nature biotechnology

Article

https://doi.org/10.1038/s41587-025-02710-z

Prime editor-based high-throughput screening reveals functional synonymous mutations in human cells

Received: 10 June 2024

Accepted: 21 May 2025

Published online: 24 June 2025

Check for updates

Xuran Niu[1,5], Wei Tang[1,2,5], Yongshuo Liu[1,3,5], Binrui Mo[1], Ying Yu[1], Ying Liu[1,4] ✉ & Wensheng Wei[1,4] ✉

Synonymous mutations are generally considered neutral, while their roles in the human genome remain largely unexplored. Here we use the PEmax system to create a library of 297,900 engineered prime-editing guide RNAs and perform extensive screening to identify synonymous mutations affecting cell fitness. Unlike recent findings in yeast, group-level analyses show that synonymous mutations diverge from nonsynonymous mutations in fitness effects yet exhibit similar phenotypic distributions relative to negative controls. Following rigorous quality control, only a small subset demonstrated measurable effects. For these functional mutations, we develop a specialized machine learning tool and uncover their impact on various biological processes such as messenger RNA splicing and transcription, supported by multifaceted experimental evidence. We find that synonymous mutations can alter RNA folding and affect translation, as demonstrated by PLK1_S2. By integrating screening data with our model, we predict clinically deleterious synonymous mutations. This research deepens our understanding of synonymous mutations, providing insights for clinical disease studies.

Because of the degeneracy of the genetic codon system, not all single-base mutations lead to changes in the amino acid sequence. Such mutations are termed synonymous and are traditionally viewed as neutral in evolutionary theory[1]. A recent study in *Saccharomyces cerevisiae* reported that synonymous and nonsynonymous mutations can similarly disrupt cell fitness, claiming that synonymous mutations may have non-neutral phenotypes[2]. However, these results have been debated[3], reigniting interest and discussion among researchers about the biological effects of synonymous mutations. Earlier studies in viruses and prokaryotes also suggested that synonymous mutations could affect the fitness of these organisms[4–7]. Yet, it remains unclear whether these findings in noneukaryotic organisms and yeast are applicable to mammals, especially humans.

Previous research has linked a small number of synonymous mutations to human diseases[8] and identified them as potential drivers in cancer through bioinformatics analyses[9]. Despite the development of tools for predicting deleterious synonymous mutations[10,11], experimentally confirmed cases in humans remain scarce, highlighting the need for a standardized experimental method for large-scale studies of synonymous mutations in human cells.

The advent of CRISPR–Cas gene-editing technology has provided a powerful tool for studying the genome[12,13]. An important derivative of CRISPR–Cas, the prime editor (PE), combines dCas9 with a reverse transcriptase to introduce various types of edits into the genome through a reverse transcription template (RTT) in the prime-editing guide RNA (pegRNA)[14]. Enhancements such as 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 Gene Function and Modulation Research, School of Life Sciences, Peking University, Beijing, China. [2]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China. [3]Department of Clinical Laboratory, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan, China. [4]Changping Laboratory, Beijing, China. [5]These authors contributed equally: Xuran Niu, Wei Tang, Yongshuo Liu. ✉e-mail: ying_liu@pku.edu.cn; wswei@pku.edu.cn

engineered pegRNA (epegRNA)[15] and the more efficient PEmax[16] have further improved the use of PE for precise genetic manipulations. In this study, using PE technology, we created a library of epegRNAs targeting 3,644 human protein-coding genes to screen for potentially functional synonymous mutations affecting human cell fitness. Our findings confirm that, while most synonymous mutations in human are likely neutral, a minority can produce phenotypic changes. Using machine learning, we identified how these mutations influence a range of biological processes, including messenger RNA splicing, folding, transcription and translation. Furthermore, we also predicted clinically deleterious synonymous mutations, thereby enhancing our understanding of these mutations and their importance in clinical research on human diseases.

Results

Development of a synonymous mutation screen method using PE

To precisely and efficiently generate synonymous mutations across different types of nucleotide substitutions, we used the PEmax system alongside epegRNA for targeted genetic edits. To facilitate high-throughput screening at a high multiplicity of infection (MOI), we integrated barcodes into the external region of the epegRNA, termed eBARs, as reported previously[17,18]. Each epegRNA was labeled with three independent eBARs, effectively creating three biological replicates for our screening process (Extended Data Fig. 1a). For the screening, we selected the human colon cancer cell line HCT116, which is beneficial for PE editing because of its naturally homozygous nonsense mutation in the *MLH1* gene[16].

Following a set of specific criteria (Extended Data Fig. 1b), we constructed an epegRNA library. Initially, we sourced potential pathogenic synonymous mutations from two human disease databases, ClinVar and SynMICdb[19]. Additionally, our goal was to investigate the function of synonymous mutations in a relatively unbiased manner rather than focusing solely on clinically recognized mutations. Consequently, we selected 67 genes for saturated synonymous mutation design on the basis of their mutation load and expression levels, including human homologous genes previously studied in yeast[2] (Extended Data Fig. 1c). Within this group, 11 essential human genes were selected for complete saturation tiling mutation design to thoroughly evaluate the biological impacts of synonymous mutations alongside various nonsynonymous mutations (Extended Data Fig. 1c). The library also included designs for single-base insertions or the introduction of premature stop codons for gene knockouts and incorporated *AAVS1*-targeting and nontargeting epegRNAs as negative controls. Each mutation site was targeted by an average of 2.2 epegRNAs (Extended Data Fig. 1d). For the 11 genes designed for saturation mutagenesis, all possible types of amino acid substitutions within the range of point mutations were included (Extended Data Fig. 1e). Ultimately, the library contained 297,900 epegRNAs targeting 94,993 synonymous mutations and 39,336 nonsynonymous mutations across 3,644 protein-coding genes (Supplementary Tables 1 and 2).

PEmax was stably integrated into HCT116 cells through lentiviral transduction and a single clone was selected for subsequent studies. The expression profiles of HCT116-PEmax cells were almost identical to those of wild-type (WT) cells and the HCT116-PEmax cells infected with nontargeting or *AAVS1*-targeting epegRNAs, which served as negative controls, showed minimal changes (Extended Data Fig. 1f). This indicates that the stable cell line closely resembles the characteristics of WT HCT116 cells. We also assessed the editing efficiency of PEmax in HCT116 cells. Over a 28-day culture period with periodic sampling, the editing of the high-efficiency *FANCF* + 5 G-to-T mutation showed a gradual slowdown after 14 days, whereas the low-efficiency *RNF2* + 1 C-to-A mutation exhibited a gradual increase in editing efficiency over time (Extended Data Fig. 1g). Given that phenotypic changes often lag behind genotypic alterations, we extended the screening duration to 35 days. We termed this methodology PRESENT (prime editor-based screen technology) (Fig. 1a).

Quality control and analysis of screening data

Because mutations were introduced through the RTT, decoding the information from the RTT and eBAR regions was sufficient for analysis (Extended Data Fig. 2a). To process the data, we developed a tailored bioinformatics algorithm called ZFC-eBAR (Methods), enabling comprehensive evaluation of the effects from the eBAR level to the mutation level (Extended Data Fig. 2b). For each epegRNA, we calculated the zLFC (z score of $\log_2$ fold change) across three eBAR replicates and applied a robust rank aggregation (RRA) algorithm[20] to assess the significance of epegRNA enrichment in either the enriched or the depleted direction. We defined the screen score by $-\log_{10}$(RRA) and direction (Extended Data Fig. 2b).

As each mutation in the library is targeted by an average of 2.2 epegRNAs, we defined the fitness score for mutations as the Top2 mean screen score. To ensure the accuracy of data analysis, we included *AAVS1*-targeting and nontargeting epegRNAs as negative controls. On the basis of their performance, we established a threshold of RRA ≤ 0.001 (|screen score| ≥ 3) for this screening (Supplementary Table 3).

To evaluate the reliability of the screening results, we assessed the performance of essential genes subjected to knockout mutations using epegRNA, which introduced nonsense or frameshift mutations through single-base insertions. Using the Chronos score, we established two evaluation thresholds: a stringent criterion (<−2.5) and a more relaxed criterion (<−0.5). The area under the curve (AUC) for the receiver operating characteristic (ROC) curves was 0.72 for epegRNAs targeting highly essential genes and 0.65 for a broader range of essential genes, reflecting the overall performance of the screening (Fig. 1b and Extended Data Fig. 2c).

Analyzing results at the epegRNA level may partially underestimate the biological effects of the mutations. To address this, we performed additional mutation-level analyses. By selecting the epegRNA with the highest screen score for a given mutation, the mutation-level AUCs for the two evaluation criteria were 0.75 and 0.68, respectively (Fig. 1c and Extended Data Fig. 2d). Further evaluation of highly essential genes using different scoring methods (Top1 score, Top2 mean score and all mean score) yielded consistent AUCs of 0.75, 0.76 and 0.75, respectively (Extended Data Fig. 2e). Additionally, zLFC values of epegRNAs introducing the same mutation in these genes showed strong correlation with the top two ranked epegRNAs (Extended Data Fig. 2f).

Given the high MOI used in this screening, subtle phenotypic changes (low LFC) may lead to random epegRNA distributions, potentially introducing background noise and affecting overall eBAR parallelism analyses (Extended Data Fig. 3a). To more precisely assess the data quality, we sought to minimize the influence of potential experimental noise by analyzing the correlation between different eBARs representing replicates as the LFC threshold increased. This approach revealed progressively stronger zLFC correlations among the three eBAR replicates at higher LFC levels, highlighting the effectiveness of the ZFC-eBAR algorithm in capturing strongly correlated epegRNAs from raw data. To intuitively illustrate data reproducibility, we applied an absolute LFC threshold of ≥1 (Fig. 1d and Extended Data Fig. 3b), with direct LFC correlation analyses exhibiting comparable consistency (Extended Data Fig. 3c–e). Furthermore, reproducibility assessment of the enriched epegRNAs subset identified by our screening threshold (n = 2,134) revealed good correlations regardless of mutation type (synonymous or nonsynonymous mutations; Extended Data Fig. 3f,g). Collectively, these results confirm the reliability and robustness of our screening results under stringent filtering criteria.

Using the defined screening threshold, we identified 1,914 mutations impacting cell fitness, encompassing both depletion and enrichment trends (Fig. 1e and Supplementary Table 3). This dataset

https://doi.org/10.1038/s41587-025-02710-z

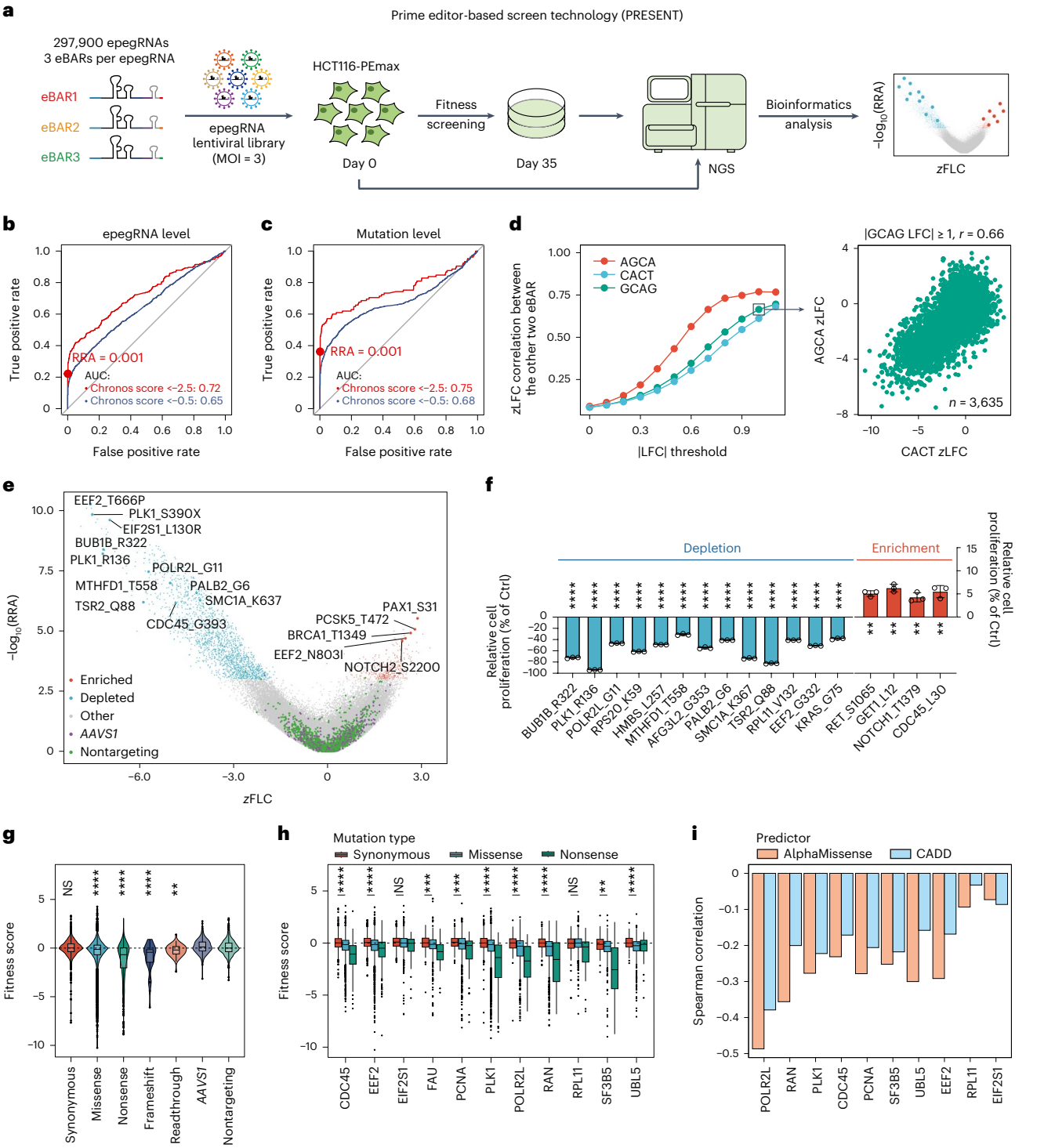


Fig. 1 | High-throughput screening unveils functional synonymous mutations in the human genome. a, Schematic of the PRESENT workflow. **b**, ROC analysis based on gene knockout effects at the epegRNA level. Red point, selected threshold. Red curve, epegRNAs targeting genes with a Chronos score <−2.5; dark-blue curve, epegRNAs targeting genes with a Chronos score <−0.5. **c**, ROC analysis based on gene knockout effects at the mutation level. Red point, selected threshold. Red curve, the highest-scoring epegRNA in the screen for each mutation targeting genes with a Chronos score <−2.5; dark-blue curve, the highest-scoring epegRNA in the screen for each mutation targeting genes with a Chronos score <−0.5. **d**, Pearson correlation analysis of zLFC values between two eBARs across varying thresholds of absolute LFC values from the third eBAR. The *x* axis indicates the absolute LFC threshold applied to each eBAR while the *y* axis represents the Pearson correlation coefficient between the remaining two eBARs. Left: specific data points enclosed within a square. Right: magnified view for a closer examination (*n* indicates the number of epegRNAs shown in the figure). **e**, Volcano plot illustrating the results of screening for functional synonymous and nonsynonymous mutations affecting cell fitness. **f**, Relative cell proliferation rates for validated mutations. Data are presented as the mean ± s.d. ($n = 3$ biological replicates). *P* values were calculated using a two-tailed Student's *t*-test. $^{**}P < 0.01$ and $^{****}P < 0.0001$. **g**,**h**, Distribution of fitness scores for the 11 genes designed by saturation mutagenesis, categorized by mutation type: combined (**g**) or shown separately by gene (**h**). *P* values were calculated using a two-sided Wilcoxon test with Benjamini–Hochberg correction. $^{**}P < 0.01$, $^{***}P < 0.001$ and $^{****}P < 0.0001$; NS, not significant. **i**, Spearman correlation between the top screen scores of missense mutations and prediction scores from AlphaMissense and CADD. For the box plots in **g** and **h**, the center line represents the median, the box limits denote the upper and lower quartiles and whiskers extend to 1.5 times the interquartile range.

underwent the aforementioned stringent quality control procedures to ensure its reproducibility. Of these, 409 were synonymous mutations and 1,505 were nonsynonymous mutations. To validate the reliability of the results, we randomly selected synonymous mutations above the threshold for experimental validation (Fig. 1f and Supplementary Fig. 1a–d). All tested mutations demonstrated precise and clean editing (Supplementary Fig. 2).

Additionally, we randomly selected 20 synonymous mutations below the threshold and measured their editing outcomes. These mutations generated pure editing products; however, their overall editing efficiency was lower compared to mutations exceeding the threshold (Supplementary Fig. 3). While the stringent threshold applied ensures a low false positive rate (Fig. 1b,c and Extended Data Fig. 2e), variations in prime-editing efficiency may result in some mutations with low editing rates or subtle phenotypic effects being overlooked.

To assess the sensitivity of the screening, we simulated the dropout process of deleterious mutations that impair cell proliferation and evaluated whether these mutations could be identified using our analysis method. The results indicate that, when the mean count decreased by 21%, some epegRNAs were detectable by the algorithm, aligning with the weakest LFC (~−0.4) observed in the screening. When the mean count was reduced by 50%, approximately half of the epegRNAs were detectable above the threshold. By relaxing the threshold to |screen score| = 2.2 (approximately 3 s.d.), our analysis method was able to detect epegRNA count changes as small as 10% (Extended Data Fig. 3h,i). These findings demonstrate that our analysis could identify functional mutations with reliable detection sensitivity.

