Preprints.org

Article

Single-Cell and Bulk RNA Sequencing Analysis Reveals CENPA as a Potential Biomarker and Therapeutic Target in Cancers

Hengrui Liu , Miray Karsidag , Kunwer Chhatwal , Panpan Wang [*] , Tao Tang [*]

Article

Single-Cell and Bulk RNA Sequencing Analysis Reveals CENPA as a Potential Biomarker and Therapeutic Target in Cancers

Tao Tang [1],*, Hengrui Liu [2,3], Panpan Wang [4],*, Miray Karsidag [5] and Kunwer Chhatwal [6]

[1]   Sun Yat-Sen University Cancer Center, Guangzhou, Guangdong, China
[2]   Cancer Research Institute, Jinan University, Guangzhou, Guangdong, China
[3]   Yinuo Biomedical Co., Ltd, Tianjin, China
[4]   The First Affiliated Hospital of Jinan University, Guangzhou, Guangdong, China
[5]   Canyon Crest Academy, San Diego, CA, USA
[6]   Hopkinton High School, Hopkinton, MA, USA
*   Corresponding: tangtao@sysucc.org.cn; wangpp@jnu.edu.cn

Abstract: Background: Cancer remains one of the most significant public health challenges worldwide. A widely recognized hallmark of cancer is the ability to sustain proliferative signaling, which is closely tied to various cell cycle processes. Centromere Protein A (CENPA), a variant of the standard histone H3, is crucial for selective chromosome segregation during the cell cycle. Despite its importance, a comprehensive pan-cancer bioinformatic analysis of CENPA has not yet been conducted. **Methods:** Data on genomes, transcriptomes, and clinical information were retrieved from publicly accessible databases. We analyzed CENPA's genetic alterations, mRNA expression, functional enrichment, association with stemness, mutations, expression across cell populations and cellular locations, link to the cell cycle, impact on survival, and its relationship with the immune microenvironment. Additionally, a prognostic model for glioma patients was developed to demonstrate CENPA's potential as a biomarker. Furthermore, drugs targeting CENPA in cancer cells were identified and predicted using drug sensitivity correlations and protein-ligand docking. **Results:** CENPA exhibited low levels of gene mutation across various cancers. It was found to be overexpressed in nearly all cancer types analyzed in TCGA, relative to normal controls, and was predominantly located in the nucleus of malignant cells. CENPA showed a strong association with the cancer cell cycle, particularly as a biomarker for the G2 phase. It also emerged as a valuable diagnostic and prognostic biomarker across multiple cancer types. In glioma, CENPA demonstrated reliable prognostic potential when used alongside other prognostic factors. Additionally, CENPA was linked to the immune microenvironment. Drugs such as CD-437, 3-Cl-AHPC, Trametinib, BI-2536, and GSK461364 were predicted to target CENPA in cancer cells. **Conclusion:** CENPA serves as a crucial biomarker for the cell cycle in cancers, offering both diagnostic and prognostic value.

Keywords CENPA; cell cycle; pan-cancer; diagnostic; prognostic; glioma; immune; protein-ligand docking

Introduction

Cancer is one of the most concerning public health issues in the world[1,2]. Many common molecular pathological mechanisms shared across different neoplastic diseases have been identified to facilitate clinical cancer diagnosis, prognosis, and therapies. Cancer databases, such as The Cancer Genome Atlas (TCGA) [3], Genotype-Tissue Expression (GTEx)[4], the Chinese Glioma Genome Atlas (CGGA)[5], and the International Cancer Genome Consortium (ICGC)[6], provide gene alteration, gene expression, and clinical information on different cancer types, facilitating pan-cancer studies for identification and understanding of targets or biomarkers that exert common effects across cancer types. Although there is biases and limitations[7], these databases have been wildly used in many previous studies[8–29].

Six hallmarks of cancer have been proposed to constitute an organizing principle that provides a logical framework for understanding the remarkable diversity of cancers [30]. One of the cancer

hallmarks wildly accepted is sustaining proliferative signaling [30], which involves most of the cell cycle biological activities [31]. Centromeric histone, Centromere Protein A (CENPA), a variant of canonical histone H3, plays an essential role in selective chromosome segregation in the cell cycle. Loading of CENPA protein at centromeres is closely associated with the cell cycle phases. When the cell proliferates, parental CENPA protein is deposited at centromeres in the S phase, whereas newly synthesized CENPA protein is deposited during the G2/M phase of the cell cycle [32–34]. A study reported that cell cycle-dependent deposition of CENPA was mediated by the Dos1/2–Cdc20 complex [35]. Although the cell cycle mechanisms involved in CENPA in cancer remain poorly studied, the function of CENPA in the cell cycle might be universal across all proliferating cells, regardless of their malignancy and tissue types, which inferred a potential common molecular pathological mechanism of CENPA shared across different cancer types.

Previous studies have reported the involvement of CENPA in a few cancer types. The overexpression of CENPA in prostate cancer has been demonstrated by a study with both in vivo and in vitro evidence [36]. In ovarian cancer, CENPA was found associated with the proliferation of cancer cells and survival of patients, which might be directly regulated by the MYBL2 [37]. In colonial cancer, CENPA was reported to recruit histone acetyltransferase general control of amino acid synthesis (GCN)-5 to the promoter region of the karyopherin α2 subunit gene (KPNA2), thereby boosting KPNα2 activation, which facilitated proliferation and glycolysis in cancer cells [38]. In clear cell renal cell carcinoma, the function of CENPA was reported to promote metastasis of cancer via the Wnt/β-catenin signaling pathway [39]. In addition, studies also suggested the prognostic value of CENPA for a few cancer types, such as ovarian cancer [37], liver cancer [40], breast cancer [41,42], and lung cancer [43]. However, so far, a systematic pan-cancer bioinformatic analysis has not been done yet. Therefore, this study aimed to systematically investigate CENPA in multiple cancer types, regarding the potential of CENPA as a pan-cancer biomarker. Furthermore, we developed strategies for the application of CENPA in glioma prognosis as an example of the future development of CENPA as a clinical cancer biomarker.

Methods

1. The acquisition of mRNA sequencing data

The mRNA data, along with clinical information, were obtained from The TCGA [3], GTEx[4], CGGA[5], and the ICGC[6]. All data acquisition and usage adhered to the guidelines and policies of the respective databases. For TCGA, mRNA sequencing data across 33 cancer types were obtained via the TCGA portal. The CGGA data, which comprises three glioma patient cohorts, were also accessed through its portal. Corresponding normal tissue mRNA sequencing data for TCGA cancer types were downloaded from the GTEx portal.

2. Gene alteration analysis

Mutation analyses were performed using cBioPortal [44]with data from the "Pan-Cancer Analysis of Whole Genomes (ICGC/TCGA, Nature 2020)" [45]. Mutation and variant data were sourced from the TCGA PanCancer Atlas Studies and UniProt. Single-nucleotide variant (SNV) and copy number variant (CNV) data were retrieved from the NCI Genomic Data Commons (GDC) for TCGA datasets. SNV visualization was performed using the maftools package [46] which facilitated mutation frequency and variant type analysis. while CNV data were processed using GISTIC2.0 [47] to identify significant regions of amplification and deletion.

3. RNA-seq data analysis and plotting

All statistical analyses and visualizations were conducted using R version 4.0.3 (R Foundation for Statistical Computing, 2020). Nomogram construction, used to predict patient survival probabilities, was implemented with the rms package, which enabled the visualization of individualized risk scores. Kaplan-Meier (KM) survival analysis was performed to assess survival differences across groups, utilizing the survival package to generate survival curves and estimate

hazard ratios with confidence intervals. Receiver Operating Characteristic (ROC) curves were constructed with the pROC package to evaluate the predictive accuracy of the biomarker, with area under the curve (AUC) values used as a measure of model performance. All plots, including survival curves and nomograms, were generated with ggplot2 (v3.3.2) for clear, publication-quality visualizations.

4. Associated genes enrichment analysis.

The top correlated genes were identified using GEPIA [48], a tool that facilitates gene expression profiling and correlation analysis based on TCGA and GTEx data. A protein-protein interaction (PPI) network was then constructed using STRING [49], with a high-confidence interaction threshold (interaction score >0.9) to ensure robust connections between genes. Enrichment analyses, including Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis, were conducted using the clusterProfiler[50] package in R, which enabled the identification of significantly enriched biological processes, molecular functions, cellular components, and pathways associated with the gene set of interest.

5. Immunohistochemistry staining

Immunohistochemistry (IHC) staining was conducted using antibody CAB008371 on microarray slides to assess protein expression across cancerous and non-cancerous tissues. Representative images were sourced from the Human Protein Atlas (HPA) [51], which provided high-quality, standardized IHC-stained samples from various tissue types on microarray slides. This setup allowed for a precise comparative analysis of protein expression, making it possible to observe differential expression patterns between cancerous and corresponding normal tissues, thereby facilitating insights into protein distribution and intensity across tissue types.

6. Immunofluorescence staining of cancer cells

Representative immunofluorescence staining images showing the subcellular distribution of the protein within the nucleus, endoplasmic reticulum (ER), and microtubules across three cancer cell lines were retrieved from the Human Protein Atlas (HPA) database [51],. These images illustrate the localization patterns of the protein within key cellular compartments, providing insights into its potential functional roles within the cell.

7. The cell cycle association analysis

The Human Protein Atlas (HPA) obtained and analyzed expression data plots from individual FUCCI U-2 osteosarcoma cells. Temporal mRNA expression patterns in these cells were characterized using the Fluorescent Ubiquitination-based Cell Cycle Indicator (FUCCI) U-2 OS cell line, which allows for precise tracking of cell cycle phases. This method enabled the observation of dynamic mRNA expression changes associated with distinct stages of the cell cycle, providing insights into cell cycle-dependent regulation of gene expression.

8. Stemness association analysis

The One-Class Logistic Regression (OCLR) algorithm [52] was employed to calculate the mRNA stemness index (mRNAsi) for TCGA pan-cancer mRNA sequencing data. This algorithm, specifically designed for single-class classification problems, works by learning a boundary that separates 'normal' data points (in this case, stemness-related features) from potential outliers or non-stemness signals in a high-dimensional space. By training on a set of stem cell-related genes, the OCLR algorithm identifies a boundary within the mRNA expression data that best characterizes stem-like properties across different cancer types. In the context of TCGA data, the OCLR model was trained on stem cell expression signatures and then applied to each tumor sample, assigning an mRNAsi score. This score reflects the degree of similarity between the tumor's gene expression profile and stem cell-like expression patterns, where higher mRNAsi values indicate stronger stemness characteristics. The mRNAsi thus serves as a quantitative measure of stemness, allowing for the

comparison of stem-like properties across different cancer types and facilitating the exploration of how stemness contributes to tumor progression and heterogeneity.

