Preprints.org

CENPA Was a Cell Cycle Biomarker in Cancers With Diagnostic Prognostic, and Therapeutic Value

Hengrui Liu and Tao Tang [*]

Article

CENPA Was a Cell Cycle Biomarker in Cancers with Diagnostic Prognostic, and Therapeutic Value

Hengrui Liu [2] and Tao Tang [1]

[1] Department of Molecular Diagnostics, Sun Yat-Sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangzhou, Guangdong, China

[2] Yinuo Biomedical Co., Ltd., Tianjin, China

* Correspondence: tangtao@sysucc.org.cn

Abstract: Background Cancer is one of the most concerning public health issues in the world. One of cancer hallmarks wildly accepted is sustaining proliferative signaling, which involved most of the cell cycle biological activities. Centromeric histone, Centromere Protein A (CENPA), a variant of canonical histone H3, plays an essential role in selective chromosome segregation in the cell cycle. However, so far, a systematic pan-cancer bioinformatic analysis has not been done yet. **Methods** We accessed genome, transcriptome, and clinical information from open databases. The genetic alteration, mRNA expression, functional enrichment, stemness, mutation association, expression in cell populations and cellular locations, cell cycle association, survival association of CENPA, and immune association were analyzed. A prognostic model for glioma patients was constructed as an example application of CENPA as a biomarker. Drugs targeting CENPA in cancer cells were also screened and predicted by the CENPA correlation of drug sensitivity and protein-ligand docking. **Results** CENPA had low gene mutation in cancers. CENPA was overexpressed in almost all cancer types in TCGA compared to their normal control. CENPA was highly expressed in the nucleus of malignant cells. CENPA was associated with the cell cycle of cancer cells. CENPA is a biomarker for the cell cycle G2 phase in cancer cells. CENPA was a diagnostic and prognostic biomarker across multiple cancer types. The prognosis of glioma with CENPA was reliable and can be applied with other prognostic factors. CENPA was associated with the immune microenvironment. Drugs CD-437, 3-Cl-AHPC, Trametinib, BI-2536, and GSK461364 were predicted to target CENPA in cancer cells. **Conclusion** CENPA was a cell cycle biomarker in cancers with diagnostic and prognostic value.

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

1. Introduction

Cancer is one of the most concerning public health issues in the world. In the past three years, the COVID-19 restrictions impeded and deleted the diagnosis of many cancer patients, accounting for a short-term reduction in new cancer cases discovered, yet, a potential rise in advanced-stage cancer cases and the increase in cancer-related death are expected to occur in the next few years [1]. Thus, the issue of cancer study is still urgent and required more effort. 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) [2], Genotype-Tissue Expression (GTEx) [3], the Chinese Glioma Genome Atlas (CGGA) [4], and the International Cancer Genome Consortium (ICGC) [5], provide gene alteration, gene expression, and clinical information of different cancer types, facilitating pan-cancer studies for identification and understanding of targets or biomarkers that exert common effects across cancer types.

Six hallmarks of cancer have been proposed to constitute an organizing principle that provides a logical framework for understanding the remarkable diversity of cancers [6]. One of cancer

hallmarks wildly accepted is sustaining proliferative signaling [6], which involved most of the cell cycle biological activities [7]. 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 at the S phase, whereas newly synthesized CENPA protein is deposited during the G2/M phase of the cell cycle [8–10]. A study reported that cell cycle-dependent deposition of CENPA was mediated by the Dos1/2–Cdc20 complex [11]. 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 [12]. 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 [13]. 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 [14]. In clear cell renal cell carcinoma, the function of CENPA was reported to promote metastasis of cancer via the Wnt/β-catenin signaling pathway [15]. In addition, studies also suggested the prognostic value of CENPA for a few cancer types, such as ovarian cancer [13], liver cancer [16], breast cancer [17,18], and lung cancer [19]. 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.

2. Methods

2.1. The acquisition of mRNA sequencing data

The mRNA data with clinical information were downloaded from The Cancer Genome Atlas (TCGA) [2], Genotype-Tissue Expression (GTEx) [3], the Chinese Glioma Genome Atlas (CGGA) [4], and the International Cancer Genome Consortium (ICGC) [5] in March 2022, in which the method of acquisition and application complied with the guidelines and policies.

2.2. Gene alteration analysis

Mutation analyses were conducted using the cBioPortal [20] using "pan-cancer analysis of whole genomes (ICGC/TCGA, Nature 2020)" [21]. The mutation or variant data were obtained from the TCGA PanCancer Atlas Studies and the UniProt. Single-nucleotide variant (SNV) and copy number variant (CNV) data were from NCI Genomic Data Commons (TCGA). SNV plots were generated by the maftools [22]. CNV data were processed with GISTICS2.0 [23].

2.3. RNA-seq data analysis and plotting

All the analyses and plotting, including receiver operating characteristic curve (ROC) plot, survival Kaplan-Meier (KM) plot, nomogram construction, etc., were implemented by R foundation for statistical computing (2020) version 4.0.3 and ggplot2 (v3.3.2).

2.4. Associated genes enrichment analysis

The top correlated genes were identified using the GEPIA [24]. The protein-protein interaction network was constructed using the STRING [25]. The minimum required interaction score was set at high confidence (>0.9). All the enrichment analyses were conducted using clusterprofiler [26].

2.5. Immunohistochemistry staining

Representative images of the immunohistochemistry staining of cancer and non-cancer tissues were accessed from Human Protein Atlas (HPA) [27]. Antibody CAB008371 was used to stain.

2.6. Immunofluorescence staining of cancer cells

Representative images of the immunofluorescence staining of the subcellular distribution of protein within the nucleus, endoplasmic reticulum (ER), and microtubules of three cancer cell lines were obtained from the HPA database.

2.7. The cell cycle association analysis

Plots of single-cell RNA-sequencing data from the FUCCI U-2 osteosarcoma cell line were accessed and analyzed using the HPA. The temporal mRNA expression patterns were characterized in individual cells using the Fluorescent Ubiquitination-based Cell Cycle Indicator (FUCCI) U-2 OS cell line.

2.8. Stemness association analysis

The OCLR algorithm [28] was used to calculate the mRNAsi for the evaluation of stemness.

2.9. Mutation association analysis

The tumor mutational burden (TMB) [29] and microsatellite instability (MSI) [30] were used to evaluate the mutation levels of samples.

2.10. Immune cell infiltration analysis

The immune cell infiltration level was calculated using the TCGA cohort. The CIBERSORT algorithms [31] were used to estimate the immune cell infiltration levels [32].

2.11. Single-cell sequencing data acquisition and analysis

The single-cell data were accessed and analyzed using the TISCH [33], the CHARTS [34], and the CancerSEA [35]. Data sets included GSE70630 [36], GSE123814 [37], GSE142213 [38], GSE143423, GSE131928 [39], GSE117988 [40], etc.

2.12. Immune therapy prediction analysis

The immune therapy prediction analysis was conducted using the TIDE [41]. 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 tested biomarker on the ICB response status.

2.13. Drug screening and prediction

Drugs were screened based on their CENPA correlation of drug sensitivity with a cut-off of remarkable significance (p<1e-5). The GSCALite [42] was used to evaluate the area under the dose-response curve (AUC) values for drugs and gene expression profiles of CENPA in different cancer cell lines. Drug sensitivity and gene expression profiling data of cancer cell lines in the GDSC [43] and the CTRP [44] are integrated for investigation. The expression of each gene in the gene set was performed by Spearman correlation analysis with the small molecule/drug sensitivity. The predictive protein structural model of CENPA from the Alphafold database [45] and the protein-ligand cavity-detection guided blind docking was conducted using the "AutoDock Vina" (1.1.2) [46].