Characterization of synonymous and nonsynonymous mutations

The inclusion of various mutation types in the library enables us to conduct a comprehensive characterization analysis on the basis of the screening results. In our screen, 0.43% of synonymous mutations demonstrated measurable effects on cell fitness, compared to 3.83% of nonsynonymous mutations. While the fitness impact of synonymous mutations did not significantly differ from that of the negative controls, nonsynonymous mutations, including missense, frameshift and nonsense types, exhibited clear effects on cell fitness (Fig. 1g). This indicates that, in general, most synonymous mutations in the human genome are likely neutral, diverging from results seen in yeast[2]. Significant fitness differences between synonymous and missense mutations were noted within the 11 genes targeted for saturated mutation design (Fig. 1h). The results of our analysis suggest that findings from yeast studies may not be directly applicable to the human genome. To further explore the differences between our screening results and those from yeast studies, we examined the distribution of saturated synonymous mutations across various gene sets: all 67 genes with saturated synonymous mutation designs, 11 highly essential genes in HCT116 and 12 human homologous genes of the yeast (*RPL39* was excluded as it was designed with only 11 epegRNAs, insufficient for statistical analysis). The distribution patterns of synonymous mutations in these datasets showed no significant differences, with all centered around a median of 0 (Extended Data Fig. 4a). This indicates that synonymous mutations in these human genes do not exhibit the fitness distribution shifts observed in yeast. Despite variations in the essentiality of these genes in HCT116 cells (Extended Data Fig. 4b), synonymous mutations consistently displayed a neutral effect compared to nonsense and frameshift mutations (Extended Data Fig. 4c).

Given that we performed full saturation mutagenesis on 11 genes, we also gathered extensive data to assess the biological effects of nonsynonymous mutations, such as missense mutations, in the human genome. Variant effect predictors, such as AlphaMissense[21], provide insights into the impact of missense mutations. To evaluate these, we compared missense mutations in the fully saturated genes with predictions from AlphaMissense and CADD[11]. Because AlphaMissense lacks predictions for *FAU*, the comparison was conducted on the remaining ten genes. Although these prediction results do not directly correspond to cell fitness, the Spearman correlations showed reasonable consistency, especially for highly essential genes (Fig. 1i). Further analysis of these nonsynonymous mutations demonstrated a direct relationship between the probability of amino acid substitution and the effect on cell fitness. Amino acids with lower substitution probabilities are more likely to exhibit more significant fitness impacts after mutation (Extended Data Fig. 5a).

We highlighted two specific cases to demonstrate the value of gene saturation design for studying different mutation types and key sites. RAN, a member of the small G-protein family with GTP hydrolase activity, showed a hotspot effect in the G3 switch II domain, with mutations impacting the enzyme active center Q69 being particularly prominent (Extended Data Fig. 5b). Narrowing the analysis to important structural domains and active sites revealed more pronounced biological effects of missense mutations (Extended Data Fig. 5c). Interestingly, the yeast homolog of RAN, Gsp1/Ran, has also been studied using saturation mutagenesis[22] but the correlation between our results and the yeast findings was relatively low (Extended Data Fig. 5d). Similarly, for POLR2L, our screen identified significant enrichment of mutations at the protein's Zn^{2+}-binding sites (Extended Data Fig. 5e)[23].

At the level of point mutations, missense mutations primarily disrupted protein function, consistent with general understanding. Nonsense mutations typically produced the most pronounced biological effects, with amino acid substitutions such as L > P, L > R and R > P frequently resulting in phenotypes (Extended Data Fig. 5f, g).

Although synonymous mutations in the human genome are predominantly neutral, the 409 enriched mutations identified in our study demonstrated biological effects on cell proliferation. Statistical analysis of these mutations revealed that the highest levels of enrichment occurred in alanine, glycine and leucine following synonymous changes. Nevertheless, the observed patterns differed when compared to those in clinical synonymous mutations (Extended Data Fig. 5h). Furthermore, C•G base pairs in these synonymous mutations were more likely to generate fitness effects after mutation compared to A•T base pairs, a trend consistent with both saturation synonymous mutations and clinical synonymous mutations (Extended Data Fig. 5i). This consistency exists regardless of the preferences associated with prime editing[24,25]. As mentioned earlier, these functional synonymous mutations were successfully validated through random selection (Fig. 1f). Over an 18-day period of continuous cell culture, significant phenotypic effects were observed for all synonymous mutations categorized under depletion, while those under enrichment showed relatively mild effects. On the basis of these observations, our subsequent research efforts concentrated on exploring the effects of depletion direction synonymous mutations.

DS Finder reveals deleterious synonymous mutation determinants

To elucidate the mechanisms behind deleterious synonymous mutations, we analyzed them from three perspectives, at the gene, mRNA and nucleotide levels. Deleterious mutations are more likely to cause aberrant splicing, disrupt RNA secondary structure and use infrequent codons. These mutations often occur in conserved nucleotides, highly expressed genes and essential genes (Extended Data Fig. 6a–f).

To identify the most influential features associated with these mutations, we developed a machine learning model named DS Finder (deleterious synonymous mutations finder) based on the CatBoost framework[26] and trained it as a binary classifier with our screening data (Fig. 2a). Using the extensive data generated from our screen, DS Finder demonstrated exceptional performance in predicting deleterious mutations within the given cellular context. It surpassed two existing models: CADD, which predicts general mutations[11], and SilVA,

https://doi.org/10.1038/s41587-025-02710-z

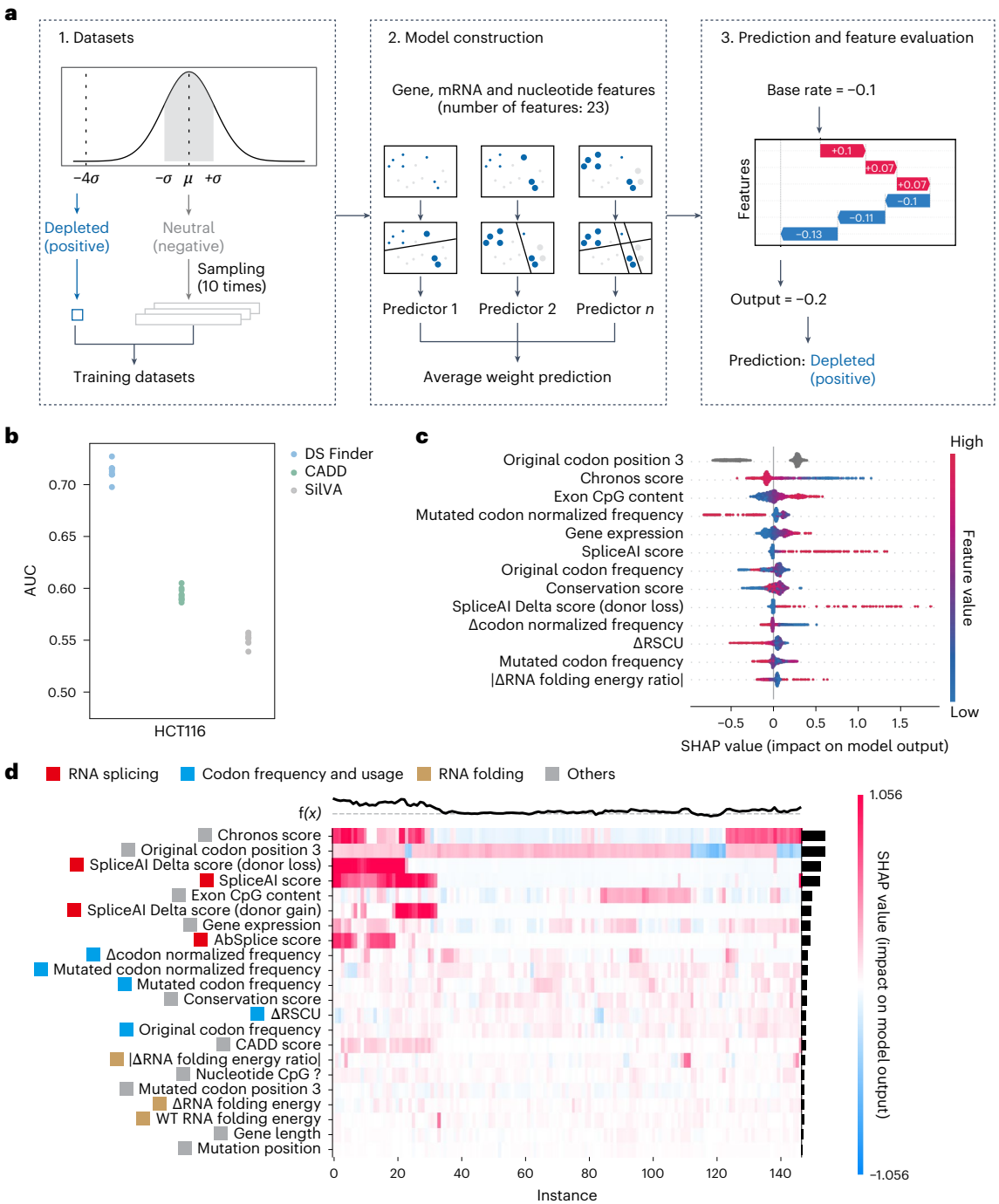


Fig. 2 | Machine learning model for analyzing determinants of deleterious synonymous mutations. a, Workflow diagram of DS Finder. The dataset was selected using a 10:1 ratio (neutral to deleterious). The model was developed using CatBoost and SHAP was used to determine feature importance. **b**, Performance of DS Finder in the HCT116 cell line compared with CADD and SilVA. Ten datasets were created by randomly sampling a 10:1 ratio of neutral to deleterious mutations, with this process repeated ten times. Each dataset was evaluated using tenfold cross-validation to obtain AUC values, represented by the ten points per group. **c**, The relative importance of various features in predicting the effects of synonymous mutations. **d**, Heat map of SHAP values, illustrating the supervised clustering method that categorizes data points on the basis of their feature explanations; f(*x*) corresponds to the predicted scores.

a model for synonymous mutations trained on a smaller dataset of 33 deleterious synonymous mutations[10] (Fig. 2b).

Key features identified by DS Finder include the third position of the codon being C•G or A•T base pairs, with C•G pairs being more likely to impact cell fitness (Extended Data Fig. 5i and Fig. 2c). Factors such as gene essentiality, expression levels and exon CpG content also have crucial roles (Fig. 2c), emphasizing the need to consider these features in the model as they vary across different tissues. Alterations in splicing, especially the loss of splice donor sites (Fig. 2c), likely represent the primary mechanism driving deleterious effects (Fig. 2d). Moreover, synonymous mutations can adversely affect cell fitness by altering codon usage and mRNA folding (Fig. 2c,d).

Altogether, our machine learning model shows that the deleterious effects of synonymous mutations are driven by a combination of factors including splicing disruptions, codon bias and nucleotide conservation.

Synonymous mutations generate aberrant splicing events

DS Finder identified disruption of splicing as a key mechanism through which synonymous mutations exert their deleterious effects. SpliceAI[27], a widely used tool for predicting erroneous splicing events, revealed that 23.13% of synonymous mutations in the depletion direction could be deleterious, affecting essential gene expression (Supplementary Fig. 4a). Most of these mutations alter splice donor sites, with a few mutations, such as KIF11_A133, introducing new splice acceptors in transcripts (Supplementary Fig. 4b–e). Mutations causing splice donor loss accounted for 12% of the total, followed by mutations that create new splice donors, which accounted for 7% (Fig. 3a,b).

Additionally, a new model called AbSplice was developed to predict aberrant splicing caused by mutations in specific human tissues[28]. Using AbSplice, we predicted new mutations, such as PLK1_R136 (AGG > AGA) and TSR2_Q88 (CAG > CAA), that could lead to aberrant splicing in colon tissue (Fig. 3c). Overall, nearly one quarter of the synonymous mutations from the depletion screening could potentially lead to aberrant splicing events.

We selected a subset of these mutations for validation, using epegRNAs with RTTs matching the reference sequence at each mutation site as controls to eliminate editing process effects. The BUB1B_R322 (AGG > AGA) mutation, a top-ranked synonymous mutation from the depletion results (Fig. 1e) located at the junction of exon 7 and intron 7 of *BUB1B*, disrupts the original splice donor site (donor loss type), resulting in intron retention (Fig. 3d). Interestingly, this mutation led to two abnormal transcript variants: one with complete intron 7 retention and another with a newly created splice donor site within intron 7, leading to a truncated intron retention (Fig. 3d). Both variants could lead to mRNA degradation because of the generated premature stop codon. We further validated the impact of this mutation on cell fitness and mRNA abundance (Fig. 3e–g). Furthermore, BUB1B_R322 (AGG > AGA) is also noted in ClinVar in persons with mosaic variegated aneuploidy syndrome 1 but labeled with 'uncertain significance', highlighting our screen can reveal synonymous mutations that may be misannotated in clinical databases. Another mutation, RPL11_V132 (GTG > GTC), disrupts a splice donor, causing retention of the intronic region and altering cellular phenotypes (Extended Data Fig. 7a–d).

EEF2_G332 (GGC > GGT) represents a different impact on splicing by generating a new splice donor within an exon (donor gain type). This new splice donor precedes the original, resulting in partial exon excision and a frameshift in the downstream sequence (Fig. 3h). This abnormal transcript significantly reduces cell fitness and mRNA levels (Fig. 3i–k). A similar case is observed in KRAS_G75 (GGG > GGT) (Extended Data Fig. 7e–h). These findings suggest synonymous mutations near exon–intron boundaries can induce aberrant splicing, affecting gene expression and cellular phenotypes.

Synonymous mutations impact RNA stability and translation

In addition to aberrant RNA splicing, synonymous mutations can impact RNA folding and translation within cells (Fig. 2d). A specific mutation, PLK1_S2 (AGT > AGC), found early in the coding sequence of the essential gene *PLK1*, enhanced RNA stability (Fig. 4a). Despite both codons encoding serine, AGC is used more frequently than AGT (codon frequencies of 19.5 per thousand versus 12.1 per thousand, respectively), implying that changes in cell fitness might stem from alterations in mRNA folding rather than codon usage. Predicted analysis of mRNA structure before and after mutation revealed that the PLK1_S2 (AGT > AGC) mutation enhanced the stability of the local mRNA structure near the start codon (Fig. 4b and Supplementary Fig. 5a–c).

To verify whether this mutation affected *PLK1* expression, we conducted a western blot analysis to measure PLK1 protein levels. Results indicated that the WT RTT had no impact on PLK1 protein levels, whereas the PLK1_S2 (AGT > AGC) mutation led to decreased protein abundance (Fig. 4c). Additionally, this mutation did not impact on neighboring codons either upstream or downstream of the mutation site (Extended Data Fig. 8a), and the observed decrease in cell fitness was solely attributable to this single synonymous mutation (Fig. 4d).

Considering that the mutation-induced stem structure disrupted the original loose structure near the start codon of WT PLK1 mRNA (Fig. 4b), potentially impeding translation initiation and conferring a translation disadvantage, we further used ribosome sequencing (Ribo-seq) to assess translation at this site. The results indicated that the PLK1_S2 (AGT > AGC) mutation made ribosome binding at the start codon more challenging, without affecting transcription levels (Fig. 4e and Extended Data Fig. 8b). The difficulty in initiating translation consequently led to reduced protein expression. This scenario underscores the relationship between changes in RNA folding induced by synonymous mutations and their effects on the translation process, ultimately influencing cellular functions through altered protein levels.

Single-cell sequencing reveals gene expression effects of synonymous mutations

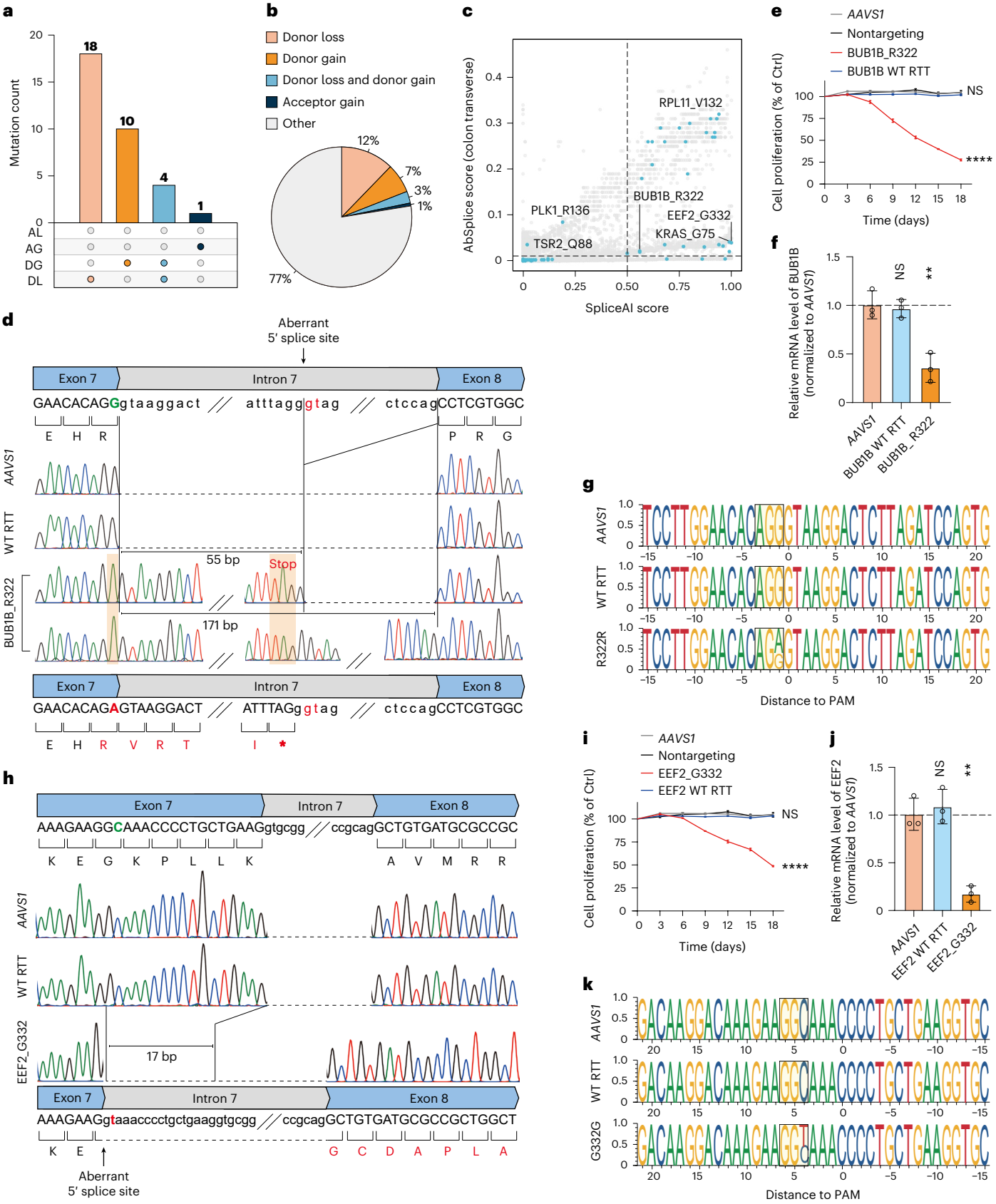
Through these studies, we elucidated potential mechanisms through which functional synonymous mutations influence biological processes. Recognizing prior studies suggesting that synonymous mutations can alter intracellular RNA abundance[2], we aimed to systematically assess the impact of synonymous mutations identified from our screening on gene expression at a high-throughput level. To achieve this, we integrated a single-cell screening approach with our PRESENT, calling it DIRECTED-seq (direct epegRNA capture and targeted sequencing). We constructed a DIRECTED-seq epegRNA library targeting each synonymous mutation using the PE to systematically investigate their effects on gene expression. We selected nearly all synonymous mutations that were either enriched or depleted, excluding those within genes expressed at low levels, resulting in a total of 370 mutations. Additionally, the library also included several negative controls: 15 nontargeting epegRNAs, 15 *AAVS1*-targeting epegRNAs and 10 epegRNAs targeting synonymous mutations that were not enriched.