9. Mutation association analysis

The mutation levels in the samples were assessed by calculating the Tumor Mutational Burden (TMB) [53] and evaluating Microsatellite Instability (MSI) status [54]. TMB, defined as the total number of mutations per megabase of a sequenced genome, was used as a quantitative measure of mutational load, providing insights into the genomic instability within each tumor sample. MSI status, an indicator of defects in DNA mismatch repair (MMR) mechanisms, was determined to identify tumors with high MSI, a characteristic often associated with increased mutation rates and potential immunogenicity. Together, TMB and MSI analyses enabled a comprehensive assessment of mutation levels and mutational signatures across the cancer samples.

10. Immune cell infiltration analysis

Using the TCGA cohort, immune cell infiltration levels within tumor samples were estimated. This analysis was conducted using the CIBERSORT algorithm[55], a computational tool that quantifies the relative abundance of various immune cell types in complex tissues based on gene expression data. CIBERSORT deconvolutes bulk tumor transcriptomic data to infer the proportions of 22 distinct immune cell types, including T cells, B cells, macrophages, and natural killer cells. By applying this algorithm, we obtained a comprehensive profile of immune cell infiltration in each sample, allowing for further exploration of the tumor microenvironment's immune landscape and its potential association with clinical outcomes and cancer progression.

11. Single-cell sequencing data acquisition and analysis

Single-cell data were accessed and analyzed through the CancerSEA [56], CHARTS[57], and TISCH [58]. The datasets utilized included GSE117988 [59], GSE142213 [60], GSE143423, GSE131928 [61], GSE123814 [62], GSE70630 [63], etc.

12. Immune therapy prediction analysis

The Tumor Immune Dysfunction and Exclusion (TIDE) algorithm was used to perform immune therapy prediction analysis [64]. TIDE is a computational framework that evaluates the potential for immune evasion by simulating dysfunction in T cells and exclusion mechanisms within the tumor microenvironment. To assess the biomarker relevance of CENPA compared to standardized cancer immune evasion biomarkers, we examined its expression across immune checkpoint blockade (ICB) sub-cohorts and visualized these comparisons in a bar plot. The predictive performance of CENPA and other biomarkers regarding ICB response status was further evaluated by calculating the area under the receiver operating characteristic curve (AUC), which provided a quantitative measure of their accuracy in distinguishing responders from non-responders to ICB therapies.

13. Drug screening and prediction

Drug screening was conducted by evaluating the correlation between CENPA expression and drug sensitivity, applying a stringent significance cutoff of $p < 1e-5$. The area under the dose-response curve (AUC) values, reflecting drug efficacy, were analyzed alongside CENPA expression profiles across various cancer cell lines using GSCALite[65]. For this analysis, drug sensitivity data from the Genomics of Drug Sensitivity in Cancer (GDSC) [66] and Cancer Therapeutics Response Portal (CTRP) [67] databases were integrated, providing a comprehensive dataset for evaluating the sensitivity of different drugs in relation to CENPA expression. Spearman correlation analysis was then applied to determine the association between the expression levels of genes in the selected gene set and the sensitivity of small molecules/drugs.

To support the investigation of drug interactions with CENPA, a predictive structural model of the CENPA protein was retrieved from the AlphaFold database [68]. Protein-ligand docking was conducted using AutoDock Vina (version 1.1.2) [69], employing cavity-detection-guided blind docking to identify potential binding sites within the protein structure. This approach enabled the

prediction of interaction sites and binding affinities, providing insights into potential therapeutic targets involving CENPA.

14. Statistical analysis

Gene expression differences were compared using either the Wilcox test or the Kruskal-Wallis test. Survival analysis was performed using Kaplan-Meier analysis, along with the log-rank test and Cox regression test. Pearson's correlation test was applied to assess the relationship between two variables. Statistical significance was determined with a threshold of $P<0.05$.

Results

1. Genomic alteration of CENPA in cancers

The initial analysis of this study focused on investigating CENPA genomic alterations in various cancers. The alteration frequency bar plot revealed that the total alteration frequencies in most cancer types were below 10%. Non-small cell lung cancer exhibited the highest frequency, at 15.2% (7 out of 46 cases). The majority of gene alterations were amplifications (S-Figure 1A). To further explore CENPA mutations in cancers, TCGA mutation data was plotted, indicating that CENPA harbored only a low number of single-nucleotide variants across cancers (S-Figure 1B), which is consistent with the previous findings. The analysis of copy number variation demonstrated that nearly all copy number alterations of CENPA were heterozygous. Most cancer types exhibited 20-40% of CENPA heterozygous amplification, and approximately half had 5-10% heterozygous deletion samples. Lung squamous cell carcinoma (LUSC) showed the highest percentage of CENPA heterozygous amplification with no instances of heterozygous deletion, aligning with the earlier observation of a high frequency of gene amplification in lung cancer. In contrast, kidney chromophobe (KICH) had about 60% heterozygous deletions of CENPA with no amplification (S-Figure 1C). Overall pan-cancer data indicated that CENPA copy number could influence mRNA expression (S-Figure 1D). Therefore, while CENPA gene mutations may not be the primary driver in most cancers, its copy number alterations might influence cancer development through changes in mRNA expression.

2. The overexpression of CENPA in cancers

The analysis demonstrated that CENPA was overexpressed in the majority of cancer types compared to normal tissues in both females and males (Figure 1A). To streamline data presentation, abbreviations were used to denote cancer types, which are listed in S-Table 1. The mRNA expression analysis of CENPA, utilizing data from TCGA and GTEx, revealed significant overexpression in 30 out of the 33 cancer types examined. Notably, mesothelioma (MESO) and uveal melanoma (UVM) lacked comparable normal tissue, while acute myeloid leukemia (LAML) was the only cancer type where CENPA expression was lower in cancerous tissues than in normal tissues (Figure 1B). To achieve better control in the comparison between cancerous and non-cancerous tissues, paired samples from the same patients were analyzed. This comparison indicated that CENPA was significantly overexpressed in 16 cancer types (Figure 1C). To further investigate CENPA overexpression in cancers, protein staining of CENPA in cancerous versus corresponding normal tissues was examined in representative cancer types. The staining images generally showed that, although CENPA staining intensity was slightly stronger in cancer tissues, the overall staining intensity in both cancerous and normal tissues was low, potentially due to the properties of the antibody used (Figure 1D).
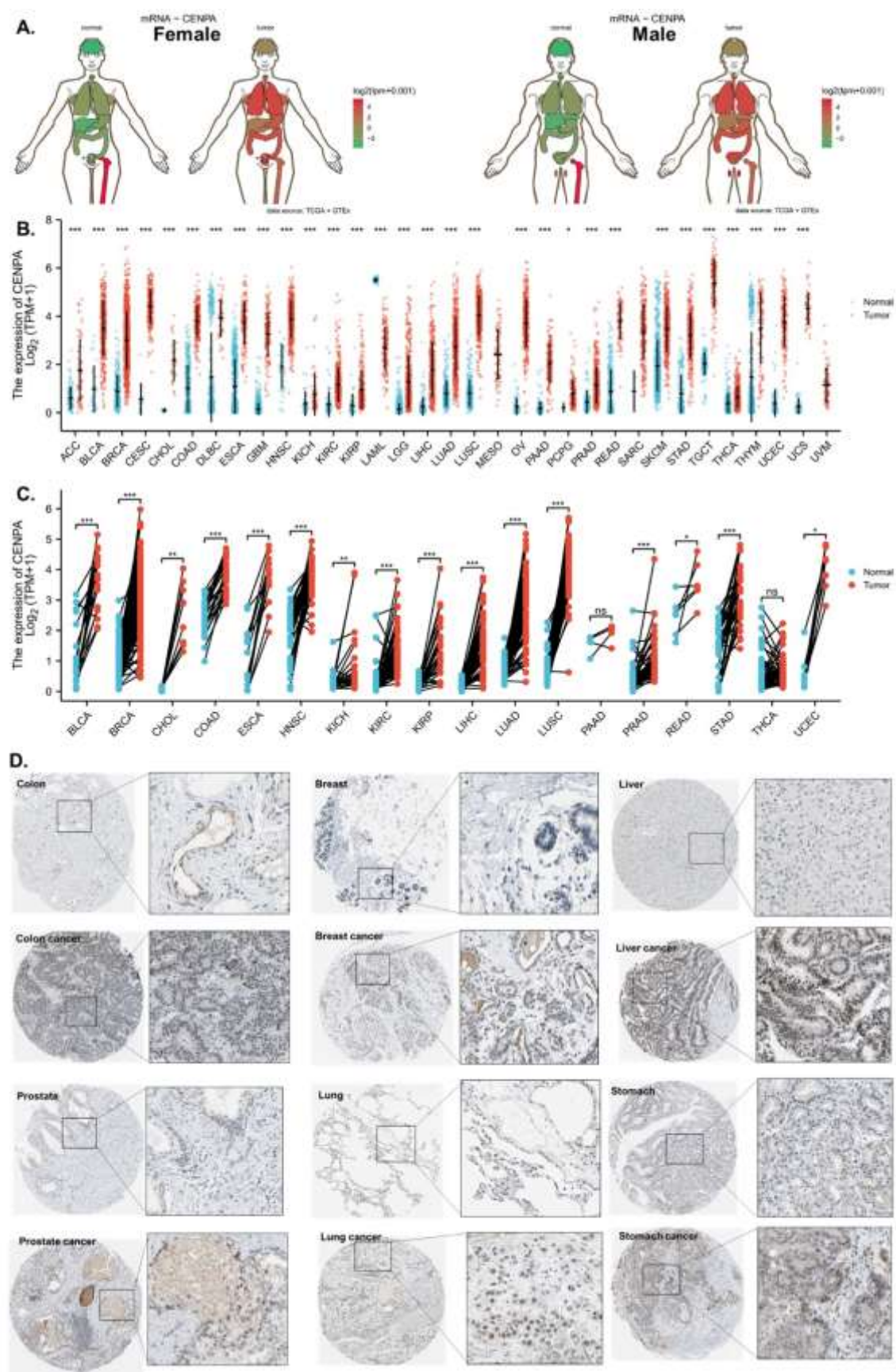
Figure 1. The overexpression of CENPA in cancers. **A.** Anatomy plot of the gene expression profile of CENPA across all tumor samples and paired normal tissues in females and males. TCGA data were plotted. **B.** The gene expression profile of CENPA across all tumor samples and normal tissues. TCGA and GTEx data were plotted. **C.** Paired sample expression profile of CENPA across all tumor samples

and normal tissues. TCGA data were plotted. **D.** Representative protein staining images of CENPA in cancers and corresponding normal tissues. The images were downloaded from the Human Protein Atlas (HPA). *p<0.05; **p<0.01; ***p<0.001.