2.14. Statistical analysis

Wilcox test or Kruskal-Wallis test was applied to compare gene expression differences. Kaplan-Meier analysis, log-rank test, and Cox regression test were used to conduct survival analysis. Pearson's correlation test was conducted to evaluate the correlation between two variables. A P<0.05 was considered to be statistically significant.

3. Results

3.1. Genomic alteration of CENPA in cancers

The first analysis of this study was to investigate CENPA genomic alteration in cancers. The alteration frequency bar plot showed that the total alteration frequencies of most of the cancer types were lower than 10%. Non-small cell lung cancer had the highest frequency, which was 15.2% (7 in 46 cases). Most of the gene alterations were amplification (S-Figure 1A). To further investigate the mutation of CENPA in cancers, the TCGA mutation data was plotted and data suggested that CENPA had only a low number of single-nucleotide variants in cancers (S-Figure 1B), which is consistent with the results above. The copy number variation analysis showed that almost all the copy number alterations of CENPA were heterozygous. Most of the cancer types had 20-40% CENPA heterozygous amplification samples and about half of the cancer types had 5-10% CENPA heterozygous deletion samples. LUSC had the highest percentage of CENPA heterozygous amplification samples with no CENPA heterozygous deletion sample, which was consistent with the above results that lung cancer had a high frequency of gene amplification. KICK had about 60% of CENPA heterozygous deletion samples with no CENPA heterozygous amplification sample (S-Figure 1C). Overall pan-cancer data also showed that the copy number of CENPA could affect the mRNA expression of CENPA (S-Figure 1D). Thus, gene mutation of CENPA might not be the major reason that drives most cancer types, but the copy number alteration of CENPA might potentially affect cancers through mRNA expression of CENPA.

3.2. The overexpression of CENPA in cancers

The analysis revealed that CENPA was overexpressed in most of the cancer types compared to normal tissues in both females and males (Figure 1A). To facilitate the display of the data, we used abbreviations to represent cancer types, which were listed in S-Table 1. The mRNA expression of CENPA from TCGA and GTEx data showed that CENPA was significantly overexpressed in 30 types of cancer among the 33 cancer types analyzed. MESO and UVM have no comparable normal tissue, while LAML is the only cancer type that expressed lower CENPA in cancer than in normal tissues (Figure 1B). To further compare cancer noncancer at a better control, paired cancer noncancer samples from the same patients of available cancer types were also compared. Results showed that 16 cancer types were found to significantly overexpress CENPA (Figure 1C). To further observe the overexpression of CENPA in cancers, we compared the protein staining of CENPA in cancers and corresponding normal tissues in representative cancer types. Generally, the staining images showed that although cancer had slightly stronger CENPA staining intensity, the staining intensity in both cancers and normal tissues was low, which might be subjected to the antibody properties (Figure 1D).
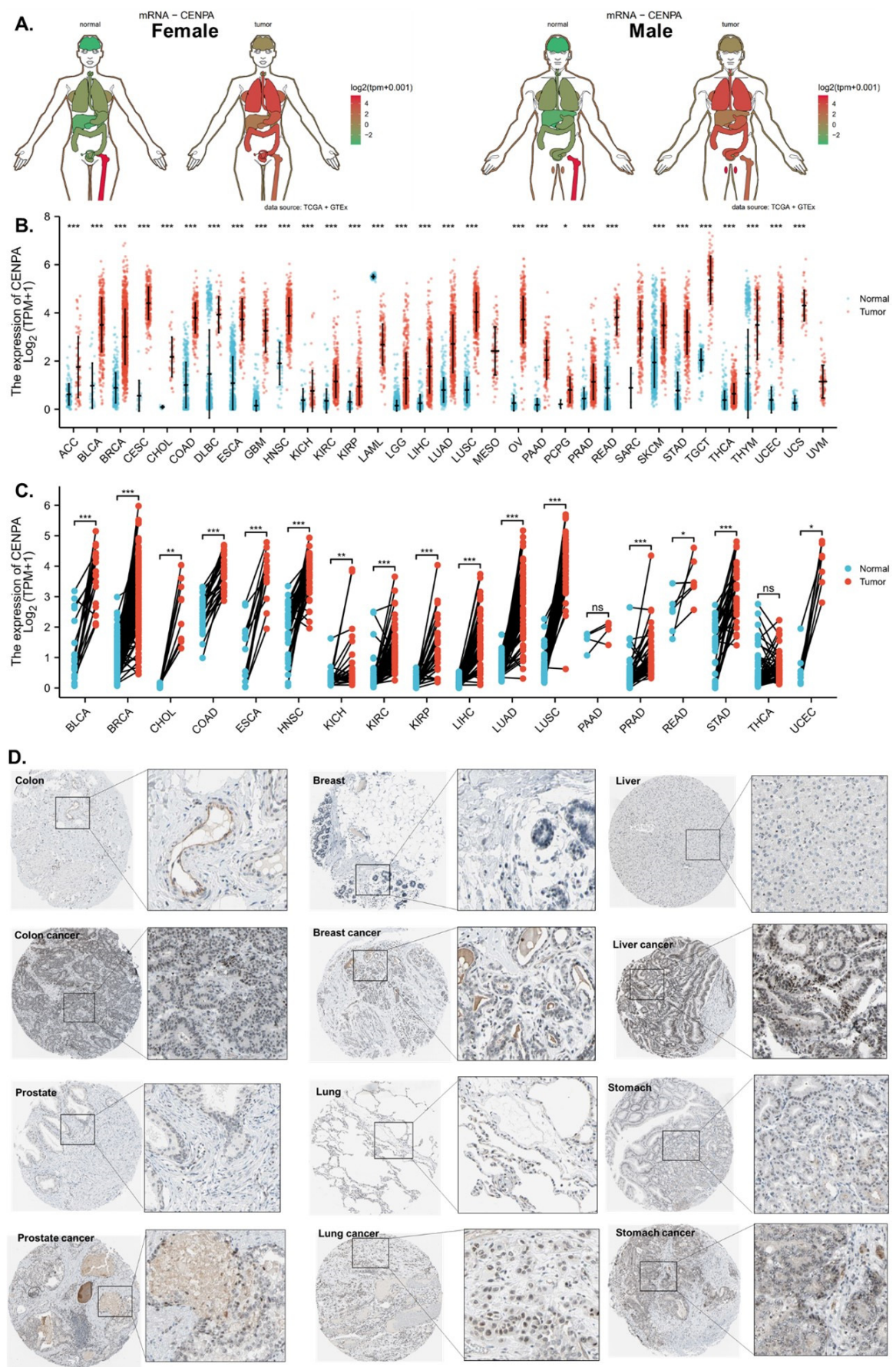
Figure 1. The overexpression of CENPA in cancers. (**A**). Anatomy plot of the gene expression profile of CENPA across all tumor samples and paired normal tissues in females and males. TCGA data were plotted. (**B**). The gene expression profile of CENPA across all tumor samples and normal tissues.

