nature communications

Article

https://doi.org/10.1038/s41467-025-57858-8

Deep learning prioritizes cancer mutations that alter protein nucleocytoplasmic shuttling to drive tumorigenesis

Received: 13 March 2023

Accepted: 5 March 2025

Published online: 14 March 2025

Check for updates

Yongqiang Zheng[1,5], Kai Yu[1,2,5], Jin-Fei Lin[1,3,5], Zhuoran Liang[1,5], Qingfeng Zhang[1], Junteng Li[1], Qi-Nian Wu[1], Cai-Yun He[1], Mei Lin[1], Qi Zhao[1], Zhi-Xiang Zuo[1], Huai-Qiang Ju[1], Rui-Hua Xu[1,4] ✉ & Ze-Xian Liu[1] ✉

Genetic variants can affect protein function by driving aberrant subcellular localization. However, comprehensive analysis of how mutations promote tumor progression by influencing nuclear localization is currently lacking. Here, we systematically characterize potential shuttling-attacking mutations (SAMs) across cancers through developing the deep learning model pSAM for the ab initio decoding of the sequence determinants of nucleocytoplasmic shuttling. Leveraging cancer mutations across 11 cancer types, we find that SAMs enrich functional genetic variations and critical genes in cancer. We experimentally validate a dozen SAMs, among which R14M in PTEN, P255L in CHFR, etc. are identified to disrupt the nuclear localization signals through interfering their interactions with importins. Further studies confirm that the nucleocytoplasmic shuttling altered by SAMs in PTEN and CHFR rewire the downstream signaling and eliminate their function of tumor suppression. Thus, this study will help to understand the molecular traits of nucleocytoplasmic shuttling and their dysfunctions mediated by genetic variants.

Many shuttling proteins are transported between different subcellular compartments in response to external signals and regulate a broad spectrum of biological processes[1–3]. This aberrant shuttling of proteins is often driven by binding to specific molecules that recognize distinct targeting signals. The dynamic nucleocytoplasmic trafficking process is functionally and mechanistically diversified and is mostly regulated by the nuclear localization signal (NLS) and nuclear export signal (NES)[4,5]. The NLS is recognized by the corresponding nuclear transporters, classically members of the importin superfamily, which can interact with nucleoporins to help NLS-containing proteins entry the nucleus through the nuclear pore complex (NPCs)[6]. However, many nuclear proteins do not contain classical NLSs, and must either use alternate entry mechanisms including non-classical NLSs, binding with other nuclear localized proteins, direct interaction with NPCs and nuclear retention mediated by DNA or RNA affinity[7–9]. In recent years, an increasing number of proteins initially identified in the cytoplasm have also been observed to localize to the nucleus and perform different functions, especially in tumors and disease states[10–14]. These discoveries further underscore the importance of additional investigations on the regulation of protein subcellular localization.

Cancer-driving mutations contribute to abnormal cellular functions that trigger tumorigenesis through a variety of mechanisms. To date, the pathological mechanisms of only a fraction of known mutations have been elucidated. In particular, for protein nuclear localization, only several investigations have described protein nucleocytoplasmic shuttling due to cancer-driving mutations[15–18].

[1]State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, China. [2]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston 77030, USA. [3]Department of Clinical Laboratory, Sun Yat-Sen University Cancer Center, Guangzhou 510060, China. [4]Research Unit of Precision Diagnosis and Treatment for Gastrointestinal Cancer, Chinese Academy of Medical Sciences, Guangzhou 510060, China. [5]These authors contributed equally: Yongqiang Zheng, Kai Yu, Jin-Fei Lin, Zhuoran Liang. ✉e-mail: xurh@sysucc.org.cn; liuzx@sysucc.org.cn

However, all of these studies reported a limited number of mutations in individual proteins. An urgent need is to systematically identify cancer mutations altering protein nucleocytoplasmic shuttling that drives tumorigenesis, and an effective in silico predictor is necessary. Numerous studies have contributed to predicting protein nuclear localization[19–29], and these efforts have substantially improved our understanding of proteins localized in specific organelles. However, limitations still exist. Most predictors yield prediction scores only for nuclear localization and do not pinpoint which sequences potentially modulate transport into the nucleus. Other tools[30–33] build models and conduct predictions based on existing targeting peptides, which are unable to predict proteins with unconventional localization sequences. Furthermore, most of these models are based on classical machine learning algorithms and sequence alignment methods, and there is still room for performance improvement. Thus, although disrupting nuclear localization is a critical mechanism by which mutations cause protein dysfunction, a computational tool for identifying shuttling-attacking mutations (SAMs) is lacking; therefore, systematic studies on this topic are lacking.

In this work, to elucidate SAMs, we construct a deep-learning framework to decipher the impact of single amino acid mutations on protein nuclear localization. First, we build a deep learning model, named prediction of SAMs (pSAM), based on a hybrid convolutional neural network (CNN) and bidirectional long short-term memory (BiLSTM) architecture with an attention mechanism to accurately predict the potential nuclear localization of proteins. Since deep learning methods can independently identify helpful information in sequences without relying on knowledge-based biological features and solve scientific problems of interest well[34–36], we avoid using known targeting peptides to detect potential vital regulatory regions in the protein sequence. Instead, the full-length sequence of the protein is used to predict specific nuclear localization, and a residue-level contribution analysis (RCA) method is utilized for model interpretation to score the residue-level contribution. The rigorous evaluation convincingly reveals that the pSAM model built in this study accurately predicts nuclear localization probability with state-of-the-art performance, and ab initio prediction without any prior NLS or NES information confers the ability to pinpoint key nuclear localization-related sequence determinants beyond the canonical presequences and internal signals. While investigating somatic mutations across 11 variant-abundant cancer types, we find that SAMs constitute approximately 6.7% of the mutations in NLSs and that cancer mutations in the sequence determinant regions are more frequent and more deleterious. The potential disruption of nucleocytoplasmic shuttling induced by single-nucleotide or truncated mutations is analyzed based on this predictor and partially experimentally validated. Overall, since most predicted disruptions are confirmed by experiments, our proposed model is shown to be accurate at the residue level and can facilitate investigations into the molecular mechanism of protein nuclear localization.

Results

Mutations might disrupt protein nuclear localization but are largely unknown

Cancer-causing mutations lead to vicious consequences through numerous molecular mechanisms[37]. The elucidation of protein SAMs requires potential protein sequence determinants that are essential for nuclear localization to be decoded (Fig. 1A). We collected all cancer mutations across 11 cancer types in The Cancer Genome Atlas (TCGA) and compared the occurrence of known missense mutations in the targeting peptides and other regions. We integrated several well-known databases, including UniProt, SeqNLS, ValidNESs and NESbase, to annotate experimentally validated NLSs/NESs in proteins. The frequency of mutations was significantly greater in the targeting peptide regions of NLSs/NESs (Fig. 1B and Supplementary Fig. 1A, B). We further used simulated mutations (i.e., signature-corrected randomized mutations) in the mutation enrichment analyses; using this, we found that putative mutations were significantly depleted in the validated NLS/NES regions (Supplementary Fig. 1C, D). We also analyzed the nuclear localization of cancer driver and non-driver genes. The results indicated that nearly 60% of the cancer driver genes are located in the nucleus, while only approximately 25% of non-driver genes are located in the nucleus (Supplementary Fig. 1E-F). We investigated whether mutations in targeting peptide regions are more likely to disrupt protein functions by introducing a deleterious score to measure their deleteriousness. Mutations in targeting peptide regions were more deleterious (Fig. 1C and Supplementary Fig. 1G, H), suggesting that these mutations significantly predispose patients to protein dysfunction. Because transcript factors (TFs) and transcriptional repressors perform transcriptional regulatory functions upon entering the nucleus, we explored the transcriptional regulatory effect of mutations in the experimentally determined NLS region of TFs and transcriptional repressors on their substrates in the TCGA-UCEC cohort. BACH2 is a transcriptional repressor that may play a role in cancer and neuronal differentiation[38]. Interestingly, we noticed a significant upregulation of its substrates when the mutations were located at the experimentally validated NLS regions, for example, those of *TXNRD1* (Fig. 1D) and *CLIP1* (Fig. 1E), two well-known oncogenic drivers that promote tumorigenicity[39,40]. These results suggest that mutations in the targeting peptide region may affect the nucleocytoplasmic shuttling of the protein, thereby regulating its biological function.

We collected eukaryotic protein sequences and localization annotations from the UniProt database[41] and summarized the distributions of nuclear localization (Supplementary Fig. 1I). Amino acid frequency statistics indicated that nucleus-localized proteins contain a greater number of positively charged residues (Supplementary Fig. 1I), consistent with previous reports that the classical NLS motif consists of lysine and arginine[7]. In addition, we found that serine was also enriched in nucleus-localized proteins and that proline and glutamate were enriched in extracellular proteins.

Ab initial identification of sequence determinants with a deep learning approach

We constructed a deep neural network, pSAM, to predict protein nuclear localization and analyzed the residue-level contribution of the nuclear localization probability calculation to dissect the decision-making process of pSAM and identify the sequence determinants for nuclear localization (DNLs). We collected eukaryotic protein sequences and localization annotations from the UniProt database[41] (Supplementary Fig. 2A). We randomly divided the whole dataset (20-2000 amino acids) into a training set, validation set, and testing set at a ratio of 7:2:1. Then, we use a combined deep neural network based on CNN, BiLSTM, and attention mechanisms to construct the model to predict nuclear localization probability (Fig. 1A). The input protein sequences were one-hot encoded into a (2000 ×21) embedded matrix, which was used as the input for the multi-subnetwork CNN layers. Then, the output was passed to the BiLSTM layer, followed by the attention layer based on Luong-style attention to assign an importance score to each hidden state of the BiLSTM layer. The Adam algorithm was used to optimize the model, and batch normalization, dropout, and early stopping (Supplementary Fig. 2B) were employed to avoid overfitting the model. Hyperparameters were carefully tuned for the residual block combination, CNN node size, dropout rate, and recurrent gate size (Supplementary Data 1).

To assess the predictive power of the constructed deep neural network model, we evaluated various aspects of performance. First, the structures of the merged model and other individual models were compared, and the robust performance of the combined deep neural network based on CNN, BiLSTM, and attention mechanisms showed its superiority (Supplementary Fig. 2C). The optimized model named pSAM presented satisfactory robustness in the ten-fold cross-

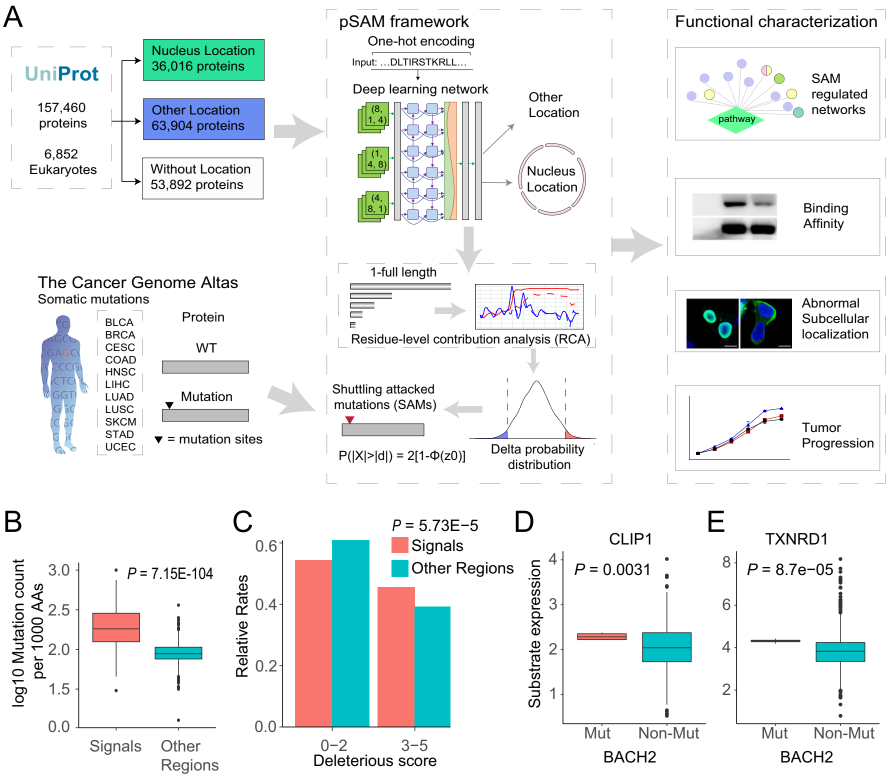


Fig. 1 | Study overview. **A** Somatic mutations from 11 cancer types in TCGA were analyzed to determine their potential effect on nucleocytoplasmic shuttling and revealed a significant enrichment in targeting peptides. A deep learning model, pSAM, was then constructed to precisely predict the protein nuclear localization probability and ab initio inference of protein sequence determinants without knowledge of known targeting peptides based on deep learning coupled with residue-level contribution analysis. The pSAM model and somatic mutation dataset were subsequently used for downstream analyses. **B** Mutation frequency distribution of the targeting peptide regions and other regions from 11 cancer types in TCGA. The proteins with validated NLS/NES region were included ($n = 564$ proteins), and the NLS/NES regions were compared with the rest regions. The data are presented as a box-and-whisker graph (bounds of box: first to third quartile, bottom and top line: minimum to maximum, central line: median). **C** The deleteriousness score distribution of mutations localized in targeting peptide regions and other regions from 11 cancer types in TCGA. The target regions include valid NLS and NES regions. **D**, **E** Mutations in NLS regions of BACH2 affect the expression of its transcriptional substrates, (**D**) *CLIP1* and (**E**) *TXNRD1*, in the TCGA-UCEC cohort (Mutated group: $n = 4$ samples; Non-Mutated group: n = 579 samples). The data are presented as a box-and-whisker graph (bounds of box: first to third quartile, bottom and top line: minimum to maximum, central line: median). Two-sided Wilcoxon test was used for (B), two-sided Chi-square test was used for (**C**), two-sided Student's *t* test was used for (**D**) and (**E**). Source data are provided as a Source Data file.

validations, with area under the receiver operating characteristic (ROC) curve (AUC) values of 0.9726–0.9848 (Supplementary Fig. 2D). Finally, pSAM achieved AUC values of 0.9932, 0.9876, and 0.9865 in the training dataset, validation dataset, and testing dataset, respectively (Supplementary Fig. 2E). With a median nuclear localization probability of 0.5 as the cutoff, the true negative rate was 99.06%, and the true positive rate was 88.49% (Supplementary Fig. 2F). After screening the best threshold for obtaining the optimal sensitivity and specificity, the true negative and true positive rates were 98.76% and 93.54%, respectively (Supplementary Fig. 2G).