3.   The diagnostic value of CENPA in cancers

To assess the diagnostic value of CENPA in various cancers, single-variable receiver operating characteristic (ROC) curves were plotted for different cancer types, and the area under the curves (AUC) was calculated using data from TCGA and GTEx. The results demonstrated that 19 cancer types had AUCs exceeding 0.9, indicating an outstanding diagnostic power of CENPA. Five cancer types had AUCs ranging from 0.8 to 0.9, supporting the excellent diagnostic capability of CENPA. Additionally, three cancer types had AUCs between 0.7 and 0.8, reflecting an acceptable diagnostic power of CENPA [70] (Figure 2). These results suggested that CENPA is a promising diagnostic molecular biomarker that can be developed for multiple cancer types.
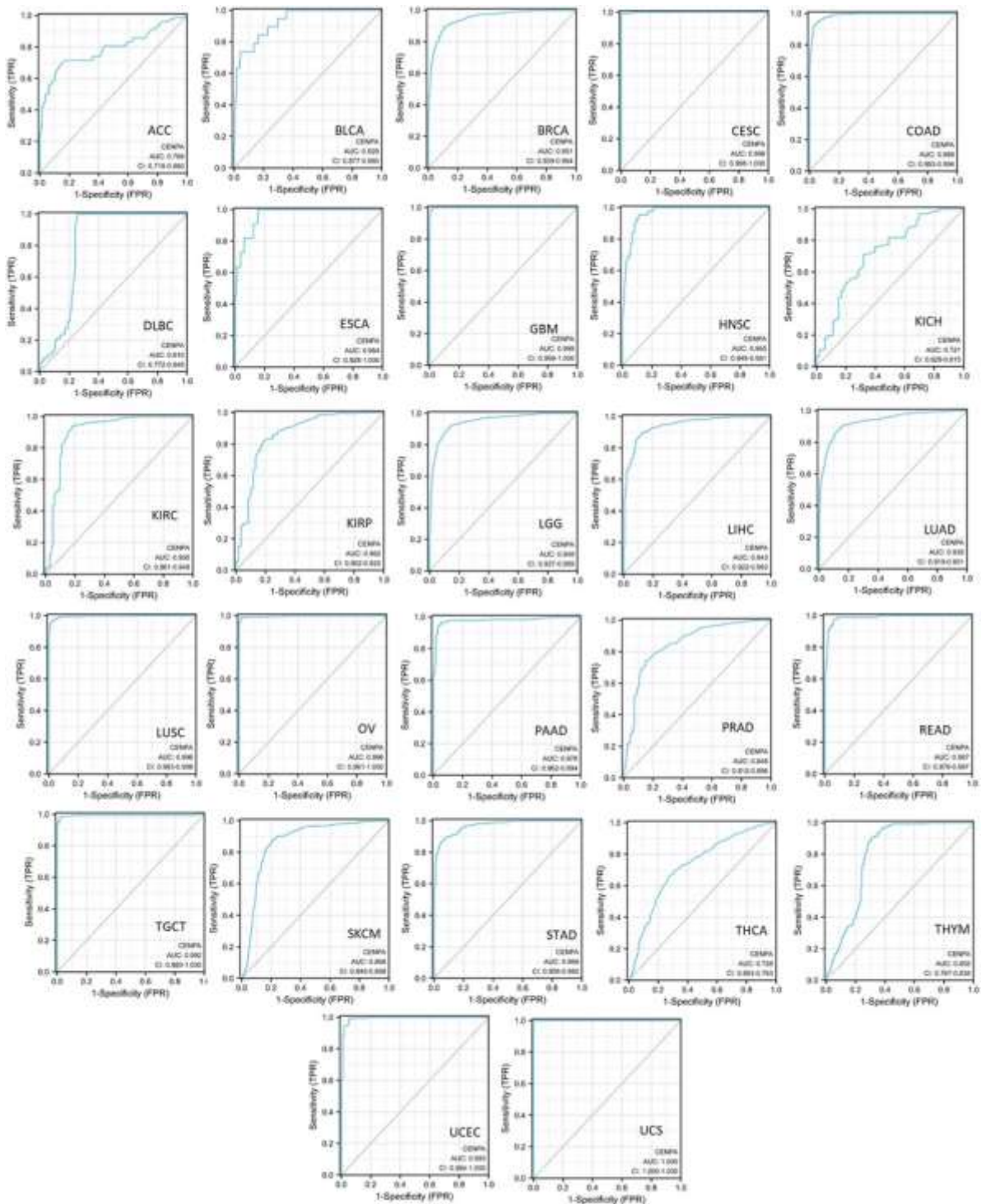
Figure 2. The pan-cancer diagnostic value of CENPA. The diagnostic receiver operating characteristic (ROC) curve of different cancer types. TCGA and GTEx data were used to calculate the ROC. The area under the curves (AUC) and the corresponding 95% confidential interval (95%CI) was shown.

4. The prognostic value of CENPA in cancers

This study also aimed to explore the prognostic value of CENPA in various cancers. To this end, univariate overall survival Cox regression analysis was performed for CENPA across 33 cancer types using TCGA data. The results revealed that CENPA was significantly associated with worse overall survival in 13 cancer types, while it was linked to better overall survival in one cancer type, thymoma (THYM) (Figure 3A). To further investigate the association between CENPA and overall survival,

Kaplan-Meier (KM) plots and log-rank analyses were conducted for the cancer types that showed significance in the Cox regression analysis. The results indicated that 12 cancer types remained significant in the log-rank analysis (Figure 3B, first panel for each cancer type).
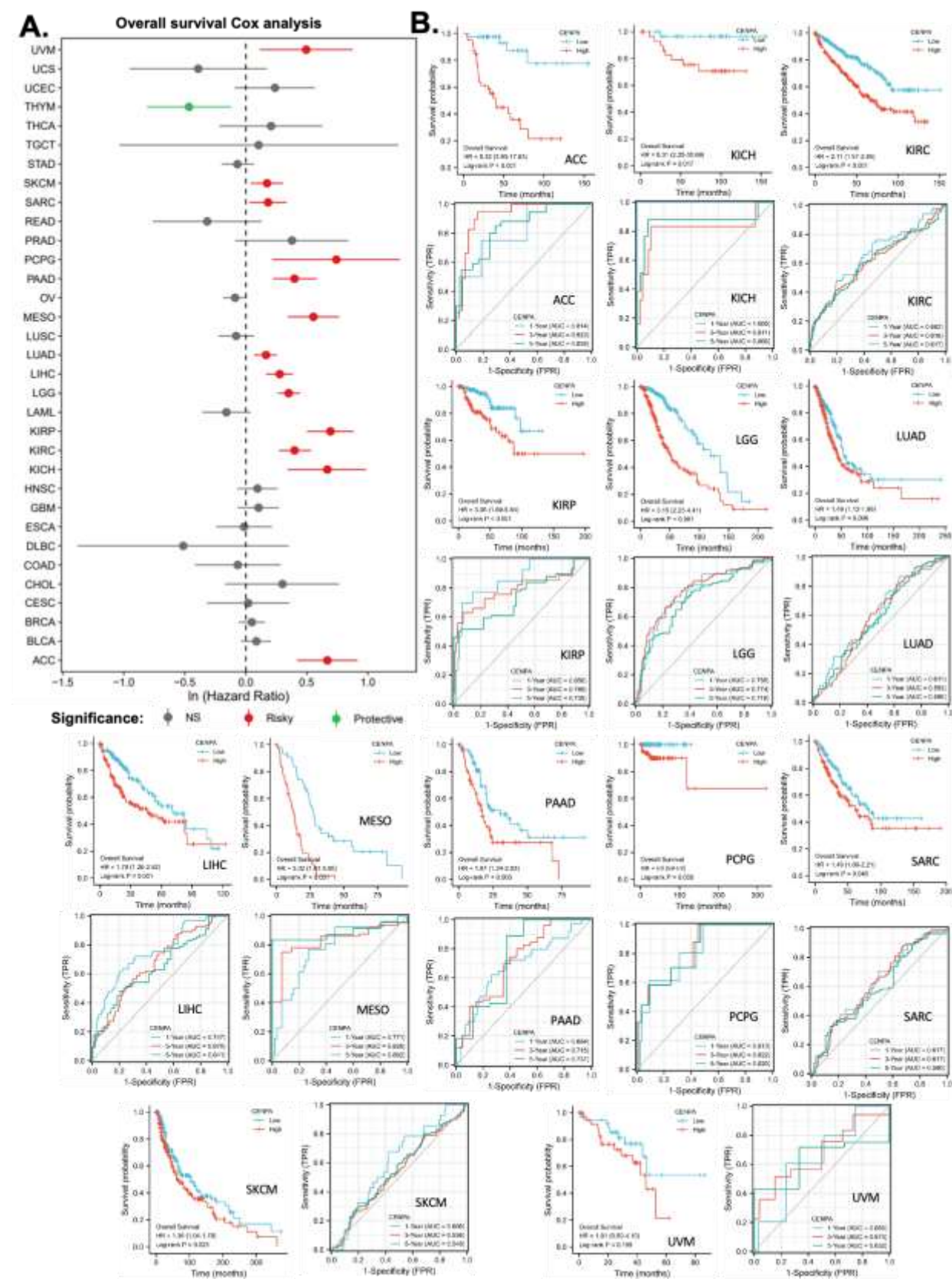

Figure 3. The pan-cancer prognostic value of CENPA. TCGA data were analyzed. **A.** Univariate Cox regression analysis of CENPA for overall survival in different cancer types. **B.** The overall survival

Kaplan-Meier (KM) plot and log-rank analysis of high (50-100%) and low (0-50%) CENPA patients with time-dependent (1-, 3-, and 5-year) overall survival prognostic receiver operating characteristic curve (ROC). Only cancer types with significance in Cox regression were plotted.