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). * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

3.3. CENPA was highly expressed in the nucleus of malignant cells

To investigate the cell populations and cellular locations where CENPA was expressed, we analyzed single-cell sequencing data and observed immunofluorescence staining of the subcellular distribution of CENPA within three cancer cell lines. We analyzed a single-cell sequencing data set of three cancer types including acute erythroid leukemia (AEL), breast cancer (BRCE), glioma, and Merkel cell carcinoma (MCC). Results showed that CENPA was expressed by a small population of malignant cells while the immune cells expressed relatively low levels of CENPA (Figure 2A). The immunofluorescence staining of the subcellular distribution of CENPA revealed that in prostate cancer cell PC-3, Rhabdomyosarcoma cell RH30, and osteosarcoma cell U2OS, CENPA was expressed mainly in the nucleus, although U2OS had a relatively lower fluorescence intensity (Figure 2B). To be mentioned, rhabdomyosarcoma is a type of sarcoma. Prostate cancer (PRAD) and sarcoma (SARC) have been demonstrated above in this study to be overexpressing CENPA, while osteosarcoma was not included in the cancer types of TCGA data.
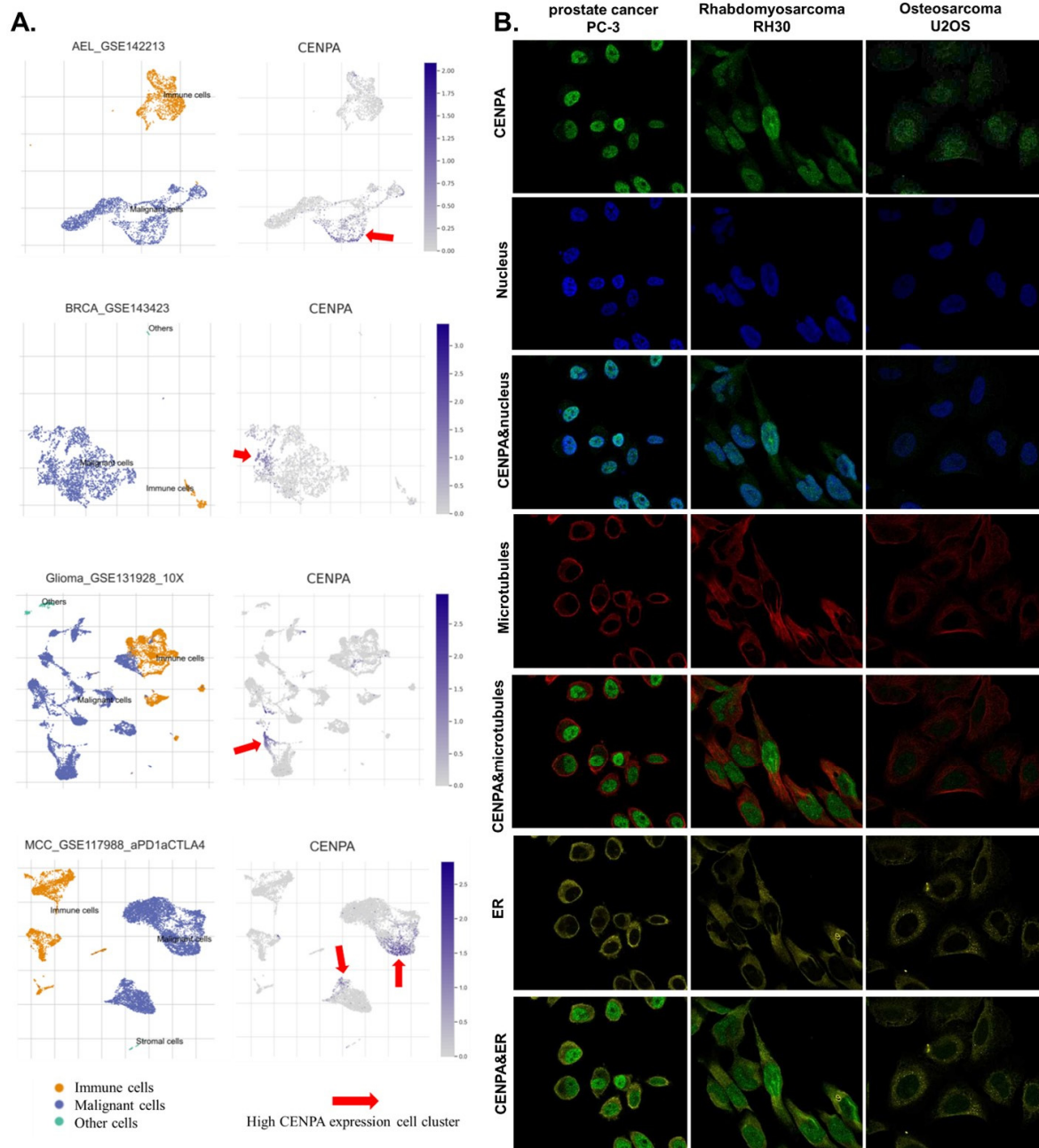
Figure 2. The expression of CENPA in cell populations and cellular locations. (**A**). The expression of CENPA in cell populations in cancer tissues. Single-cell mRNA expression cohorts were accessed and analyzed using the TISCH. (**B**). Immunofluorescence staining of the subcellular distribution of CENPA within the nucleus, endoplasmic reticulum (ER), and microtubules of three cancer cell lines.

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

Since CENPA had been demonstrated to be mainly detected in the nucleus of cancer cells, we proposed two possible roles of CENPAN in cancers: 1) CENPA might impact the mutation of other genes, as the transcription for gene expression takes place in the nucleus, and 2) CENPA might regulate the cell cycle because the duplication of DNA during the cell cycle takes place in the nucleus. Therefore, for the first hypothesis, we analyzed the correlation between CENPA expression and two mutation indicators: tumor mutation burden (TMB) and microsatellite instability (MSI). TMB is the approximate amount of gene mutation that occurs in the cancer genome, while MSI is the condition of genetic hypermutability (predisposition to mutation) that results from impaired DNA mismatch repair (MMR).

The presence of MSI represents phenotypic evidence that MMR is not functioning normally. The analysis revealed that CENPA expression was positively correlated with the TMB and MSI of most of the cancer types with different coefficients, but with weak significances (S-Figure 2A,B). These results indicated that CENPA was not commonly associated with genome instability in cancers.

On the other hand, to explore the potential common functional effects of CENPA in cancers, we identified the top CENPA-correlated genes using data from all 33 TCGA cancer types as a single cohort and conducted the further analysis. The top 30 CENPA-correlated genes were used to construct a protein-protein interaction (PPI) network to display the potential association of CENPA and these genes (S-Figure 2C). The top 200 correlated genes were further analyzed in the GO and KEGG enrichment study. The KEGG pathways enrichment showed that the top two human relative pathways associated with CENPA were "DNA replication" and "Cell cycle". The top GO molecular function (MF) ATPase activities. The top GO cellular components (CC) were chromosome regions. The top GO biological process (BP) were "organelle fusion", "mitotic nucleus division", and "nucleus division". The GO-enriched terms were all associated with cancer proliferation and cell cycle (S-Figure 2D).

To further validate the potential association of CENPA and cancer proliferation and cell cycle, we analyzed the correlation of CENPA expression and cancer functional signals using multiple single-cell data sets with different cancer types. These correlation results were summed up (as shown in the top bar plot of S-Figure 2E) to overview the potential common roles of CENPA in cancers. Results showed that the top two most striking positive correlations were "cell cycle" and "proliferation", which was consistent with the hypothesis that CENPA might regulate the cell cycle and proliferation. In addition, CENPA might negatively associate with "apoptosis", "DNA repair", and "metastasis" (S-Figure 2E). These data supported that CENPA might regulate cancer growth.