As mentioned above, there are a handful of in silico tools applicable to nucleus-specific or general prediction of subcellular localization. Considering both predictor availability and classification system diversity, we compared pSAM with the general localization prediction tools YLoc[22], PSORT_II[20], MultiLoc[26], DeepLoc[27] and MULocDeep[28] and the nuclear localization-specific prediction tool NucPred[29]. We also trained another model using 8,421 experimentally verified nuclear proteins and 18,278 non-nuclear proteins from NLSdb based on the same architecture as pSAM[42]. Comparing all tools, pSAM outperformed the other tools, although the 1000 proteins used for evaluation were not included in the training dataset of pSAM but were probably in the other tools mentioned above (Fig. 2A). Taken together, these findings indicate that pSAM shows state-of-the-art performance for predicting nuclear localization.

Since pSAM does not utilize any prior knowledge about nuclear localization, we further assessed its performance by interrogating the consistencies between the predictions and current knowledge. First, we performed predictions for specific proteins known to localize in the nucleus, including proteins containing ankyrin repeat domains, histones, proteins with validated NLS annotations, TFs and their cofactors. These proteins were predicted to have a significantly higher nuclear localization probability when compared against the whole proteome, especially the widely studied TFs and histones (Supplementary Fig. 2H). Furthermore, we predicted the nuclear localization score of the NLS peptides generated by Kosugi et al.[30], from which the relative NLS activities were detected according to the localization phenotype of the green fluorescent protein (GFP) reporter (Supplementary Data 2). The scores predicted by pSAM and the measured NLS activity showed satisfactory agreement (Fig. 2B). We then calculated the delta score, which refers to the alteration in nuclear localization probability upon simulated region loss or simulated mutation. As expected, the residues located in the experimentally validated NLSs exhibited higher delta scores than the residues in randomly selected regions (Fig. 2C and Supplementary Data 3).

We then performed residue-level contribution analysis (RCA) to dissect the decision-making process of pSAM and identify the DNLs (Supplementary Fig. 3A and Supplementary Data 4). We also calculated the relative positions of these DNLs in the proteins, observing that

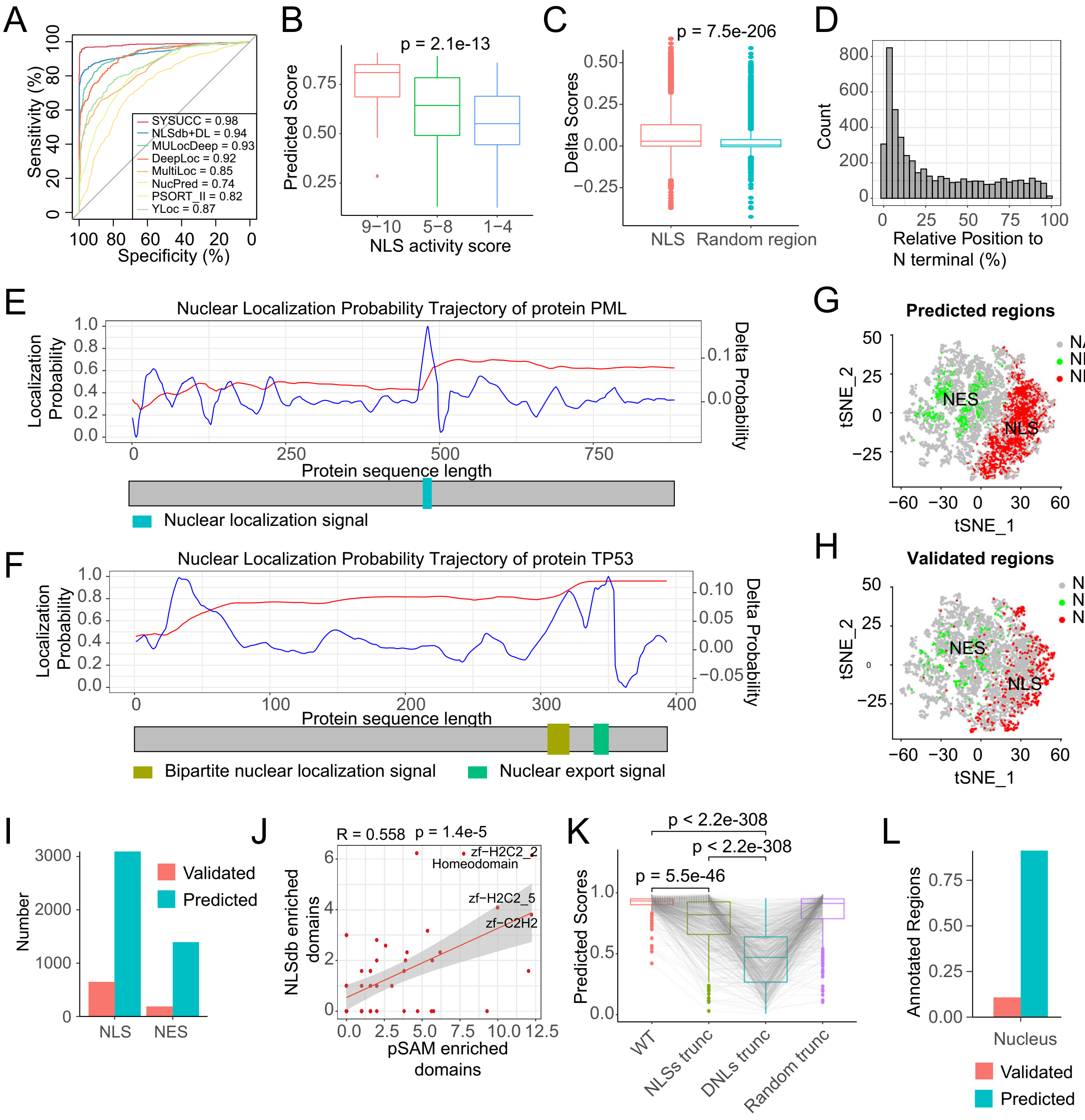


Fig. 2 | Evaluation of the performance of the pSAM model and identification of DNLs with residue-level contribution analysis. A pSAM shows significantly better predictive performance than existing tools for 1000 randomly selected proteins. **B** The pSAM-predicted probabilities on nuclear-localized peptides whose NLS activity scores were measured previously. The data are presented as a box-and-whisker graph (bounds of box: first to third quartile, bottom and top line: minimum to maximum, central line: median). NLS activity score classes: scores 9–10 (n = 57 peptides), scores 5-8 (n = 168 peptides) and scores 1-4 (*n* = 149 peptides). **C** The delta scores between amino acids located in known NLS regions and randomly selected regions. Alterations in nuclear localization probability was calculated for 8920 amino acids in NLSs and randomly selected regions, respectively. The data are presented as a box-and-whisker graph (bounds of box: first to third quartile, bottom and top line: minimum to maximum, central line: median). **D** The position relative to the terminal terminus of the predicted DNL regions. **E** The residue-level contribution of the nuclear localization probability of PML. **F** The residue-level contribution of the nuclear localization probability of TP53. **G** The t-SNE distribution of predicted NLS-like and NES-like regions. **H** The t-SNE distribution of known NLSs and NESs. **I** The number of known and predicted NLSs and NESs. **J** Domain analysis of NLS regions and NLS-like regions retrieved from NLSdb and pSAM. The error bands represent 95% confidence intervals. **K** Shuffling analysis of known NLSs and predicted determinants of nuclear localization (DNLs). For the 1752 selected proteins, the matched wildtype sequence (WT), validated NLS-truncated sequence, DNL-truncated sequence and randomly truncated sequence were compared. The data are presented as a box-and-whisker graph (bounds of box: first to third quartile, bottom and top line: minimum to maximum, central line: median). **L** Coverage of known and predicted targeting peptides for nuclear localization. Two-sided Kruskal-Wallis test was used for (**B**), two-sided Student's *t* test was used for (**C**), two-sided Pearson correlation test was used for (**J**), two-sided Wilcoxon test was used for (**K**). Source data are provided as a Source Data file.

although DNLs can be located at almost any part of the protein sequence, there is a striking enrichment at the N-terminus (Fig. 2D), which allows the NLS to be recognized by importin proteins[7]. We further compared the known NLSs and the newly defined DNLs. For example, promyelocytic leukemia protein (PML) is a protein that exhibits antiviral activity against both DNA and RNA viruses[43,44] and was annotated to be able to localize to the nucleus in the UniProt database; its NLS is the sequence 476-KRKCSQTQCPRKVIK-490. We observed that the delta score and nuclear localization probability rapidly increased from residue T473 to K490 (Fig. 2E), which caused the final nuclear localization probability of PML to reach 0.625, a moderate score. These findings for PML present high concordance with the findings of a previous study[45]. Furthermore, we performed a detailed analysis of another representative protein, cellular tumor antigen p53 (TP53), a tumor suppressor expressed in many tumor types that induces growth arrest or apoptosis, depending on the physiological circumstances[46,47]. Three DNLs on the p53 protein identified by pSAM were predicted to be significantly associated with nuclear localization (Fig. 2F). The second DNL, 319-KKK-321, overlapped with the known NLS, and the third DNL, 339-EMFRELNEALE-LADKQ-354, overlapped with the known NES, indicating that both regions are critical residues regulating p53 nuclear localization. Although p53 was annotated to have only one NLS and one NES according to the UniProt database, it was reported to have an NLS at positions 305-321 and two NESs at both terminals[48–50], consistent with the positioning of the first DNL, 28-ENNVLSPLPSQAMDDLM-44, at the N-terminus. The results indicated high concordance between the identified DNLs and known NLSs/NESs.

The predicted DNLs were then classified into NLS-like, NES-like and NA types to further interpret the predictions from pSAM, according to the residue composition of NLS and NES motifs[42]. To increase their credibility, we further calculated the false discovery rates (FDRs) of NLS-like and NES-like DNLs based on the sequence similarity with experimentally validated NLSs and NESs, respectively. The DNLs with an FDR < 0.05 were annotated as the predicted NLSs or NESs. The predicted NLS-like and NES-like DNLs were very close to the known NLSs and NESs (Fig. 2G, H), verifying the reliability of predicted NLSs and NESs. Our approach has defined 3,095 NLSs and 1,394 NESs, largely extending the knowledge base of NLSs and NESs (Fig. 2I and Supplementary Data 4).

The amino acid composition and sequence structure were similar between predicted NLSs/NESs and validated NLSs/NESs (Supplementary Fig. 3B). The motif analysis detected the regions significantly enriched in two classical NLS motifs and other regions enriched in DNA binding-related motifs (Supplementary Data 5). Interestingly, the domain analysis revealed that the NLS and NLS-like DNLs were enriched in DNA-binding domains such as homeodomains and helix-loop-helix DNA-binding domains (Supplementary Data 6-7). We compared the domain enrichment results for NLS-like regions with known NLS regions in NLSdb (Supplementary Data 6-7). The two datasets were consistent, and both contained enriched zinc finger structural domains (Fig. 2I), while previous studies revealed that DNA-/RNA-binding domains in diverse nucleic acid-binding proteins potentially function as NLSs/NESs[51,52]. We shuffled the annotated regions for proteins with known targeting peptides, and the significant decrease in predicted nuclear localization probabilities for the truncated proteins confirmed the accuracy of pSAM (Fig. 2J); the further decrease observed for the proteins with predicted truncated regions indicated that the current annotation of sequences critical for nuclear localization was insufficient (Fig. 2K), which is intuitive. In addition, we performed shuffling analysis of the N-terminus, middle region, and C-terminus of the human proteome. A striking decrease in nuclear localization probability was observed in N-terminally truncated proteins, consistent with the previous discovery that the N-terminus is enriched in key sequences critical for nuclear localization (Supplementary Fig. 3C).

Many DNLs have not been classified into NLSs/NESs due to a lack of sequence features, and a tempting speculation is that these proteins are transported into the nucleus through specific pathways. For example, gelsolin family members enter the nucleus by binding directly to NTF2 without the involvement of importin[53]. This suggested that some DNLs might be involved in protein interactions, but an approach for validating these functions and the nuclear proteins with which they interact remains elusive. Annotation of the determinants revealed that the potential targeting peptide annotation covers approximately 75% of human proteins (Fig. 2L), and further analysis on these regions will improve our understanding of subcellular localization.

Systematic analyses of SAMs across cancers

With this state-of-the-art predictor of protein nuclear localization, we speculate that abnormal protein nucleocytoplasmic shuttling is driven by cancer-related mutations. First, we compared the occurrence of known missense mutations in the determinants and other regions, and the results showed that the frequency of mutations was significantly greater in DNLs (Fig. 3A). The results of signature-corrected randomized mutation simulation showed that simulated mutations were significantly depleted in the predicted NLS/NES regions (Supplementary Fig. 3D). We investigated whether the mutations in DNLs were more likely to disrupt protein functions by introducing a deleterious score to measure their deleteriousness. Mutations in DNL regions were more deleterious (Fig. 3B), suggesting that these mutations significantly predispose patients to protein dysfunction. We calculated the relative delta nucleus probability according to the percentage of simulated mutation sites. We selected a two-tailed alpha > 5% as the statistical significance of mutation rates. When the mutation rate was 5%, the delta score was approximately 0.1 (exactly, 0.1022), so we used 0.1 as the threshold for determining whether a site was significant (Fig. 3C). Then, we calculated the potential changes in protein nuclear localization during mutagenesis for all cancer mutations in TCGA (Supplementary Data 8). Approximately 6.7% of the mutations in the NLS regions affected protein nuclear localization (Fig. 3D). Several protein functional clusters, including enzymes, TFs, and cancer-related proteins, were selected to inspect the effects of mutagenesis on protein nuclear localization. The prediction results from pSAM were consistent with most deleterious predictors (Supplementary Data 8). Pathway enrichment analysis revealed that the SAM-affected proteins were enriched in pathways related to mTORC1 signaling, IL2-STAT5 signaling, Myc signaling and PI3K-akt-mTOR signaling, all of which are involved in the occurrence and development of tumors, indicating that these SAMs would regulate the occurrence and development of tumors through various biological processes by affecting the nucleocytoplasmic shuttling of proteins (Fig. 3E).

As an approach to expanding the molecular mechanisms of SAMs, we reasoned that TFs might be a bridge connecting SAMs and biological functions because the mRNA expression levels of the substrates might serve as indices for TF protein function. Considering that TFs function after entering the nucleus, SAMs might block this process and lead to a decrease in the mRNA expression of their targets (Fig. 3F). As a proof of principle, differentially expressed genes were analyzed for tumors harboring mutations in TP63, which is a known tumor suppressor TF. TP63 upregulates the expression of *BLCAP*, encoding a protein that reduces growth arrest at the G(1)/S checkpoint and reduces cell growth by stimulating apoptosis[54]. When TP63 carries SAMs, the nuclear localization of its encoded protein decreases, which leads to decreased expression of *BLCAP*, therefore promoting the progression of tumors (Fig. 3G). ATF3 and *NOTUM* are another example of a TF-substrate pair that also reflects the loss of the nuclear localization ability of TFs due to mutation, regulating tumor cell stemness and growth[55].