To assess the prognostic value of CENPA in these cancer types, time-dependent prognostic ROC curves were plotted. For 1-year overall survival, the AUC for kidney chromophobe (KICH) exceeded 0.9, indicating outstanding predictive power. The AUCs for adrenocortical carcinoma (ACC), kidney renal papillary cell carcinoma (KIRP), and pheochromocytoma and paraganglioma (PCPG) ranged between 0.8 and 0.9, suggesting excellent predictions. The AUCs for lower-grade glioma (LGG), liver hepatocellular carcinoma (LIHC), and mesothelioma (MESO) were between 0.7 and 0.8, indicating acceptable predictions. For 3-year overall survival, the AUC for adrenocortical carcinoma (ACC) was over 0.9, indicating outstanding predictive accuracy. The AUCs for kidney chromophobe (KICH), mesothelioma (MESO), and pheochromocytoma and paraganglioma (PCPG) ranged between 0.8 and 0.9, suggesting excellent predictions, while the AUCs for kidney renal papillary cell carcinoma (KIRP), lower-grade glioma (LGG), and pancreatic adenocarcinoma (PAAD) were between 0.7 and 0.8, indicating acceptable predictions. For 5-year overall survival, the AUCs for adrenocortical carcinoma (ACC), kidney chromophobe (KICH), mesothelioma (MESO), and pheochromocytoma and paraganglioma (PCPG) were between 0.8 and 0.9, indicating excellent predictive power, while the AUCs for kidney renal papillary cell carcinoma (KIRP), lower-grade glioma (LGG), and pancreatic adenocarcinoma (PAAD) were between 0.7 and 0.8, suggesting acceptable predictions (Figure 5B, second panel for each cancer type). These findings suggest that CENPA is a promising prognostic molecular biomarker with potential applicability in multiple cancer types, such as adrenocortical carcinoma (ACC), kidney chromophobe (KICH), kidney renal papillary cell carcinoma (KIRP), lower-grade glioma (LGG), mesothelioma (MESO), and pheochromocytoma and paraganglioma (PCPG).

5. The application of CENPA for glioma prognosis

To demonstrate the practicable clinical application of CENPA, we focused on one cancer type, glioma, where CENPA was demonstrated to have promising prognostic value. The World Health Organization (WHO) defined glioma into four grades based on histology and clinical criteria: G1, G2, G3, and G4 [71]. The G1 glioma is generally benign and has a very good prognosis, which has been distinguished from the G2, G3, and G4 glioma. In TCGA cohort, G2 and G3 glioma together are referred to as "low-grade glioma (LGG)", while G4 glioma is referred to as "glioblastoma multiforme (GBM)" (highest grade glioma) [72]. In this context, this study combined LGG and GBM and analyzed the prognostic value of CENPA for overall glioma.

To validate the prognostic accuracy of CENPA for overall survival in glioma patients, we examined its prognostic association across five independent glioma datasets: TCGA (LGG+GBM) (n=703), CGGA mRNAseq693 (n=693), CGGA mRNAseq325 (n=325), CGGA mRNA-array301 (n=301), and ICGC (pediatric brain tumor) (n=120). Kaplan-Meier (KM) plots and Cox regression analyses demonstrated that high CENPA expression was significantly associated with worse survival across all five datasets. The hazard ratios (HR) ranged from 2.95 to 7.21. ROC analysis revealed that for 1-year overall survival prediction, four datasets indicated acceptable accuracy. For 3-year overall survival prediction, three datasets suggested excellent accuracy, while two indicated acceptable accuracy. For 5-year survival prediction, three datasets showed excellent accuracy, and two indicated acceptable accuracy (Figure 4A).
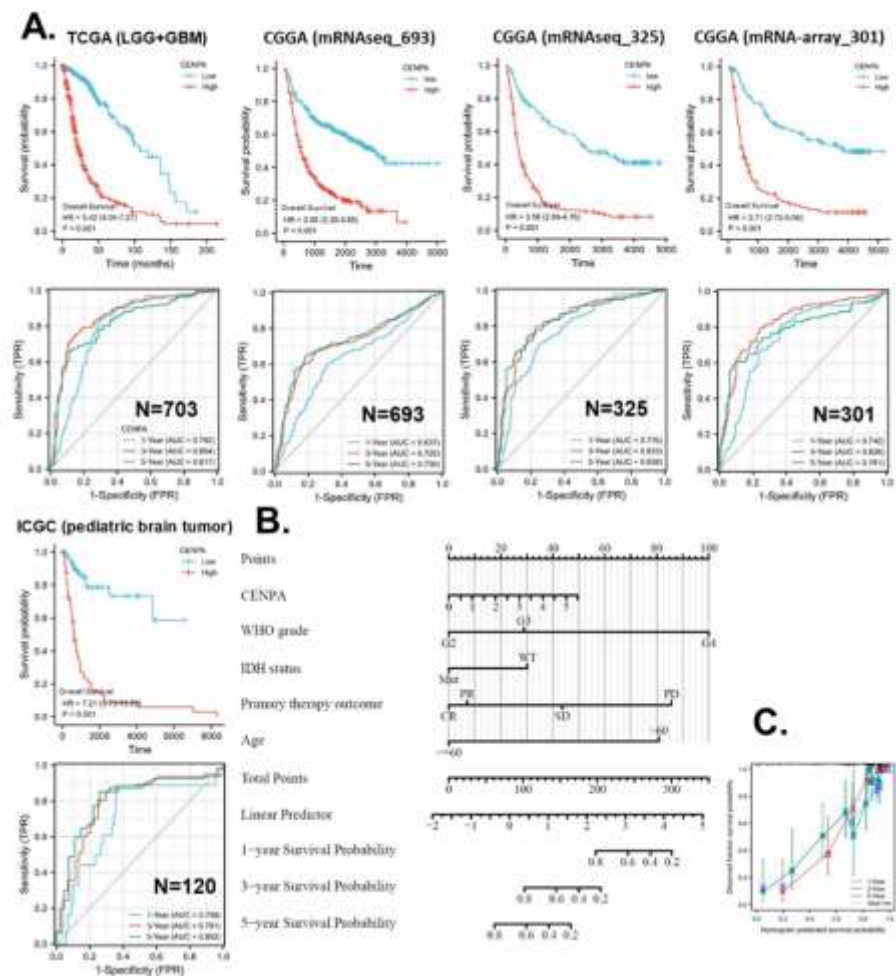
Figure 4. Application of CENPA for glioma prognosis. **A.** validation of the survival association of CENPA in five independent glioma cohorts. TCGA (LGG+GBM), CGGA (mRNAseq 693), CGGA (mRNAseq 325), CGGA (mRNA-array 301), and ICGC (pediatric brain tumor) were analyzed. The overall survival Kaplan-Meier (KM) plot and Cox analysis of high (50-100%) and low (0-50%) CENPA patients with time-dependent (1-, 3-, and 5-year) overall survival prognostic receiver operating characteristic curve (ROC) were shown. **B.** Nomogram for the prediction of 1-, 3-, and 5-year overall survival of glioma patients. The TCGA (LGG+GBM) cohort was used to construct the prognostic model of CENPA for glioma. **C.** Calibration plots of the nomogram for estimation of overall survival of glioma patients at years 1, 3, and 5.

In this study, we developed strategies for applying CENPA in glioma prognosis, illustrating its potential as a clinical prognostic biomarker for cancer. To identify variables for the CENPA-based prognostic model in glioma patients, we performed Cox regression analysis to evaluate prognostic factors. Univariate Cox regression results indicated that CENPA level, 1p/19q codeletion, primary therapy outcome, IDH status, and age were significantly associated with overall survival in glioma patients. Multivariate Cox regression showed that CENPA level, primary therapy outcome, IDH status, and age remained significant after adjustment for other variables, suggesting they provide additional prognostic power as independent factors in the model (S-Table 2). Consequently, these factors, along with WHO grade (G2-4), were included in the prognostic model for overall survival in glioma patients. Based on this model, a nomogram was constructed to predict the survival probability of glioma patients at 1, 3, and 5 years (Figure 4B). The calibration curves of the nomogram predictions generally aligned with the observed outcomes in patients (Figure 4C).

