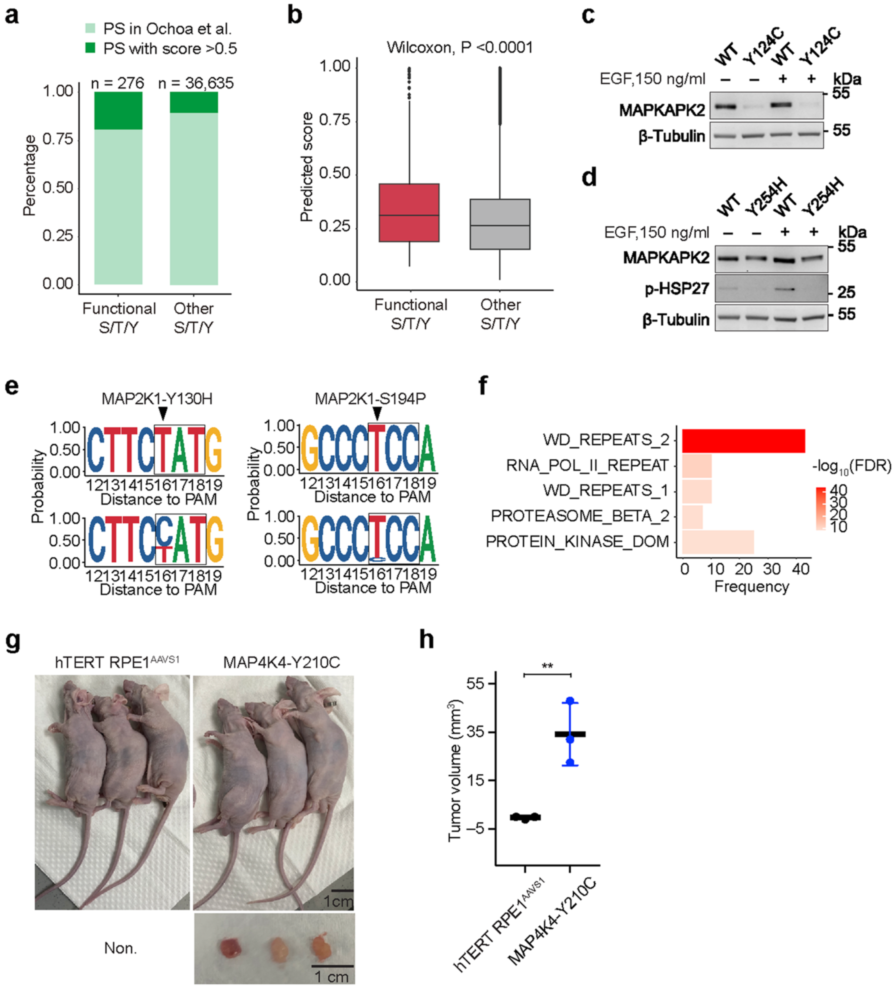
Extended Data Fig. 6 | Comparison of the cell fitness effects between S/T/Y mutation and gene activation. a, Scatter plot showing the distribution of mutants RRA in S/T/Y mutagenesis screens and phenotype score in CRISPRa screening. The red dots represent S/T/Y with depleted sgRNAs and the blue dots represent S/T/Y with enriched sgRNAs. The amino acids with depleted and enriched sgRNAs are labelled as lowercased 'd' and 'e'. The genes inhibiting cell growth, not affecting cell growth and promoting cell growth while overexpressed are labelled as superscript 'D', 'N', and 'E'. The right table showing the number of sites in each category. **b**, Venn plots showing the number of overlapped dN (top) and eN (bottom) sites based on gene KO and activation screenings.



Extended Data Fig. 7 | GO enrichment for functional S/T/Y belonging genes and typical functional sites in enriched GO terms. a and **b**, GO enrichment analysis of genes with identified residues leading to cell death or growth inhibition (**a**) and promoting cell growth (**b**) respectively. The statistics were performed with one-sided Fisher's exact test without adjustment. **c**, Top 20 mutants presented as examples involved in top two biological processes, including 'metabolism of RNA', 'Ribonucleoprotein complex biogenesis', 'MAPK signaling pathway' and 'Renal cell carcinoma'. Lists longer than 20 mutants were truncated, indicated by jagged lines.