Recently, an increasing number of proteins initially identified in the cytoplasm have also been observed to localize to the nucleus and perform different functions, especially in tumors and disease states[10–14]. We summarized the SAMs that increased the nuclear import ability of tumor suppressor genes and constructed a protein–protein interaction network for those genes that acquired nuclear localization capability (Fig. 4A). The cytoplasmic proteins that interacted with those genes were enriched in housekeeping pathways, such as those of the

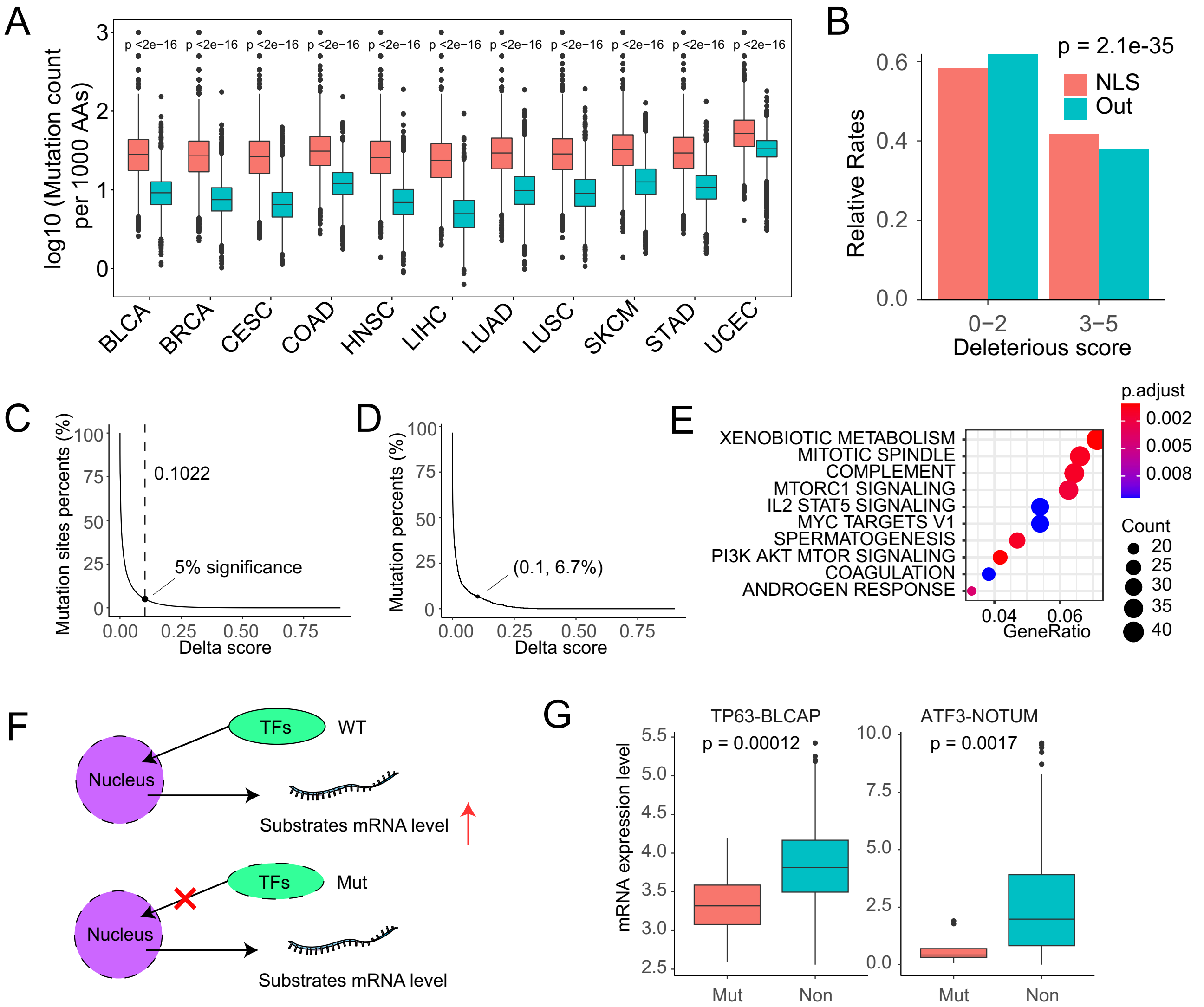


Fig. 3 | The discovery of abnormal nuclear localization driven by cancer-related mutations. **A** Mutation frequency distribution of predicted DNL regions and other regions. Proteins with annotated somatic mutation from 11 cancer types in TCGA were included (protein number: BLCA, $n = 3934$ proteins; BRCA, $n = 3746$ proteins; CESC, $n = 3582$ proteins; COAD, $n = 4210$ proteins; HNSC, $n = 3606$ proteins; LIHC, $n = 2812$ proteins; LUAD, $n = 3948$ proteins; LUSC, $n = 3909$ proteins; SKCM, $n = 4079$ proteins; STAD, $n = 4092$ proteins; UCEC, $n = 4417$ proteins). The data are presented as a box-and-whisker graph (bounds of box: first to third quartile, bottom and top line: minimum to maximum, central line: median). **B** The deleteriousness score distribution of mutations located in predicted DNL regions and other regions. **C** Relative delta scores according to the percentage of simulated mutation sites. We selected a two-tailed alpha > 5% as the level of statistical significance of mutation rates. When the mutation rate was 5%, the delta score was approximately 0.1 (0.1022), so we used 0.1 as the threshold for determining whether a site was significant. **D** The percentage of SAMs with different delta score cutoffs in NLS regions. **E** Enrichment analysis of mutated genes affecting nuclear localizations. **F** Diagram depicting the strategy for associating SAMs with putative transcription factor substrates. **G** COAD patients ($n = 411$ patients) carry SAMs in TP63 and ATF3, affecting the expression of their transcriptional substrates *BLCAP* and *NOTUM*. 15 SMAs in TP63 and 9 SAMs in ATF3 were detected, which divided the COAD patients into SAM-carried patients and non-SAM patients. The data are presented as a box-and-whisker graph (bounds of box: first to third quartile, bottom and top line: minimum to maximum, central line: median). Two-sided Wilcoxon test was used for (**A**) and (**G**), two-sided Chi-square test was used for (**B**) and two-sided hypergeometric test was used to calculate p values and Benjamini-Hochberg method was used to adjusted the p values for (**E**). TF, transcript factor, WT, wild type, Mut, mutated type. Source data are provided as a Source Data file.

tricarboxylic acid cycle, focal adhesion and cytokine-cytokine receptor interactions (Fig. 4B). However, proteins located in the nucleus play important roles in tumor-associated pathways, such as ubiquitin-mediated proteolysis and the cell cycle, providing a potential explanation for mutations in tumor suppressor genes leading to tumor progression. TNK1 is known to be localized in the cytoplasm and may serve as an oncogene or a tumor suppressor depending on the cell context[56,57]. However, few studies have determined how TNK1 gene mutations cause cancer, and none of them have considered the potential effect of nuclear localization. We discovered numerous mutations that could result in TNK1 nuclear localization (Figs. 4A and 4C). 14-3-3 proteins interact with and inhibit the kinase activity of TNK1 in the cytoplasm[56]. TNK1 gained nuclear localization capability while harboring the detected SAMs. Tyrosine kinases can regulate a wide range of TFs, and we discovered that some carcinogenic TFs, such as JUN, could act downstream of TNK1. The above analysis provides a possible explanation for the duality of TNK1 function.

We attempted to determine the relationship between phosphorylation and nuclear localization in the DNL region. We first mutated the Ser/Thr sites on DNLs to the phosphomimetic Asp and the inactive mimic Ala. After further enrichment analysis of proteins with significantly altered nuclear localization potential, we found that

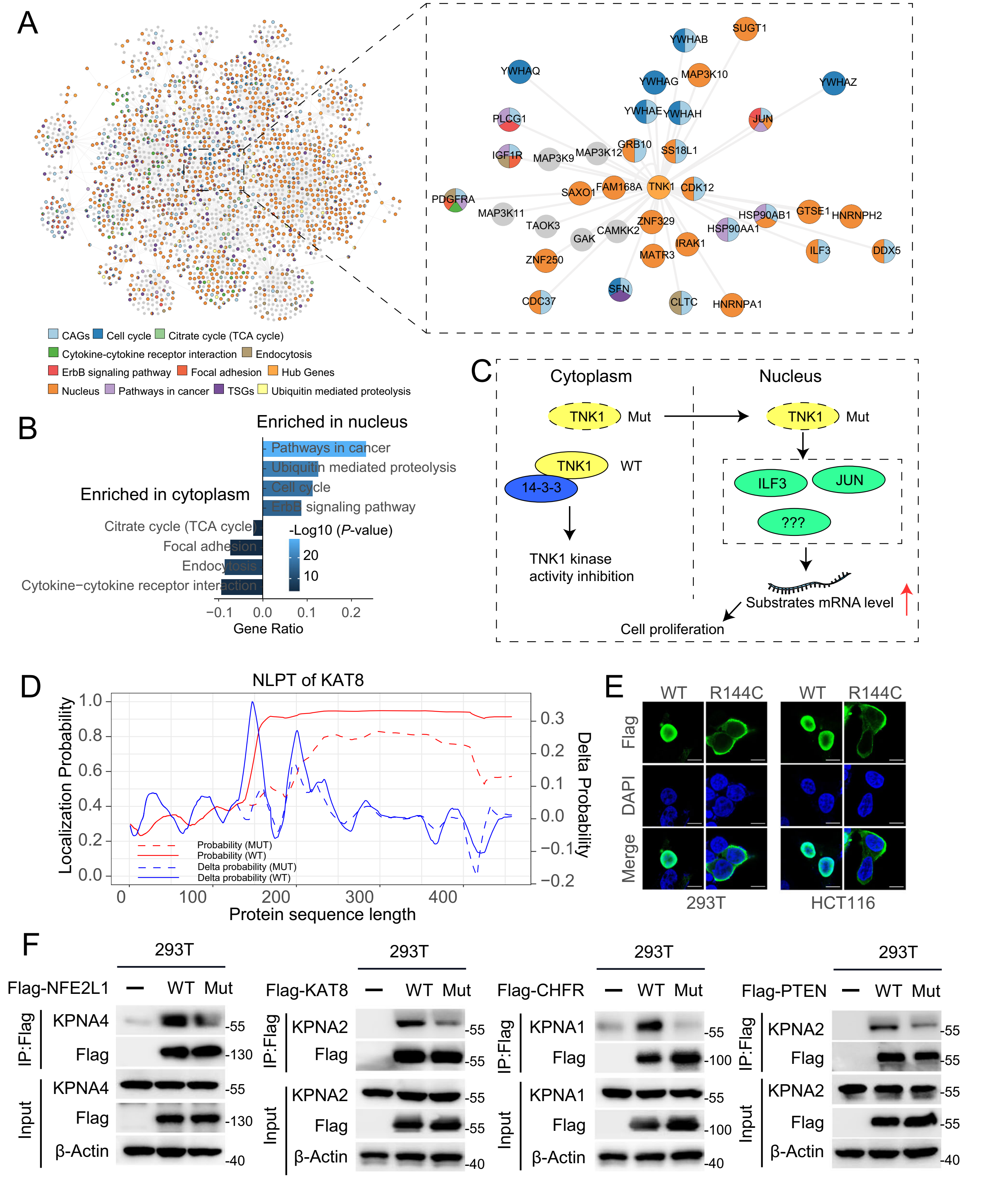


Fig. 4 | Validation of nuclear localization affected by mutations. A PPI network of tumor suppressor genes harboring SAMs and their experimentally validated interacting proteins. **B** KEGG enrichment of the interacting proteins located in different nucleocytoplasmic contexts. **C** Diagram of the mechanism by which TNK1 functions as a tumor suppressor in the cytoplasm and SAM blocks this process by entering the nucleus, which induces an oncogenic function. **D** The residue-level contribution of the nuclear localization probability of KAT8 and its R144C mutant. **E** Immunofluorescence staining for KAT8 and its R144C mutant. Scale bar = 10 μm. Representative experiment out of $n = 3$ biological replicates. **F** Affinity validation of selected wild-type and mutant proteins with importin. Representative experiment out of $n = 3$ biological replicates. Two-sided hypergeometric test was used to calculate p values and Benjamini-Hochberg method was used to adjusted the p values for (**B**). WT, wild type; Mut, mutated type; NLPT, nuclear localization probability trajectory. Source data are provided as a Source Data file.

phosphorylated PLK3 substrates were more likely to have altered nuclear localization ability (Supplementary Fig. 4A), and *PLK3* expression was associated with the overall survival of patients (Supplementary Fig. 4B). PLK3 is a mediator of apoptosis and cellular stress that regulates the cell cycle and functions as an oncogene or tumor suppressor in a variety of cancers. However, researchers have not elucidated the function of the phosphorylation of its substrate by PLK3, and our analysis provides a possible molecular mechanism by which PLK3-mediated modification of its substrates affects nuclear localization and thus regulates downstream biological networks.

Validation of the predicted perturbation of nucleocytoplasmic shuttling driven by SAMs

Due to the lack of standard datasets to assess the performance of pSAM, we selected ten SAMs (KAT8 p.R144C, DEAF1 p.K304N, NFE2L1 p.R674C, CMAS p.R201W, PTEN p.R14M, NHLRC1 p.G274W, CHFR p.P255L, SYNPO2 p.G4E, MAPK1 p.G136E and TP63 p.R393Q), which were predicted to have a significant impact on nuclear localization probability (Fig. 4D and Supplementary Fig. 5A), to further validate the performance of pSAM. Among all 10 predicted SAMs, eight (KAT8 p.R144C, DEAF1 p.K304N, NFE2L1 p.R674C, CMAS p.R201W, PTEN p.R14M, NHLRC1 p.G274W, CHFR p.P255L and SYNPO2 p.G4E) were proven to exhibit a considerable degree of decreased nuclear localization, and only two SAMs failed to be demonstrated with positive results (Fig. 4E and Supplementary Fig. 5B). Previously, the KAT8 lysine acetyltransferase was reported to be essential for cancer development, and several studies have shown that KAT8 is a potential therapeutic target for cancer[58–60]. Nuclear localization probability trajectory (NLPT) analysis revealed that KAT8 underwent a significant change in nuclear localization after mutation at R144 (Fig. 4D), and immunofluorescence staining revealed that the KAT8 R144C mutation significantly abolished the nuclear localization ability of the protein (Fig. 4E).