3.5. 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 [28] has been wildly applied for the estimation of the stemness in a tissue sample. This study calculated the mRNAsi (a score of stemness) for the 33 cancer types in TCGA and analyzed the correlation between CENPA and pan-cancer stemness. The calculation indicated that CENPA expression was positively correlated with stemness in most of the cancer types (Figure 3A), suggesting that the stemness association might be a common mechanism of CENPA in cancer. Based on the above analyses, we proposed that CENPA might be a novel cell cycle biomarker, we conducted the GSEA enrichment of CENPA-correlated genes in "REACTOME CELL CYCLE CHECKPOINTS". Not surprisingly, results showed that CENPA-correlated genes were significantly enriched in "REACTOME CELL CYCLE CHECKPOINTS" (Figure 3B).

To gain further insight into the exact role of CENPA in the cell cycle's different phases in cancer cells, we accessed the single-cell expression data of CENPA in different cell cycle phases of U2OS cells which had been demonstrated to express CENPA mainly in their nucleus. Results revealed that CENPA was lowly expressed in the G1 phase and highly expressed in the S&G2 phase (Figure 3C). Therefore, we hypothesized that CENPA might closely associate with the G2 phase of the cell cycle. To validate this hypothesis, we observed the CENPA expression in several single-cell cancer data sets and also compared the single-cell signals of the G2/M checkpoint (a hallmark cell proliferation-related gene set in GSEA) [47]. 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 3D).
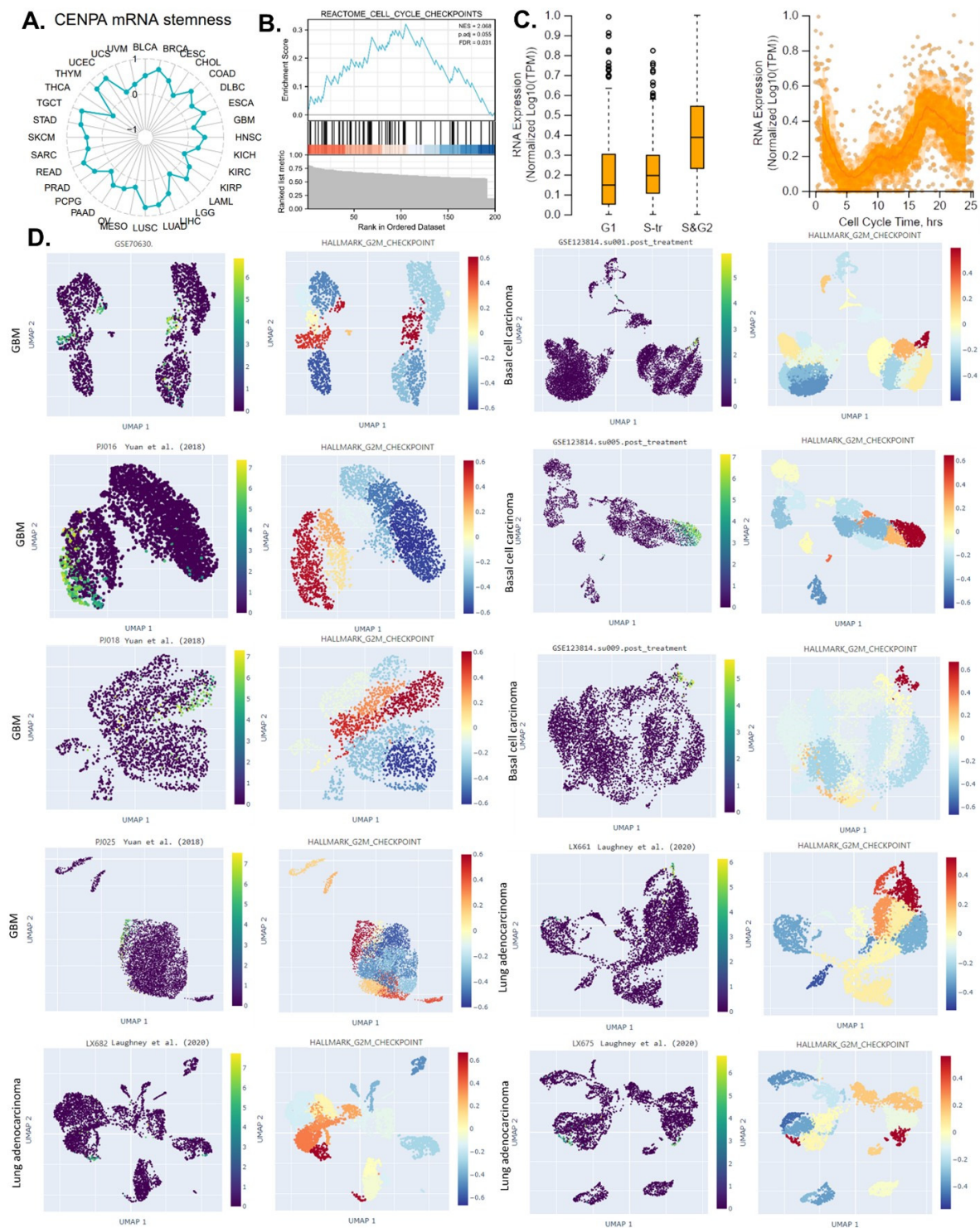
Figure 3. The potential of CENPA as a cell-cycle biomarker for the M/G2 phase in cancers. (**A**). The correlation of OCLR scores and CENPA in TCGA cancer data. The OCLR algorithm was used to calculate the mRNAsi (OCLR scores) for the evaluation of stemness. (**B**). The GSEA enrichment of CENPA-correlated 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.