HCT116-PEmax cells were transduced with this pooled library at a high MOI of about 5, followed by selection with puromycin.

Fig. 3 | Impact of synonymous mutations on aberrant RNA splicing. a,**b**, The statistical analysis of synonymous mutations causing various aberrant RNA splicing events is presented using bar charts and pie charts. Different colors indicate different splicing impact events based on the prediction by SpliceAI. AL, acceptor loss; AG, acceptor gain; DG, donor gain; DL, donor loss. **c**, Integration of predictions from SpliceAI and AbSplice tools. **d**, Schematic depiction of the splicing alterations caused by the BUB1B_R322 (AGG > AGA) mutation. The transcript sequence information was obtained by sequencing the cDNA from the experimental and control groups. **e**, Validation of the effect of the BUB1B_R322 (AGG > AGA) mutation on cell proliferation in HCT116 cells. **f**, Relative mRNA expression levels of *BUB1B* in experimental and control groups. The mRNA level of each sample was quantified by real-time qPCR and normalized by *GAPDH*, and the indicated relative mRNA level was normalized to that of *AAVS1*-targeting control cells. **g**, Analysis of editing outcomes for epegRNA targeting BUB1B_R322 and controls by genome sequence amplification and NGS. **h**, Schematic depiction of the splicing alterations caused by the EEF2_G332 (GGC > GGT) mutation. The transcript sequence information was obtained by sequencing the cDNA from the experimental and control groups. **i**, Validation of the effect of the EEF2_G332 (GGC > GGT) mutation on cell proliferation in HCT116 cells. **j**, Relative mRNA expression levels of *EEF2* in the experimental and control groups. The mRNA level of each sample was quantified by real-time qPCR and normalized by *GAPDH*, and the indicated relative mRNA level was normalized to that of *AAVS1*-targeting control cells. **k**, Analysis of editing outcomes for epegRNA targeting EEF2_G332 and controls by genome sequence amplification and NGS. Data are presented as the mean ± s.d. ($n = 3$ biological replicates for cell proliferation assay; $n = 3$ technical replicates for real-time qPCR). *P* values were calculated using a two-tailed Student's *t*-test. $^{**}P < 0.01$ and $^{****}P < 0.0001$.

https://doi.org/10.1038/s41587-025-02710-z

After 14 days of culture, we simultaneously captured the epegRNAs and transcriptomes from single cells (Fig. 5a). A custom primer designed from the evopreQ$_1$ motif achieved the reverse transcription of epegRNAs and the transcriptome library was generated using an oligo-dT reverse transcription primer (Extended Data Fig. 9a). Furthermore, bulk RNA sequencing (RNA-seq) was performed on cells from the same batch to evaluate the editing efficiency (Fig. 5a). The majority of epegRNAs were effective, showing an average editing efficiency of about 15% for mutations detectable at sufficient sequencing depth (Fig. 5b).



https://doi.org/10.1038/s41587-025-02710-z

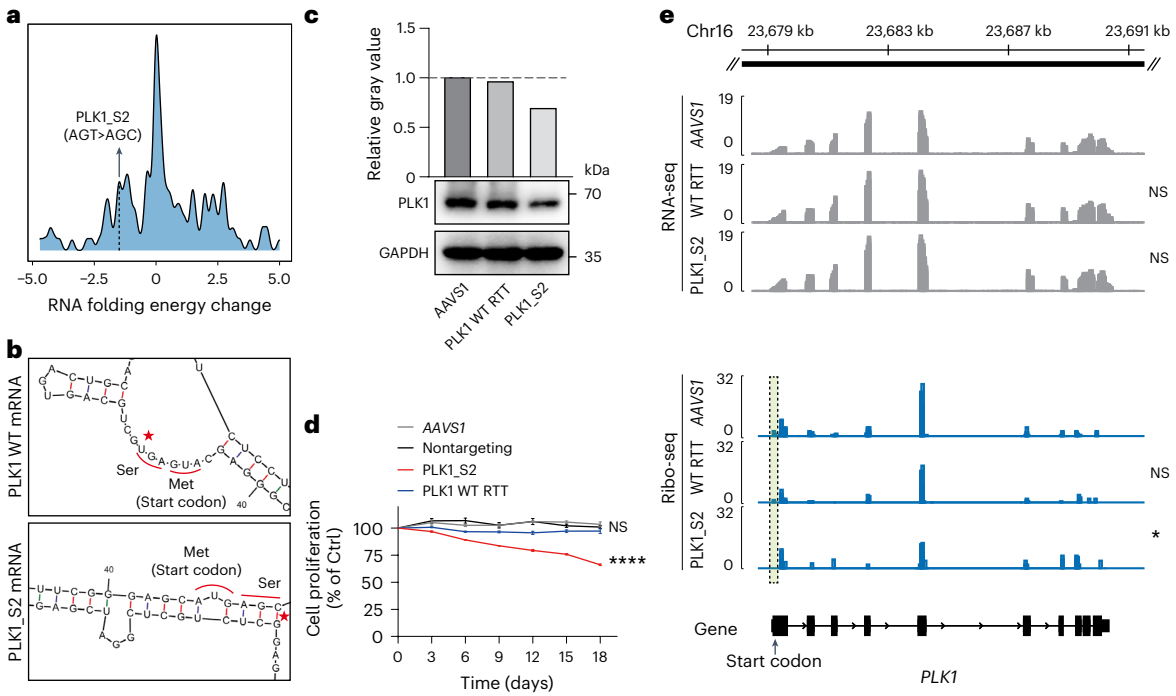


Fig. 4 | Influence of synonymous mutations on protein translation through RNA folding. a, Mapping of RNA folding energy changes for synonymous mutations in the depletion direction. **b**, Comparison of local mRNA structure of *PLK1* before and after the PLK1_S2 (AGT > AGC) mutation, as predicted using RNA folding form[49]. The pentagram symbol indicates the mutation site. **c**, Western blot analysis comparing PLK1 protein levels in PLK1_S2 mutant HCT116 cells versus control cells. The vertical axis of the upper bar chart represents the relative gray value of the PLK1 protein band. **d**, Validation of the effect of the PLK1_S2 (AGT > AGC) mutation on cell proliferation in HCT116 cells. Data are presented as the mean ± s.d. ($n = 3$ biological replicates). *P* values were calculated using a two-tailed Student's *t*-test. ****$P < 0.0001$. **e**, RNA-seq and Ribo-seq analyses of PLK1_S2 mutant HCT116 cells and control cells, with two biological replicates per group. The positional relationship between peaks and corresponding exons is shown, with the chromosomal region of *PLK1* labeled above. The dotted box indicates the peaks at the start codon. Gray represents the RNA-seq result peak chart and blue represents the Ribo-seq result peak chart. *P* values were calculated using two-tailed Student's *t*-test. *$P < 0.05$.

We recovered 129,193 single cells, averaging 5.5 epegRNAs per cell, with each targeted synonymous mutation represented in approximately 1,692 cells (Fig. 5c,d). Consistent with previous findings[29], a positive correlation was observed between the number of epegRNAs and unique molecular identifiers per cell (Extended Data Fig. 9b), suggesting sequencing depth as a potential confounding factor. To address this, we used a conditional resampling approach (SCEPTRE) for our differential expression analysis[29]. Nontargeting epegRNAs served as negative controls, showing no observable effects (Extended Data Fig. 9c,d). Ultimately, we identified 40 synonymous mutations that significantly influenced gene expression, at a 10% false discovery rate (FDR) (Fig. 5e and Supplementary Table 4).

A notable finding was the substantial decrease in *EEF2* gene expression caused by the synonymous mutation EEF2_G332 (GGC > GGT), where expression dropped to 32% of its original level ($\log_2$FC = −1.63, adjusted *P* value = 3.39×10^{-248}). Given EEF2's essential role in cell fitness, this significant reduction correlates with the mutation's presence in the depletion direction of the screen. Similar effects were observed with synonymous mutations in other essential genes. For instance, substitution of GTG with GTC, GTT and GTA at RPL11_V132 ($\log_2$FC = −0.488, −0.312 and −0.401, respectively, with adjusted *P* values of 2.52×10^{-141}, 2.36×10^{-135} and 8.33×10^{-77}, respectively) and substitution of AGG with AGA at BUB1B_R322 ($\log_2$FC = −0.702, adjusted *P* value = 5.16×10^{-11}) significantly lowered gene expression, which corresponded with proven aberrant splicing events (Fig. 3d–k and Extended Data Fig. 7a–d). Conversely, some synonymous mutations, such as BRCA1_V863 (GTT > GTG), were found to increase gene expression ($\log_2$FC = 0.389, adjusted *P* value = 0.0713). Given BRCA1's critical roles in DNA damage repair, cell-cycle checkpoint control and genomic stability maintenance, these mutations might substantially impact cellular functions and tumor suppression mechanisms (Fig. 5f and Supplementary Table 4).

These results from DIRECTED-seq effectively link specific mutations to changes in gene expression, unveiling their potential biological impacts. Although changes in transcript abundance could be driven by several mechanisms, such as misregulated RNA splicing, not all are linked to such changes. For instance, BRCA1_V863 (GTT > GTG) may affect RNA stability independently of splicing disruptions (Supplementary Table 5). Mutations could destabilize the global RNA structure, leading to increased transcript degradation; conversely, they may enhance the stability of transcripts[30]. Additionally, transcripts containing infrequently used codons during the early stage of translation elongation, referred to as the ramp, may also be subjected to degradation because of the slow translation speed of such transcripts[31,32]. These potential impacts highlight the need for further studies to explore how mutations influence transcript abundance and their broader biological effects.

Identifying new disease-linked synonymous mutations

We systematically studied functional synonymous mutations in our library using PRESENT and DIRECTED-seq. However, relying solely on screening and experimentation to uncover potential deleterious synonymous mutations in clinical databases has its limitations. To overcome this, we used our novel machine learning model, DS Finder, trained on our screening data, to identify novel functional clinical mutations across a broader dataset (Fig. 6a). We focused on mutations related to colonic diseases in the ClinVar database, such as lactose intolerance and inflammatory bowel disease, which share similar genetic backgrounds

https://doi.org/10.1038/s41587-025-02710-z

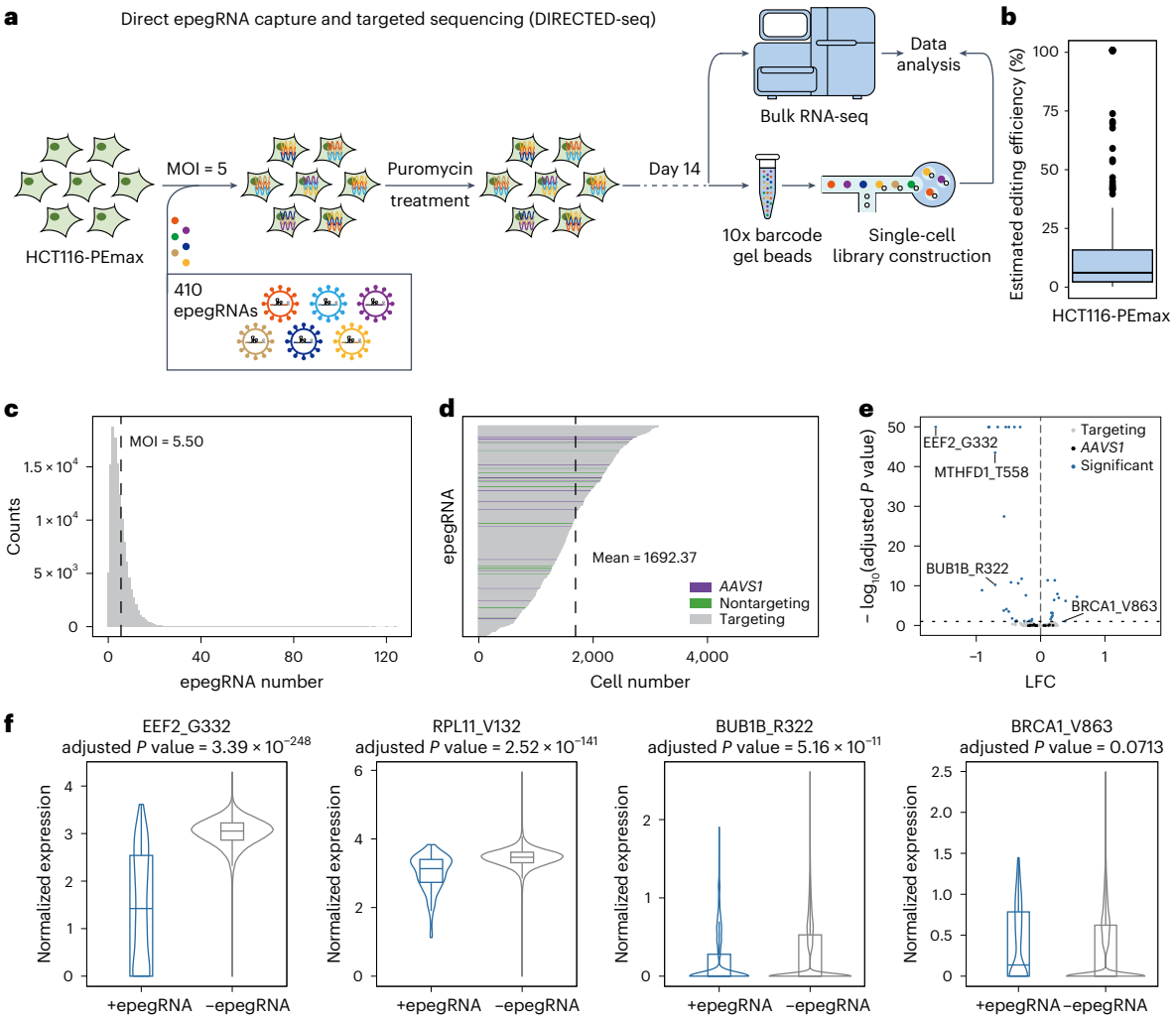


Fig. 5 | Investigating the impact of synonymous mutations on gene expression using DIRECTED-seq. a, Workflow diagram for DIRECTED-seq, conducted in HCT116-PEmax cells. Cells were transduced with the epegRNA library at an MOI of 5 and subsequently treated with puromycin. On day 14, both bulk RNA-seq and single-cell sequencing were performed. **b**, Estimated editing efficiency at sites where mutations were predicted to be detectable using bulk RNA-seq data (n = 207 mutation sites with nonzero reads in sequencing). **c**, Histogram showing the distribution of the number of epegRNAs per cell, with the mean value indicated by a dashed line. **d**, Distribution of the number of cells detected for each epegRNA, categorized by coverage. The dashed line represents the mean value, with purple, green and gray lines representing *AAVS1*-targeting, nontargeting and mutation-targeting epegRNAs, respectively. **e**, Volcano plot displaying the effects of synonymous mutations on the expression of corresponding genes. **f**, Single-cell gene expression profiles for cells harboring epegRNAs targeting EEF2_G332, RPL11_V132, BUB1B_R322 and BRCA1_V863. From left to right, the values of n (number of cells expressing or not the corresponding epegRNA) for each figure are 142 and 108,225 (EEF2_G332), 158 and 108,160 (RPL11_V132), 559 and 108,056 (BUB1B_R322) and 224 and 108,755 (BRCA1_V863), respectively. For **e** and **f**, P values were calculated using the SCEPTR method and adjusted with the Benjamini–Hochberg correction. For the box plots, the center line represents the median, the box limits denote the upper and lower quartiles and whiskers extend to 1.5 times the interquartile range.

to the HCT116 cell line used in our studies. Using DS Finder, we scored 585 clinically recorded synonymous mutations, most of which were previously annotated as benign or likely benign (Supplementary Table 6).

One notable example is G6PC3 c.G399A, associated with autosomal recessive severe congenital neutropenia. Although annotated as 'likely benign' in ClinVar, this mutation received the highest DS Finder score in our analysis and was supported by SilVA and CADD (Fig. 6b and Supplementary Table 6). This mutation disrupts RNA splicing, impairing the gene's normal function in a manner consistent with the disease mechanism (Fig. 6c). To validate the pathogenic potential of predicted synonymous mutations, we tested G6PC3 c.G399A in HCT116 cells. This mutation produced an incorrect transcript (Fig. 6d) and significantly reduced *G6PC3* expression (Fig. 6e).

To determine DS Finder's predictive thresholds and test its discriminatory power, we selected 45 confirmed pathogenic mutations from the SilVA training set[10] and other clinically relevant synonymous mutations with reported pathogenic effects (Supplementary Table 6). Alternative substitutions for these 45 mutations (e.g., if a pathogenic mutation was C > T, we also included C > G and C > A as controls) were included alongside 1,439 synonymous mutations without phenotypes from our screening as negative controls. DS Finder effectively distinguished among these three groups of mutations (Fig. 6f), particularly unreported alternative substitutions at the same site, demonstrating the sensitivity of our model.