6. **CENPA was highly expressed in the nucleus of malignant cells**

To explore the cell populations and cellular locations where CENPA is expressed, we conducted an analysis of single-cell sequencing data and observed the subcellular distribution of CENPA through immunofluorescence staining in three cancer cell lines. The single-cell sequencing data set included three cancer types: acute erythroid leukemia (AEL), breast cancer (BRCE), glioma, and Merkel cell carcinoma (MCC). The analysis revealed that CENPA was expressed by a small subset of malignant cells, whereas immune cells exhibited relatively low levels of CENPA expression (Figure 5A). Immunofluorescence staining of the subcellular distribution of CENPA in prostate cancer cell line PC-3, rhabdomyosarcoma cell line RH30, and osteosarcoma cell line U2OS demonstrated that CENPA was predominantly localized in the nucleus, although U2OS exhibited relatively lower fluorescence intensity (Figure 5B). It is worth noting that rhabdomyosarcoma is a type of sarcoma. Prostate cancer (PRAD) and sarcoma (SARC) were shown earlier in this study to overexpress CENPA, while osteosarcoma was not included among the cancer types in the TCGA data set.
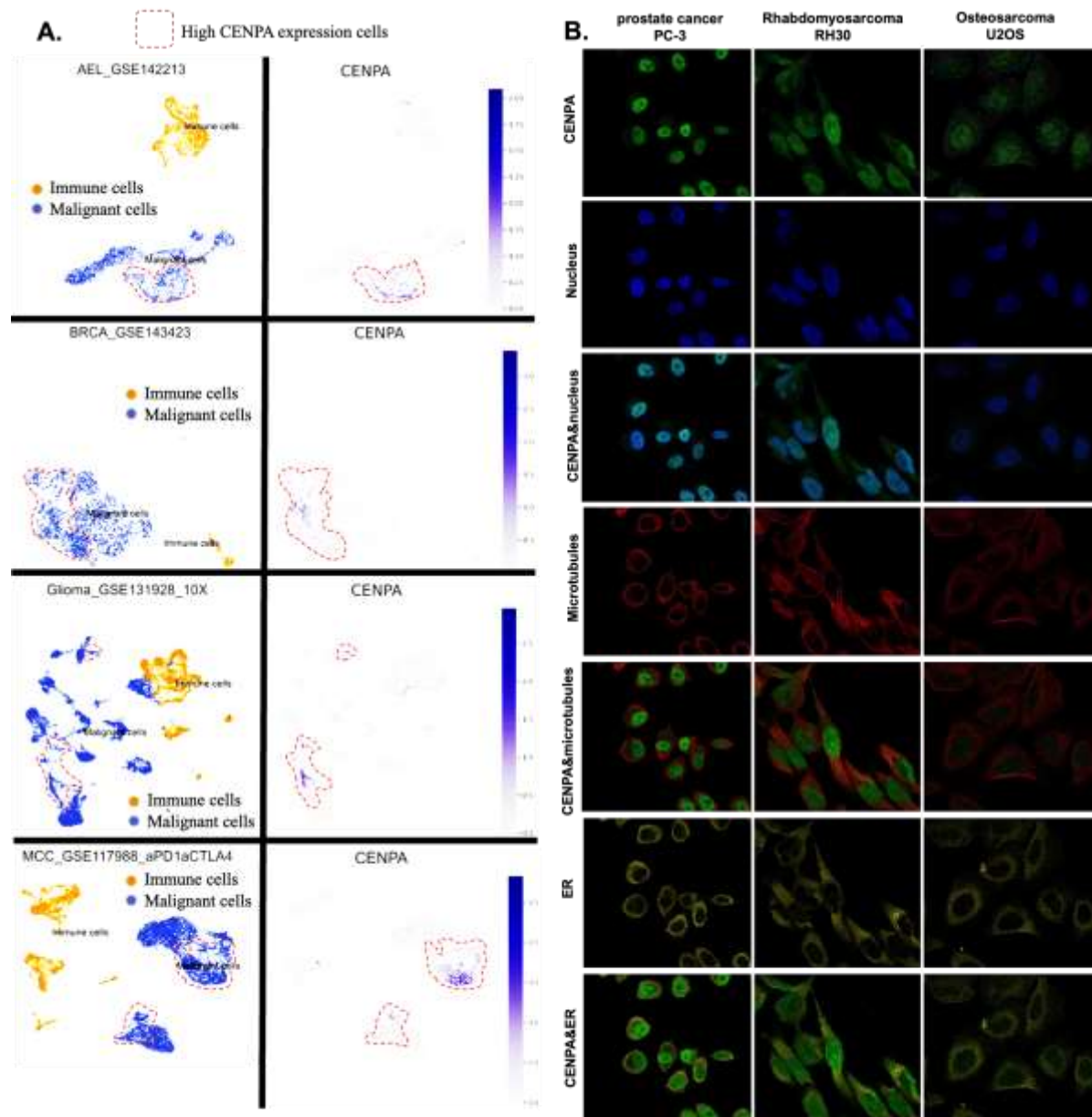

Figure 5. The expression of CENPA in cell populations and cellular locations. **A.** The expression of CENPA in cell populations in cancer tissues. Single-cell mRNA expression cohorts were accessed and analyzed using the TISCH. **B.** Immunofluorescence staining of the subcellular distribution of CENPA within the nucleus, endoplasmic reticulum (ER), and microtubules of three cancer cell lines.

7. CENPA was associated with the cell cycle of cancer cells

Since CENPA was predominantly detected in the nucleus of cancer cells, we hypothesized two potential roles for CENPA in cancers: 1) CENPA may influence the mutation of other genes, given that gene transcription occurs in the nucleus, and 2) CENPA could regulate the cell cycle, as DNA replication during the cell cycle also takes place in the nucleus. To test the first hypothesis, we analyzed the correlation between CENPA expression and two mutation indicators: tumor mutation burden (TMB) and microsatellite instability (MSI). TMB quantifies the approximate number of gene mutations within the cancer genome, while MSI reflects a state of genetic hypermutability resulting from impaired DNA mismatch repair (MMR). The presence of MSI serves as phenotypic evidence that MMR is not functioning correctly. The analysis indicated that CENPA expression was positively correlated with TMB and MSI across most cancer types, though the correlations were weakly significant (S-Figure 2A-B). These findings suggest that CENPA is not generally associated with genomic instability in cancers.

To explore the potential common functional effects of CENPA in cancers, we identified the top CENPA-correlated genes by analyzing data from all 33 TCGA cancer types as a single cohort. The top 30 CENPA-correlated genes were used to construct a protein-protein interaction (PPI) network, highlighting the possible associations between CENPA and these genes (S-Figure 2C). Further analysis of the top 200 correlated genes was conducted through GO and KEGG enrichment studies. KEGG pathway enrichment revealed that the top two pathways associated with CENPA were "DNA replication" and "Cell cycle." The top GO molecular function (MF) was related to ATPase activities, the top GO cellular component (CC) was chromosome regions, and the top GO biological processes (BP) included "organelle fusion," "mitotic nucleus division," and "nucleus division." These GO-enriched terms were all linked to cancer proliferation and the cell cycle (S-Figure 2D).

To further validate the potential association between CENPA and cancer proliferation and cell cycle regulation, we analyzed the correlation between CENPA expression and cancer functional signals using multiple single-cell data sets across various cancer types. These correlation results were summarized (as shown in the top bar plot of S-Figure 2E) to provide an overview of CENPA's potential common roles in cancers. The results indicated that the most significant positive correlations were with "cell cycle" and "proliferation," supporting the hypothesis that CENPA may regulate the cell cycle and proliferation. Additionally, CENPA appeared to be negatively associated with "apoptosis," "DNA repair," and "metastasis" (S-Figure 2E). These data support the notion that CENPA may play a role in regulating cancer growth.

8. CENPA is a biomarker for the cell cycle G2 phase in cancer cells

The ability of a tumor to proliferate and propagate relies on a small population of stem-like cells, the OCLR algorithm [52] has been wildly applied for the estimation of the stemness in a tissue sample. In this study, the mRNAsi (a measure of stemness) was calculated for 33 cancer types in the TCGA, and the correlation between CENPA expression and pan-cancer stemness was analyzed. The results indicated that CENPA expression was positively correlated with stemness across most cancer types (Figure 6A), suggesting that the association with stemness might be a common mechanism of CENPA in cancer. Building on these findings, we proposed that CENPA could serve as a novel cell cycle biomarker and conducted a GSEA enrichment analysis of CENPA-correlated genes in the "REACTOME CELL CYCLE CHECKPOINTS" pathway. As expected, the analysis showed that CENPA-correlated genes were significantly enriched in "REACTOME CELL CYCLE CHECKPOINTS" (Figure 6B).
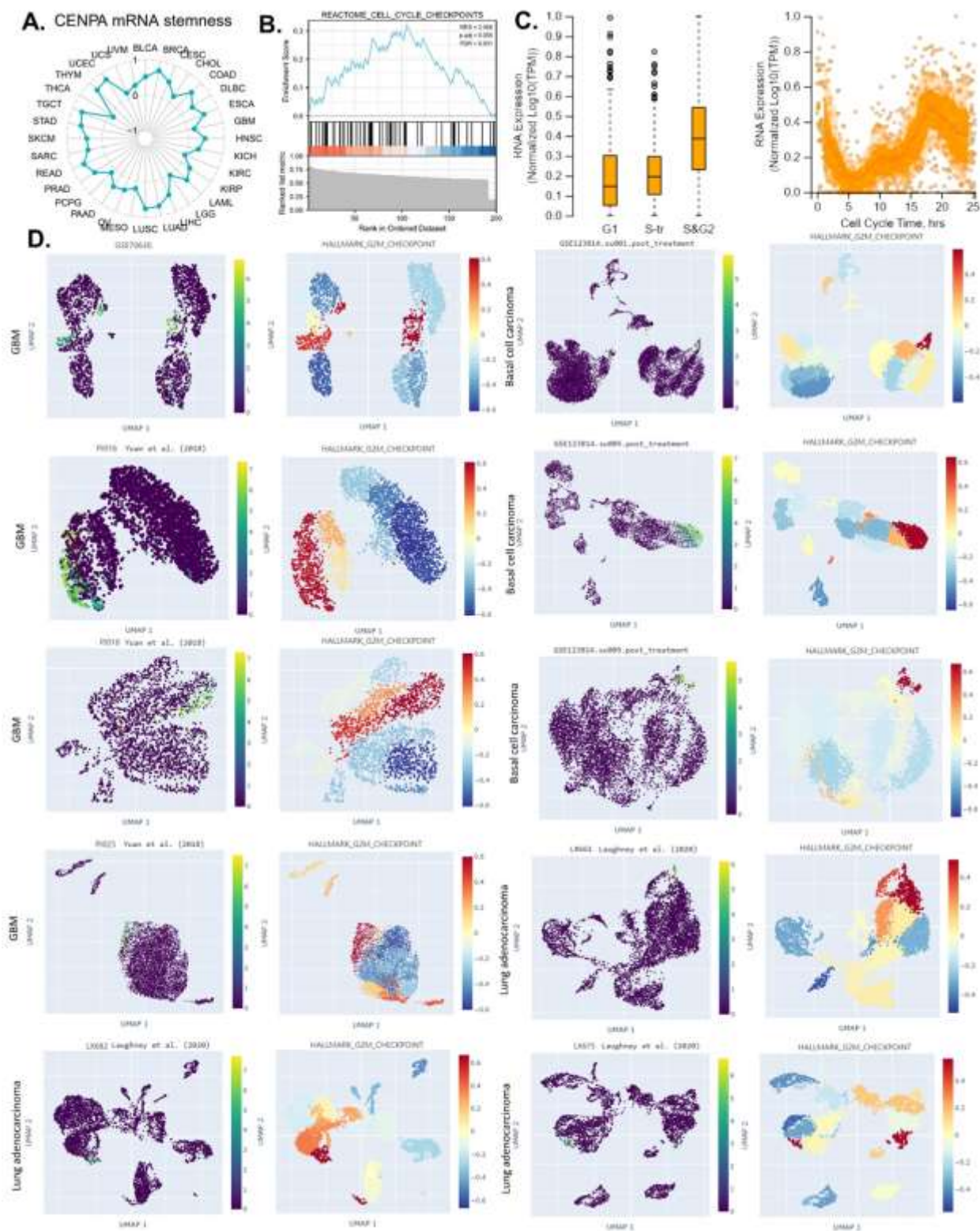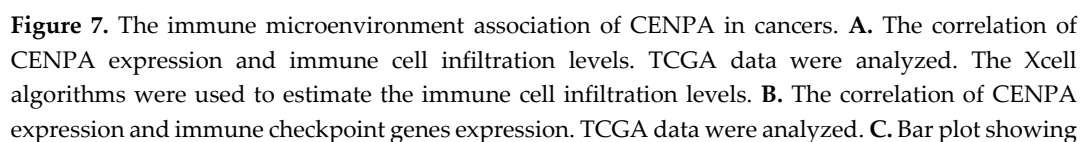
Figure 6. The potential of CENPA as a cell-cycle biomarker for the M/G2 phase in cancers. **A.** The correlation of OCLR scores and CENPA in TCGA cancer data. The OCLR algorithm was used to calculate the mRNAsi (OCLR scores) for the evaluation of stemness. **B.** The GSEA enrichment of CENPA-correlated genes in "REACTOME CELL CYCLE CHECKPOINTS". The top 200 CENPA-correlated genes were identified using the GEPIA based on all TCGA cancer data and used for the GSEA enrichment analysis. **C.** Plots of single-cell RNA-sequencing data from the FUCCI U-2 OS osteosarcoma cell line, showing the correlation between CENPA mRNA expression and cell cycle progression. **D.** The expression of CENPA in single cells and the G2M checkpoint hallmark signals in cancer tissues. Single-cell data were accessed and analyzed using the CHARTS.