3.6. The diagnostic value of CENPA in cancers

To evaluate the diagnostic value of CENPA in cancers, single-variable diagnostic receiver operating characteristic (ROC) curves of different cancer types were plotted and the area under the curves (AUC) was calculated using TCGA and GTEx data. Results showed that the AUCs of 19 cancer types were over 0.9, which indicates an outstanding diagnostic power of CENPA. The AUCs of 5 cancer types were between 0.8-0.9, which supported an excellent diagnostic power of CENPA. The AUCs of 3 cancer types were between 0.7-0.8, including an acceptable diagnostic power of CENPA [48] (Figure 4). These results suggested that CENPA is a promising diagnostic molecular biomarker that can be developed for multiple cancer types.
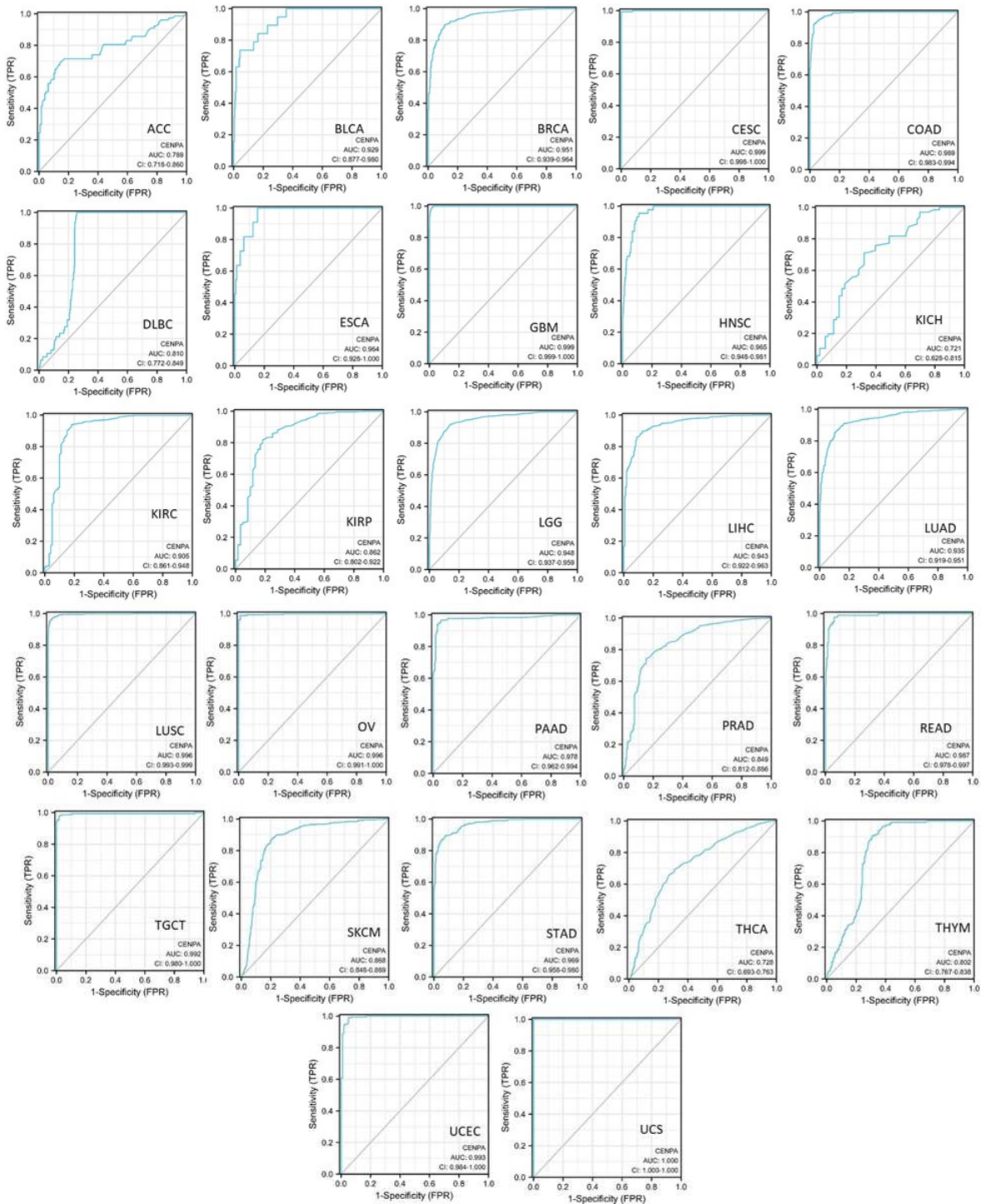
Figure 4. The pan-cancer diagnostic value of CENPA. The diagnostic receiver operating characteristic (ROC) curve of different cancer types. TCGA and GTEx data were used to calculate the ROC. The area under the curves (AUC) and the corresponding 95% confidential interval (95%CI) was shown.

3.7. The prognostic value of CENPA in cancers

This study was also interested in the prognostic value of CENPA in cancers. Thus, univariate overall survival Cox regression analysis of CENPA was conducted across 33 cancer types using TCGA data. Results showed that CENPA was significantly associated with worse overall survivals in 13 cancer types, while it was also associated with better overall survivals in one cancer type, THYM (Figure 5A). KM plot and log-rank analysis of the significant cancer types in Cox regression were also used to further observe the association of CENPA and the overall survival of patients. Results showed that 12 cancer types remained significant in the log-rank analysis (Figure 5B panel one for each cancer type).

To evaluate the prognostic value of CENPA in these cancer types, time-dependent prognostic ROCs were plotted. Results showed that, for year 1 overall survival, the AUC of KICH was over 0.9, which indicated an outstanding prediction. The AUCs of ACC, KIRP, and PCPG were between 0.8 to 0.9, which indicated excellent predictions. The AUCs of LGG, LIHC, and MESO, were between 0.7 to 0.8, which indicated acceptable predictions. For year 3 overall survival, the AUC of ACC was over 0.9, which indicated outstanding predictions. The AUCs of KICH, MESO, and PCPG were between 0.8 to 0.9, which indicated excellent predictions. The AUCs of KIRP, LGG, and PAAD were between 0.7 to 0.8, which indicated acceptable predictions. For year 5 overall survival, the AUC of ACC, KICH, MESO, and PCPG were between 0.8 to 0.9, which indicated excellent predictions. The AUCs of KIRP, LGG, and PAAD were between 0.7 to 0.8, which indicated acceptable predictions (Figure 5B panel two for each cancer type). These results suggested that CENPA is a promising prognostic molecular biomarker that can potentially be developed for multiple cancer types, such as ACC, KICH, KIRP, LGG, MESO, and PCPG.
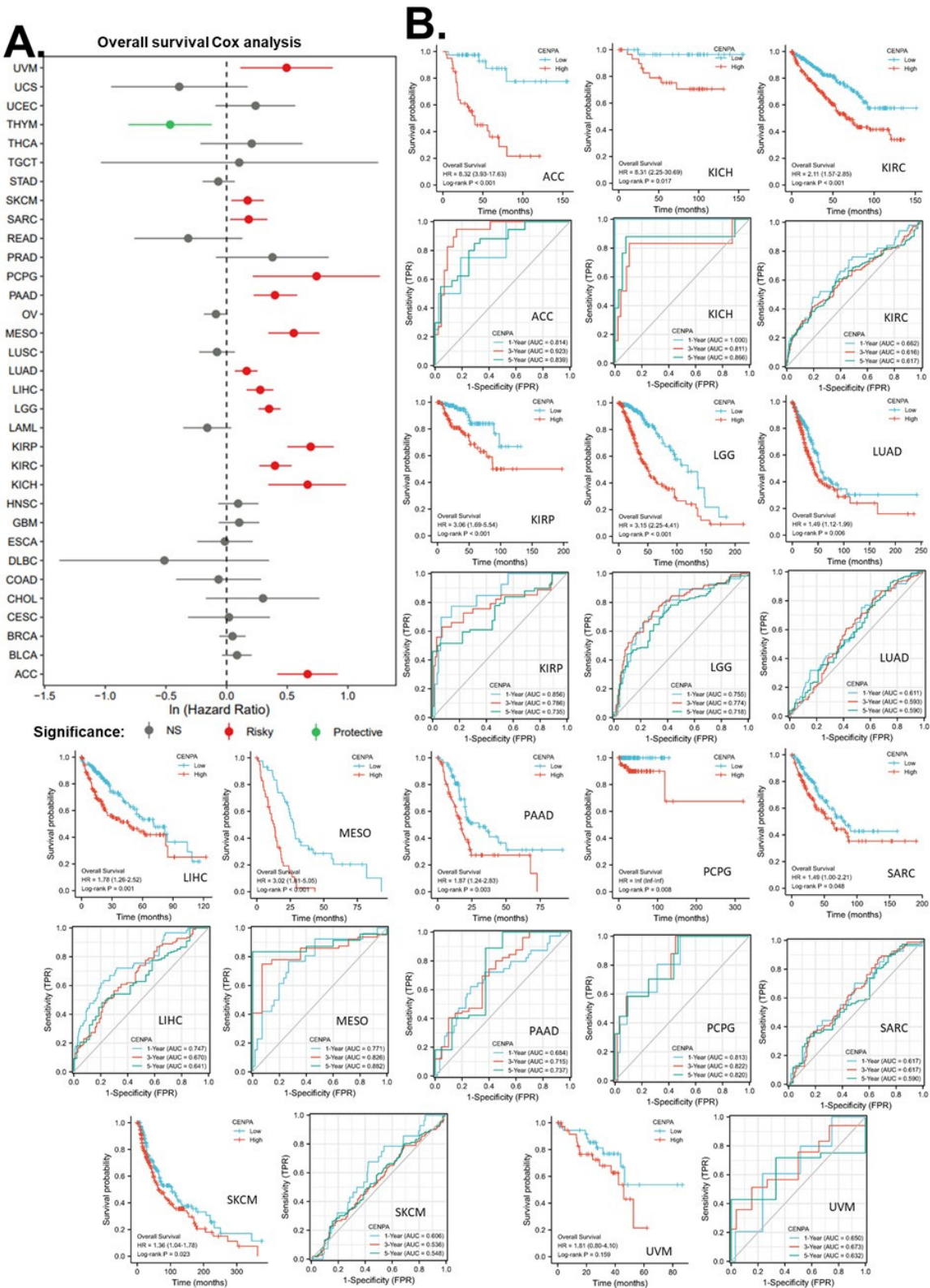
Figure 5. The pan-cancer prognostic value of CENPA. TCGA data were analyzed. (**A**). Univariate Cox regression analysis of CENPA for overall survival in different cancer types. (**B**). The overall survival Kaplan-Meier (KM) plot and log-rank analysis of high (50-100%) and low (0-50%) CENPA patients 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.