Using the screening's negative control group as a benchmark, a DS Finder score threshold of 0.138 yielded a recall of 46.7% (21/45) at a false positive rate of 5%, which was on par with other methods (Extended Data Fig. 10a,b). On the basis of this threshold, 31 potentially deleterious synonymous mutations were predicted. This threshold provides a high recall rate, allowing mutations above this threshold to be classified as potentially pathogenic, while those below it are considered likely benign. When the threshold was increased to 0.370, the false positive

https://doi.org/10.1038/s41587-025-02710-z

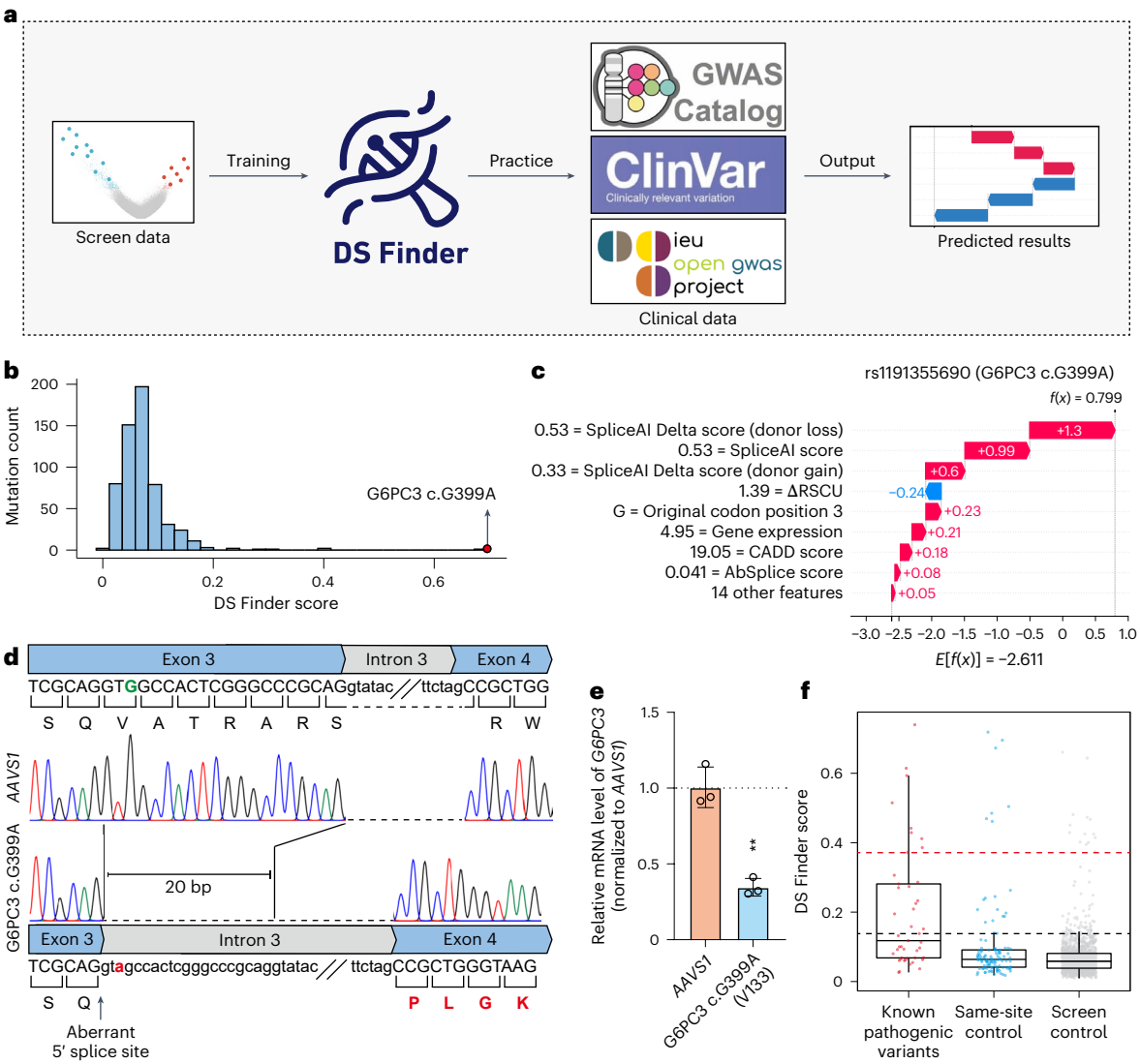


Fig. 6 | Predicting deleterious synonymous mutations in clinical databases. **a**, Flowchart illustrating the process of training the DS Finder using screening data and its application to clinical data for identifying novel deleterious synonymous mutations. **b**, Prediction scores from DS Finder for synonymous mutations associated with colonic diseases. The horizontal axis represents the DS Finder score, while the vertical axis represents the number of mutations corresponding to each score. **c**, SHAP value waterfall plots for the top prediction result, G6PC3 c.G399A, highlighting how individual features influence the prediction outcomes. The black values on the left are the feature values, while red and blue bars represent features with positive and negative effects on the predicted deleterious synonymous mutations, respectively. The numbers next to the bars indicate the magnitude of each feature's influence. **d**, Schematic depiction of the splicing alterations caused by the G6PC3 c.G399A mutation. The transcript sequence information was obtained by sequencing the cDNA from the experimental and control groups. **e**, Relative mRNA expression levels of *G6PC3* in the experimental and control groups. The mRNA level of each sample was quantified by real-time qPCR and normalized by *GAPDH*, and the indicated relative mRNA level was normalized to that of *AAVS1*-targeting control cells. Data are presented as the mean ± s.d. ($n = 3$ technical replicates). The *P* value was calculated using a two-tailed Student's *t*-test. $^{**}P < 0.01$. **f**, DS Finder scores for known pathogenic variants, other substitutions at the same site of known pathogenic variants and nonphenotypic mutation controls from the screening. The black dashed line represents the selected threshold with a false positive rate of 5% and the red dashed line represents the selected threshold with a false positive rate of 1%. For the box plots, the center line represents the median, the box limits denote the upper and lower quartiles and whiskers extend to 1.5 times the interquartile range.

rate dropped to 1%, resulting in predictions with greater accuracy and reliability classified as likely pathogenic. At this stricter threshold, three mutations were identified as likely pathogenic, G6PC3 c.G399A, ADAM17 c.C1632T and ADAM17 c.C315T (Supplementary Table 6).

Encouraged by these results and to expand the scope beyond colonic diseases, we extended our study to other tissues and organs. Considering the importance of genetic background, we applied PRESENT to A549 and KYSE-30 cell lines using the same epegRNA^{eBAR} library (Extended Data Fig. 10c–f and Supplementary Tables 7 and 8) and trained a prediction model (Extended Data Fig. 10g). The analysis of the screening data and the evaluation of model performance validated the reliability of the results, providing additional insights into synonymous mutations. Notably, consistent with findings in HCT116 cells, essential synonymous mutations constituted a small proportion in other cell lines (1.04% of A549 and 0.56% of KYSE-30). This was particularly evident in the context of inhibiting cell proliferation, where nonsynonymous mutations continued to dominate.

Additionally, we launched a website called Hearing Silence (https://search-synonymous-mutations.streamlit.app/), enabling open access to screening results from three cell lines and offering the DS Finder algorithm for free use. Overall, our model effectively identified novel pathogenic synonymous mutations in clinical databases, showing

potential for broad generalizability across diverse genetic disease backgrounds. Our findings also suggest that current clinical databases may underreport deleterious synonymous mutations, underscoring the value of predictive models for more precise annotations. This research underscores the potential of exploring additional functional synonymous mutations within the human genome, which is vital for understanding disease etiology and delving deeper into their molecular genetic mechanisms.

Discussion

In this study, we developed a high-throughput screening technique called PRESENT, using PEs to investigate functional synonymous mutations within the human genome. Overall, our findings indicate that the majority of synonymous mutations in the human genome are likely neutral, including those occurring in essential genes. This aligns with the prevailing understanding among geneticists, further supported by recent human genetics research[33]. Similar conclusions can also be drawn from screening data generated using base-editing approaches[34,35]. This contrasts with previous studies in yeast, where synonymous mutations appeared more consequential. Our library, which includes human homologs of the genes studied in yeast, did not show significant enrichment of these mutations, potentially because of differences between haploid yeast and diploid human cells and the greater complexity and larger intron regions in the human genome[36]. The low correlation between our saturation analysis results for RAN and those from a similar study in yeast[22] may be partially attributed to differences in the genetic backgrounds of the two species. Nevertheless, mutation data from the yeast Gsp1/Ran protein still suggest that synonymous mutations are generally neutral (Extended Data Fig. 5d). Therefore, we propose that the non-neutral effects of synonymous mutations observed in yeast cannot be broadly extrapolated to the human genome, where most synonymous mutations are likely to be neutral or nearly neutral.

Our study effectively addresses and avoids shortcomings associated with previous work[3]. For our large-scale screening, we implemented a library design that included appropriate positive and negative controls, with replicates established using the eBAR methodology. For validation, we used *AAVS1*-targeting and nontargeting epegRNAs as negative controls and specifically designed WT RTT epegRNAs to eliminate the influence of the PE process on the results. Moreover, our use of PE without relying on additional sgRNA to produce nicks greatly minimizes the risk of indels and genomic structural variations often associated with Cas9-based mutation methods. In terms of statistical power, our library includes a larger number of genes, mutations and mutation types compared to the yeast study. As a result, our large-scale screening is well equipped to uncover the genuine biological effects of synonymous mutations in the human genome.

Through comprehensive population-level screening incorporating multiple embedded controls and stringent quality control analyses, we identified and further characterized 409 synonymous mutations that impact cell fitness in HCT116 cells. Among these synonymous mutations, changes involving C•G base pairs were particularly impactful on cell fitness. This highlights the importance of codons ending in C or G (also named GC3) in genomic structure and their specific responses to synonymous mutations[37]. These mutations may affect CpG content within gene sequences, a feature incorporated into our machine learning model, DS Finder. Additionally, high-frequency codons often are GC3, which may relate to the stability of anticodon pairing during translation. The pronounced biological effects of GC3 mutations suggest that they are subject to negative selection, maintaining genomic stability over evolution time.

Our observations also show that synonymous mutations can alter gene expression levels, echoing findings from yeast studies[2]. These effects were systematically analyzed in human cells using DIRECTED-seq, primarily because of aberrant splicing events on mRNA. Dominant effects were observed at splice donor sites, likely because of the more conserved sequences upstream of these sites compared to downstream of splice acceptor sites[38]. Whether predicted or identified by DS Finder, the impact on RNA splicing is a substantial contributor to the biological effects of synonymous mutations. One possible explanation is that synonymous mutations undergo evolutionary selection closely tied to the transcription process, helping to prevent the generation of unwanted transcripts within the cell[39]. When gene expression is not altered, translation-coupled biological mechanisms may have a role. For example, we observed that increased mRNA stability near the start codon could impede translation initiation, reducing protein expression levels, consistent with phenomena observed in prokaryotes[40] and various genomic studies[41]. Changes in protein levels might also be linked to variations in codon usage frequency. In this study, we outlined a summary of possible explanations. While there are many other potential biological mechanisms through which synonymous mutations could affect human cell functions, further experimental investigations are necessary.

Additionally, we developed DS Finder, a prediction model for deleterious synonymous mutations. DS Finder's training set is more extensive than that of SilVA[10], which includes only 33 experimentally verified deleterious mutations and offers greater specificity than CADD[11], which predicts all mutation types. We demonstrated that DS Finder effectively distinguishes between known pathogenic mutations and mutations without phenotypes, further emphasizing its strong sensitivity. DS Finder also considers cell type, tissue type and gene background in its predictions, enhancing its accuracy and use in clinical data analysis, making it invaluable for future clinical research and diagnostics.

However, we recognize certain inherent limitations in our screening approach. Firstly, the scarcity of sufficient synonymous mutations known to influence cell fitness in prior genetic research poses challenges in establishing robust positive controls, complicating the precise quantification of potential false negatives in the screen. Secondly, despite optimizing editing efficiency by using naturally MMR-deficient HCT116 cells with epegRNA and extending the screening duration to 35 days, the overall editing level of PE still somewhat restricts detection sensitivity. The stringent threshold selected for this screening was designed to ensure a low false positive rate, but inherently low editing efficiency means that mutations with subtle fitness impacts may remain undetected. Therefore, mutations with less pronounced effects might be overlooked and should be further explored in subsequent studies. Adjusting the detection threshold could potentially increase sensitivity and recall rates. In addition, we emphasize that population-level screening alone cannot conclusively characterize the effects of individual synonymous mutation without experimental validation. Nevertheless, these limitations do not affect the statistical validity of our group-level analyses, as these comparisons systematically evaluated fitness effects across all mutation types under uniform detection conditions. Future implementation of more efficient PEs, such as the recently reported PE7 system[42], may further improve sensitivity for detecting subtle phenotypic variants.

Overall, this study provides new insights and experimental evidence on the impact of synonymous mutations in the human genome and their potential biological mechanisms. It underscores the precision of the PE over traditional CRISPR–Cas and base editors for high-throughput genomic studies[43–48]. Looking ahead, PRESENT and DIRECTED-seq provide useful tools in characterizing and exploring mechanisms behind clinical drug-resistant mutations and other genetic phenomena.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-025-02710-z.

References

1. Kimura, M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**, 275–276 (1977).
2. Shen, X., Song, S., Li, C. & Zhang, J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature* **606**, 725–731 (2022).
3. Kruglyak, L. et al. Insufficient evidence for non-neutrality of synonymous mutations. *Nature* **616**, E8–E9 (2023).
4. Cuevas, J. M., Domingo-Calap, P. & Sanjuan, R. The fitness effects of synonymous mutations in DNA and RNA viruses. *Mol. Biol. Evol.* **29**, 17–20 (2012).
5. Kristofich, J. et al. Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme. *PLoS Genet.* **14**, e1007615 (2018).
6. Walsh, I. M., Bowman, M. A., Soto Santarriaga, I. F., Rodriguez, A. & Clark, P. L. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl Acad. Sci. USA* **117**, 3528–3534 (2020).
7. Lebeuf-Taylor, E., McCloskey, N., Bailey, S. F., Hinz, A. & Kassen, R. The distribution of fitness effects among synonymous mutations in a gene under directional selection. *eLife* **8**, e45952 (2019).
8. Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12**, 683–691 (2011).
9. Supek, F., Minana, B., Valcarcel, J., Gabaldon, T. & Lehner, B. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* **156**, 1324–1335 (2014).
10. Buske, O. J., Manickaraj, A., Mital, S., Ray, P. N. & Brudno, M. Identification of deleterious synonymous variants in human genomes. *Bioinformatics* **29**, 1843–1850 (2013).
11. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
12. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
13. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
14. Anzalone, A. V. et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
15. Nelson, J. W. et al. Engineered pegRNAs improve prime editing efficiency. *Nat. Biotechnol.* **40**, 402–410 (2022).
16. Chen, P. J. et al. Enhanced prime editing systems by manipulating cellular determinants of editing outcomes. *Cell* **184**, 5635–5652 (2021).
17. Zhu, S. et al. Guide RNAs with embedded barcodes boost CRISPR-pooled screens. *Genome Biol* **20**, 20 (2019).
18. Zhu, S. et al. Genome-wide CRISPR activation screen identifies candidate receptors for SARS-CoV-2 entry. *Sci. China Life Sci.* **65**, 701–717 (2022).
19. Sharma, Y. et al. A pan-cancer analysis of synonymous mutations. *Nat. Commun.* **10**, 2569 (2019).
20. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
21. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
22. Mathy, C. J. P. et al. A complete allosteric map of a GTPase switch in its native cellular network. *Cell Syst.* **14**, 237–246 (2023).
23. He, Y. et al. Near-atomic resolution visualization of human transcription promoter opening. *Nature* **533**, 359–365 (2016).
24. Kim, H. K. et al. Predicting the efficiency of prime editing guide RNAs in human cells. *Nat. Biotechnol.* **39**, 198–206 (2021).
25. Mathis, N. et al. Predicting prime editing efficiency and product purity by deep learning. *Nat. Biotechnol.* **41**, 1151–1159 (2023).
26. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. CatBoost: unbiased boosting with categorical features. In *Proc 32nd International Conference on Neural Information Processing Systems* (eds Bengio, S. et al.) 6639–6649 (NIPS, 2018).
27. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
28. Wagner, N. et al. Aberrant splicing prediction across human tissues. *Nat. Genet.* **55**, 861–870 (2023).
29. Barry, T., Wang, X., Morris, J. A., Roeder, K. & Katsevich, E. SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. *Genome Biol* **22**, 344 (2021).
30. Nackley, A. G. et al. Human catechol-*O*-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* **314**, 1930–1933 (2006).
31. Tuller, T. et al. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344–354 (2010).
32. Bae, H. & Coller, J. Codon optimality-mediated mRNA degradation: linking translational elongation to mRNA stability. *Mol. Cell* **82**, 1467–1476 (2022).
33. Dhindsa, R. S. et al. A minimal role for synonymous variation in human disease. *Am. J. Hum. Genet.* **109**, 2105–2109 (2022).
34. Cuella-Martin, R. et al. Functional interrogation of DNA damage response variants with base editing screens. *Cell* **184**, 1081–1097 (2021).
35. Hanna, R. E. et al. Massively parallel assessment of human variants with base editor screens. *Cell* **184**, 1064–1080 (2021).
36. Dujon, B. Basic principles of yeast genomics, a personal recollection. *FEMS Yeast Res.* **15**, fov047 (2015).
37. Chamary, J. V., Parmley, J. L. & Hurst, L. D. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**, 98–108 (2006).
38. Liu, Y. et al. Genome-wide screening for functional long noncoding RNAs in human cells by Cas9 targeting of splice sites. *Nat. Biotechnol.* **36**, 1203–1210 (2018).
39. Radrizzani, S., Kudla, G., Izsvak, Z. & Hurst, L. D. Selection on synonymous sites: the unwanted transcript hypothesis. *Nat. Rev. Genet.* **25**, 431–448 (2024).
40. Kudla, G., Murray, A. W., Tollervey, D. & Plotkin, J. B. Coding-sequence determinants of gene expression in Escherichia coli. *Science* **324**, 255–258 (2009).
41. Gu, W., Zhou, T. & Wilke, C. O. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.* **6**, e1000664 (2010).
42. Yan, J. et al. Improving prime editing with an endogenous small RNA-binding protein. *Nature* **628**, 639–647 (2024).
43. Ren, X. et al. High-throughput PRIME-editing screens identify functional DNA variants in the human genome. *Mol. Cell* **83**, 4633–4645 (2023).
44. Gould, S. I. et al. High-throughput evaluation of genetic variants with prime editing sensor libraries. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02172-9 (2024).
45. Cirincione, A. et al. A benchmarked, high-efficiency prime editing platform for multiplexed dropout screening. *Nat. Methods* **22**, 92–101 (2025).
46. Belli, O., Karava, K., Farouni, R. & Platt, R. J. Multimodal scanning of genetic variants with base and prime editing. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02439-1 (2024).
47. Kim, Y., Oh, H. C., Lee, S. & Kim, H. H. Saturation profiling of drug-resistant genetic variants using prime editing. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02465-z (2024).