Furthermore, we scanned SAMs on NES regions, and then predicted their impacts on the nuclear localization probability of proteins. The protein CHP1 has two validated NES regions: 138-147 and 176-185 (Supplementary Data 2). The first NES region was predicted to have a SAM, namely CHP1 p.R140L (Supplementary Data 8). Therefore, we experimentally examined the effects of the p.R140L mutation on the nuclear localization probability of CHP1. The results revealed that the wild type of CHP1 was excluded from the nucleus, while the mutated type of CHP1 was retained in the nucleus, demonstrating that the p.R140L mutation would significantly impair the nuclear export function of the NES motif (Supplementary Fig. 5C, D).

In addition, we observed interesting cases in which individual point mutations had little effect, but synergistic mutation of two loci significantly reduced the nuclear localization of the protein (Supplementary Fig. 5E, F). Through a literature search, we selected six proteins with the potential to bind to the nuclear import factor importin. We identified the impact of mutations on the binding of these proteins to importins through affinity tests. Among these six proteins, four exhibited decreased binding to importin due to mutation effects (Fig. 4F).

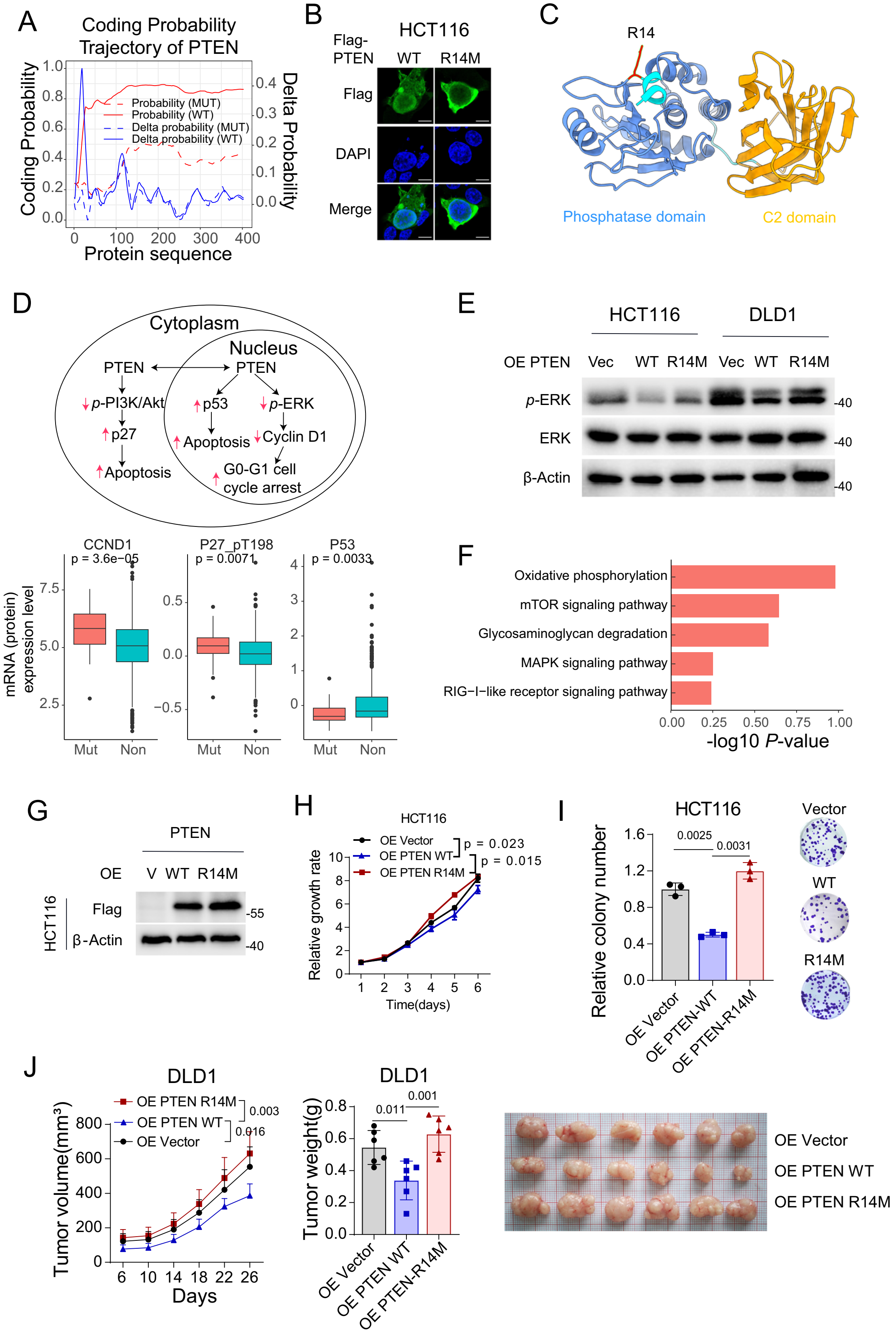
Since the current NLS annotations were incomplete, we performed a truncation analysis of the identified DNLs and screened proteins for experimental validation (Supplementary Data 9). As shown in Supplementary Fig. 6, after truncating the known NLS signals of QKI and METTL3, the nuclear entry ability of these proteins did not decrease significantly. However, the truncation of the predicted regions significantly reduced the nuclear localization level. METTL3 is an N6-methyltransferase that methylates adenosine residues at the N(6) position of RNAs, thus regulating various cellular processes, such as the response to DNA damage, the circadian clock and primary miRNA processing[61–63]. According to the prediction from pSAM, METTL3 contains three potential DNLs, the third of which is a known NLS[64]. Interestingly, truncation of the NLS did not lead to a significant decrease in the nuclear localization fluorescence signal. After truncating the three predicted regions, the nuclear localization ability of METTL3 decreased significantly, which further verified the reliability of the prediction from pSAM. In addition, both truncating the known NLS signal and identifying DNLs led to a decrease in the nuclear accumulation of NTH. According to the NLPT analysis, the known NLS sequence exhibited the highest delta score, suggesting that this region plays a key role in regulating NTH nuclear localization. HIPK3 and CDK8, which have no known NLSs, also showed a significant decrease in nuclear entry after the predicted DNLs were truncated (Supplementary Fig. 6). Taken together, these findings show that pSAM can precisely predict the sequence determinants critical for nuclear localization at the residue level.

R14M disrupted the nuclear localization and tumor-suppressive function of PTEN

PTEN has been reported to be a tumor suppressor with lipid phosphatase activity and shown to be a potential therapeutic target for cancer[65,66]. PTEN localizes to both the nucleus and the cytoplasm and shuttles between each compartment through numerous mechanisms[67]. Thus, we selected the PTEN R14M mutant for subsequent experimental and functional validation. NLPT analysis revealed that PTEN underwent a significant change in nuclear localization after mutation at R14 (Fig. 5A and Supplementary Fig. 5A), and immunofluorescence staining revealed that the PTEN R14M mutation significantly abolished the nuclear localization of the protein (Fig. 5B and Supplementary Fig. 5B). R14M is located at the junction where the phosphatase and an N-terminal α-helix region are connected in the crystal structure of PTEN (Fig. 5C), which is exposed at the surface of PTEN. Previous studies have emphasized the critical role of the N-terminal sequence in promoting the nuclear entry of PTEN[68]. PTEN modulates cell survival by antagonizing the PI3K/Akt/p27 signaling pathway in the cytoplasm[69,70]; however, when it is located in the nucleus, it increases cyclin D1 levels and G0-G1 cell cycle arrest through the downregulation of ERK[71] and mediates p53-dependent apoptosis[72]. Nuclear PTEN exerts a stronger tumor-suppressive effect than cytoplasmic PTEN[67]. Our analysis suggested that the *CCND1* mRNA and p27 phosphorylation levels were significantly increased in patients carrying PTEN mutations, while the p53 protein level was significantly decreased (Fig. 5D). We confirmed that PTEN could regulate the phosphorylation of ERK and that the R14M mutation disrupted this process (Fig. 5E and Supplementary Fig. 7A). To identify the biological functions of the PTEN R14M mutation, we performed transcriptome analysis and found that several oncogenic pathways, such as the oxidative phosphorylation, mTOR and MAPK signaling pathways, were dysregulated (Fig. 5F and Supplementary Fig. 7B). Further cytological experiments showed that the PTEN R14M mutant significantly increased tumor growth and colony formation (Fig. 5G–I and Supplementary Fig. 7C). In vivo tumorigenesis experiments showed that the PTEN R14M mutant significantly reduced the tumor-suppressive function of PTEN (Fig. 5J and Supplementary Fig. 7D). Based on these results, the PTEN R14M mutation disrupts the function of the PTEN protein by regulating its nuclear localization. In addition, we randomly selected a tumor suppressor gene, the E3 ubiquitin-protein ligase CHFR, and observed similar biological phenomena (Supplementary Fig. 5A, B and Supplementary Fig. 7D–G).

The SAM-carrying gene network connects nucleocytoplasmic shuttling and carcinogenesis

Understanding how SAMs interact with tumorigenesis is a longstanding challenge. The fundamental question is what effect the mutation has on the structure or function of the protein and which pathways are targeted[73,74]. We constructed a tumorigenesis-related



network by integrating the genes that harbor nuclear SAMs and cancer hallmark pathways to identify the position of nuclear SAMs in human cancer (Fig. 6). We classified these genes into eleven classes: hypoxia, myogenesis, spermatogenesis, PI3K/AKT/mTOR, complement, MYC targets, epithelial-mesenchymal transition, cholesterol homeostasis, mitotic spindle and androgen response. According to the network, most genes are known cancer-associated genes, drug targets, enzymes and metastasis-related genes. We further mapped the cancer hallmark genes from a well-curated database named HOC[75] to the selected genes, and enrichment analysis revealed that eight cancer hallmark traits, namely, sustained proliferation, evasion of antigrowth signaling, immune evasion, replicative immortality, invasion and metastasis, angiogenesis, cell death and metabolism, are significantly affected by nuclear SAMs (Fig. 6).

Fig. 5 | The PTEN R14M mutation affects the function of the PTEN protein by regulating its nuclear localization. A The residue-level contribution of the nuclear localization probability of PTEN and its R14M mutant. **B** Immunofluorescence staining for PTEN and its R14M mutant. Scale bar = 10 µm. **C** Locations of the phosphatase domain, C2 domain and K14 (red) in the 3D structure of PTEN. **D** Diagram of the mechanism by which PTEN functions as a tumor suppressor in the nucleus and cytoplasm. Boxplot showing the substrate (*CCND1* mRNA [mutated, *n* = 39 samples; non-mutated, *n* = 487 samples], p27 phosphorylated protein and p53 protein [mutated, *n* = 34 samples; non-mutated, *n* = 354 samples]) expression levels of PTEN in UCEC patients carrying SAMs. The data are presented as box-and-whisker graphs (bounds of box: first to third quartile, bottom and top line: minimum to maximum, central line: median). **E** Western blot showing the expression and phosphorylation level of the ERK by PTEN R14M mutation in two cell lines. **F** KEGG pathway enrichment analysis of differentially expressed genes between PTEN-overexpressing and R14M-mutated cell lines. **G** Western blot showing the protein expression by PTEN R14M mutations. **H**, **I** Cell growth assays showing the effects of PTEN and its mutant on the HCT116 cell lines. **H** Relative growth rare and (**I**) Relative colony number. Three groups were compared: OE vector (*n* = 3 replicates), OE PTEN WT (*n* = 3 replicates) and OE PTEN R14M-mutated (*n* = 3 replicates). The data are presented as the mean±S.D. **J** In vivo tumorigenesis experiments of PTEN and its mutants in the DLD1 cell line. Three groups were compared: OE vector (*n* = 6 mice), OE PTEN WT (*n* = 6 mice) and OE PTEN R14M-mutated (*n* = 6 mice). The data are presented as the mean±S.D. Representative experiment out of *n* = 3 biological replicates for (**B**, **E**, **G**). Two-sided Wilcoxon test was used for (**D**), two-sided hypergeometric test was used for (**F**), and two-sided Student's *t* test was used for (**H**–**J**). Source data are provided as a Source Data file.

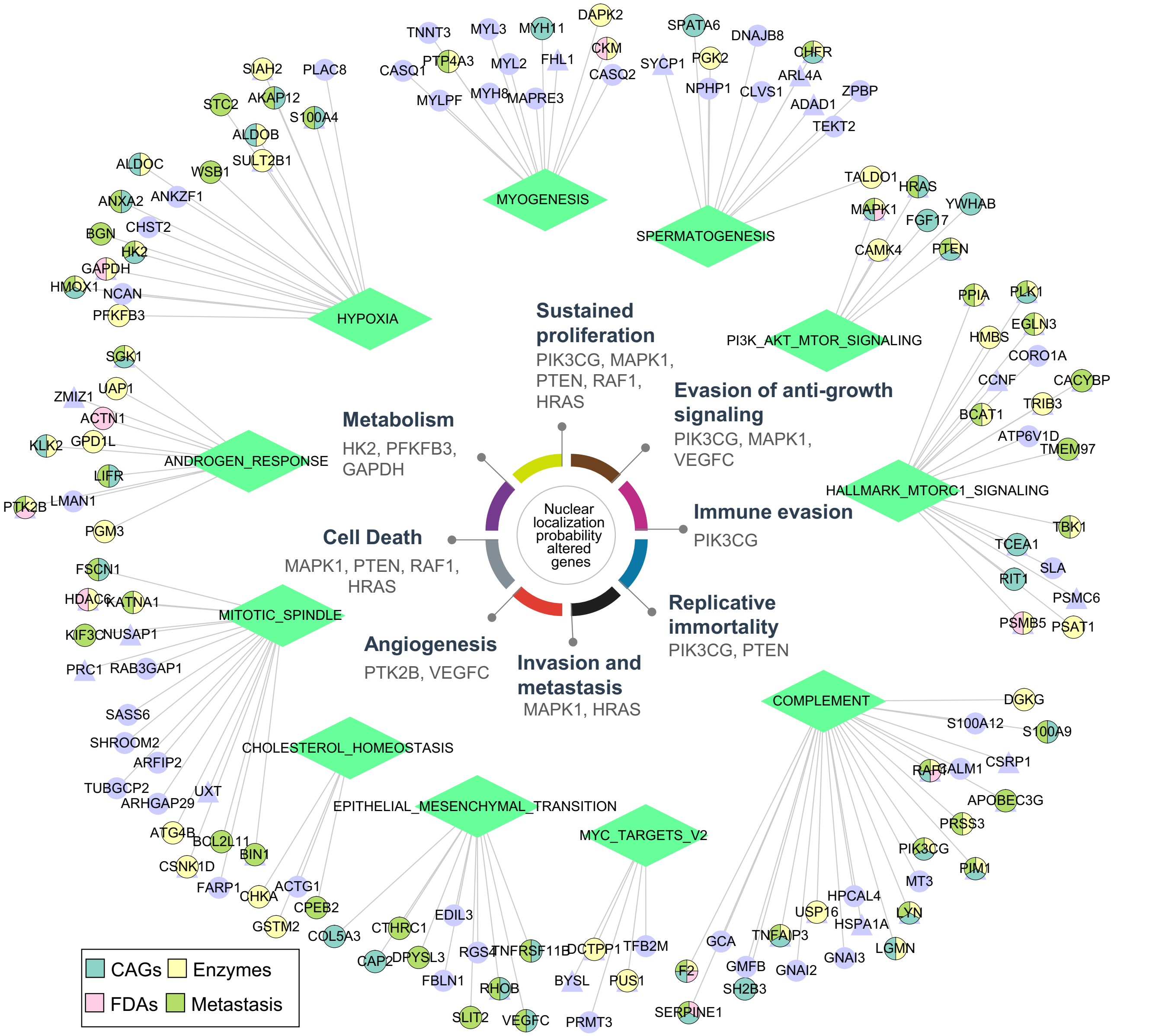


Fig. 6 | Nuclear localization probability altered genes regulating the network that connects nuclear localization and carcinogenesis. The 139 genes were classified into eleven groups based on their major biological functions. Representative hallmark genes are shown in the circle. Source data are provided as a Source Data file.