To further understand CENPA's specific role in different phases of the cell cycle in cancer cells, we analyzed single-cell expression data for CENPA across various cell cycle phases in U2OS cells, which predominantly express CENPA in the nucleus. The results revealed that CENPA expression was low during the G1 phase and high during the S and G2 phases (Figure 6C). Based on these findings, we hypothesized that CENPA might be closely associated with the G2 phase of the cell cycle. To test this hypothesis, we examined CENPA expression across several single-cell cancer datasets and compared it with single-cell signals of the G2/M checkpoint, a hallmark gene set related to cell proliferation in GSEA[73]. Among all the ten single-cell cancer data sets analyzed, CENPA was highly expressed in a population of cell clusters that had strong signals of G2M checkpoint. These results confirmed that CENPA was a biomarker for the cell cycle G2 phase (Figure 6D).

9. The immune microenvironment association of CENPA in cancers

This study also investigated the potential of CENPA as a biomarker for the immune microenvironment. Since the effectiveness of cancer immune therapy largely depends on immune cell infiltration levels and the presence of immune checkpoints, we explored the value of CENPA as a predictive biomarker for immune therapy from these two perspectives.

Earlier analyses revealed that CENPA was predominantly expressed in a small population of malignant cells, with relatively low expression in immune cells. However, whether CENPA expression in tumors affects immune cells has not been previously examined. To address this, we calculated immune cell infiltration levels in cancers and analyzed their correlation with CENPA expression. The analysis identified T cell CD4+ as the most notable immune cell type correlated with CENPA; Th2 cells were positively correlated with CENPA across all cancer types, and Th1 cells were positively correlated in the majority of cancers. Additionally, common lymphoid progenitors showed a positive correlation with CENPA in most cancer types. CENPA was closely associated with multiple immune cells across different cancers, particularly in lung squamous cell carcinoma (LUSC), lung adenocarcinoma (LUAD), glioblastoma multiforme (GBM), and thymoma (THYM) (Figure 7A).

Figure 7. The immune microenvironment association of CENPA in cancers. **A.** The correlation of CENPA expression and immune cell infiltration levels. TCGA data were analyzed. The Xcell algorithms were used to estimate the immune cell infiltration levels. **B.** The correlation of CENPA expression and immune checkpoint genes expression. TCGA data were analyzed. **C.** Bar plot showing

the biomarker relevance of CENPA compared to standardized cancer immune evasion biomarkers in immune checkpoint blockade (ICB) sub-cohorts. The area under the receiver operating characteristic curve (AUC) was applied to evaluate the predictive performances of the biomarkers on the ICB response status.

We also examined the correlation between CENPA and several commonly used immune checkpoints in current immune therapies. The results indicated that CENPA was positively associated with most immune checkpoints in thyroid carcinoma (THCA), lung adenocarcinoma (LUAD), liver hepatocellular carcinoma (LIHC), lower-grade glioma (LGG), kidney renal clear cell carcinoma (KIRC), breast invasive carcinoma (BRCA), and bladder urothelial carcinoma (BLCA), while it showed a negative correlation with most immune checkpoints in thymoma (THYM), lung squamous cell carcinoma (LUSC), glioblastoma multiforme (GBM), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), and adrenocortical carcinoma (ACC) (Figure 7B).

To compare the predictive performance of CENPA for immune checkpoint blockade (ICB) treatment with other standardized biomarkers, we assessed the relevance of CENPA and other biomarkers based on their ability to predict ICB response outcomes in various sub-cohorts. The results showed that CENPA expression had an AUC greater than 0.5 in 11 out of 25 ICB sub-cohorts, which is higher than the number of cohorts where microsatellite instability (MSI) score, tumor mutational burden (TMB), T cell clonality (T.Clonality), and B cell clonality (B.Clonality) achieved an AUC over 0.5 (seven, nine, and six cohorts, respectively). However, the predictive value of CENPA was lower than that of CD27A, tumor immune dysfunction and exclusion (TIDE), interferon-gamma (IFNG), CD8, and Merck 18 (Figure 7C). These comparisons highlight the potential value of CENPA as a predictive biomarker for immune therapy.

10. Computational drug predictions of CENPA in cancers

Given that our study demonstrated a close association between CENPA and the cancer cell cycle, patient survival, and the immune microenvironment, we proposed CENPA as a potential therapeutic target for cancer drug treatment. To explore this, we screened and predicted potential drugs targeting CENPA using cancer drug databases and computational methods. Drug sensitivity data were obtained from the GDSC and CTRP databases, and the correlation between CENPA expression and the sensitivity of cancer cell lines to various small molecules and drugs was analyzed. Data from multiple cancer cell lines in GDSC and CTRP were integrated for these calculations. We applied a significance cutoff of $p<1e-5$ to identify relevant drugs. The screening identified 8 drugs with sensitivities negatively correlated with CENPA levels in cancer cells and 4 drugs with sensitivities positively correlated with CENPA levels (Figure 8A and S-Table 3). We hypothesized that these 12 drugs might directly interact with the CENPA protein.
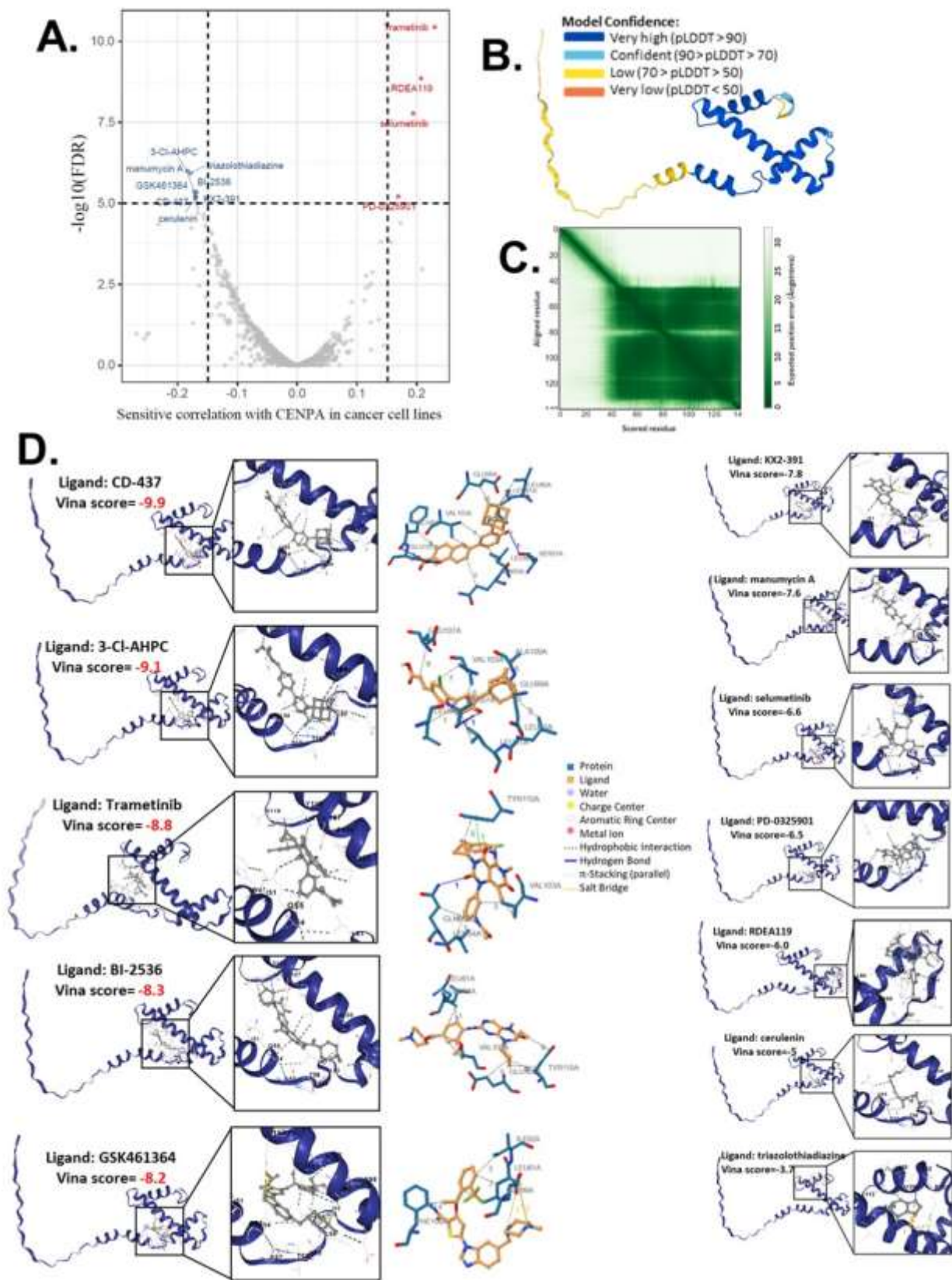
.

Figure 8. The computational drug prediction of CENPA in cancers. **A.** The volcano plot of the correlation of CENPA expression and small molecule/drug sensitivity of cancer cell lines. GDSC and CTRP data were analyzed. Drug sensitivity and gene expression profiling data of multiple cancer cell lines in GDSC and CTRP were integrated for investigation. The expression of CENPA was performed by Spearman correlation analysis with the small molecule/drug sensitivity (area under the IC50

curve). **B.** Predictive protein structural model of CENPA. **C.** Predicted aligned error of the CENPA protein structure model. **D.** Protein-ligand docking models of CENPA and identified drugs. The names of the ligands and the docking vina scores were shown. For models with a vina score of lower than -8.0 (indicates a binding affinity), the protein-ligand molecular interaction profiles were displayed on the right.