48. Herger, M. et al. High-throughput screening of human genetic variants by pooled prime editing. *Cell Genom.* **5**, 100814 (2025).
49. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).

Methods

Cell lines and cell culture

The HCT116 cell line (CCL-247, American Type Culture Collection (ATCC)) and HEK293T cell line (CRL-3216, ATCC) were obtained from EdiGene, the A549 cell line was purchased from ATCC (CRM-CCL-185) and the KYSE-30 cell line (ACC 351, German Collection of Microorganisms and Cell Cultures) was obtained from Z. Liu's laboratory at Peking Union Medical College. The HCT116-PEmax, A549-PEmax and KYSE-30-PEmax cell lines were generated in this study. HCT116, HCT116-PEmax, KYSE-30 and KYSE-30-PEmax cells were cultured in RPMI 1640 medium (Gibco, C11875500BT). HEK293T, A549 and A549-PEmax cells were cultured in DMEM (Gibco, C11995500BT). All cultures were supplemented with 10% FBS (Biological Industries, 04-001-1ACS) and 1% penicillin–streptomycin. Cultures were maintained at 37 °C in a 5% CO_2 environment and regularly checked for *Mycoplasma* contamination using the *Mycoplasma* detection kit (Yeasen, 40612ES60).

Plasmids

The pLenti-PEmax-P2A-EGFP expression plasmid was constructed by cloning the PEmax-P2A-EGFP sequence from pCMV-PEmax-P2A-GFP (Addgene, 180020). The target sequence was amplified by PCR using *TransStart FastPfu* fly DNA polymerase (TransGen Biotech, AP231-22) during the cloning procedure. The epegRNA expression vectors (pLenti-U6-epegRNA-SV40-puro or pLenti-U6-epegRNA-SV40-mCherry) were derived from the pLenti-sgRNA(lib)-puro vector (Addgene, 119976). All candidate epegRNAs for validation (Supplementary Table 9) were cloned into the vector by Golden Gate assembly and all plasmids were verified through Sanger sequencing by Tsingke Biotech (Beijing).

Design and construction of the epegRNA library

Library design. The design of epegRNAs involved identifying all potential spacers from protein-coding genes, followed by generating and adjusting RTTs on the basis of genomic annotation to incorporate the desired mutations and then designing linker sequences. The GRCh38.p14 genome and MANE[50] transcript annotations were downloaded from the National Center for Biotechnology Information (NCBI), with the search scope including entire transcripts and 30 nt upstream and downstream of exons. Quality control was based on three criteria: (1) G+C content between 20% and 80%; (2) absence of poly(T) sequences; and (3) unique genomic alignment to prevent potential off-target effects, analyzed using the Bowtie software (version 1.2.1.1)[51] with the parameters set to '-k 2 -v 0'. Following spacer design, PBS and RTT lengths were determined. Given that a 13-nt PBS length is associated with optimal efficiency, this was the chosen length[24]. Considering coverage and efficiency, the initial RTT length was set to 17 nt, subject to trimming if the sequence commenced with a C, accepting a minimum RTT length of 13 nt. At the design's inception, all synonymous mutations were considered, with adjustments made to the original RTT to accommodate all possible synonymous variations by moving along the codon. The design principle for linkers aimed to minimize secondary structural interactions with other components, accomplished using the pegLIT algorithm (version 1.0.1)[15]. Sequences containing BsmBI sites were excluded.

Six epegRNAs were designed for each gene to achieve knockout, either by replacing a codon with TAG at 30%, 40% and 50% of the gene's coding region length or by inserting an A to induce a premature stop codon or frameshift mutation. Negative controls included nontargeting and *AAVS1*-targeting epegRNAs.

Library construction. Oligonucleotides were synthesized by Synbio Technologies and epegRNA sequences were PCR-amplified using Phusion Hot Start Flex 2× master mix (New England Biolabs, M0536S) with primers targeting the flanking sequences of the oligos. After purification with a DNA clean and concentrator 25 kit (Zymo Research, D4034), epegRNA sequences were respectively cloned into three types of pLenti-U6-epegRNA^{eBAR}-SV40-puro vectors through Golden Gate assembly (eBAR1, CACT; eBAR2, GCAG; eBAR3, AGCA). The Golden Gate product of each group was separately purified with a DNA Clean and Concentrator 5 kit (Zymo Research, D4014) and electroporated into *E. coli* HST08 Premium Electro-Cells (TaKaRa, 9028). The plasmids of each epegRNA^{eBAR} library were extracted using EndoFree Plasmid Maxi Kit (Qiagen, 12362) and further mixed in a 1:1:1 molar ratio. Next, the epegRNA^{eBAR} library plasmids were cotransfected with lentiviral packaging plasmids pMD2.G (Addgene, 12259) and pCMVR8.74 (Addgene, 22036) into HEK293T cells using the X-tremeGENE HP DNA transfection reagent (Roche, 6366244001) to generate the lentiviral library.

Functional screening of synonymous mutations

HCT116-PEmax, A549-PEmax and KYSE-30-PEmax cells were transduced with the lentiviral library at a high MOI of 3 with a high coverage for each epegRNA (1,500-fold; 500-fold for each eBAR). Then, 48 h after transduction, the library cells were cultured with 1 μg ml^{-1} (KYSE-30-PEmax), 1.5 μg ml^{-1} (A549-PEmax) or 2 μg ml^{-1} (HCT116-PEmax) puromycin (Solarbio, P8230) for 2 days. The day that puromycin treatment ended was denoted as day 0 of the screening and some of the viable cells were harvested as the reference group. During the screening, one library size of cells was maintained and passaged every 3 days and harvested on day 35 as the experimental group.

Genomic DNA isolation and sequencing

Genomic DNA was extracted from reference cells and experimental cells using the DNeasy blood and tissue kit (Qiagen, 69506). All extracted genomes were used as PCR templates and the epegRNA sequences with eBAR were PCR-amplified using a KAPA HiFi HotStart ReadyMix PCR kit (Roche, KK2631) with five pairs of primers (Supplementary Table 10). Then, the PCR products were mixed together and purified with a DNA clean and concentrator 25 kit (Zymo Research, D4034), followed by next-generation sequencing (NGS) with paired-end 150-bp reads on the Illumina HiSeq X TEN platform. The three different cell libraries were performed separately as above.

Computational analysis of screens

The first step was to extract information from sequencing data. PANDAseq (version 2.11)[52] was used to assemble two FASTQ files, capturing the target sequences based on the sequences at both ends of the PCR product. To further improve the success rate of extraction, we relaxed the search requirements in the presence of mismatches, using BLASTn (version 2.11.0+)[53] to search for scaffold sequences and structural motif sequences, and then extracted the corresponding sequences after determining the coordinates. The sequences extracted in the first two steps were combined and analyzed statistically.

In the second step of statistical analysis, improvements were made on the basis of an algorithm previously developed by our laboratory[54]. On the basis of the ZFC algorithm, each eBAR was treated as an independent experiment and adopted a smoother fitting method (locally weighted scatterplot smoothing regression)[55]. To elevate the statistical level from epegRNA^{eBAR} to epegRNA, RRA was used for rank aggregation analysis as an indicator of significance statistics. This method was referred to as ZFC-eBAR.

The screen score and fitness score were defined as follows:

$$\text{Screen score} = -\log_{10}(\text{RRA}) \times \begin{cases} 1, \text{if } \text{RRA}_{\text{up}} < \text{RRA}_{\text{down}} \\ -1, \text{else} \end{cases}$$

$$\text{Fitness score} = \frac{\text{Screen score}_{x_1} + \text{Screen score}_{x_2}}{2}$$

where $\{x_1, x_2, \ldots, x_n\}$ are epegRNAs ranked in ascending order according to RRA values. If a mutation corresponds to only one epegRNA (that is, $n = 1$), then $x_2 = x_1$.

Validation of candidate mutations identified from the screen

All epegRNAs for validation were cloned into the pLenti-U6-epegRNA-SV40-mCherry vector individually. The *AAVS1*-targeting and nontargeting epegRNAs were served as negative controls. The lentivirus was transduced into HCT116-PEmax cells at an MOI > 1. The percentage of mCherry-positive cells was quantified through flow cytometry analysis (BD LSRFortessa SORP, Becton Dickinson) every 3 days. The first flow cytometry analysis started 3 days after infection (labeled as day 0), serving as a baseline for data normalization. FlowJo version 10 was used for flow cytometry data analyses.

Detection of the prime-editing outcomes by NGS analysis

Genome preparation. All epegRNAs were cloned into the pLenti-U6-epegRNA-SV40-puro vector individually. The lentivirus was transduced into HCT116-PEmax cells at an MOI > 1 and were further treated with 2 μg ml^{-1} puromycin for 2 days. Then, 14 days after transduction, epegRNA-infected cells were collected and subjected to genome DNA isolation using the DNeasy blood and tissue kit (Qiagen, 69506). For all the experimental and reference cells, sequences of approximately 200 bp near the target site of each epegRNA were amplified using specific primers (Supplementary Table 10) by PrimeSTAR GXL premix (TaKaRa, R051A), followed by NGS analysis on the Illumina HiSeq X TEN platform.

NGS analysis. After sequencing data quality control with FastQC (version 0.11.9; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and adaptor removal with fastp (version 0.12.4)[56], PANDAseq (version 2.11)[52] was used to assemble paired-end reads. Using sequence information from both ends, we precisely identified and extracted DNA regions containing the expected editing mutations. The self-developed ZFC-eBAR algorithm was then applied to analyze the screening data. To assess sensitivity, we simulated deleterious variants on the basis of the counts of 492 effective but nonenriched *AAVS1*-targeting epegRNAs. In this simulation, the control group remained unchanged, while the experimental group reduced counts by a specified percentage to replicate a known degree of dropout under free-riding effects. Additionally, efficiency analysis was visualized using WebLogo (version 0.0.0)[57].

Features analysis of identified mutations from the screen

We obtained 75 features (Supplementary Table 5) and used the Wilcox test to examine whether there were significant differences among features of synonymous mutations. When genomic inconsistencies occurred, LiftOver (online version)[58] was used for conversion.

Real-time qPCR analysis

RNA of the cultured cells infected by candidate epegRNA lentivirus was extracted using the Quick-RNA miniprep kit (Zymo Research, D1054) and complementary DNA (cDNA) was synthesized using HifairIII first-strand cDNA synthesis supermix (Yeasen, 11137ES60). Real-time qPCR was performed using TB Green Premix Ex Taq II (TaKaRa, RR820A) on Roche LightCycler96 real-time PCR system. All cDNA samples were assayed in triplicate and the relative RNA expression level of each sample was normalized by *GAPDH*. All the primers used for real-time qPCR are listed in Supplementary Table 10.

Bulk RNA-seq and data analysis

cDNA library construction and sequencing. The total RNA of each sample for RNA-seq was extracted using Quick-RNA miniprep kit (Zymo Research, D1054) and the RNA-seq libraries were prepared as previously described[59]. All samples were sequenced on the Illumina HiSeq X TEN platform.

Data analysis. Quality control of the data was performed using FastQC (version 0.11.9; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), followed by the removal of adaptors with fastp (version 0.12.4)[56]. The STAR aligner (version 2.5.2b)[60] was used for mapping to the GRCh38.p14 reference genome and gene expression levels were extracted using RSEM (version 1.2.28)[61].

The same quality control and alignment procedures were applied to the single-cell sequencing library. Subsequently, to estimate editing efficiency, SAMtools (version 1.7)[62] mpileup was used to extract base information at specific sites and calculate the mutation proportion. The estimated editing efficiency was calculated as $\text{maximum}\{\frac{\text{MUT reads}\times\text{library size}}{\text{WT reads}\times\text{MOI}}, 1\}$.

Single-cell RNA library preparation and sequencing

When the cell library for single-cell sequencing was cultured to day 14, we used the 10x Genomics 5′ HT v2 kit for single-cell isolation and cDNA library construction[63], following the standard procedures outlined in the manual. Briefly, approximately 120,000 cells were loaded on the 10x Genomics chip (Chromium Next GEM Chip N single-cell kit, PN-1000357). After cell capture and lysis, the epegRNA and transcripts were captured subsequently and followed by cDNA synthesis through reverse transcription (Chromium Next GEM single-cell 5′ HT kit v2, PN-1000356). Notably, the reverse transcription primer sequence used for epegRNA capture was 5′-CGTAACTAGATAGAACCGCG-3′ (5 pmol). Samples were sequenced on the Illumina NovaSeq 6000 platform.

Processing of single-cell RNA-seq data

Firstly, mRNA and epegRNA within cells were extracted using Cell Ranger (version 7.0.0), with the genome reference set to refdata-gex-GRCh38-2020-A. For epegRNA searching, the pattern was specified as R2's (BC)GCACCG. Given that sequencing depth poses a confounding factor, the single-cell data were analyzed using SCEPTRE (version 0.9.1)[29]. In the process of assigning epegRNA, we opted for the mixture method. Synonymous mutations considered to impact expression were those with an FDR < 0.1 following Benjamini–Hochberg adjustment.

Expression plots for each epegRNA and target gene, alongside LFC calculations, were conducted using Seurat (v4.4.0)[64]. The adopted filtering criteria were nFeature_RNA > 1,000, nFeature_RNA < 8,000, nCount_RNA < 40,000 and percent.mt < 20. The choice of threshold for assigning epegRNA influenced the LFC, generally showing a trend where stricter thresholds resulted in more pronounced absolute values of LFC. Consequently, for each pair, the threshold corresponding to the maximum absolute LFC was selected, provided the cell number was adequate (cell number ≥ 100), indicating the 'cleanest' state for the with epegRNA group.

Ribo-seq and data analysis

The ribosome profiling service was provided by Cloud-Seq Biotech by following the manual for the GenSeq Ribo profile kit (GenSeq, GS-LC-026). Briefly, after treatment with cycloheximide, cells were lysed with lysis buffer and then digested with nucleases. The digested samples were purified with size-exclusion chromatography to obtain ribosome footprints. The ribosome-protected RNA fragments were selected by PAGE and subjected to ribosomal RNA removal. The purified RNA was end-repaired and ligated with a 3′ adaptor and reverse-transcribed to cDNA. The cDNA was purified by PAGE, circularized and then amplified by PCR. The amplified library was purified and sequenced on Illumina NovaSeq 6000 platform.

Paired-end reads were obtained from the sequencer. First, the quality of the raw data was controlled by Q30. Adaptors were removed and low-quality reads were trimmed by cutadapt software (version 1.9.3) to obtain high-quality clean reads. The clean reads were aligned to the reference genome using Tophat2 software (https://ccb.jhu.edu/software/tophat/index.shtml). Then, HTSeq software (version 0.9.1)[65] was used to get the raw count and edgeR[66] was used to perform

normalization. Finally, differentially expressed mRNAs were identified by *P* value and fold change.

Machine learning model construction and testing

For dataset construction, we extracted samples at a 1:10 ratio of positive to negative instances, repeating the random extraction process ten times. Considering 22 features that showed significant differences across various types of synonymous mutations, nine different machine learning models were evaluated using fivefold cross-validation, with the AUC as the evaluation metric, aggregating performance across ten dataset subsets. The best-performing model, CatBoost, was further optimized through grid search, using the same evaluation metric to identify the most effective model configuration. The implementation was performed using Python (version 3.8.10) with scikit-learn (version 1.3.2) and CatBoost (version 1.2.2). Shapley additive explanations (SHAP) values were calculated and visualized using the SHAP package (version 0.44.1) to understand the importance of different features. Our methodology was compared against SilVA (version 1.1.1, without UNAfold or ViennaRNA)[10] and CADD (version 1.7)[11], using their provided scores to compute AUC for comparison. Further details are described below.

Dataset construction. Deleterious synonymous mutations (defined as those with values ≤ mean − 4 s.d.) were fully included to construct the dataset. Neutrality (defined as those within mean ± 1 s.d.) was randomly sampled at a ratio of 1:10, with random seeds fixed at 1 to 9 and 42.