Discussion

Nuclear localization enables and determines protein function and then orchestrates intracellular processes and responses to external signals. Protein shuttling between different subcellular compartments is driven by distinct targeting signals and regulates a broad spectrum of biological processes, particularly in tumor and disease states[3,11,13]. However, until now, due to the lack of a robust tool and method to infer the impact of amino acid perturbations on protein nuclear localization at the residue level, there has been a lack of systematic analysis of how mutations affect the development of tumors by influencing protein nuclear localization. Here, we first systematically identified the sequence determinants for distinct nuclear localization by performing ab initio analysis and visualized the residue-level contribution of nuclear localization probability across human proteins. Moreover, we investigated the potential disruption of nucleocytoplasmic shuttling induced by single-nucleotide or truncated mutations, followed by experimental validation, which highlighted the general mechanism of tumorigenesis driven by mutations that alter the subcellular shuttling of proteins. While validating all mutations is beyond our capabilities, the predicted disruptions confirmed by the experiments support that our proposed model might facilitate investigations into the molecular mechanisms of protein nucleocytoplasmic shuttling and provide candidates for further experimental discoveries.

The RCA analysis helped us identify the determinants for protein nuclear localization. PML can localize to the nucleus and exhibits antiviral activity against both DNA and RNA viruses[43,44]. The identified DNL was consistent with previous studies[45]. The tumor suppressor protein TP53, which is expressed in many tumor types, induces growth arrest or apoptosis, depending on the physiological circumstances[46,47], and was annotated to have only one NLS and one NES in the UniProt database. These known NLSs and NESs are consistent with the predictions from pSAM. In addition, pSAM suggests that the N-terminal sequence of TP53 may be associated with nuclear localization. METTL3 is an N6-methyltransferase that regulates various processes, such as the response to DNA damage, the differentiation of stem cells, the circadian clock and primary miRNA processing[61–63]. Interestingly, truncating the known NLS region did not significantly decrease the nuclear localization fluorescence signal. After truncating the predicted determinants (comprising the known NLS), the nuclear localization ability of METTL3 decreased significantly. Taken together, pSAM provides comprehensive annotation of localization signals on proteins and can identify both classical and unconventional localization signals.

The interpretation of our results should be of interest to researchers in several fields. We discovered that the DNL regions are highly correlated with DNA-binding domains, indicating the potential role of zinc finger domains in protein nuclear localization (Supplementary Data 6), which was proven by previous studies showing that DNA-/RNA-binding domains in diverse nucleic acid-binding proteins function as NLSs/NESs[51,52]. We speculated that some proteins that do not contain potential NLS-like/NES-like DNLs may be transported into the nucleus through relatively specific pathways. For example, gelsolin family members enter the nucleus by binding directly to NTF2 without the involvement of importin[53]. These findings suggested that a variety of DNLs could perform protein-interacting functions, but the approach for validating their protein-interacting functions and the nuclear proteins with which they interact remains elusive. After the annotation of determinants, the potential targeting signal annotation covered approximately 75% of the human proteins, which will greatly improve our understanding of nuclear localization.

Currently, only a few experimental studies have investigated this topic, and most studies on cancer mutations have focused on functional domains, such as catalytic domains, rather than on nuclear localization signals. We identified many somatic mutations that may regulate protein function by affecting nuclear localization, and these functional annotations of mutations may help researchers further understand the pathogenesis of disease. PTEN has been reported to shuttle between the nucleus and the cytoplasm via numerous mechanisms[67]. Our analysis showed that PTEN underwent a significant change in nuclear localization after mutation at R14, and we further confirmed that the PTEN R14M mutation significantly abolished the nuclear localization ability of the protein and reduced the tumor-suppressive function of PTEN. PTEN modulates cell survival by antagonizing the PI3K/Akt/p27 signaling pathway in the cytoplasm[69,70], whereas when it is located in the nucleus, it increases cyclin D1 levels and G0-G1 cell cycle arrest through the downregulation of ERK[71] and mediates p53-dependent apoptosis[72]. We summarized the SAMs that increase the nuclear importing ability of tumor suppressor genes and found that the cytoplasmic proteins that interact with those genes are enriched in housekeeping pathways, while proteins located in the nucleus play important roles in oncogenic pathways. However, few studies have determined how TNK1 functions as an oncogene or a tumor suppressor depending on the cell context. We proposed a potential mechanism by which the kinase activity of TNK1 is inhibited by 14-3-3 proteins in the cytoplasm and regulates some carcinogenic TFs in the nucleus when mutated. We elucidated the potential mechanisms of some proteins in tumors from the perspective of mutations leading to abnormal nuclear localization, resulting in abnormal protein function. The molecular mechanisms of many of these genes deserve further investigation.

The existing knowledge posed limitations to the prediction model constructed in this study. On the one hand, positive datasets for nuclear localization should be reliable, but negative datasets are challenging. Although some proteins were not detected in a certain subcellular compartment, this lack of detection may be due to research limitations, and their existence and function in certain subcellular compartments have not been elucidated. For example, in recent years, an increasing number of proteins that were originally detected in the cytoplasm have also been observed to localize in the nucleus, and some of them have been proven to perform completely different functions upon entering the nucleus, especially in tumor and disease states[10–14]. This limitation will be overcome by determining the comprehensive nuclear localization of more proteins. On the other hand, the proportion of targeting peptides of proteins, particularly internal signals such as NLSs/NESs, has been experimentally confirmed to be limited, which leads to a lower concordance between known NLSs/NESs and identified DNLs than between known presequences and known DNLs. Lastly, environmental conditions could affect the subcellular location and shuttling of proteins, but the environmental conditions across diverse organisms and histopathological settings are complicated and are hard to be assessed. Nevertheless, such influences are not directly caused and are attained through intracellular mechanisms such as mutations, protein posttranslational modifications or protein cleavage, which change the natures of proteins. Thus, although environmental conditions could provide some clues, our study focused on the characteristics of the proteins and how genetic variants alter their subcellular location. Further detailed and comprehensive investigation about the impacts of environmental conditions will better corroborate the reliability of our results.

Although we only verified the SAMs and DNLs in several proteins, we anticipate that the modeling methods and results mentioned in this article will provide opportunities to dissect the molecular landscape of nucleocytoplasmic shuttling and reveal the protein variants mediating dysregulated localization and dysfunction in cancers and diseases. Moreover, we constructed the iNuLoC database (http://inuloc.omicsbio.info) and prediction tools based on the nuclear localization prediction model to facilitate researchers' use, which might provide guidance to researchers studying the nuclear localization of proteins.

Methods

Ethical statement

Our study was approved by the Medical Ethics Committee of Sun Yat-sen University Cancer Center (G2023-356) and complied with the Declaration of Helsinki. All animal experiments were performed based on the protocol approved by Institutional Ethics Committee for Clinical Research and Animal Trials of the Sun Yat-sen University Cancer Center (No. L102012023220Y), and all mice were housed in a temperature-controlled room under pathogen-free conditions on a 12 h light–dark cycle. Transplanted tumors were not to exceed a diameter of 2.0 cm or 10% of body weight as permitted by the Institutional Ethics Committee for Clinical Research and Animal Trials of the Sun Yat-sen University Cancer Center.

TCGA mutation distribution and deleteriousness analysis

The mutation data of 11 cancer types (BLCA, BRCA, CESC, COAD, HNSC, LIHC, LUAD, LUSC, SKCM, STAD, and UCEC) were downloaded from the Xena Browser (https://xenabrowser.net/datapages/) for further analysis. Only missense variants were retained. For each protein, we counted the number of mutations falling within the NLS region and other regions as $N_{(nls)}$ and $N_{(other)}$, respectively, and then calculated the number of mutations that occurred per thousand amino acids in the NLS region and other regions and defined them as $Mut_{(nls)}$ and $Mut_{(other)}$, respectively.

$$Mut_{(nls)} = \frac{N_{(nls)}}{l1} \times 1000 \quad (1)$$

$$Mut_{(other)} = \frac{N_{(other)}}{l2} \times 1000 \quad (2)$$

where $l1$ and $l2$ are the amino acid numbers of NLS regions and other regions, respectively.

We introduced a deleteriousness score to measure their deleteriousness and investigated whether mutations in NLS regions are more likely to impair specific protein functions, as reported by Chen et al.[76]. We defined a mutation as deleterious to protein function if a given SNV disrupted a functional domain or regulatory region of a specific protein. Thus, five software programs, namely, SIFT[77], PolyPhen2 HDIV[78], PROVEAN[79], FATHMM[80] and AlphaMissense[81], were used to predict the functional consequences of our identified NLS region mutations. To ensure the accuracy of the predictions, we further defined the deleteriousness score by integrating the predictions of the above five software programs. Specifically, the deleteriousness score was calculated by counting the number of methods described above that considered the mutation to be deleterious. A deleteriousness score of 0 means that the predicted mutations are tolerated in all methods. In contrast, a deleteriousness score of 5 means that the corresponding mutations are predicted to be harmful by all five predictors. Thus, the deleteriousness score can range from 0 to 5, where a higher score indicates a greater probability of deleteriousness. Next, a two-tailed proportionality test was applied to assess the differences in deleteriousness between mutations that occurred in NLS regions and those that occurred in other regions.

Gene set collection

The TFs and cofactors were downloaded from AnimalTFDB[82]. The proteins with experimentally validated NLS annotations were downloaded from the SeqNLS[31] and UniProtKB databases. The proteins with NES annotations were downloaded from the ValidNESs[83], NESbase[84], and UniProtKB databases. Proteins that contained ankyrin repeat domains were retrieved from the Pfam database[85]. The list of histones was downloaded from the UniProtKB database. We also extracted 8421 experimentally verified nuclear proteins and 18,278 non-nuclear proteins from NLSdb[42]. NLSdb provided a large number of NLS motifs generated in silico, including 2253 NLSs, which were also used in the validation of our data[42].

Signature-corrected randomized mutations

Signature-corrected randomized mutations were generated as described[86]. To compute the signature-corrected randomized mutations, we used the *MOAT* (Mutations Overburdening Annotations Tool; https://github.com/gersteinlab/MOAT), which is a computational system for identifying significant mutation burdens in genomic elements with an empirical, nonparametric method[87]. We used the variant distribution simulator, namely *MOAT-s*, to produce shuffled mutation sets preserving mutational signatures associated with cancer while removing the local position-specific effects of cancer mutations. TCGA mutation data of 11 cancer types was selected as the input of variant file. Here, observed mutations were randomly shuffled within a 50,000-base pair (bp) window. During the shuffling process, the trinucleotide context of a mutation was preserved. Mutations affecting known cancer driver genes were excluded from the generation of these randomized mutation sets. As the reference genome file of the MOAT was based on GRCh37 (hg19), we used the UCSC *liftover* software to transform the gene location files.

Driver gene analysis

We analyzed whether these genes were significantly mutated in human cancers to ascertain which genes might promote tumor development or repress tumor growth. The 20/20+ method[88] is based on a random forest that scores each gene as a cancer driver gene. We performed the 20/20+ method using default parameters for each of the 11 cancer types individually and aggregated the results of all cancers together. A significant putative driver gene was determined with a cutoff P value less than 0.05.

CERES dependency score

CERES dependency scores were based on data from CRISPR technology to knock down target genes and perform cell depletion assays[89]. Target gene dependency scores were obtained from the DepMap web portal (https://depmap.org/portal/download/). A lower CERES score for a particular gene in a specific cell line indicates that the gene of interest has an essential function in that cell line. A score of 0 indicates that the gene is not essential; accordingly, a score of -1 is the median of all panessential gene function scores.

Dataset preparation

We downloaded all reviewed protein sequences of eukaryotes from the UniProt database[41] (UniProt release 2021_01) for a total of 157,460 proteins from 6,852 organisms. To ensure high data quality, the proteins with no nuclear localization information or uncertain annotations with keywords such as 'by similarity', 'potential', and 'probable' were classified into the "without localization" subgroup. Then, we extracted their localization information based on UniProt annotations and classified the remaining proteins into "nuclear localization" and "nonnuclear localization" subgroups based on whether they contained nuclear localization information. Since CNNs take fixed-size inputs, protein sequences longer than 2000 amino acids or shorter than 20 amino acids, accounting for 2.75% of sequences, were excluded from the dataset. As a result, 36,016 proteins with nuclear localization and 63,904 proteins with nonnuclear localization were ultimately extracted. For some protein sequences containing nonstandard amino acids (i.e., B, O, U, and Z), we designated these amino acids as pseudo amino acids, i.e., "-". We then split the whole dataset into a training set, a validation set, and a testing set at a ratio of 7:2:1. The training dataset was used for cross-validation and hyperparameter tuning to generate the optimal model. The validation dataset was used to adjust the model weights, determine the number of training session epochs, and evaluate the model performance. In addition, we randomly selected 1000 protein sequences (500 nuclear localization proteins and 500 nonnuclear localization proteins) from the testing dataset to compare the performance of the model with that of other protein nuclear localization prediction tools.