3.8. 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 [49]. The G1 glioma is generally benign and has a very good prognosis, which has been distinguished from 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) [50]. In this context, this study combined LGG and GBM and analyzed the prognostic value of CENPA for overall glioma.

To confirm the prognostic accuracy of CENPA for the overall survival of glioma patients, we analyzed the prognostic association in five independent data sets of glioma, including 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). The KM plot and Cox analysis showed that, among all five independent data sets, the high expression of CENPA was significantly associated with worse survival. The hazard ratios (HR) calculated with the five data sets varied from 2.95 to 7.21. The ROC calculations showed that, for year 1 overall survival prediction, 4 data sets suggested an acceptable accuracy. For year 3 overall survival prediction, 3 data sets suggested an excellent accuracy and 2 data sets suggested an acceptable accuracy. For year 5 survival prediction, 3 data sets suggested an excellent accuracy and 2 data sets suggested an acceptable accuracy (Figure 6A).

In this study, we developed strategies for the application of CENPA in glioma prognosis as an example of the future development of CENPA as a clinical prognostic biomarker of cancer. To screen variables for the prognostic model of CENPA for glioma patients, we conducted a Cox regression analysis to evaluate the prognostic factors. The univariate Cox regression results showed that the CENPA level, the 1p/19q codeletion, primary therapy outcome, IDH status, and age were significantly associated with the overall survival of glioma patients. The multivariate Cox regression results showed that the CENPA level, primary therapy outcome, IDH status, and age remained significant after being adjusted for other variables, indicating that they could provide additional prognostic power as independent factors in the prognostic model (S-Table 2). Therefore, these factors as well as the WHO grade (G2-4) were included to calculate the prognostic model for the overall survival of glioma patients. Based on the model, a nomogram was constructed for the prediction of the survival probability of glioma patients at years 1, 3, and 5 (Figure 6B). The prediction results of the nomogram calibration curves of the nomogram were generally consistent with patients' observation results (Figure 6C).
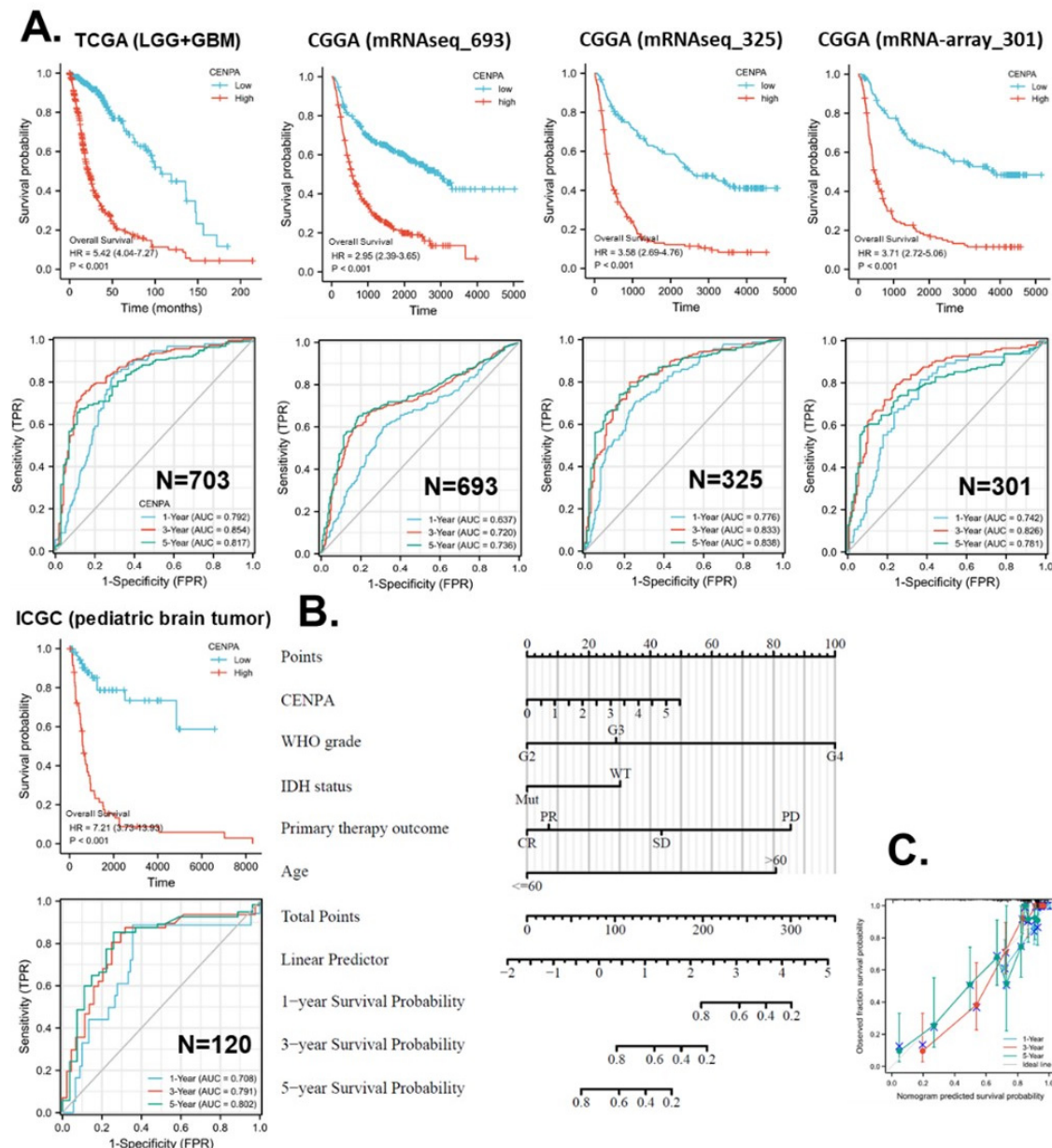
Figure 6. Application of CENPA for glioma prognosis. (**A**). validation of the survival association of CENPA in five independent glioma cohorts. TCGA (LGG+GBM), CGGA (mRNAseq 693), CGGA (mRNAseq 325), CGGA (mRNA-array 301), and ICGC (pediatric brain tumor) were analyzed. The overall survival Kaplan-Meier (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.