Feature construction. We selected feature candidate sets from several different sources:

(1) DepMap data: We downloaded Chronos Score, CERES Score, gene effect, dependency and gene expression from DepMap (22Q2).
(2) HCT116 data: We calculated TPM (transcripts per million) for *AAVS1*, PEmax, WT and nontargeting using bulk RNA-seq in this study. Quality control was performed on raw data using FastQC (version 0.11.9; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and adaptors were trimmed using fastp[56]. The sequencing data were aligned to the genome using STAR[60] and TPM was obtained using RSEM[61].
(3) Codon frequencies: We obtained codon frequencies online (http://www.kazusa.or.jp/codon/)[67].
(4) Splice site prediction: We used SpliceAI to predict splice site changes[27], generating eight values: DS_AG (delta score, acceptor gain), DS_AL (delta score, acceptor loss), DS_DG (delta score, donor gain), DS_DL (delta score, donor loss), DP_AG (delta position, acceptor gain), DP_AL (delta position, acceptor loss), DP_DG (delta position, donor gain) and DP_DL (delta position, donor loss). The highest delta score among the four was taken as the splicing score. Delta scores ranged from 0 to 1, indicating the probability of splice changes. SpliceAI reference thresholds were 0.2 (high recall), 0.5 (recommended) and 0.8 (high precision).
(5) Splicing scores: We obtained splicing scores for corresponding tissues from AbSplice.
(6) Synonymous mutation analysis: We used SilVA to analyze synonymous mutations[10], providing SilVA prediction scores and rankings, along with RSCU, dRSCU, GERP scores, CpG sites, exon CpG scores, SR, FAS6, MES, MEC, PESE and PESS.
(7) Transcript information: We recorded transcript length and the relative position of the site within the transcript.
(8) RNA folding: We predicted RNA folding energy using RNAfold[68], calculating free energy for mutated and WT transcripts, the change in energy and the proportion of energy change.
(9) CADD scores: We recorded RawScore and PHRED from CADD[11].

Feature selection. We calculated the significance of differences in screen scores between groups of synonymous mutations using a *t*-test, selecting a union of significant features across different cell lines, totaling 22 features. Numeric features were standardized using StandardScaler and missing values were imputed with zero, assuming no specificity for these features. CatBoost handled categorical features directly while, for other models, 'mutated codon position 3' and 'original codon position 3' were transformed using one-hot encoding.

Model testing. We tested SVM, random forest, gradient boosting, logistic regression, *k*-nearest neighbors, naive Bayes, LDA, XGBoost, LightGBM and CatBoost, initially using default parameters. Fivefold cross-validation was used for evaluation, calculating the AUC across ten dataset subsets.

CatBoost parameter tuning. Focusing on AUC as the primary performance metric, the model was set to run in silent mode to minimize output during training. The loss function was 'logloss', with a depth of 6, 'uniform' feature_border_type and 'depthwise' grow policy. The parameter search range included iterations, subsample size, random strength, column sampling rate and L2 regularization strength, chosen to explore potential performance improvements. A grid search with fivefold stratified cross-validation identified the optimal model configuration on the basis of roc_auc scoring.

Result presentation. Final results were visualized using the R package ggplot2 (version 3.3.6).

Statistical analysis

GraphPad Prism 9 was used for statistical analyses. The statistical tests, exact values and descriptions for *n* are provided in figure legends. Unless otherwise stated, *n* represents the number of biological replicates of the samples. Statistical significance was evaluated using a two-tailed Student's *t*-test, with significance levels indicated as follows: $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ and $^{****}P < 0.0001$.

Reporting summary

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

Data availability

The plasmids used in this study were deposited to Addgene or are available upon request. All raw sequencing data were deposited to the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation and Beijing Institute of Genomics, Chinese Academy of Sciences under accession number HRA007615. The human reference genome used in this study is GRCh38.p14 from the NCBI (GCF_000001405.40). Databases involved in this study included ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), SynMICdb (https://synmicdb.dkfz.de/rsynmicdb/), DepMap (https://depmap.org/portal/), GWAS Catalog (https://www.ebi.ac.uk/gwas/) and the Medical Research Council Integrative Epidemiology Unit OpenGWAS (https://gwas.mrcieu.ac.uk/). Source data are provided with this paper.

Code availability

The ZFC-eBAR (version 0.2.0) algorithm and DS Finder (version 0.1.0), implemented in Python 3, can be downloaded from GitHub (https://github.com/UronicAcid/ZFC-eBAR and https://github.com/UronicAcid/DS-Finder). Other processed data and code can be found on Zenodo (https://doi.org/10.5281/zenodo.14639522)[69].

References

50. Morales, J. et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* **604**, 310–315 (2022).

51. 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).
52. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for Illumina sequences. *BMC Bioinformatics* **13**, 31 (2012).
53. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
54. Xu, P. et al. Genome-wide interrogation of gene functions through base editor screens empowered by barcoded sgRNAs. *Nat. Biotechnol.* **39**, 1403–1413 (2021).
55. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74**, 829–836 (1979).
56. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
57. Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14**, 1188–1190 (2004).
58. Hinrichs, A. S. et al. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34**, D590–D598 (2006).
59. Ding, B. et al. Noncoding loci without epigenomic signals can be essential for maintaining global chromatin organization and cell viability. *Sci. Adv.* **7**, eabi6020 (2021).
60. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
61. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
62. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
63. Replogle, J. M. et al. Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing. *Nat. Biotechnol.* **38**, 954–961 (2020).
64. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
65. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
66. Robinson, M. D., McCarthy, D. J. & Smyth, G. K.edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
67. Nakamura, Y., Gojobori, T. & Ikemura, T. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* **28**, 292 (2000).
68. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res.* **36**, W70–W74 (2008).
69. Tang, W. et al. Data and code related to 'Prime editor-based high-throughput screening reveals functional synonymous mutations in human cells'. *Zenodo* https://doi.org/10.5281/zenodo.14639522 (2025).

Acknowledgements

We thank the National Center for Protein Sciences at Peking University for their assistance with fluorescence-activated cell sorting and analysis, especially H. Lv and H. Yang for their technical support. We thank the High-Performance Computing Platform of Peking University for enabling our data analysis. We thank the National Center for Protein Sciences PKU-EdiGene High-Throughput Screening Core at Peking University. We thank Tsingke Biotech for providing primer synthesis and sequencing services. We thank Cloud-Seq Biotech for providing the Ribo-seq service and Emei Tongde Technology Development for their technical support in single-cell RNA library preparation. This research was funded by the National Natural Science Foundation of China (31930016, to W.W.), the Peking-Tsinghua Center for Life Sciences (to W.W.), Changping Laboratory (to W.W.) and the Taishan Scholarship (tsqn202312362, to Yongshuo Liu).

Author contributions

W.W., X.N., W.T. and Ying Liu conceptualized the project, with W.W. supervising it. W.W., X.N., W.T., Ying Liu and Yongshuo Liu designed the experiments. X.N. and Yongshuo Liu conducted the library screens. X.N. performed the single-cell screens, subsequent validations and experimental data analysis with assistance from B.M. W.T. handled all bioinformatics analysis. Y.Y. constructed the NGS library. W.T. developed the machine learning model and designed the web page. X.N., W.T. and Ying Liu wrote the paper, which W.W. revised.

Competing interests

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

Additional information

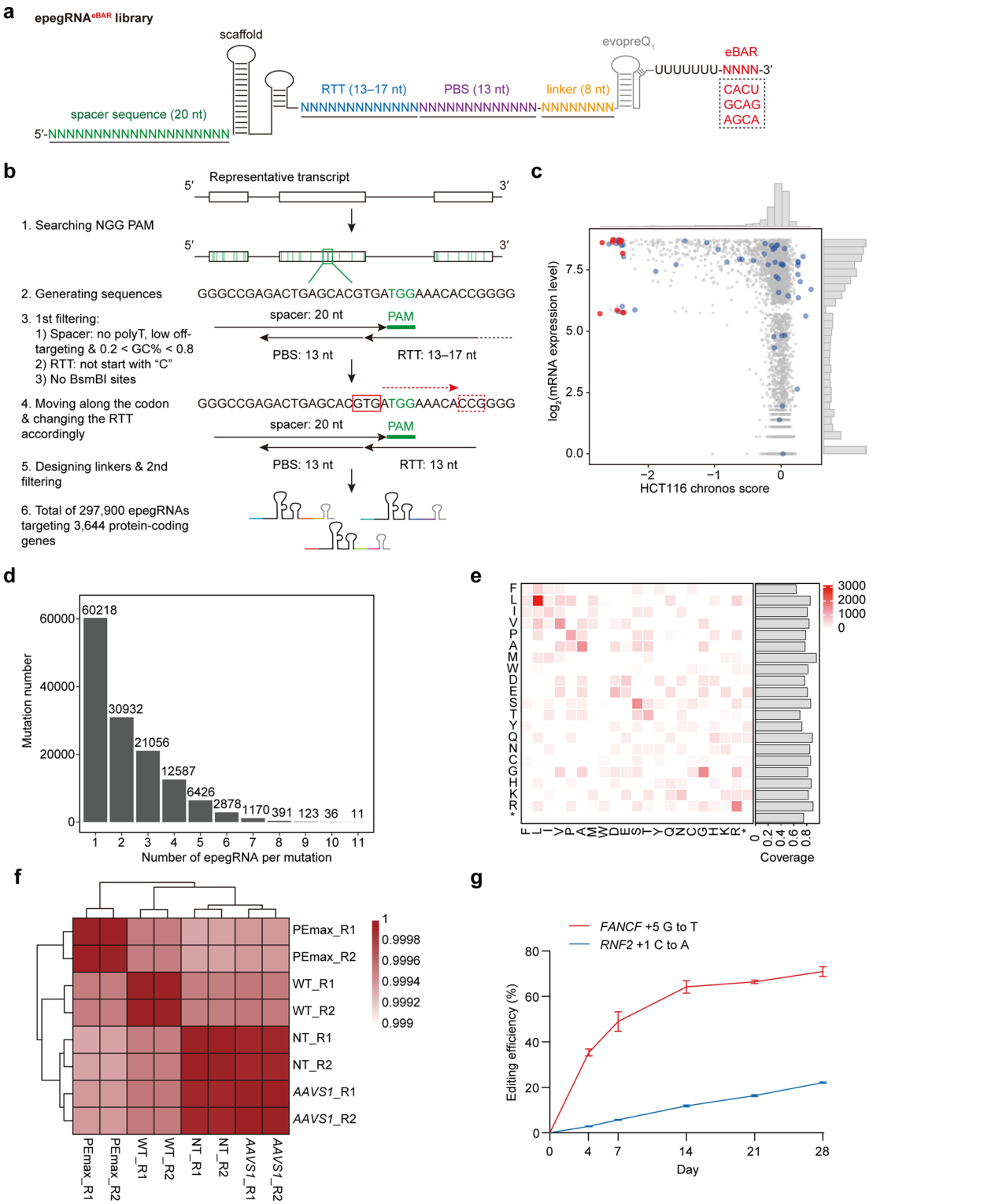
Extended data is available for this paper at https://doi.org/10.1038/s41587-025-02710-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-025-02710-z.

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

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

https://doi.org/10.1038/s41587-025-02710-z

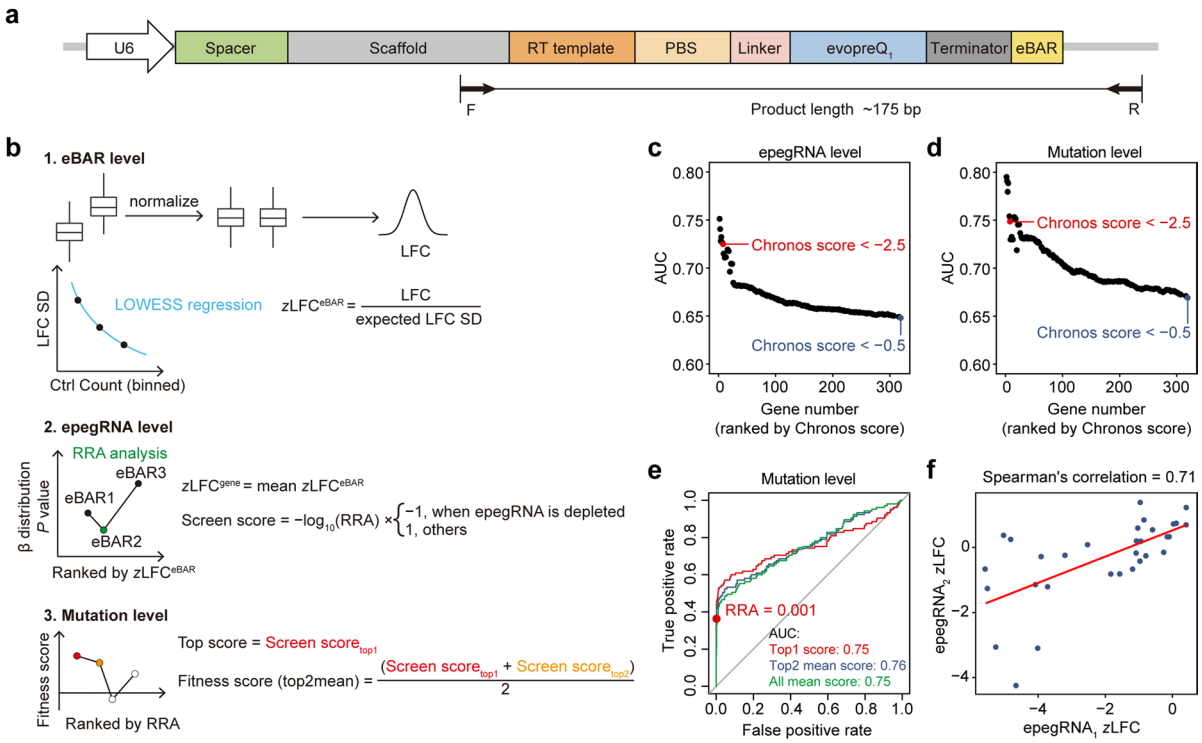


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41587-025-02710-z

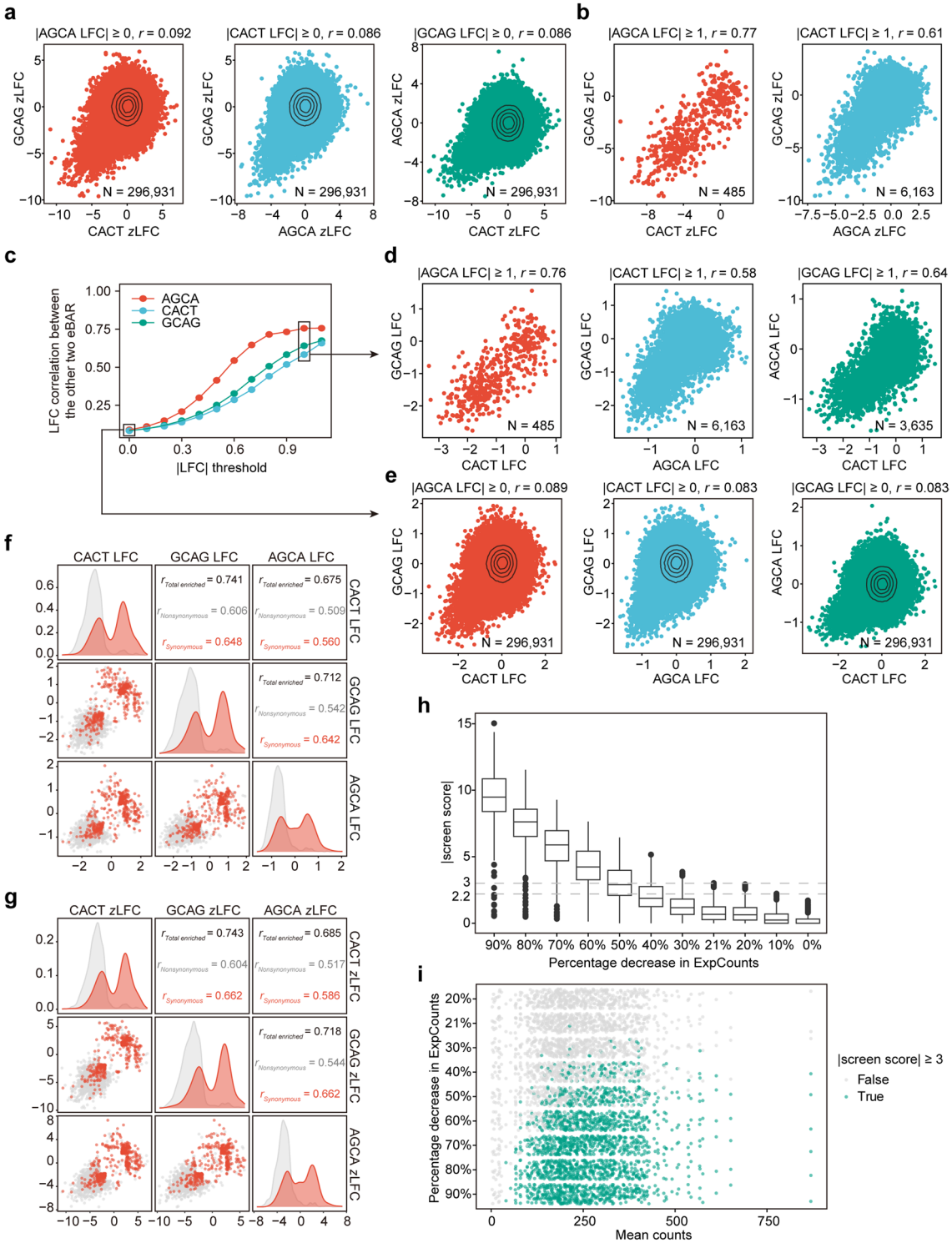
Extended Data Fig. 1 | Design of the epegRNA library for screening functional synonymous mutations and development of the experimental system using PEmax. a, Diagram of the epegRNA^{eBAR} library structure. RTT: reverse transcription template, PBS: primer binding site. **b**, Principles and workflow for designing epegRNAs within the library. Searches for NGG PAMs are conducted on the representative transcript of each gene to determine the spacer sequence, PBS sequence, and RTT, followed by a first filtering step. Comprehensive searches along the coding region are then performed to achieve saturation mutagenesis of synonymous mutations, with linker sequences subsequently designed and a secondary filter applied, resulting in the final epegRNA sequences. **c**, Distribution of expression levels and essentiality of genes within the library in HCT116 cells. Red dots represent the 11 genes targeted for complete saturation mutagenesis, blue dots represent the 56 genes targeted for synonymous saturation mutagenesis, and gray dots represent all other genes. **d**, Histogram of the number of epegRNA designed for mutations in the library. **e**, Statistics of the types of amino acid substitutions for the 11 genes targeted for saturation mutagenesis. The y-axis represents the original amino acid, the x-axis represents the amino acid post-PE editing, and the values represent the number of epegRNAs designed for each substitution. The histogram on the right illustrates the coverage of the corresponding amino acids, indicating the proportion of amino acids that can be targeted among all amino acids in the 11 genes. **f**, Clustered heatmap displaying Pearson correlation coefficients of the whole transcriptome between HCT116-PEmax cells and wild-type (WT) HCT116 cells, as well as after the addition of nontargeting (NT) epegRNA and *AAVS1*-targeting epegRNA. **g**, Graph showing the changes in editing efficiency over a 28-day period of continuous culture in HCT116-PEmax cells with the addition of two test epegRNAs targeting different sites. The data is presented as the mean ± s.d. (n = 3 biological replicates).