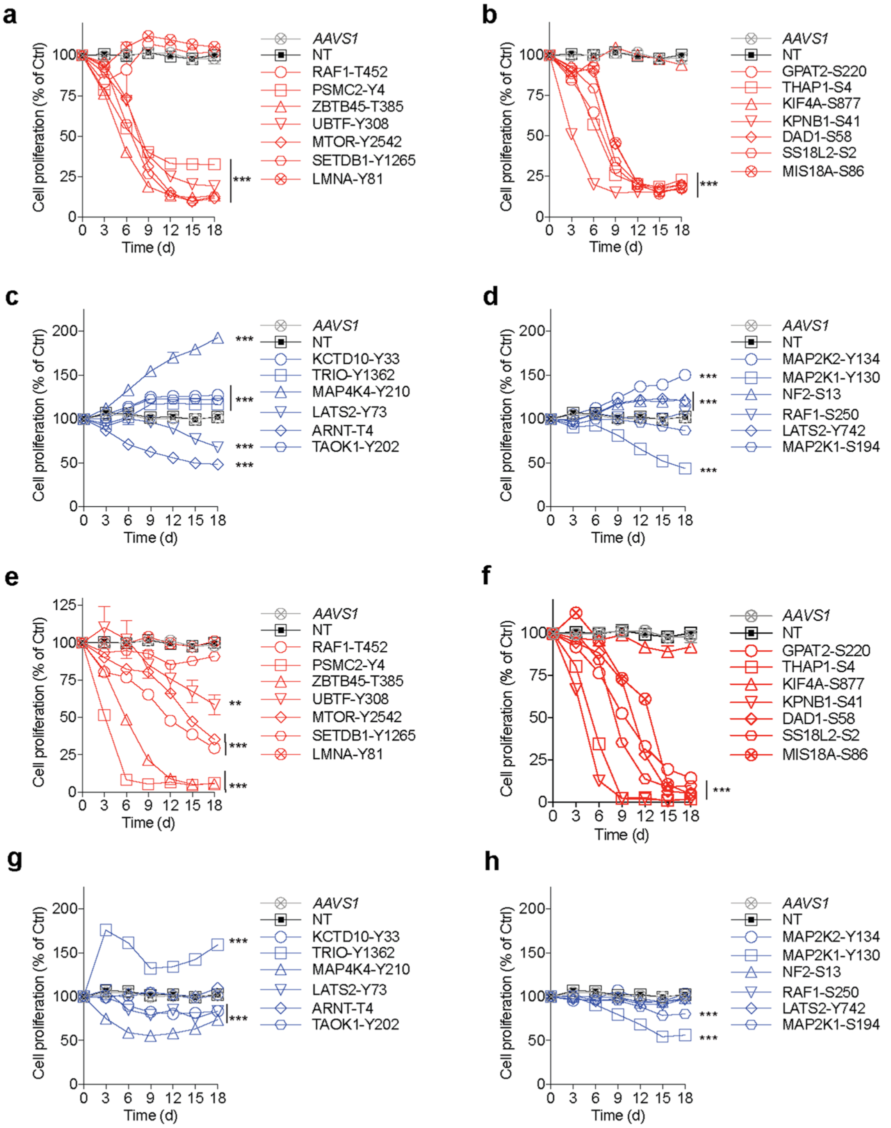


Extended Data Fig. 8 | Functional S/T/Y regulating MAPK signaling pathway in RPE1 cell. Schematic showing the functional S/T/Y residues in MAPK signaling cascades, in which the regulators are indicated in ellipses, positively selected S/T/Y mutations are represented in blue rectangles while negatively selected mutations are represented in red rectangles. Proteins acted as drug targets in DrugBank database were colored in yellow. Pathway map is modified from the KEGG-MAPK signaling pathway (map04010).