Model construction

We designed a series of deep learning-based models to predict the potential nuclear localizations of proteins and then evaluated model performance to select the optimal prediction model. We collected eukaryotic protein sequences and localization annotations from the UniProt database[41] (Supplementary Fig. 2A). Approximately 30% of the proteins lacked clear subcellular localization annotations, and the ratio of nuclear to non-nuclear protein localization was approximately 1:2. To avoid too many outliers, protein sequences with extreme lengths of less than 20 amino acids or more than 2000 amino acids were excluded. The input protein sequences are one-hot encoded into a (2000 ×21) embedded matrix, and this matrix is later used as the input for the constructed CNN layers. We randomly divided the whole dataset into a training set, validation set, and testing set at a ratio of 7:2:1. Then, we use a combined deep neural network based on CNN, BiLSTM, and attention mechanisms to construct the model to predict nuclear localization probability (Fig. 1A). The input protein sequences were one-hot encoded into a (2000 ×21) embedded matrix, which was used as the input for the CNN layers. Different convolutional kernel sizes (e.g., 1, 2, 4, 8, 12, and 16) were combined to form a multisubnetwork CNN layer. Batch normalization and rectified linear unit (ReLU) activation are applied before each convolutional layer. In addition, all padding methods in the CNN layers are valid padding methods, which means that the convolution kernel is not allowed to go beyond the original image boundaries. Subsequently, we used the concatenate layer to merge the output of the multilayer residual block into one layer, followed by an additional batch normalization layer and ReLU activation function.

For subcellular localization, the precise identification of targeting signals often captures the interest of biologists. LSTM coupled with attention mechanisms has proven to be effective in addressing this issue. Here, we compute the activation of LSTM memory blocks (hidden states) at each position in the protein sequence and apply an attention function to determine the significance of each hidden state, which represents the importance of the underlying positions in the input sequence. This enables us to assign importance to each position in the protein sequence to decode essential amino acids that influence subcellular localization. Subsequently, we calculate the attention-weighted sum of all hidden states, using it as input to a dense feed-forward neural network with a hidden layer and Adam output for the final prediction of nuclear localization.

Hence, the model's output is passed to the BiLSTM layer, which uses the L2 Regularizer to apply a penalty to the layer's kernel, bias, and output with a value of 0.01. After these two layers, a dropout layer is used to prevent further model overfitting. Finally, we used the attention layer constructed based on Luong-style attention to assign an importance score to each hidden state of the BiLSTM layer. The attention-weighted sum of all hidden states was used as the input of the final dense feed-forward neural network with one hidden layer and output as a two-dimensional classification probability score. The optimizer of the model is the Adam optimizer, the loss calculation function is a categorical cross-entropy loss function, the evaluation metric is accuracy, and the model training batch size is set to 64. In addition, we introduce the early stop method to prevent further model overfitting, and the evaluation metric is that the loss in the validation set does not decrease after more than two successive epochs.

The rationale for selecting this architecture is grounded in the following conceptual flow of information from distinctive subsequences to categorization. The CNN filters address the challenge of aligning sequences and offer a trained filter-bank attuned to specific sequence patterns. If such a pattern appears anywhere in the sequence, it results in an elevated absolute input to the LSTM at that specific position. The bidirectional LSTM layers then locally integrate this information both forward and backward in sequence. The attention function, leveraging its acquired contextual awareness, conveys discriminative information from these RNN hidden states to the dense layer and subsequently to the Adam classifier.

Hyperparameter tuning

Hyperparameter tuning was performed over the residual block combination, CNN node size, dropout rate, and recurrent gate size. The alternative residual block combination consists of three or four parallel subnetworks with four combination modes: (4×1, 8×1, 12×1), (1×1, 2×1, 4×1, 8×1), (4×1, 8×1, 1×1) and (4×1, 8×1, 12×1, 16×1). After comparing the performances on the training and validation sets, we selected combinations of (4×1, 8×1, 1×1), (1×1, 4×1, 8×1), and (8×1, 1×1, 4×1) as the optimal model combinations. To define the best model architecture, the CNN node size, dropout rate, and recurrent gate size were initially set to 64, 0.5, and 128, respectively. Then, we determined the CNN node size, dropout rate, and recurrent gate size one by one, keeping two of the parameters unchanged and adjusting one of the remaining parameters. The CNN node ranges were 32, 64, 128, and 256. The dropout rate range was 0.3, 0.4, 0.5 and 0.6. The range of the recurrent gate was 64, 128, 256, and 512. We selected a CNN node size = 64, dropout rate range = 0.4, and recurrent gate size = 128 for further analysis because these values yielded the highest validation accuracy.

Model performance evaluation

As previously described[90], we adopted two measurements to evaluate the model's performance: sensitivity (Sn) and specificity (Sp). To evaluate the model performance, we performed 10-fold cross-validation and calculated the AUC values. An AUC value of 1 indicates that the model can predict each series accurately, while an AUC value of 0.5 indicates that the model makes random predictions. A higher AUC indicates that the model is more effective.

Comparison of pSAM with existing tools

We compared the prediction accuracy of pSAM with that of several other tools to further evaluate its performance. Although dozens of in silico tools have been developed for nuclear localization prediction, some of them are no longer accessible. Considering both code and webserver availability and classification system diversity, we selected the general localization prediction tools YLoc[22], PSORT_II[20], MultiLoc[26], DeepLoc[27] and MULocDeep[28] and the nuclear localization-specific prediction tool NucPred[29]. We also trained another model using the 8,421 experimentally verified nuclear proteins and 18,278 non-nuclear proteins from NLSdb based on the same architecture as pSAM (CNN +BiLSTM)[42]. These classification systems involve three widely used prediction algorithms, sequence alignment, machine learning, and deep learning, and show satisfactory performance in their domains. For this comparison, we randomly selected 1000 protein sequences (500 proteins with specific localization and 500 proteins with other localization) from the testing dataset. With pSAM, we predicted the nuclear localization probability based on the model constructed with the training dataset. The 1000 protein sequences must not be used in the training dataset in this situation. Since the other tools are primarily web-based tools and limit the number of protein sequences submitted at a time, we uploaded the protein sequences individually and obtained the prediction results. The results for the 1000 sequences from each tool were downloaded and merged to retrieve the nuclear localization score. Finally, we calculated the AUC for each tool and presented these results as ROC curves.

Prediction of nuclear localization probabilities of reported peptides with NLS activity

Kosugi et al.[30] selected peptides bound by importin alpha and identified six classes of NLSs by screening random peptide libraries via mRNA display (Supplementary Data 2). The relative NLS activities of these altered sequences were examined in yeast and classified into ten classes according to the localization phenotype of the GFP reporter. We added ten pseudo amino acids ("-") at both termini since the altered sequences were less than 20 amino acids in length. Then, the sequences were encoded and predicted by the pSAM model to obtain their nuclear localization probabilities. Finally, we used a boxplot to show the consistency of the predicted and experimental results.

Shuffling analysis

We assessed the relative importance of the three protein regions—the terminal-terminus, the middle region, and the C-terminus—by shuffling each region independently in the human proteome. The truncated sequences and the wild-type proteins were encoded by one-hot encoding and transferred as the input of pSAM. Then, the protein nuclear localization probabilities were calculated for these sequences. A boxplot was used to illustrate the variation in nuclear localization probability from wild-type to random shuffling and terminal shuffling for each protein. For proteins with known NLSs, we shuffled the annotated NLS regions. We performed random shuffling of same-length regions and computed the localization probability for each sequence. A paired-sample boxplot was used to show the variation in nuclear localization probability from the wild type to random shuffling and NLS shuffling for each protein.

Analysis of residue-level contributions via deep learning

We then studied the learning process of pSAM by analyzing the trajectory of the nuclear localization probability as a function of sequence position for the best model. A detailed description of the procedure used to determine the nuclear localization probability trajectory is provided below.

First, the truncated protein sequence was defined from position 1 to position i as $Pro_{(i)}$. The nuclear localization probability of this sequence calculated using pSAM was defined as $P_{trunc}(i)$. Then, we proposed $\widetilde{P}_{trunc}(i)$ to represent the smoothed nuclear localization probability of position i using unweighted sliding-average smoothing with a window size of length $w = 16$,

$$\widetilde{P}_{trunc}(i) = \frac{\sum_{j=-w/2}^{w/2} P_{trunc}(i+j)}{\min\left(L, i+\frac{w}{2}\right) - \max\left(1, i-\frac{w}{2}\right) + 1} \quad (3)$$

where L is the entire length of the protein sequence. For each protein, we calculated the smoothed nuclear localization probabilities from position 1 to L and visualized them together as the nuclear localization probability trajectory, as shown in Fig. 3. $\widetilde{P}_{trunc}(i)$ represents the cumulative probability of nuclear localization of the protein.

We calculated the probability change $\Delta P_{trunc}(i)$ as a function of position for each sequence using the same window size $w = 16$ as mentioned above:

$$\Delta P_{trunc}(i) = \widetilde{P}_{trunc}\left(i+\frac{w}{2}\right) - \widetilde{P}_{trunc}\left(i-\frac{w}{2}\right) \quad (4)$$

to determine at which point in this sequence a decision is made by the pSAM.

Given the cumulative effect of probability scores, if a protein sequence has obtained a high $\widetilde{P}_{trunc}(i)$ at position i, the score $\widetilde{P}_{trunc}(j)$ at subsequent position j becomes nonsignificant relative to $\widetilde{P}_{trunc}(i)$. However, this position may also significantly affect protein nuclear localization. Therefore, when calculating $\Delta P_{trunc}(i)$, we want to exclude the aforementioned factors as much as possible. When constructing the distribution of $\Delta P_{trunc}(i)$ at each position of all proteins, we found that $|\Delta P_{trunc}(i)| > 0.1$ accounted for approximately 5% of the total number of loci. The interval containing $|\Delta P_{trunc}(i)| > 0.1$ sites significantly enhances the probability of protein nuclear localization, resulting in a nonsignificant $\Delta P_{trunc}(i)$ at subsequent sites. Hence, we designed a stepwise truncation method to recalculate $\Delta P_{trunc}(i)$.

We calculated $\widetilde{P}_{trunc}(i)$ and $\Delta P_{trunc}(i)$ for each position $i (1 \le i \le L)$ of each protein. Then, we counted all regions with $\Delta P_{trunc}(i) > 0.1$ and obtained the position j of the C-terminus of the last region. The truncated protein was generated by shuffling position 1 to position i. For the truncated protein, we repeated the operation described above for the full-length protein until there were no regions with $\Delta P_{trunc}(i) > 0.1$. Next, the calculated $\Delta P_{trunc}(i)$ replaces $\Delta P_{trunc}(i)$ at the corresponding position in the previous section. Hence, we obtained the same $\widetilde{P}_{trunc}(i)$ and updated $\Delta P_{trunc}(i)$ for each protein relative to the score obtained from the first calculation at the corresponding position.

With the calculated $\Delta P_{trunc}(i)$, we extracted all regions with $\Delta P_{trunc}(i) > 0.1$ for each protein and defined them as significant regions. These regions have a strong influence on nuclear localization, and their higher level of probability change reflects that the model assigns a higher weight to this part of the region, which is worth focusing on, and its sequence structure may be related to nuclear localization.

Subsequently, we computed the probability score changes for all positions on the protein and constructed a normal distribution curve. We calculated the proportion of anomalous sites under different cutoffs by tallying the number of abnormal sites and set the cutoff corresponding to a 5% proportion of anomalous sites for subsequent mutation analysis.

Classification of DNLs into NLS-like and NES-like types

Although we identified various significant regions, which we defined as sequence DNLs, their characteristics and action mechanisms remain unknown. There is no doubt that these regions are related to nuclear localization, either nuclear importing or exporting. We further classified the DNLs into NLS-like and NES-like according to the description provided by Bernhofer et al.[42]. Regions with at least three positively charged residues (H, K, or R) and an overall positive charge were defined as potential NLS-like DNLs. Regions with at least three hydrophobic residues (A, F, I, L, M, or V) and more than 30% hydrophobic amino acid residues were defined as potential NES-like DNLs. The regions that could not be classified into these two groups were defined as the unknown type (NA).

To increase their credibility, we further calculated the FDR of NLS-like and NES-like DNLs based on the sequence similarity with experimental validated NLS and NES, respectively. The *stringdist* package was used to obtain the cosine similarity of NLS-like DNLs with experimentally validated NLSs, as well as NES-like DNLs with experimentally validated NESs. The p values of NLS-like and NES-like DNLs were calculated to determine whether the given cosine similarities were significantly higher than those of randomly selected background sequences of the same size. The Benjamini-Hochberg method was further used to adjust the p values and obtain FDR. The FDRs of all NLS-like and NES-like DNLs are provided in supplementary Data 4.

We calculated the cosine similarity with the whole dataset for each region and generated a sequence similarity matrix. The sequence similarity visualization of our defined NLSs/NESs and validated NLSs/NESs is shown by the t-distributed stochastic neighbor embedding (t-SNE) plot.

In addition, considering that a known NLS does not have a strong motif, to refine the features of the NLS/NES, we performed unsupervised hierarchical clustering of the regions in the NLS-like/NES-like/NA groups based on the sequence similarity matrix and divided each group into six subgroups for further analysis.

Construction of the iNuLoC webserver

To facilitate basic research, we generated a webserver, iNuLoC, to calculate the protein nuclear localization probability and predict critical regions. iNuLoC is a platform developed to identify potentially critical regions that facilitate protein nuclear localization in five model organisms, namely, *Homo sapiens, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae*, and *Drosophila melanogaster*. It provides the prediction results from pSAM and integrates several well-known databases, including UniProt, NLSdb, SeqNLS, ValidNESs, and NESbase, to display known and candidate NLSs/NESs in proteins, which might provide helpful information for research on protein nuclear localization. The platform provides experimentally determined NLSs/NESs for 1530 proteins through database integration. Using the RCA method, the NLS/NES annotations were extended to 13,481 proteins, and the final dataset matched more than 93% of all known nuclear proteins. The database also contains 14,585/23,208/27,551 predicted nuclear proteins for the proteomes of five model organisms with high/

medium/low thresholds. Users can simply search for the name of their gene of interest and jump to the corresponding result interface. By clicking on the "Detail" button, users can search for detailed information about the relevant protein, for example, potential nuclear localization probabilities, predicted significant regions, and NLS and NES annotations from other databases. We visualized the residue-level contribution of nuclear localization and highlighted the critical regions.

Furthermore, information on the structure and physicochemical properties of the protein sequences, including disorder, exposure status, polarity, charge, secondary structure, surface accessibility, and hydropathy, is also shown. The disorder information of the query sequence was calculated using IUPred[91]. Surface accessibility and secondary structure information were predicted by NetSurfP[92]. We also integrated the pSAM model and constructed a prediction interface. Users can submit sequence information for proteins of interest and obtain their corresponding nuclear localization probabilities and other biological features. The framework for the deep learning model was Keras, with TensorFlow as its backend implementation. Finally, the webserver was constructed in HTML, PHP, and Python and can be accessed at http://inuloc.omicsbio.info. To provide a robust service, we tested the website of iNuLoC on various web browsers, such as the Internet Explorer, Google Chrome, and Mozilla Firefox, and found that it operates normally.