https://doi.org/10.1038/s41587-025-02710-z



Extended Data Fig. 2 | Decoding and ROC analysis of the epegRNA^{eBAR} library. **a**, Diagram illustrating the genomic PCR process performed on the cell library post-screening. **b**, Algorithm description of ZFC-eBAR (see Methods for details). **c**, The x-axis represents the number of genes corresponding to the positive controls (epegRNAs that introduces nonsense and frameshift mutations on essential genes), while the y-axis represents the AUC calculated using these positive controls as well as negative controls (*AAVS1*-targeting and nontargeting epegRNA). The score used to calculate AUC is the screen score. **d**, Similar to **c**, but nonsense/frameshift mutations at the same site are regarded as the same mutation. The score used to calculate AUC is the Top score. **e**, ROC analysis of epegRNAs in highly essential genes (Chronos score < −2.5) based on different scores. The red point represents the threshold we selected. The red line represents the ROC curve for the top1 score of epegRNA, the dark blue line represents the ROC curve for the top2 mean score of epegRNA, and the green line represents the ROC curve for the all mean score of epegRNA. **f**, Spearman correlation of zLFC for the top 2 ranked epegRNAs in highly essential genes (Chronos score < −2.5).

https://doi.org/10.1038/s41587-025-02710-z

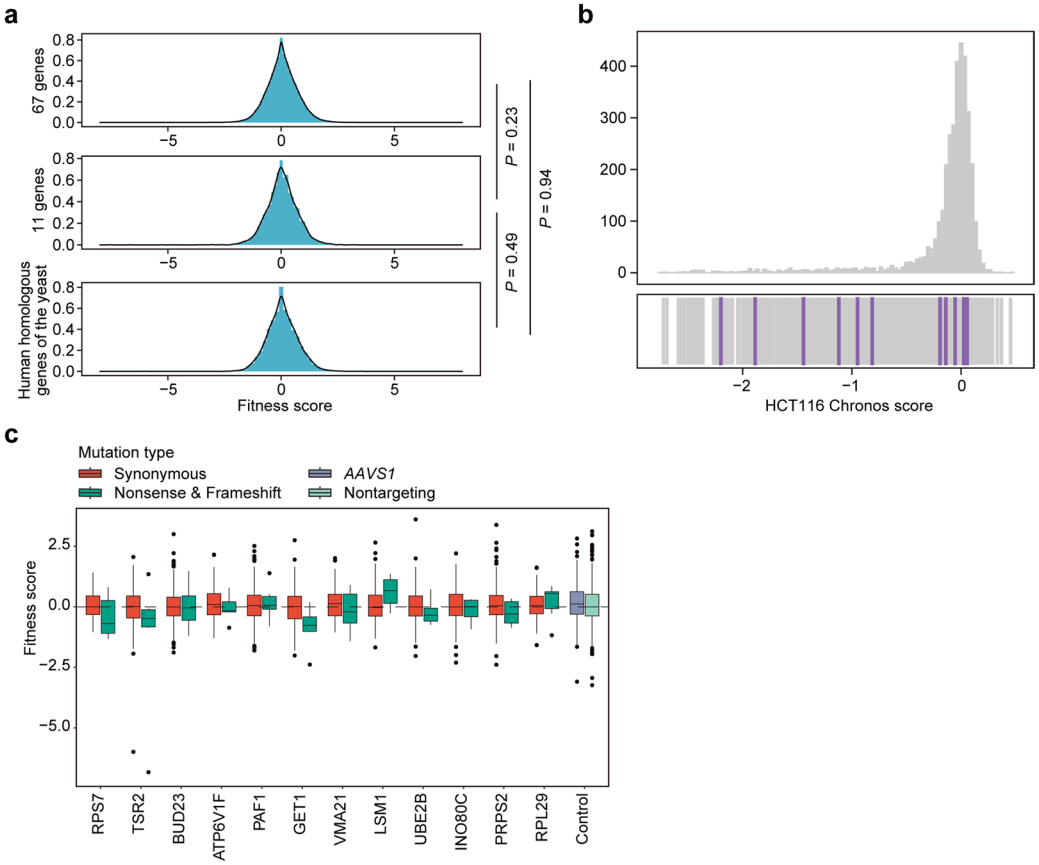


Extended Data Fig. 3 | See next page for caption.

https://doi.org/10.1038/s41587-025-02710-z

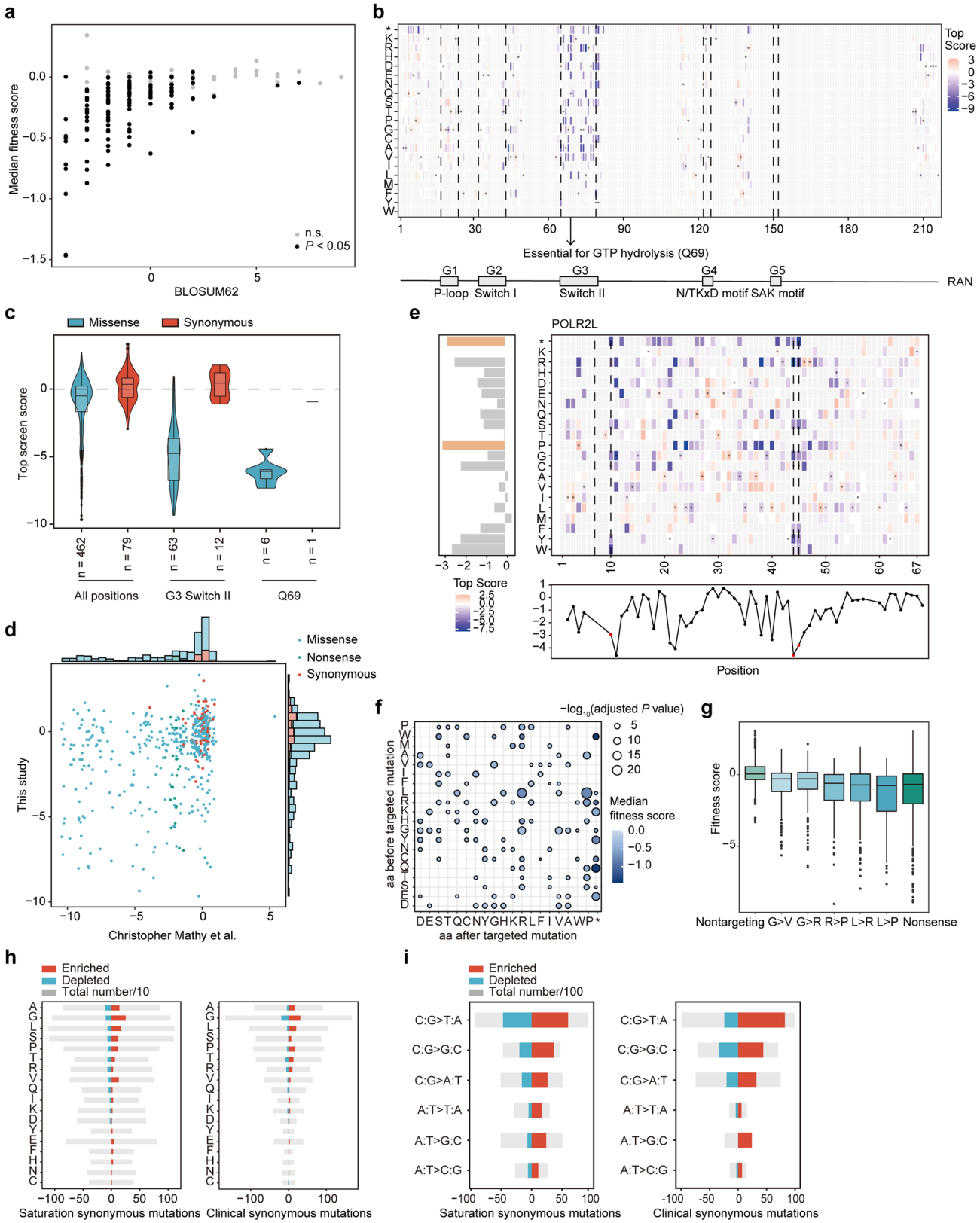
Extended Data Fig. 3 | Evaluation of screening reproducibility and detection sensitivity. a, Correlation of *z*LFC values between two eBARs replicates for epegRNA with the third eBAR's |LFC| ≥ 0. **b**, Correlation of *z*LFC values between two eBARs for epegRNAs with |AGCA LFC| ≥ 1 and |CACT LFC| ≥ 1. **c**, Pearson correlation analysis of LFC values between two eBARs across varying thresholds of absolute LFC values from the third eBAR. The x-axis indicates the absolute LFC threshold applied to each eBAR, while the y-axis represents the Pearson correlation coefficient between the remaining two eBARs. **d**, Correlation of LFC values between two eBARs for epegRNAs with the third eBAR's |LFC| ≥ 1. **e**, Correlation of LFC values between two eBARs for epegRNAs with the third eBAR's |LFC| ≥ 0. For panels **a**, **b**, **d**, and **e**, Pearson correlation coefficient (*r*) is labeled. N indicates the number of epegRNAs shown in these figures (for panels **a** and **e**, excluding those with zero sequencing read counts and black lines indicate density-based contours). **f**–**g**, Correlation analysis of enriched mutations in the HCT116 screen across three eBARs, based on LFC values (**f**) or *z*LFC values (**g**). Diagonal panels show histograms of value distributions, lower left panels show scatter plots of pairwise correlations, and upper right panels present Pearson correlation coefficient (*r*). The dataset includes 1,717 epegRNAs with nonsynonymous mutations and 417 epegRNAs with synonymous mutations. **h**, Boxplot of the relationship between the degree of reduction in counts of simulated deleterious variants in the experimental group (x-axis) and the calculated |screen score| (y-axis). $n = 492$ per group (number of epegRNAs). The gray dashed lines correspond to thresholds of 3 and 2.2. Boxplots are depicted as follows: the center line represents the median, the box limits denote the upper and lower quartiles, and whiskers extend to 1.5 times the interquartile range. **i**, Corresponding to **h**, the x-axis represents the average counts of these simulated deleterious variants in day 0, and the y-axis indicates the percentage reduction in day 35. Green labels highlight points where the |screen score| after reduction exceeds the selection threshold (3).

https://doi.org/10.1038/s41587-025-02710-z



Extended Data Fig. 4 | Performance of synonymous mutations in human homologous genes of the yeast. a, Comparison of fitness score distributions for synonymous mutations across three groups: 67 saturation-mutated genes (synonymous mutations only), 11 saturation-mutated genes (including all mutation types), and human homologous genes of the yeast (excluding *RPL39* due to limited data). *P* values were calculated using a two-sided Wilcoxon test. **b**, Distribution of essentiality for human homologous genes of yeast in HCT116 cells, data of *RPL29* are missing in DepMap. The corresponding positions are labeled in purple. **c**, Shown in order as in **b**, distribution of fitness scores for synonymous, nonsense and frameshift mutations for each gene. From left to right, the mutation counts for each group are 46/5 (*RPS7*), 269/6 (*TSR2*), 459/6 (*BUD23*), 167/6 (*ATP6V1F*), 791/6 (*PAF1*), 233/6 (*GET1*), 131/6 (*VMA21*), 162/6 (*LSM1*), 193/6 (*UBE2B*), 323/6 (*INO80C*), 487/6 (*PRPS2*) and 80/6 (*RPL29*), with $n = 494$ for *AAVS1* and $n = 981$ for Nontargeting. Boxplots are depicted as follows: the center line represents the median, the box limits denote the upper and lower quartiles, and whiskers extend to 1.5 times the interquartile range.

https://doi.org/10.1038/s41587-025-02710-z

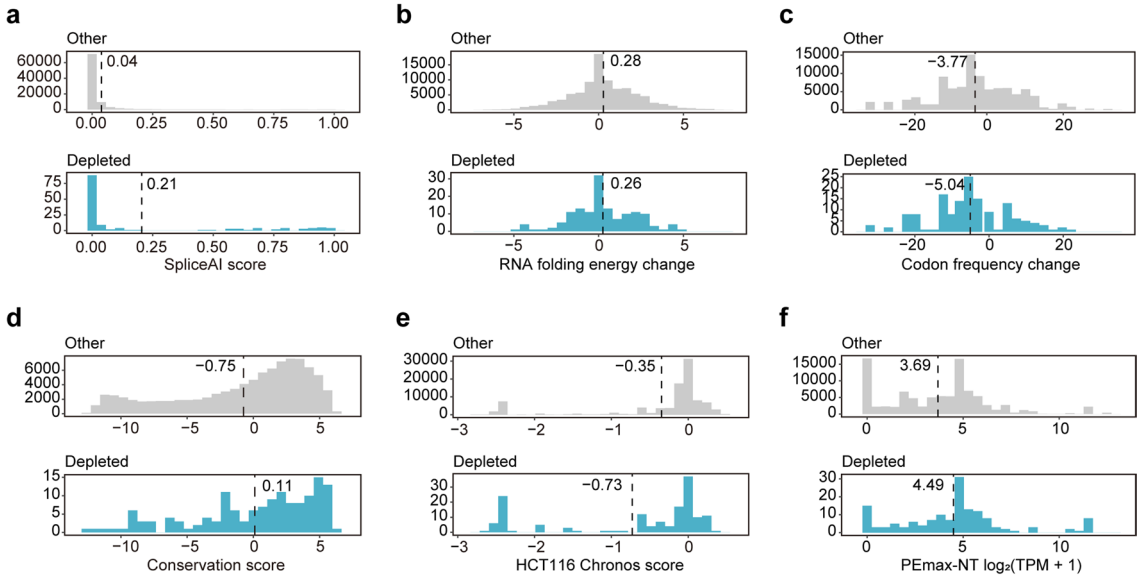


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41587-025-02710-z

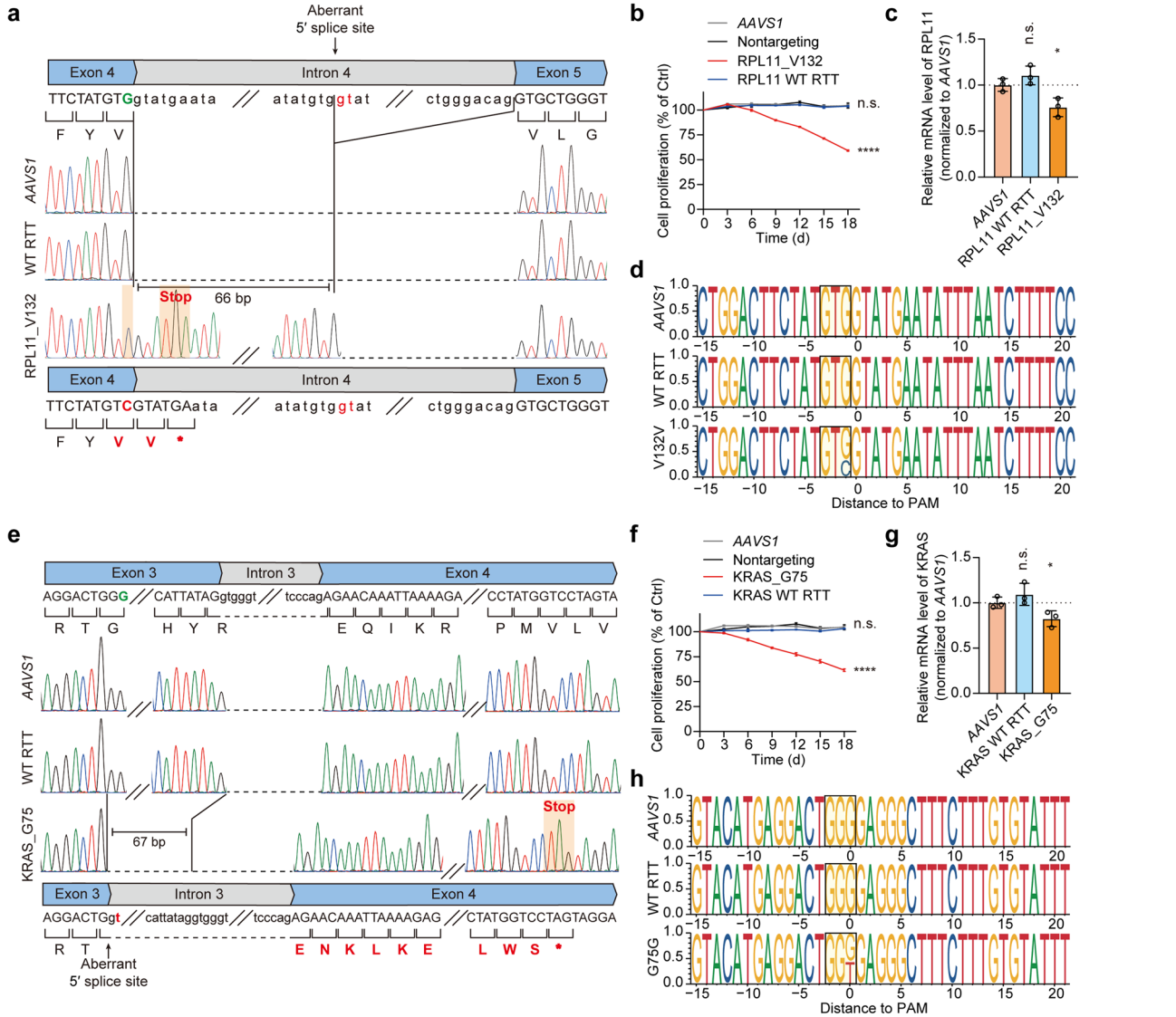
Extended Data Fig. 5 | Comprehensive analysis of different mutation types and their effects in saturation mutagenesis. a, Correlation of BLOSUM62 scores with the screening results for 11 genes. Mean scores were compared to nontargeting controls, with mutations showing adjusted $P < 0.05$ (two-tailed Student's *t* test with Benjamini–Hochberg correction) marked in black. Only mutations with counts ≥10 were included. **b**, Upper panel displays a heatmap of screen scores for all mutations in RAN. Light gray indicates amino acids not included in the library. Synonymous mutations are marked with dots, and dashed lines denote domain boundaries. The lower panel illustrates domain annotations for RAN. **c**, Top scores for missense and synonymous mutations are compared across all regions of RAN, including G3 Switch II and Q69, n indicates the number of mutations. **d**, Comparison of top screen scores from our study with scores from ref. 22. Spearman correlation coefficient = 0.3076786. **e**, Detailed Mutation Analysis in POLR2L. Middle panel: Heatmap of screening scores for all mutations in POLR2L. Light gray indicates amino acids not designed in the screening library. Dots indicate synonymous mutations. Dashed lines represent four Zn^{2+} binding sites. Lower panel: Average top scores across each position, with three designed Zn^{2+} binding sites highlighted in red. Left panel: Average top scores by amino acid, with the two most depleted ones shown in orange. **f**, Cell fitness effects of different amino acid (aa) substitutions (statistical methods as **a**). **g**, The distributions of fitness scores for mutation types with a significant impact on cell fitness. From left to right, the mutation counts are 1,475, 182, 258, 113, 233, 230, and 850 respectively. **h**–**i**, Distribution of enriched mutations for various amino acid substitutions (**h**) and among different types of base pair substitutions (**i**) in both saturation synonymous mutations and clinical synonymous mutations. Top score represents the maximum |screen score| value for different epegRNAs corresponding to each mutation in the screen. For **c** and **g**, boxplots are depicted as follows: the center line represents the median, the box limits denote the upper and lower quartiles, and whiskers extend to 1.5 times the interquartile range.