Extended Data Fig. 9 | Selected S/T/Y sites function in either phosphorylation-dependent or independent way. a, Bar plot showing the percentage of predicted functional phosphorylation site in S/T/Y library grouped by our screening result. **b**, Boxplot showing predicted score distribution of sgRNAs targeting functional sites (n = 276) and non-functional sites (n = 36,635) according to our screening. The centerline representing the median and the whiskers showing the minimum to the maximum. The boundaries of the box indicate the first and third quantiles. The statistics was performed using two-sided Wilcoxon rank sum test, p-value = 5.012e-07. **c** and **d**, Western blot analysis of MAPKAPK2 and Hsp27 expression in cell clones with wild type, MAPKAPK2-Y124C (**c**) and MAPKAPK2-Y254H (**d**) mutants upon EGF (150 ng/ml) stimulation. The experiment was repeated three times independently with similar results. **e**, NGS showing the editing outcomes of sgRNA targeting MAP2K1-Y130 and MAP2K1-S194. **f**, Bar plot showing domain enrichment analysis of functional S-P mutations, the length of each bar represents number of functional S-P in top 5 domains corresponding to Fig. 5d, the color represents enrichment significance relative to designed S-P number in this domain (Fisher's exact test, one-sided, FDR was the adjusted p-value using Benjamini-Hochberg method). **g**, Images of mice 28 days following 3 × 10^6 cells injection subcutaneously into nude mice compared with *AAVS1* controls. Scale bars, 1 cm. **h**, Tumor volumes 28 days with injection. Tumor volumes were presented in three individual mice and there were never detected at eight mice of control group. Data are presented as the mean ± SD of three individual mice. The comparison of tumor volumes at the end point (day 28) in wild type group and mutant group was calculated using one-sided Student's t-test without adjustment (p-value = 5.046e-03), *p-value < 0.05, **p-value < 0.01, ***p-value < 0.001. The centerlines represent the median, and the whiskers show the minimum to maximum values. The dots indicate the tumor volumes of each sample.

https://doi.org/10.1038/s41589-024-01731-0



Extended Data Fig. 10 | Validation of functional S/T/Y mutations in A375 and SW620 cell lines. a to **d**, Effects of top hit sgRNAs in sense library (**a**, **c**) and anti-sense library (**b**, **d**) on cell proliferation in A375 cell line. **e** to **h**, Effects of top hit sgRNAs in sense library (**e**, **g**) and anti-sense library (**f**, **h**) on cell proliferation in SW620 cell line. Data are presented as the mean ± SD of three independent experiments. *p*-values represent comparisons with sgRNA targeting *AAVS1* at the end point (day 18), calculated using one-sided Student's t-test without adjustment (n = 3, biologically independent experiments), **p*-value < 0.05, ***p*-value < 0.01, ****p*-value < 0.001. NT, non-targeting.

nature portfolio

Corresponding author(s): Wensheng Wei

Last updated by author(s): Jul 31, 2024

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: No software was used for data collection

Data analysis: R (v.3.6.0) script (ggplot2 (v.3.4.4), dplyr (v.1.1.3), clusterProfiler (v.4.8.2)) and Python (v.3.7.1) script (biopython (v.1.81), pandas (v.2.1.0), numpy (v.1.25.2), scipy (v.1.11.3)) were used for statistical analysis and data visualization. Statistical methods were included in the related figure legends. Customized bash scripts (with AWK v.4.0.2) were used to generate raw counts from FASTQ files. Bowtie (v.1.3.1) software was used to calculate sgRNA off-target sites. ZFC (v.0.1.6) software (https://github.com/wolfsonliu/zfc) was used to analyze cell fitness screens with iBARs. Metascape (http://metascape.org/gp/index.html#) was used for gene GO enrichment analysis. PyMOL 3.0 was used for protein structure visualization. FlowJo 10 and GraphPad Prism 7 were used for basic statistical analysis and graph production. Custom code used in the analysis is available on GitHub (https://github.com/yzjoyceli/functional-STY/)