Domain and motif enrichment

The HMMER3 format domain information was downloaded from the Pfam database (V3.1b2). We extracted the sequence based on the UniProt database (UniProt release 2021_01), prepared the FASTA format file for each identified NLS, and validated the NLS. Domain enrichment analysis was performed using HMMER (v3.3.2)[93] with the parameters --noali -E 1.0. The motif enrichment analysis was performed using meme (v5.3.3)[94] with the parameters -minw 5 -maxw 10, 15.

Point and paired mutation analysis

The mutation data of 11 cancer types were downloaded from the Xena Browser, and only missense variants were retained for further analysis. We first merged all mutation files and removed duplicate mutations, and the retained mutations were considered possible human-derived mutations. For each mutation site, we calculated the nuclear localization probability score of its wild-type and mutant proteins and used the difference between the mutant protein score and the wild-type protein score as the effect of the mutation on nuclear localization potential.

$$\Delta P(i, mut) = P(i, mut) - P(i, wt) \tag{5}$$

For paired mutation analysis, we first counted all possible mutations in a protein, combined these mutated sites two by two, and constructed the mutated sequences. These sequences were predicted with pSAM for potential nuclear localization probabilities. Based on the predicted single point mutation scores and wild-type scores produced as described above, we obtained the change after mutation combination relative to the single point mutation.

$$\Delta P_{comp}(i,j) = \Delta P(i, mut, j, mut) - \Delta P(i, mut) - \Delta P(j, mut) \tag{6}$$

If $\Delta P(i, mut) < 0$, $\Delta P(j, mut) < 0$ and $\Delta P_{comp}(i,j) < 0$, the two mutation sites synergistically affect protein nuclear localization, and a smaller $\Delta P_{comp}(i,j)$ represents a stronger synergistic effect.

Truncated mutation analysis

We assessed the relative importance of the identified NLSs critical for protein nuclear localization by shuffling these regions in the human proteome. The truncated sequences and the wild-type proteins were encoded by one-hot encoding and transferred as the input of pSAM. Then, the protein nuclear localization probabilities were calculated for these sequences. The difference between the truncated protein score and the wild-type protein score was defined as the effect of truncation on the nuclear localization potential.

Protein iteration network

The experimentally confirmed PPI dataset was downloaded and curated from BioGRID[95], IID[96], I2D[97], BioPlex[98] and IntAct[99]. In total, 377,796 interactions among 20,379 proteins were collected. The visualization of the network was generated with Cytoscape[100].

Phosphomimetic and inactivation mutation analysis

We assessed the relative importance of the phosphorylation sites located in the identified NLSs critical for protein nuclear localization by performing phosphomimetic and inactivation mutation analyses. For each protein, we mutated the Ser and Thr amino acids to D in the phosphomimetic mutation analysis and mutated the amino acid to Ala in the inactivation mutation analysis. The nuclear localization probabilities of the three types of sequences were then predicted using pSAM. We first calculated the delta probability between the phosphomimetic mutant and wild-type sequence as Delta1 and the delta probability between the inactivation mutant and wild-type sequence as Delta2. We classified the phosphorylation sites critical for nuclear localization with thresholds of Delta1 > 0.25 and Delta2 < -0.25. The selected proteins harboring these sites were then used for kinase enrichment analysis to identify the critical kinase that regulates the nuclear localization of its substrates by phosphorylation.

Cell lines and plasmid transfection

The human HCT116 and DLD1 colorectal cancer (CRC) cell lines and human embryonic kidney (HEK) 293 T cells were obtained from the American Type Culture Collection (ATCC, Rockville, MD, USA) and cultured according to standard guidelines. All cells were authenticated by short tandem repeat fingerprinting before use at the Medicine Laboratory of the Forensic Medicine Department of Sun Yat-sen University (Guangzhou, China). All plasmids with a Flag tag were obtained from Saisofi Biotechnology Co., Ltd. (Jiangsu, China) and transfected into cells using ViaFect Transfection Reagent (E4982, Promega, Madison, WI, USA) according to the recommended protocol. CRC cells were transfected with lentiviruses and selected with puromycin (HY-B1743A, MedChemExpress, NJ, USA) for 7 days. The sequences of shRNA of PTEN were as following: shPTEN#1: ACTTGAAGGCGTATACAGGA, shPTEN#2: CGACTTAGACTTGACCTATAT.

Western blot and Co-Immunoprecipitation (IP)

Proteins were extracted using RIPA buffer (P0013B, Beyotime, Jiangsu, China), separated on SDS–PAGE gels and then transferred to PVDF membranes (IPVH00010, Bio-Rad Laboratories, Hercules, CA, USA). The membranes were incubated with anti-Flag tag (14793, Cell Signaling Technology, Danvers, MA, USA), anti-β-Actin (A5441, Sigma-Aldrich), anti-phosphorylated ERK (ab76299, Abcam, Cambridge, MA, USA), anti-ERK2 (ab32081, Abcam), and anti-Importin Alpha 5 (KPNA1) (18137-1-AP, Proteintech), anti-KPNA2 (ab289858, Abcam), anti-KPNA4 (ab302556, Abcam), anti-PTEN (ab170941, Abcam) antibodies (4 °C, overnight). followed by peroxidase-conjugated secondary antibodies (ZB-2301 and ZB-2305, ZSGB-BIO, Beijing, China) (RT, 1 h). The bands were visualized using chemiluminescence (34096, Thermo Fisher Scientific, Carlsbad, CA, USA). For Co-IP, the cell lysate was incubated with Anti-Flag and protein A/G beads (MedChemExpress, NJ, USA) overnight at 4 °C. The protein-magnetic beads complexes were washed and eluted by 1× loading buffer, and the bound proteins were used in Western blot.

Immunofluorescence assay

Cells were plated on glass-bottom culture dishes (Cellvis, Mountain, CA, USA), fixed with 4% paraformaldehyde for 15 min at 37 °C, and permeabilized with 0.2% Triton X-100 in PBS for 10 min at room temperature. The samples were blocked with 5% FBS for 1 h at room temperature and incubated with an anti-FLAG antibody and Alexa Fluor® 488 conjugated antibody overnight at 4 °C and for 1 h at room temperature. 4',6-Diamidino-2-phenylindole (DAPI) (Invitrogen, Carlsbad, CA, USA)-containing medium was then used to counterstain the nuclei, and images were observed with a confocal fluorescence microscope (LSM880 with Fast Airyscan, Zeiss, Germany); scale bar = 10 µm.

Library construction, quality control and sequencing

Total RNA was used as input material for the RNA sample preparations. Sequencing libraries were generated using the NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, USA, Catalog #: E7530L) following the manufacturer's recommendations, and index codes were added to attribute sequences to each sample.

Briefly, mRNA was purified from total RNA using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5X). First-strand cDNA was synthesized using random hexamer primers and M-MuLV reverse transcriptase (RNase H). Second-strand cDNA synthesis was subsequently performed using DNA polymerase I and RNase H. The remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of the 3' ends of the DNA fragments, NEBNext adaptors with hairpin loop structures were ligated to prepare for hybridization. To preferentially select cDNA fragments 370~420 bp in length, the library fragments were purified with the AMPure XP system (Beverly, MA, USA). Then, 3 µL of USER Enzyme (NEB, USA) was used with size-selected, adaptor-ligated cDNA at 37 °C for 15 min, followed by 5 min at 95 °C before PCR. Then, PCR was performed with Phusion High-Fidelity DNA polymerase, universal PCR primers and Index (X) Primer. Finally, the PCR products were purified (AMPure XP system), and the library quality was assessed on an Agilent 5400 system (Agilent, USA) and quantified by QPCR (1.5 nM).

The qualified libraries were pooled and sequenced on Illumina platforms with the PE150 strategy at Novogene Bioinformatics Technology Co., Ltd. (Beijing, China), according to the effective library concentration and data amount needed.

Cell proliferation and colony formation assays

HCT116 cells (1×10^3 cells/well) were seeded in a 96-well plate, and cell viability was determined using MTS (Qiagen, Hilden, Germany) according to the manufacturer's instructions. A colony formation assay was performed in 12-well plates with 250 cells, which were cultured for 1 week. After being fixed, the colonies were counted and stained with 0.05% crystal violet.

In vivo tumorigenesis

All experiments were performed using a protocol approved by our institutional Animal Care and Use Committee (SYSUCC). Female BALB/c nude mice (5 weeks old) were obtained from Beijing Vital River Laboratory Animal Technology Co., Ltd. Next, 2×10^6 CRC cells overexpressing wild-type or site-specific PTEN or CHFR mutants were injected subcutaneously into the flanks of each mouse. The diameter and width of the tumors from the mice were measured every 4 days, and the tumor volumes were calculated using the formula $V = 0.5 \times D \times W^2$ (V, volume; D, diameter; W, width). All the mice were sacrificed at the appropriate time points, and the tumors were removed and weighed for further analysis.

Statistics and reproducibility

When comparing data from two groups, a Student's *t* test or Wilcoxon test was used (in R program version 4.0.3). Chi-square test was used in the analyses of contingency tables. Log-rank tests were adopted to test differences in survival. Hypergeometric test was used to calculate p values for enrichment analyses. A two-sided p value less than 0.05 was considered to indicate statistical significance. Benjamini-Hochberg method was used to adjusted the p values for the calculation of FDRs if necessary. All experiments were performed at least three times as independent experiments with similar results, and representative images are shown.

Reporting summary

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

Data availability

The RNA-seq data from the PTEN WT and R14M mutant DLD1 cell lines have been deposited in the BioProject under accession number PRJNA930129. The data used to establish the pSAM model are available on the Download pages of pSAM [http://inuloc.omicsbio.info/] and GitHub [https://github.com/lzxlab/pSAM]. In addition, we have also provided other supporting data, including experimentally validated NLSs and NESs, the shuttling-attacking mutations based on TCGA data and the predictions of the nuclear localization probability and site-specific contributions of proteins from five organisms (human, mouse, rat, yeast and fruit fly), on the Download pages of pSAM [http://inuloc.omicsbio.info/download.php]. The investigated reference proteomes are available from UniProt [https://www.uniprot.org/proteomes]. All processed mutation data and gene expression profiles from patients in TCGA were downloaded from the Xena Browser [https://xenabrowser.net/datapages/]. Source data are provided with this paper.

Code availability

The Python version of the pSAM deep-learning model is publicly available at GitHub [https://github.com/lzxlab/pSAM] (https://doi.org/10.5281/zenodo.14670783)[101], under the MIT license. A web version of pSAM is publicly available at the iNuLoC website [http://inuloc.omicsbio.info/]. The iNuLoC webserver is free for all users. The iNuLoC webserver provides the following functions: (1) the iNuLoC webserver provides predictions of the nuclear localization probability and site-specific contributions of proteins from five organisms (human, mouse, rat, yeast and fruit fly); (2) the pNuLoC section helps to the predict nuclear localization of any given peptides; and (3) the Browse section summarizes all predictions of the nuclear localization probability and site-specific contributions of proteins from five organisms. The source code of this paper is publicly available at GitHub [https://github.com/lzxlab/pSAM_paper_code].

References

1. Lu, M., Muers, M. R. & Lu, X. Introducing STRaNDs: shuttling transcriptional regulators that are non-DNA binding. *Nat. Rev. Mol. cell Biol.* **17**, 523–532 (2016).
2. Butler, G. S. & Overall, C. M. Proteomic identification of multitasking proteins in unexpected locations complicates drug targeting. *Nat. Rev. Drug Discov.* **8**, 935–948 (2009).
3. Orre, L. M. et al. SubCellBarCode: proteome-wide mapping of protein localization and relocalization. *Mol. cell* **73**, 166–182 (2019).
4. Oka, M. & Yoneda, Y. Importin α: functions as a nuclear transport factor and beyond. *Proc. Jpn. Acad. Ser. B, Phys. Biol. Sci.* **94**, 259–274 (2018).
5. Cautain, B., Hill, R., de Pedro, N. & Link, W. Components and regulation of nuclear transport processes. *FEBS J.* **282**, 445–462 (2015).
6. Miyamoto, Y., Yamada, K. & Yoneda, Y. Importin α: a key molecule in nuclear transport and non-transport functions. *J. Biochem.* **160**, 69–75 (2016).
7. Lu, J. et al. Types of nuclear localization signals and mechanisms of protein import into the nucleus. *Cell Commun. Signal.: CCS* **19**, 60 (2021).