To predict the direct interaction between CENPA and these 12 drugs, we accessed the predictive protein structural model of CENPA from the Alphafold database and performed protein-ligand docking for CENPA and the identified drugs. The predicted aligned error of the CENPA protein structure model indicated that the N-terminus had a long tail with low model confidence, while the docking was focused on regions with very high model confidence (Figure 8B-C). A protein-ligand model with a vina score lower than -8 was considered to have a very good binding affinity. The docking results revealed that CD-437, 3-Cl-AHPC, Trametinib, BI-2536, and GSK461364 had high binding affinities to CENPA (S-Table 3), suggesting that these drugs are likely to directly target CENPA in cancer cells. All docking models are displayed in Figure 8D.

Discussion

This study used bioinformatic data to support the potential values of CENPA for clinical cancer diagnosis and prognosis. Although the function of CENPA in cell growth and cell cycle has been studied [74], the association of CENPA and cancers has not been studied comprehensively and the clinical use of CENPA as a biomarker for cancer has not been developed. CENPA has been proposed as a genomic marker for centromere activity[75]. Single-cell analysis in this study suggested that CENPA was highly expressed during the S&G2 phase in the cell cycle and was closely associated with the G2/M checkpoint in cancer cells. These indicated that CENPA can be a biomarker for the G2 phase in the cell cycle.

In addition, CENPA plays a central role in the regulation of centromere activity. The inheritance of genetic material requires the faithful segregation of chromosomes during cell division, when kinetochores, a unique centromere macromolecular protein, attach chromosomes to the spindle for proper movement and segregation. CENPA directly regulates the assembly of active kinetochores, thereby regulating cell division[76]. While this process is crucial for nearly all proliferating cells, irrespective of whether they are malignant, one common characteristic of cancer cells is their significantly higher proliferation rate compared to non-cancerous cells. This suggests that cancer cells undergo more frequent cell divisions and, therefore, might require higher levels of CENPA for proper kinetochore regulation. The expression analysis supported this hypothesis, showing that cancer cells indeed express higher levels of CENPA compared to non-cancerous cells. The results revealed that almost all cancer types overexpressed CENPA, with the exception of LAML, which exhibited lower CENPA expression in cancerous tissues than in normal tissues. This finding is understandable, as LAML, being a type of leukemia, likely has distinct cell cycle regulation mechanisms compared to most other cancer types [77]. The overexpression of CENPA in cancers inferred its pan-cancer potential as a diagnostic biomarker and a therapeutic target. Nevertheless, further studies to compare the diagnostic power of CENPA with present diagnostic biomarkers are required for further development of CENPA for clinical use.

The gene alteration analysis in this study indicated that CENPA mutations are unlikely to be a major driving factor in cancer development, given the low mutation rate observed. However, changes in copy number could potentially influence cancer progression by increasing the transcription of CENPA mRNA. As a result, our study primarily focused on the expression levels of CENPA. A previous study has reported that overexpression of CENPA can promote genome instability in human cells, particularly when the retinoblastoma protein is inactivated [78]. Our TMB and MSI analysis indicated that CENPA was not associated with genome instability in all cancers. In eye cancer (UVM), CENPA was not correlated with TMB but correlated with MSI. The analysis of single-cell data (SFigure 2E) also suggested that CENPA was negatively correlated with DNA repair in eye cancer. Most of these results were consistent with the previous study.

The expression of CENPA has been reported to be associated with worse overall survival of some cancer types, such as ovarian cancer [37], liver cancer [40], breast cancer [41,42], and lung cancer [43]. Most of these studies were also using TCGA data, but they were only limited to one cancer type regardless of the common roles of CENPA across multiple cancer types. A previous study had demonstrated the potential of CENPA as a prognostic biomarker for GBM [79]. However, the conclusions of the previous study were based solely on TCGA data and focused exclusively on GBM. In contrast, this study extends those conclusions to glioma as a whole, encompassing both low-grade and high-grade gliomas. The prognostic association between CENPA and glioma patient survival was supported by five independent glioma datasets, with sample sizes of 703, 693, 325, 301, and 120, respectively. Given the relatively larger number of datasets and independent data sources, we believe that the prognostic performance of CENPA is quite reliable.

The immune association of CENPA in certain cancer types, such as lung and liver cancer [80], has been previously demonstrated using TCGA data[43]. This study broadened the scope of this association to a pan-cancer context and compared CENPA's predictive value for immune checkpoint blockade (ICB) response with that of other immune response biomarkers. While the ICB cohorts in this study were not large, the findings suggest a potential role for CENPA in predicting immune therapy outcomes, which warrants further validation. Additionally, we used computational methods to screen and predict potential drugs targeting CENPA in cancer cells. These computational predictions also require experimental validation to confirm their efficacy

Conclusions

CENPA holds promise as a biomarker in cancers linked to cell cycle regulation and stemness, with significant potential for diagnostic, prognostic, and therapeutic applications.

References

1. Siegel RL, Giaquinto AN, Jemal A: **Cancer statistics, 2024**. *CA Cancer J Clin* 2024, **74**(1):12-49.
2. Sonkin D, Thomas A, Teicher BA: **Cancer treatments: Past, present, and future**. *Cancer Genet* 2024, **286-287**:18-24.
3. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM: **The Cancer Genome Atlas Pan-Cancer analysis project**. *Nat Genet* 2013, **45**(10):1113-1120.

4. Trust W: **Sharing data from large-scale biological research projects: a system of tripartite responsibility**. In: *Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA: 2003*: Wellcome Trust London; 2003.

5. Zhao Z, Zhang KN, Wang Q, Li G, Zeng F, Zhang Y, Wu F, Chai R, Wang Z, Zhang C *et al*: **Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients**. *Genomics, proteomics & bioinformatics* 2021, **19**(1):1-12.

6. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS *et al*: **International network of cancer genome projects**. *Nature* 2010, **464**(7291):993-998.

7. Liu H, Guo Z, Wang P: Genetic expression in cancer research: Challenges and complexity. *Gene Reports* 2024:102042.

8. Ou L, Liu H, Peng C, Zou Y, Jia J, Li H, Feng Z, Zhang G, Yao M: Helicobacter pylori infection facilitates cell migration and potentially impact clinical outcomes in gastric cancer. *Heliyon* 2024(in-press).

9. Liu H, Weng J, Huang CLH, Jackson AP: **Voltage-gated sodium channels in cancers**. *Biomarker Research* 2024, **12**(1):70.

10. Liu H, Dong A, Rasteh AM, Wang P, Weng J: **Identification of the novel exhausted T cell CD8 + markers in breast cancer**. *Scientific Reports* 2024, **14**(1):19142.

11. Liu H, Tang T: MAPK signaling pathway-based glioma subtypes, machine-learning risk model, and key hub proteins identification. *Scientific Reports* 2023, 13(1):19055.

12. Liu H, Tang T: Pan-cancer genetic analysis of disulfidptosis-related gene set. *Cancer Genet* 2023, 278-279:91-103.

13. Liu H, Tang T: A bioinformatic study of IGFBPs in glioma regarding their diagnostic, prognostic, and therapeutic prediction value. *Am J Transl Res* 2023, 15(3):2140-2155.

14. Liu H, Tang T: Pan-cancer genetic analysis of disulfidptosis-related gene set. *bioRxiv* 2023:2023.2002.2025.529997.

15. Hengrui L: An example of toxic medicine used in Traditional Chinese Medicine for cancer treatment. *J Tradit Chin Med* 2023, 43(2):209-210.

16. Liu H, Weng J: A Pan-Cancer Bioinformatic Analysis of RAD51 Regarding the Values for Diagnosis, Prognosis, and Therapeutic Prediction. *Frontiers in oncology* 2022, 12.

17. Liu H, Weng J: A Comprehensive Bioinformatic Analysis of Cyclin-dependent Kinase 2 (CDK2) in Glioma. *Gene* 2022:146325.

18. Liu H, Tang T: Pan-cancer genetic analysis of cuproptosis and copper metabolism-related gene set. *Frontiers in oncology* 2022, 12:952290.

19. Liu H, Li Y: Potential roles of Cornichon Family AMPA Receptor Auxiliary Protein 4 (CNIH4) in head and neck squamous cell carcinoma. *Cancer biomarkers : section A of Disease markers* 2022.

20. Liu H, Dilger JP, Lin J: A pan-cancer-bioinformatic-based literature review of TRPM7 in cancers. *Pharmacology & Therapeutics* 2022:108302.

21. Liu H: **Pan-cancer profiles of the cuproptosis gene set**. *American journal of cancer research* 2022, **12**(8):4074-4081.

22. Li Y, Liu H: Clinical powers of Aminoacyl tRNA Synthetase Complex Interacting Multifunctional Protein 1 (AIMP1) for head-neck squamous cell carcinoma. *Cancer biomarkers : section A of Disease markers* 2022.

23. Li Y, Liu H, Han Y: Potential Roles of Cornichon Family AMPA Receptor Auxiliary Protein 4 (CNIH4) in Head and Neck Squamous Cell Carcinoma. *Research Square* 2021.

24. Liu H, Weng J, Huang CL, Jackson AP: Is the voltage-gated sodium channel β3 subunit (SCN3B) a biomarker for glioma? *Funct Integr Genomics* 2024, 24(5):162.

25. Agarwal K, Liu H: Potential Cancer Biomarkers: Mitotic Intra-S DNA Damage Checkpoint Genes. *bioRxiv* 2024:2024.2009. 2019.613851.

26. Arumilli S, Liu H: Protein Kinases in Phagocytosis: Promising Genetic Biomarkers for Cancer. *bioRxiv* 2024:2024.2010.2009.617495.

27. Chhatwal KS, Liu H: RAD50 is a potential biomarker for breast cancer diagnosis and prognosis. *bioRxiv* 2024:2024.2009. 2007.611821.

28. Dong A, Rasteh AM, Liu H: **Pan-Cancer Genetic Analysis of Mitochondrial DNA Repair Gene Set**. *bioRxiv* 2024:2024.2009. 2014.613048.

29. Liu H, Dong A, Rasteh AM, Wang P, Weng J: Identification of the novel exhausted T cell CD8 + markers in breast cancer. *Sci Rep* 2024, 14(1):19142.

30. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation**. *Cell* 2011, **144**(5):646-674.

31. Williams GH, Stoeber K: **The cell cycle and cancer**. *The Journal of pathology* 2012, **226**(2):352-364.

32. Black BE, Cleveland DW: Epigenetic centromere propagation and the nature of CENP-a nucleosomes. *Cell* 2011, 144(4):471-479.

33. Schuh M, Lehner CF, Heidmann S: Incorporation of Drosophila CID/CENP-A and CENP-C into centromeres during early embryonic anaphase. *Current biology : CB* 2007, 17(3):237-243.

34. Jansen LE, Black BE, Foltz DR, Cleveland DW: **Propagation of centromeric chromatin requires exit from mitosis**. *J Cell Biol* 2007, **176**(6):795-805.

35. Gonzalez M, He H, Sun S, Li C, Li F: **Cell cycle-dependent deposition of CENP-A requires the Dos1/2–Cdc20 complex**. *Proceedings of the National Academy of Sciences* 2013, **110**(2):606-611.

36. Saha AK, Contreras-Galindo R, Niknafs YS, Iyer M, Qin T, Padmanabhan K, Siddiqui J, Palande M, Wang C, Qian B *et al*: **The role of the histone H3 variant CENPA in prostate cancer**. *The Journal of biological chemistry* 2020, **295**(25):8537-8549.

37. Han J, Xie R, Yang Y, Chen D, Liu L, Wu J, Li S: CENPA is one of the potential key genes associated with the proliferation and prognosis of ovarian cancer based on integrated bioinformatics analysis and regulated by MYBL2. *Transl Cancer Res* 2021, 10(9):4076-4086.

38. Liang YC, Su Q, Liu YJ, Xiao H, Yin HZ: Centromere Protein A (CENPA) Regulates Metabolic Reprogramming in the Colon Cancer Cells by Transcriptionally Activating Karyopherin Subunit Alpha 2 (KPNA2). *The American journal of pathology* 2021, 191(12):2117-2132.

39. Wang Q, Xu J, Xiong Z, Xu T, Liu J, Liu Y, Chen J, Shi J, Shou Y, Yue C *et al*: CENPA promotes clear cell renal cell carcinoma progression and metastasis via Wnt/β-catenin signaling pathway. *J Transl Med* 2021, 19(1):417.

40. Zhang Y, Yang L, Shi J, Lu Y, Chen X, Yang Z: The Oncogenic Role of CENPA in Hepatocellular Carcinoma Development: Evidence from Bioinformatic Analysis. *Biomed Res Int* 2020, 2020:3040839.

41. Rajput AB, Hu N, Varma S, Chen CH, Ding K, Park PC, Chapman JA, Sengupta SK, Madarnas Y, Elliott BE *et al*: **Immunohistochemical Assessment of Expression of Centromere Protein-A (CENPA) in Human Invasive Breast Cancer**. *Cancers* 2011, **3**(4):4212-4227.

42. Zhang S, Xie Y, Tian T, Yang Q, Zhou Y, Qiu J, Xu L, Wen N, Lv Q, Du Z: High expression levels of centromere protein A plus upregulation of the phosphatidylinositol 3-kinase/Akt/mammalian target of rapamycin signaling pathway affect chemotherapy response and prognosis in patients with breast cancer. *Oncol Lett* 2021, 21(5):410.

43. Zhou H, Bian T, Qian L, Zhao C, Zhang W, Zheng M, Zhou H, Liu L, Sun H, Li X *et al*: **Prognostic model of lung adenocarcinoma constructed by the CENPA complex genes is closely related to immune infiltration**. *Pathology, research and practice* 2021, **228**:153680.

44. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E *et al*: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data**. *Cancer Discov* 2012, **2**(5):401-404.

45. Pan-cancer analysis of whole genomes. *Nature* 2020, 578(7793):82-93.

46. Mayakonda A, Koeffler HP: Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. *BioRxiv* 2016:052662.

47. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology* 2011, 12(4):R41.

48. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z: GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res* 2017, 45(W1):W98-w102.

49. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P *et al*: The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021, 49(D1):D605-d612.

50. Yu G, Wang LG, Han Y, He QY: **clusterProfiler: an R package for comparing biological themes among gene clusters**. *Omics : a journal of integrative biology* 2012, **16**(5):284-287.

51. Pontén F, Jirström K, Uhlen M: **The Human Protein Atlas—a tool for pathology**. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 2008, **216**(4):387-393.

52. Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN, Kamińska B, Huelsken J, Omberg L, Gevaert O *et al*: **Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation**. *Cell* 2018, **173**(2):338-354.e315.

53. Chan TA, Yarchoan M, Jaffee E, Swanton C, Quezada SA, Stenzinger A, Peters S: **Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic**. *Annals of oncology : official journal of the European Society for Medical Oncology* 2019, **30**(1):44-56.

54. Salipante SJ, Scroggins SM, Hampel HL, Turner EH, Pritchard CC: **Microsatellite instability detection by next generation sequencing**. *Clinical chemistry* 2014, **60**(9):1192-1199.

55. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA: **Profiling Tumor Infiltrating Immune Cells with CIBERSORT**. *Methods in molecular biology (Clifton, NJ)* 2018, **1711**:243-259.

56. Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, Xu L, Luo T, Yan H, Long Z *et al*: **CancerSEA: a cancer single-cell state atlas**. *Nucleic Acids Res* 2019, **47**(D1):D900-d908.

57. Bernstein MN, Ni Z, Collins M, Burkard ME, Kendziorski C, Stewart R: CHARTS: a web application for characterizing and comparing tumor subpopulations in publicly available single-cell RNA-seq data sets. *BMC bioinformatics* 2021, 22(1):83.

58. Sun D, Wang J, Han Y, Dong X, Ge J, Zheng R, Shi X, Wang B, Li Z, Ren P *et al*: TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res* 2021, 49(D1):D1420-d1430.

59. Paulson KG, Voillet V, McAfee MS, Hunter DS, Wagener FD, Perdicchio M, Valente WJ, Koelle SJ, Church CD, Vandeven N *et al*: **Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA**. *Nat Commun* 2018, **9**(1):3868.

60. Di Genua C, Valletta S, Buono M, Stoilova B, Sweeney C, Rodriguez-Meira A, Grover A, Drissen R, Meng Y, Beveridge R *et al*: C/EBPα and GATA-2 Mutations Induce Bilineage Acute Erythroid Leukemia through Transformation of a Neomorphic Neutrophil-Erythroid Progenitor. *Cancer Cell* 2020, 37(5):690-704.e698.

61. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, Richman AR, Silverbush D, Shaw ML, Hebert CM *et al*: **An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma**. *Cell* 2019, **178**(4):835-849.e821.

62. Yost KE, Satpathy AT, Wells DK, Qi Y, Wang C, Kageyama R, McNamara KL, Granja JM, Sarin KY, Brown RA *et al*: **Clonal replacement of tumor-specific T cells following PD-1 blockade**. *Nature medicine* 2019, **25**(8):1251-1259.

63. Tarashansky AJ, Xue Y, Li P, Quake SR, Wang B: **Self-assembling manifolds in single-cell RNA sequencing data**. *eLife* 2019, **8**.

64. Fu J, Li K, Zhang W, Wan C, Zhang J, Jiang P, Liu XS: **Large-scale public data reuse to model immunotherapy response and resistance**. *Genome Med* 2020, **12**(1):21.

65. Liu CJ, Hu FF, Xia MX, Han L, Zhang Q, Guo AY: **GSCALite: a web server for gene set cancer analysis**. *Bioinformatics* 2018, **34**(21):3771-3772.

66. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, Bindal N, Beare D, Smith JA, Thompson IR *et al*: **Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells**. *Nucleic Acids Res* 2013, **41**(Database issue):D955-961.

67. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, Javaid S, Coletti ME, Jones VL, Bodycombe NE *et al*: **Correlating chemical sensitivity and basal gene expression reveals mechanism of action**. *Nat Chem Biol* 2016, **12**(2):109-116.

68. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A *et al*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**(7873):583-589.

69. Trott O, Olson AJ: AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* 2010, 31(2):455-461.

70. Mandrekar JN: **Receiver operating characteristic curve in diagnostic test assessment**. Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer 2010, **5**(9):1315-1316.

71. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW, Kleihues P: **The 2007 WHO classification of tumours of the central nervous system**. *Acta neuropathologica* 2007, **114**(2):97-109.

72. Claus EB, Walsh KM, Wiencke JK, Molinaro AM, Wiemels JL, Schildkraut JM, Bondy ML, Berger M, Jenkins R, Wrensch M: **Survival and low-grade glioma: the emergence of genetic information**. *Neurosurg Focus* 2015, **38**(1):E6-E6.

73. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P: **The Molecular Signatures Database (MSigDB) hallmark gene set collection**. *Cell systems* 2015, **1**(6):417-425.

74. Aristizabal-Corrales D, Yang J, Li F: Cell Cycle-Regulated Transcription of CENP-A by the MBF Complex Ensures Optimal Level of CENP-A for Centromere Formation. *Genetics* 2019, 211(3):861-875.

75. Valdivia MM, Hamdouch K, Ortiz M, Astola A: **CENPA a genomic marker for centromere activity and human diseases**. *Current genomics* 2009, **10**(5):326-335.

76. Kixmoeller K, Allu PK, Black BE: The centromere comes into focus: from CENP-A nucleosomes to kinetochore connections with the spindle. *Open biology* 2020, 10(6):200051.

77. Schnerch D, Yalcintepe J, Schmidts A, Becker H, Follo M, Engelhardt M, Wäsch R: **Cell cycle control in acute myeloid leukemia**. *American journal of cancer research* 2012, **2**(5):508-528.

78. Amato A, Schillaci T, Lentini L, Di Leonardo A: **CENPA overexpression promotes genome instability in pRb-depleted human cells**. *Molecular cancer* 2009, **8**:119.

79. Chen X, Pan Y, Yan M, Bao G, Sun X: Identification of potential crucial genes and molecular mechanisms in glioblastoma multiforme by bioinformatics analysis. *Mol Med Rep* 2020, 22(2):859-869.

80. Wang D, Liu J, Liu S, Li W: Identification of Crucial Genes Associated With Immune Cell Infiltration in Hepatocellular Carcinoma by Weighted Gene Co-expression Network Analysis. *Frontiers in genetics* 2020, 11:342.