3.9. The immune microenvironment association of CENPA in cancers

Furthermore, this study also explored CENPA as an immune microenvironment biomarker. For cancer immune therapy, the response of the therapy largely depends on immune cell infiltration levels and the levels of immune checkpoints, thus we analyzed the potential value of CENPA as an immune therapy prediction biomarker from these two aspects.

In the above analysis, we have demonstrated that CENPA was mainly expressed in a small population of malignant cells while the immune cells expressed relatively lower CENPA, but

whether CENPA expression of tumor affected the immune cells had not been studied. Hence, we calculated the immune cell infiltration levels in cancers and analyzed their correlation with the CENPA expression level. The most striking immune cell type that correlated with CENPA was T cell CD4+, of which Th2 was positively correlated with all cancer types and Th1 was positively correlated with the majority of cancer types. In addition, common lymphoid progenitor was also positively correlated with most of the cancer types. CENPA was closely associated with multiple immune cells across different cancer types, especially in some cancer types, such as lung cancer (LUSC and LUAD), GBM, and THYM (Figure 7A).

On the other hand, we also calculated the correlation between CENPA and multiple commonly used immune checkpoints in current immune therapies. Results showed that CENPA was positively associated with most of the immune checkpoints in THCA, LUAD, LIHC, LGG, KIRC, BRCA, and BLCA, while negatively correlated with most of the immune checkpoints in THYM, LUSC, GBM, CESC, and ACC (Figure 7B).

To compare the prediction performance of CENPA for ICB (immune checkpoint blockade treatment) with other commonly used standardized biomarkers, we evaluated the biomarker relevance of CENPA and other standardized biomarkers based on their predictive power of response outcomes of ICB sub-cohorts. Results showed that CENPA expression had an AUC of over 0.5 in 11 of the 25 ICB subcohorts, which had more cohorts with an AUC of over 0.5 than MSI.score, TMB, T.Clonality, and B. Clonality., which had AUC values of over 0.5 in only seven, nine, and six ICB sub-cohorts, but the predictive value of CENPA was lower than CD27A, TIDE, IFNG, CD8, and Merck 18 (Figure 7C). These comparisons generally supported the potential value of CENPA for the prediction of immune therapy.
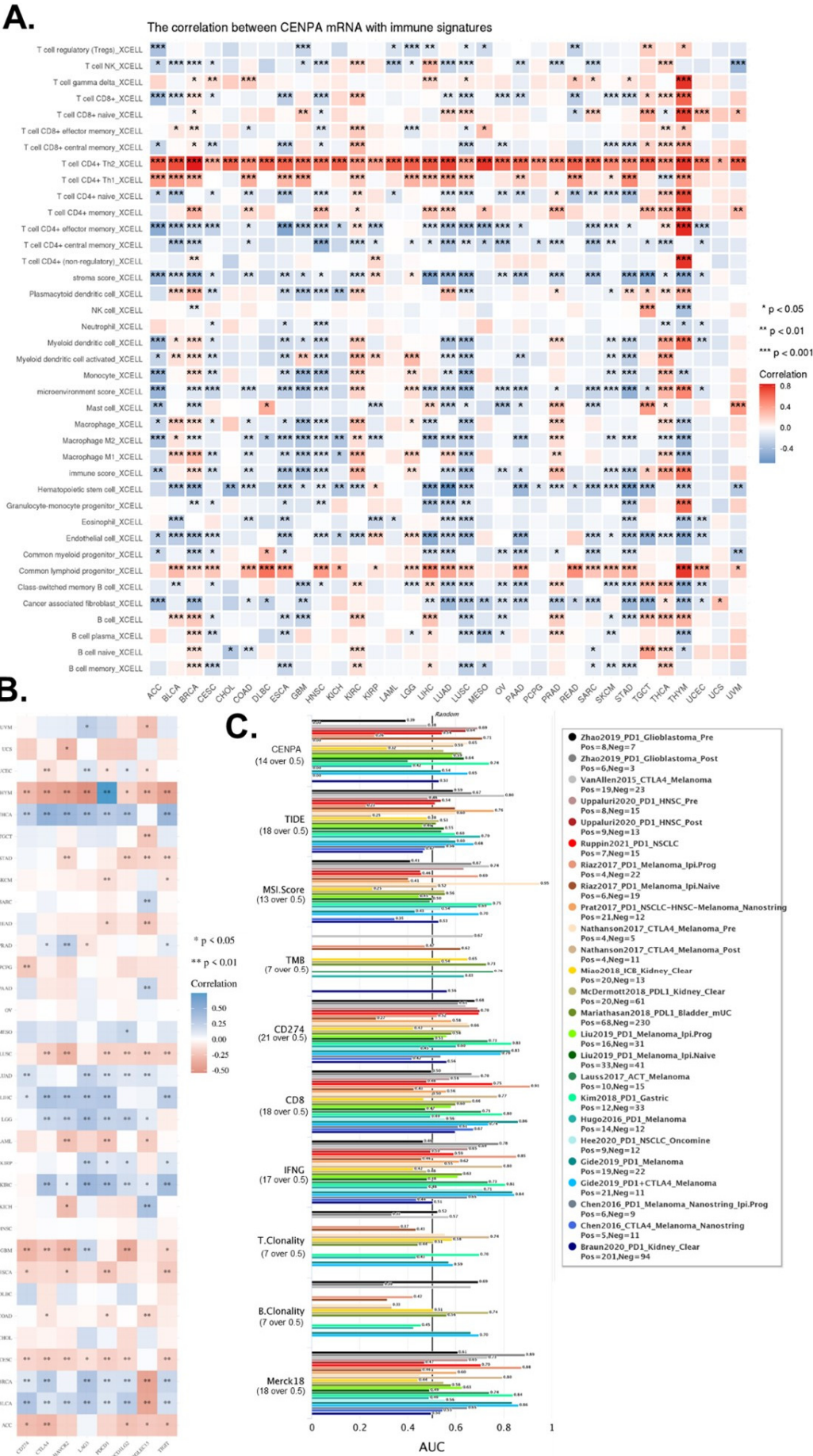
Figure 7. The immune microenvironment association of CENPA in cancers. (**A**). The correlation of CENPA expression and immune cell infiltration levels. TCGA data were analyzed. The Xcell algorithms were used to estimate the immune cell infiltration levels. (**B**). The correlation of CENPA expression 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.

3.10. Computational drug predictions of CENPA in cancers

As our study had demonstrated that CENPA was closely associated with the cancer cell cycle, cancer patient survival, and cancer immune microenvironment, we proposed CENPA as a potential cancer therapeutic target for drug treatment. Therefore, we screened and predicted potential drugs that target CENPA by cancer drug databases and computational methodology. We accessed the drug sensitivity data from GDSC and CTRP databases and analyzed the correlation of CENPA expression and small molecule/drug sensitivity of cancer cell lines. Multiple cancer cell lines in GDSC and CTRP were integrated for the calculations. We screened drugs with a cut-off of remarkable significance (p<1e-5). The screening identified 8 drugs of which sensitivities were negatively correlated with the CENPA level in cancer cells and 4 drugs of which sensitivities were positively correlated with the CENPA level in cancer cells (Figure 8A and S-Table 3). We proposed that these 12 drugs might directly interact with the CENPA protein.