https://doi.org/10.1038/s41587-025-02710-z



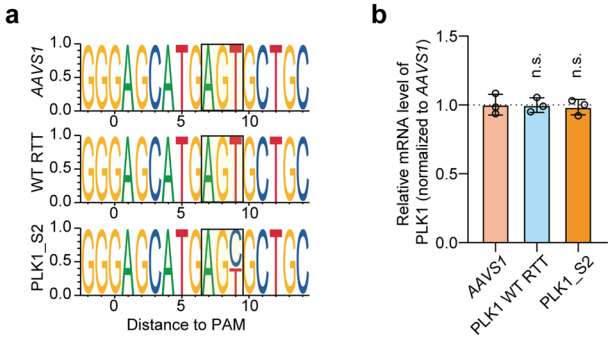
Extended Data Fig. 6 | Distribution of different key features for synonymous mutations. a, Distribution of SpliceAI scores for synonymous mutations. **b**, Distribution of RNA folding energy changes for synonymous mutations. **c**, Distribution of codon usage frequency changes for synonymous mutations. **d**, Distribution of conservation scores for synonymous mutations. **e**, Distribution of gene essentiality (Chronos score) for synonymous mutations in HCT116 cells. **f**, Distribution of expression levels for synonymous mutations in HCT116 cells.

https://doi.org/10.1038/s41587-025-02710-z



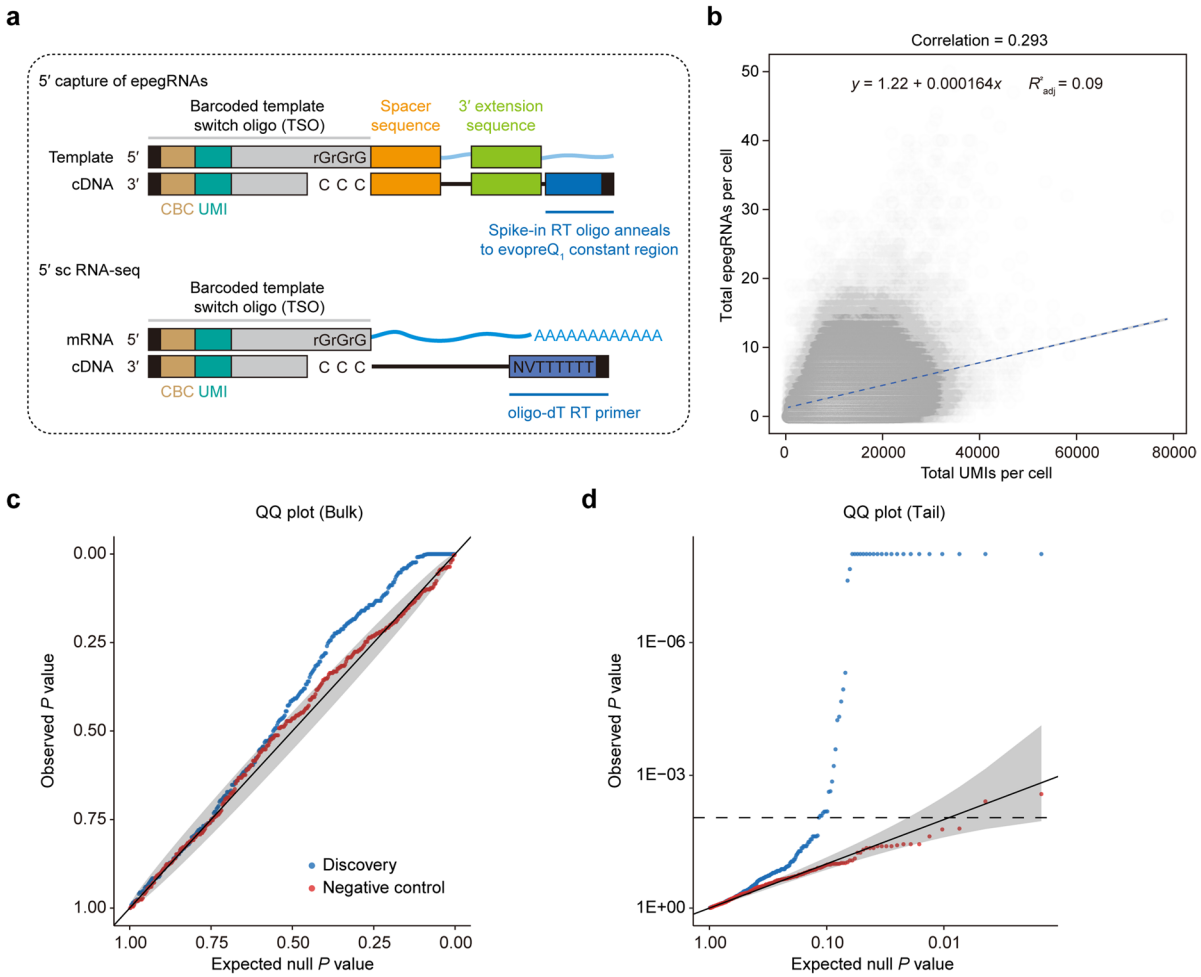
Extended Data Fig. 7 | Validation of synonymous mutations causing aberrant RNA splicing. a, Schematic depiction of the splicing alterations caused by the RPL11_V132 (GTG > GTC) mutation. The transcript sequence information was obtained by sequencing the cDNA from the experimental and control groups. **b**, Validation of the effect of the RPL11_V132 (GTG > GTC) mutation on cell proliferation in HCT116 cells. **c**, Relative mRNA expression levels of *RPL11* in the experimental and control groups. The mRNA level of each sample was quantified by real-time qPCR and normalized by *GAPDH*, and the indicated relative mRNA level was normalized to that of *AAVS1*-targeting control cells. **d**, Analysis of editing outcomes for epegRNA targeting RPL11_V132 and controls via genome sequence amplification and NGS. **e**, Schematic depiction of the splicing alterations caused by the KRAS_G75 (GGG > GGT) mutation. The transcript sequence information was obtained by sequencing the cDNA from the experimental and control groups. **f**, Validation of the effect of the KRAS_G75 (GGG > GGT) mutation on cell proliferation in HCT116 cells. **g**, Relative mRNA expression levels of *KRAS* in the experimental and control groups. The mRNA level of each sample was quantified by real-time qPCR and normalized by *GAPDH*, and the indicated relative mRNA level was normalized to that of *AAVS1*-targeting control cells. **h**, Analysis of editing outcomes for epegRNA targeting KRAS_G75 and controls via genome sequence amplification and NGS. All data are presented as mean ± s.d. ($n = 3$ biological replicates for cell proliferation assay, $n = 3$ technical replicates for real-time qPCR). *P* values were calculated using two-tailed Student's *t* test, $^{*}P < 0.05$, $^{****}P < 0.0001$; n.s., not significant.

https://doi.org/10.1038/s41587-025-02710-z



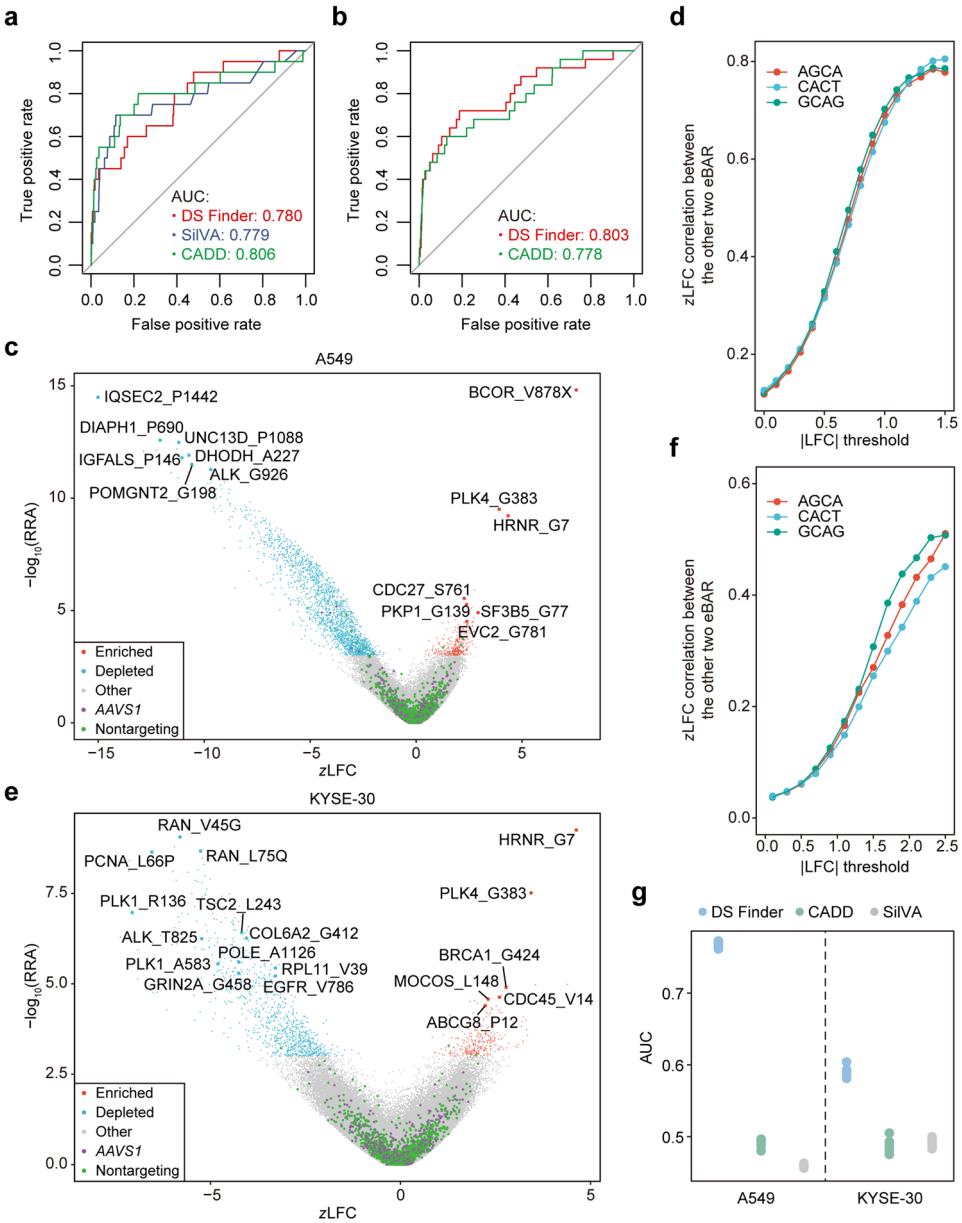
Extended Data Fig. 8 | Editing and the gene expression analysis at the PLK1_S2 site. a, Analysis of editing outcomes for epegRNA targeting PLK1_S2 and controls via genome sequence amplification and NGS. **b**, Relative mRNA expression levels of *PLK1* in the experimental and control groups. The mRNA level of each sample was quantified by real-time qPCR and normalized by *GAPDH*, and the indicated relative mRNA level was normalized to that of *AAVS1*-targeting control cells. Data are presented as mean ± s.d. ($n = 3$ technical replicates). *P* values were calculated using two-tailed Student's *t* test, n.s., not significant.

https://doi.org/10.1038/s41587-025-02710-z



Extended Data Fig. 9 | Capture and data quality control for DIRECTED-seq. **a**, Overview of the process for capturing epegRNA and single-cell transcriptomes. The cDNA of epegRNA was obtained by reverse transcription using primers targeting the constant secondary structure of evopreQ$_1$, and the cDNA of the entire transcriptome within the cells was obtained by reverse transcription using oligo-dT primers. The figure was adapted from ref. 63. **b**, Sequencing depth impact epegRNA detection probability and observed gene expression levels. The blue dashed line represents the linear regression line correlating total epegRNAs per cell (y) with total UMIs per cell (x). **c**–**d**, Quantile–quantile plots comparing 15 nontargeting epegRNAs (negative control) with 395 epegRNAs across 213 genes (discovery). Genes for the nontargeting tests were randomly selected from the entire gene set. The left-hand panel shows the QQ plot of raw *P* values, and the right-hand panel shows the QQ plot of $-\log_{10}$ transformed *P* values. In both panels, the black solid line denotes the theoretical y = x reference under the null hypothesis and the surrounding gray band represents the pointwise 95% confidence interval of that regression.

https://doi.org/10.1038/s41587-025-02710-z



Extended Data Fig. 10 | Comparison of the performance of different prediction models and the screening results in the A549 and KYSE-30 cell lines. a–**b**, ROC analysis of DS Finder, SilVA, and CADD on 45 confirmed pathogenic mutations (positive controls) and 1,439 synonymous mutations without phenotypes from our screening (negative controls). The x-axis represents the false positive rate, and the y-axis represents the true positive rate. The red, dark blue, and green curves represent DS Finder, SilVA, and CADD, respectively. The test set in **a** contains 20 positive controls, excluding the mutations used in the SilVA training set and the test set in **b** consists of the 25 mutations used in the SilVA training set. **c**–**f**, Screening results analysis of three cell lines. Volcano plot illustrating the results of screening for functional synonymous and nonsynonymous mutations affecting cell fitness in A549 (**c**) and KYSE-30 (**e**). Blue and red dots denote depleted and enriched epegRNAs, respectively. Analyzing the correlation between three eBARs using zLFC in A549 (**d**) and KYSE-30 (**f**), following the same method as in Fig. 1d. The data were filtered according to the $\log_2$ fold change of a specific eBAR and the Pearson correlation of the zLFC of another pair of eBARs was investigated. The y-axis represents the Pearson correlation of the zLFC, while the x-axis indicates the specific threshold that the absolute value of the LFC must surpass. Red, green, and blue denote three individual eBARs: AGCA, CACT, and GCAG, respectively. **g**, Performance of DS Finder in three cell lines compared with CADD and SilVA. Each point on the graph represents a different dataset, totaling 10.

nature portfolio

Corresponding author(s): Wensheng Wei

Last updated by author(s): May 16, 2025

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	Statistical analysis and data visualization were performed using R (v3.6.3) with the ggplot2 package (v3.3.6) and Python (v3.8.10). Statistical methods were included in the related figure legends. Fastp (v0.12.4), PANDAseq (v2.11), blastn (v2.11.0+), and customized scripts were used to generate raw counts from FASTQ files. Editing visualization was performed using WebLogo (v0.0.0). ZFC-eBAR software (v0.2.0, https://github.com/UronicAcid/ZFC-eBAR) was used to analyze cell fitness screens with eBARs. DS Finder (v0.1.0, https://github.com/UronicAcid/DS-Finder) was built by python (v3.8.10) with sklearn (v1.3.2), CatBoost (v1.2.2), and SHAP (v0.44.1). FlowJo10 was used for flow cytometry data analyses.

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

All raw sequencing data have been deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences, at https://ngdc.cncb.ac.cn/gsa-human/browse/HRA007615. Human reference genome used in this study is GRCh38.p14 from NCBI (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.40/). Databases involved in this study include ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), SynMICdb (https://synmicdb.dkfz.de/rsynmicdb/), DepMap (https://depmap.org/portal/), GWAS Catalog (https://www.ebi.ac.uk/gwas/), and the MRC IEU OpenGWAS (https://gwas.mrcieu.ac.uk/). Processed data are provided in https://zenodo.org/records/14639522.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	This study did not involve any human research participants.
Reporting on race, ethnicity, or other socially relevant groupings	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 all 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
☒	☐ Plants

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), PLK1 (Proteintech, 10305-1-AP, target species: human, mouse, rat). 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): The HCT116 cell line (#CCL-247, ATCC) and HEK293T cell line (#CRL-3216, ATCC) were obtained from EdiGene Inc., the A549 cell line was purchased from ATCC (#CRM-CCL-185) and the KYSE-30 cell line (#ACC 351, DSMZ) was obtained from Z. Liu's laboratory at Peking Union Medical College.

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.

Plants

Seed stocks: This study did not involve any plants.

Novel plant genotypes: This study did not involve any plants.

Authentication: This study did not involve any plants.

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 epegRNAs 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
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 FSC-A and FSC-W(P2) following SSC-A and SSC-W(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.