8. Sharma, M., Jamieson, C., Lui, C. & Henderson, B. R. The hydrophobic rich N- and C-terminal tails of β-catenin facilitate nuclear import. *J. Biol. Chem.* **290**, 18479 (2015).
9. Lyst, M. J. et al. Affinity for DNA Contributes to NLS Independent Nuclear Localization of MeCP2. *Cell Rep.* **24**, 2213–2220 (2018).
10. Gringhuis, S. I., Kaptein, T. M., Wevers, B. A., Mesman, A. W. & Geijtenbeek, T. B. Fucose-specific DC-SIGN signalling directs T helper cell type-2 responses via IKKε-and CYLD-dependent Bcl3 activation. *Nat. Commun.* **5**, 1–13 (2014).
11. Jin, X. et al. Pyruvate kinase M2 promotes the activation of dendritic cells by enhancing IL-12p35 expression. *Cell Rep.* **31**, 107690 (2020).
12. Kofuji, S. et al. IMP dehydrogenase-2 drives aberrant nucleolar activity and promotes tumorigenesis in glioblastoma. *Nat. cell Biol.* **21**, 1003–1014 (2019).
13. Pennacchio, F. A., Nastały, P., Poli, A. & Maiuri, P. Tailoring cellular function: the contribution of the nucleus in mechanotransduction. *Front. Bioengineer. Biotech.* **8**, 596746 (2020).
14. Yang, W. et al. ERK1/2-dependent phosphorylation and nuclear translocation of PKM2 promotes the Warburg effect. *Nat. cell Biol.* **14**, 1295–1304 (2012).
15. Esrig, D. et al. p53 nuclear protein accumulation correlates with mutations in the p53 gene, tumor grade, and stage in bladder cancer. *Am. J. Pathol.* **143**, 1389 (1993).
16. Pauty, J. et al. Cancer-causing mutations in the tumor suppressor PALB2 reveal a novel cancer mechanism using a hidden nuclear export signal in the WD40 repeat motif. *Nucleic acids Res.* **45**, 2644–2657 (2017).
17. Chang, C.-J. et al. PTEN nuclear localization is regulated by oxidative stress and mediates p53-dependent tumor suppression. *Mol. Cell. Biol.* **28**, 3281–3289 (2008).
18. Yang, J.-M. et al. Characterization of PTEN mutations in brain cancer reveals that pten mono-ubiquitination promotes protein stability and nuclear localization. *Oncogene* **36**, 3673–3685 (2017).
19. Hua, S. & Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinforma. (Oxf., Engl.)* **17**, 721–728 (2001).
20. Horton, P. et al. WoLF PSORT: protein localization predictor. *Nucleic acids Res.* **35**, W585–W587 (2007).
21. Briesemeister, S. et al. SherLoc2: a high-accuracy hybrid method for predicting subcellular localization of proteins. *J. proteome Res.* **8**, 5363–5366 (2009).
22. Briesemeister, S., Rahnenführer, J. & Kohlbacher, O. YLoc-an interpretable web server for predicting subcellular localization. *Nucleic acids Res.* **38**, W497–W502 (2010).
23. Chou, K. C., Wu, Z. C. & Xiao, X. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. *PloS one* **6**, e18258 (2011).
24. Yu, C. S. et al. CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. *PloS one* **9**, e99368 (2014).
25. Goldberg, T. et al. LocTree3 prediction of localization. *Nucleic acids Res.* **42**, W350–W355 (2014).
26. Höglund, A., Dönnes, P., Blum, T., Adolph, H.-W. & Kohlbacher, O. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinforma. (Oxf., Engl.)* **22**, 1158–1165 (2006).
27. Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H. & Winther, O. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinforma. (Oxf., Engl.)* **33**, 3387–3395 (2017).
28. Jiang, Y. et al. MULocDeep: A deep-learning framework for protein subcellular and suborganellar localization prediction with residue-level interpretation. *Computational Struct. Biotechnol. J.* **19**, 4825–4839 (2021).
29. Brameier, M., Krings, A. & MacCallum, R. M. NucPred-predicting nuclear localization of proteins. *Bioinforma. (Oxf., Engl.)* **23**, 1159–1160 (2007).
30. Kosugi, S., Hasebe, M., Tomita, M. & Yanagawa, H. Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc. Natl Acad. Sci. USA* **106**, 10171–10176 (2009).
31. Lin, J. R. & Hu, J. SeqNLS: nuclear localization signal prediction based on frequent pattern mining and linear motif scoring. *PloS one* **8**, e76864 (2013).
32. Nguyen Ba, A. N., Pogoutse, A., Provart, N. & Moses, A. M. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinforma.* **10**, 202 (2009).
33. Teufel, F. et al. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. biotech.* **40**, 1023–1025 (2022).
34. Angermueller, C., Pärnamaa, T., Parts, L. & Stegle, O. Deep learning for computational biology. *Mol. Syst. Biol.* **12**, 878 (2016).
35. Kim, G. B., Kim, W. J., Kim, H. U. & Lee, S. Y. Machine learning applications in systems metabolic engineering. *Curr. Opin. Biotechnol.* **64**, 1–9 (2020).
36. Zou, J. et al. A primer on deep learning in genomics. *Nat. Genet.* **51**, 12–18 (2019).
37. Thusberg, J. & Vihinen, M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum. Mutat.* **30**, 703–714 (2009).
38. Yao, C. et al. BACH2 enforces the transcriptional and epigenetic programs of stem-like CD8(+) T cells. *Nat. Immunol.* **22**, 370–380 (2021).
39. Cadenas, C. et al. Role of thioredoxin reductase 1 and thioredoxin interacting protein in prognosis of breast cancer. *Breast cancer Res.* **12**, 1–15 (2010).
40. Izumi, H. et al. The CLIP1–LTK fusion is an oncogenic driver in nonsmallcell lung cancer. *Nature* **600**, 319–323 (2021).
41. Consortium, U. UniProt: the universal protein knowledgebase in 2021. *Nucleic acids Res.* **49**, D480–D489 (2021).
42. Bernhofer, M. et al. NLSdb—major update for database of nuclear localization signals and nuclear export signals. *Nucleic acids Res.* **46**, D503–D508 (2018).
43. Moller, A. & Schmitz, M. L. Viruses as hijackers of PML nuclear bodies. *Arch. Immunol. Ther. Exp. (Warsz.)* **51**, 295–300 (2003).
44. Chen, D., Feng, C., Tian, X., Zheng, N. & Wu, Z. Promyelocytic leukemia restricts enterovirus 71 replication by inhibiting autophagy. *Front. Immunol.* **9**, 1268 (2018).
45. Duprez, E. et al. SUMO-1 modification of the acute promyelocytic leukaemia protein PML: implications for nuclear localisation. *J. cell Sci.* **112**, 381–393 (1999).
46. Miki, T., Matsumoto, T., Zhao, Z. & Lee, C. C. p53 regulates Period2 expression and the circadian clock. *Nat. Commun.* **4**, 1–11 (2013).
47. Guo, A. et al. The function of PML in p53-dependent apoptosis. *Nat. Cell Biol.* **2**, 730–736 (2000).
48. Liang, S.-H. & Clarke, M. F. The nuclear import of p53 is determined by the presence of a basic domain and its relative position to the nuclear localization signal. *Oncogene* **18**, 2163–2166 (1999).
49. Stommel, J. M. et al. A leucine-rich nuclear export signal in the p53 tetramerization domain: regulation of subcellular localization and p53 activity by NES masking. *EMBO J.* **18**, 1660–1672 (1999).
50. Nie, L., Sasaki, M. & Maki, C. G. Regulation of p53 nuclear export through sequential changes in conformation and ubiquitination. *J. Biol. Chem.* **282**, 14616–14625 (2007).
51. Black, B. E., Holaska, J. M., Rastinejad, F. & Paschal, B. M. DNA binding domains in diverse nuclear receptors function as nuclear export signals. *Curr. Biol.* **11**, 1749–1758 (2001).

52. LaCasse, E. C. & Lefebvre, Y. A. Nuclear localization signals overlap DNA-or RNA-binding domains in nucleic acid-binding proteins. *Nucleic acids Res.* **23**, 1647 (1995).
53. Van Impe, K. et al. A new role for nuclear transport factor 2 and Ran: nuclear import of CapG. *Traffic* **9**, 695–707 (2008).
54. Han, F. et al. A-to-I RNA editing of BLCAP promotes cell proliferation by losing the inhibitory of Rb1 in colorectal cancer. *Exp. Cell Res* **417**, 113209 (2022).
55. Pentinmikko, N. et al. Notum produced by Paneth cells attenuates regeneration of aged intestinal epithelium. *Nature* **571**, 398–402 (2019).
56. Chan, T.-Y. et al. TNK1 is a ubiquitin-binding and 14-3-3-regulated kinase that can be targeted to block tumor growth. *Nat. Commun.* **12**, 1–17 (2021).
57. May, W. S. Jr. et al. Tnk1/Kos1: a novel tumor suppressor. *Trans. Am. Clin. Climatological Assoc.* **121**, 281 (2010).
58. Zhang, S. et al. RNAi screening identifies KAT8 as a key molecule important for cancer cell survival. *Int. J. Clin.* **6**, 870 (2013).
59. Kim, J.-Y., Yu, J., Abdulkadir, S. A. & Chakravarti, D. KAT8 regulates androgen signaling in prostate cancer cells. *Mol. Endocrinol.* **30**, 925–936 (2016).
60. Radzisheuskaya, A. et al. Complex-dependent histone acetyltransferase activity of KAT8 determines its role in transcription and cellular homeostasis. *Mol. cell* **81**, 1749–1765.e1748 (2021).
61. Xiang, Y. et al. RNA m 6 A methylation regulates the ultraviolet-induced DNA damage response. *Nature* **543**, 573–576 (2017).
62. Lee, H. et al. Stage-specific requirement for Mettl3-dependent m 6 A mRNA methylation during haematopoietic stem cell differentiation. *Nat. cell Biol.* **21**, 700–709 (2019).
63. Zhong, X. et al. Circadian clock regulation of hepatic lipid metabolism by modulation of m6A mRNA methylation. *Cell Rep.* **25**, 1816–1828.e1814 (2018).
64. Schöller, E. et al. Interactions, localization, and phosphorylation of the m6A generating METTL3–METTL14–WTAP complex. *Rna* **24**, 499–512 (2018).
65. Madhunapantula, S. V. & Robertson, G. P. The PTEN–AKT3 signaling cascade as a therapeutic target in melanoma. *Pigment cell melanoma Res.* **22**, 400–419 (2009).
66. Vaishnave, S. BMI1 and PTEN are key determinants of breast cancer therapy: A plausible therapeutic target in breast cancer. *Gene* **678**, 302–311 (2018).
67. Planchon, S. M., Waite, K. A. & Eng, C. The nuclear affairs of PTEN. *J. cell Sci.* **121**, 249–253 (2008).
68. Gil, A. et al. Nuclear localization of PTEN by a Ran-dependent mechanism enhances apoptosis: Involvement of an N-terminal nuclear localization domain and multiple nuclear exclusion motifs. *Mol. Biol. cell* **17**, 4002–4013 (2006).
69. Vandeput, F., Backers, K., Villeret, V., Pesesse, X. & Erneux, C. The influence of anionic lipids on SHIP2 phosphatidylinositol 3, 4, 5-trisphosphate 5-phosphatase activity. *Cell. Signal.* **18**, 2193–2199 (2006).
70. Costa, H. A. et al. Discovery and functional characterization of a neomorphic PTEN mutation. *Proc. Natl Acad. Sci.* **112**, 13976–13981 (2015).
71. Trotman, L. C. et al. Ubiquitination regulates PTEN nuclear import and tumor suppression. *Cell* **128**, 141–156 (2007).
72. Mayo, L. D., Dixon, J. E., Durden, D. L., Tonks, N. K. & Donner, D. B. PTEN protects p53 from Mdm2 and sensitizes cancer cells to chemotherapy. *J. Biol. Chem.* **277**, 5484–5489 (2002).
73. Niu, B. et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* **48**, 827–837 (2016).
74. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids Res.* **39**, e118–e118 (2011).
75. Baker, S. et al. Automatic semantic classification of scientific literature according to the hallmarks of cancer. *Bioinformatics* **32**, 432–440 (2016).
76. Chen, L. et al. Pan-cancer analysis reveals the functional importance of protein lysine modification in cancer development. *Front. Genet.* **9**, 254 (2018).
77. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids Res.* **31**, 3812–3814 (2003).
78. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **76**, 7.20. 21–27.20. 41 (2013).
79. Choi, Y. & Chan, A. P. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinforma. (Oxf., Engl.)* **31**, 2745–2747 (2015).
80. Shihab, H. A. et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34**, 57–65 (2013).
81. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
82. Hu, H. et al. AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic acids Res.* **47**, D33–D38 (2019).
83. Fu, S.-C., Huang, H.-C., Horton, P. & Juan, H.-F. ValidNESs: a database of validated leucine-rich nuclear export signals. *Nucleic acids Res.* **41**, D338–D343 (2013).
84. La Cour, T. et al. NESbase version 1.0: a database of nuclear export signals. *Nucleic acids Res.* **31**, 393–396 (2003).
85. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic acids Res.* **47**, D427–D432 (2019).
86. Kumar, S. et al. Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. *Cell* **180**, 915–927.e916 (2020).
87. Lochovsky, L., Zhang, J. & Gerstein, M. MOAT: efficient detection of highly mutated regions with the Mutations Overburdening Annotations Tool. *Bioinforma. (Oxf., Engl.)* **34**, 1031–1033 (2018).
88. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl Acad. Sci.* **113**, 14330–14335 (2016).
89. Ghandi, M. et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).
90. Yu, K. et al. Deep learning based prediction of reversible HAT/HDAC-specific lysine acetylation. *Brief. Bioinforma.* **21**, 1798–1805 (2020).
91. Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434 (2005).
92. Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M. & Lundegaard, C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC Struct. Biol.* **9**, 51 (2009).
93. Potter, S. C. et al. HMMER web server: 2018 update. *Nucleic acids Res.* **46**, W200–W204 (2018).
94. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. *Nucleic acids Res.* **43**, W39–W49 (2015).
95. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Res* **43**, D470–D478 (2015).
96. Kotlyar, M., Pastrello, C., Sheahan, N. & Jurisica, I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res* **44**, D536–D541 (2016).
97. Brown, K. R. & Jurisica, I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.* **8**, R95 (2007).

98. Huttlin, E. L. et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).
99. Orchard, S. et al. The MIntAct project-IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358–D363 (2014).
100. Dokmanovic, M., Clarke, C. & Marks, P. A. Histone deacetylase inhibitors: overview and perspectives. *Mol. cancer Res.: MCR* **5**, 981–989 (2007).
101. Liu Z. et al. Deep learning prioritizes cancer mutations that alter protein nucleocytoplasmic shuttling to drive tumorigenesis. Zenodo: lzxlab/pSAM: V1.0 (V1.0). Zenodo. https://doi.org/10.5281/zenodo.14670783 (2025).

Acknowledgements

This study was supported by the Noncommunicable Chronic Diseases-National Science and Technology Major Project [2023ZD0501600], National Key R&D Program of China [2021YFA1302100], National Natural Science Foundation of China [32370698, 81972239, 82321003, 82173128], Program for Guangdong Introducing Innovative and Entrepreneurial Teams [2017ZT07S096], Tip-Top Scientific and Technical Innovative Youth Talents of Guangdong Special Support Program [2019TQ05Y351], Young Talents Program of Sun Yat-sen University Cancer Center [YTP-SYSUCC-0029], Science and Technology Program of Guangzhou [202206080011], Guangdong Basic and Applied Basic Research Foundation [2023B1515040030], CAMS Innovation Fund for Medical Sciences (CIFMS) [2019-I2M-5-036], Cancer Innovative Research Program of Sun Yat-sen University Cancer Center [CIRP-SYSUCC-0004], the Chih Kuang Scholarship for Outstanding Young Physician-Scientists of Sun Yat-sen University Cancer Center [CKS-SYSUCC-2024009] and the Postdoctoral Science Foundation of China [2024M763801, GZB20240907]. We thank Dr. Yu Xue from Huazhong University of Science and Technology for his helpful suggestions to improve our study. We thank Dr. Sushant Kumar from Princess Margaret Cancer Centre for sharing the methods to simulate signature-corrected randomized mutations. We thank Dr. Michael Bernhofer from Technische Universität München for sharing the data of NLSdb. We also thank all authors who released their data along with their article or responded to our data request.

Author contributions

ZX.L. and RH.X. designed and supervised the experiments. YQ.Z., K.Y., and JF.L. performed the experiments and data analysis; QN.W., CY.H., and HQ.J. assisted the experiments; K.Y., YQ.Z., ZR.L., QF.Z., and Q.Z. developed the predictor. YQ.Z., JT.L., ZR.L., ZX.Z., and M.L. contributed to the data analysis. YQ.Z., K.Y., and ZX.L. wrote the manuscript with contributions from all the authors. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-57858-8.

Correspondence and requests for materials should be addressed to Rui-Hua Xu or Ze-Xian Liu.

Peer review information *Nature Communications* thanks Fumihiko Nakamura and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

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

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