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

FASTQ files of raw sgRNA reads by NGS are available in NCBI BioProject PRJNA874493. Raw screening NGS data are available in NCBI BioProject PRJNA874493. Processed screen data and MS data are provided in Supplementary Data. Reference human proteome data used in this study is UP000005640.fasta from Uniprot Proteomes (https://www.uniprot.org/proteomes/). Human reference genome used is hg38 from UCSC (https://genome.ucsc.edu/). Databases involved in this study include dbPTM (https://awi.cuhk.edu.cn/dbPTM/), KEGG (https://www.genome.jp/kegg/), PhosphoSitePlus v.6.7.4 (https://www.phosphosite.org/homeAction.action), PROSITE (https://prosite.expasy.org/), DisProt (https://www.disprot.org/), Interactome INSIDER (http://interactomeinsider.yulab.org/), ICGC (https://dcc.icgc.org/).

Human research participants

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

Reporting on sex and gender	This study did not involve any human research participants
Population characteristics	This study did not involve any human research participants
Recruitment	This study did not involve any human research participants
Ethics oversight	This study did not involve any human research participants

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	The library coverage of screens in RPE1 cells was based on the recommend size from Bao et al. (2023).
Data exclusions	No data exclusions.
Replication	The numbers of replications were indicated in the text, methods or figure legends. All attempts at replication were successful.
Randomization	All samples from cultured cells were randomly allocated after mixing for experiments.
Blinding	No blinding was performed due to the involvement of several experimentators.

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

Antibodies

Antibodies used	Primary antibodies used here were GAPDH (Abcam, ab9485, target species: human), MAPKAPK2 (Proteintech, 13949-1-AP, target species: human, mouse), Phospho-Hsp27 (Ser15) (Cell Signaling Technology, #2404, target species: human,monkey), MEK1 (Abcam, ab32091,target species: human et al.), MEK1 (phosphor S298) (Abcam, ab96379, target species: human et al.), Phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) (Cell Signaling Technology, #8544, target species: human et al.), pMEK1/2(phosphor Ser217/221) (Cell Signaling Technology, #9154, target species: human et al.). Goat anti-rabbit IgG-HRP(Jackson Immunoresearch, 111035003) or goat-mouse IgG-HRP (Jackson Immunoresearch, 115035003) secondary antibodies were used.
Validation	All antibodies used in this study were validated by the manufacturers, and the western blot experiments were performed according to the manufacturer's instruction. And the western blot data were provided in manuscript.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	HEK293T cells from C. Zhang's laboratory (Peking University), hTERT PRE1 cells from Y. Sun's laboratory (Peking University). A375 cells were purchased from ATCC.
Authentication	STR analysis was used for cell line authentication.
Mycoplasma contamination	All cells were tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	The experimental animals included 7-week-old BALB/C male mice (Beijing Vital River Laboratory). Mice were housed in12-h light-dark cycle with food and water available ad libitum under SPF (specific pathogen-free) conditions in the Laboratory Animal Center of Peking University. Temperature and humidity: Temperatures of ~18-23°C, 60 ± 10% relative humidity.
Wild animals	No wild animals were used in this study.
Reporting on sex	All mice used in this study were male.
Field-collected samples	This study did not involve field-collected samples.
Ethics oversight	The animal experiments were approved by Peking University Laboratory Animal Center(Beijing). All experiment protocols were approved by the respective Laboratory Animal Care and Use Committees of Peking University, and undertaken in accordance with the National Institute of Health Guide for Care and Use of Laboratory Animals.

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 were infected by lentivirus containing sgRNAs with polybrene. About 48 to 72 h later, cells were digested with trypsin and collected for the following FACS according to the fluorescence marker.
Instrument	BD LSRFortessa and BD Aria SORP
Software	BD FACSDiva 10and FlowJo_V10
Cell population abundance	Over 10000 single cells with normal shape of each sample were analyzed for the percentage of positive fluorescence in cell proliferation assay.
Gating strategy	FSC-A and SSC-A(P1) were used to gate cells with normal shape, then SSC-W and SSC-H(P2) following FSC-W and FSC-H(P3) were used to gate single cells for further analysis.

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