To predict the direct interaction of CENPA and these 12 drugs, we accessed the predictive protein structural model of CENPA from the Alphafold database and docked the protein-ligand binding of CENPA and identified drugs. The Predicted aligned error of the CENPA protein structure model showed that the N-term had a long tail that had low model confidence, while the docking was based on the part that the models had very high confidence (Figure 8B,C). A protein-ligand model of a vina score of lower than -8 was considered a protein-ligand pair with very good binding affinity. The docking results showed that CD-437, 3-Cl-AHPC, Trametinib, BI-2536, and GSK461364 had high binding affinities to CENPA (S-Table 3), indicating that they were very likely to directly target CENPA in cancer cells. All the docking models were displayed in Figure 8D and the PBD files of the top 5 binding models of each drug were provided in the Supplementary Materials (file name "MD").
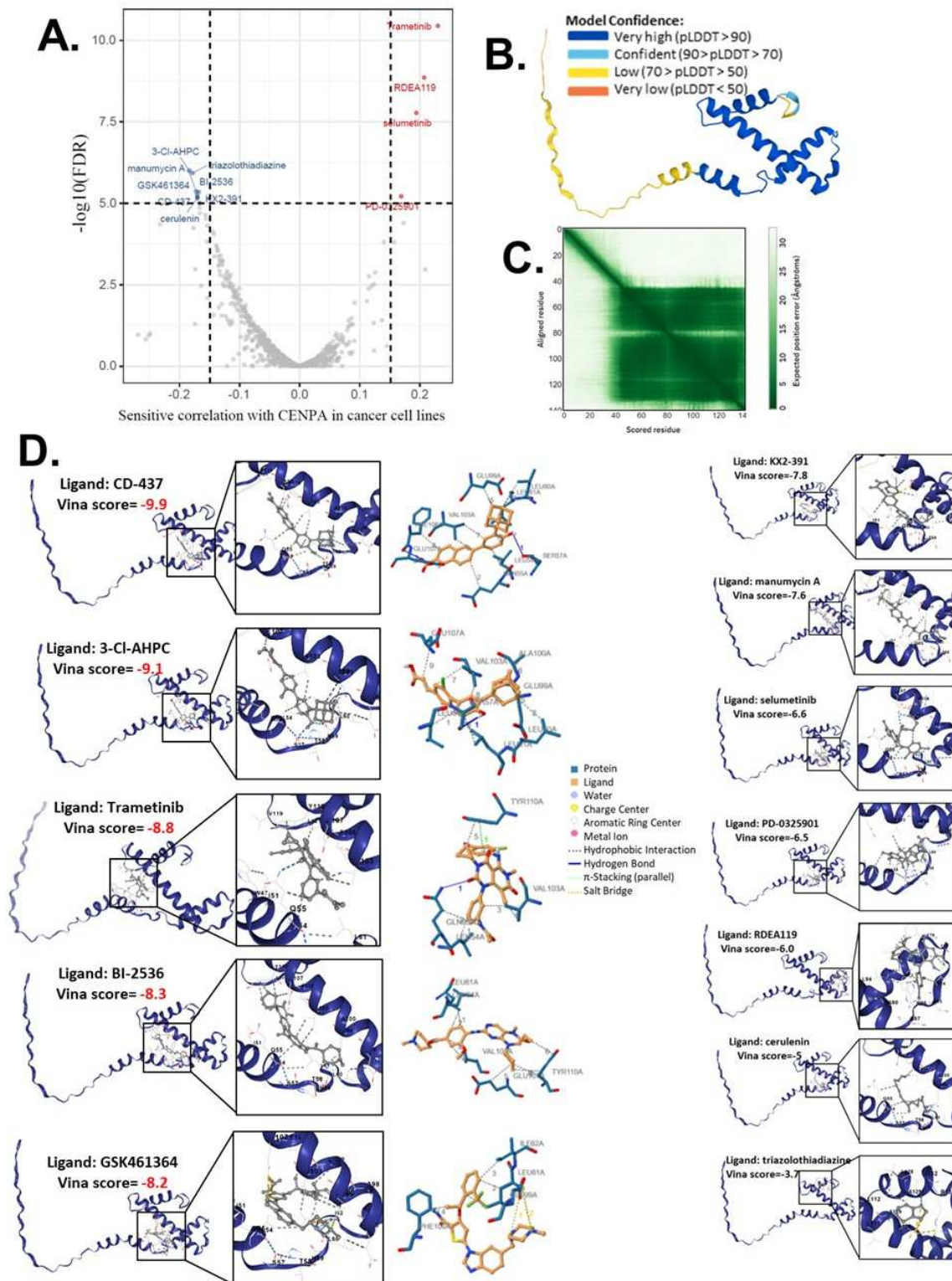
Figure 8. The computational drug prediction of CENPA in cancers. (**A**). The volcano plot of the correlation of CENPA expression and small molecule/drug sensitivity of cancer cell lines. GDSC and CTRP data were analyzed. Drug sensitivity and gene expression profiling data of multiple cancer cell lines 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.

4. 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 [51], 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 [52]. 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 regulated the assembly of active kinetochores, thereby regulating cell division [53]. Much as this process is essential for almost all proliferating cells, regardless of the malignancy of the cells, one of the common characteristics of cancer cells is their much higher proliferation rate than non-cancer cells. This indicated that the cancer cells have more cell divisions and, thus, might potentially require more CENPA for kinetochores regulation. The expression analysis validated this hypothesis that cancer requires and therefore expressed a higher level of CENPA compared with non-cancer by comparing CENPA mRNA expressions in cancer and non-cancer tissues. Results demonstrated that almost all cancer types overexpressed CENPA. LAML is the only cancer type that expressed lower CENPA in cancer than in normal tissues, yet, this made sense because LAML is a type of leukemia in which cell is supposed to have different cell cycle regulation mechanisms from most of the other cancer types [54]. 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 of this study suggested that CENPA mutations might not be the major factor that drives the development of cancers, because the mutation rate of CENPA in cancers was low. But the copy number might potentially affect cancer by increasing the transcription of CENPA mRNA. Therefore, we focused mostly on the expression level of CENPA. A previous study reported that CENPA overexpression promoted genome instability in human cells with the inactivation of the retinoblastoma protein [55]. 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 (SFig.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 associate with worse overall survival of some cancer types, such as ovarian cancer [13], liver cancer [16], breast cancer [17,18], and lung cancer [19]. 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 [56]. However, the conclusion of this study was supported by TCGA data only and was applied to GBM only. Whereas, in this study, we expand the conclusions to overall glioma, including both low-grade and high-grade glioma. In this study, the prognostic association of CENPA and the survival of glioma patients were supported by five independent data sets of glioma, which had case numbers of 703, 693, 325, 301, and 120 respectively. We relatively larger numbers of data sets and relatively larger numbers of independent data sources, we believed the prognostic performance of CENPA was rather reliable.

The immune association of CENPA in some cancer types, such as lung cancer and liver cancer [57], has been previously proven by TCGA data [19], this study expanded the immune association to

the pan-cancer range and compared its ICB response prediction with other immune response prediction biomarkers. The ICB cohorts in this study did not have a large number of patients, but it provided a hind for the potential value of CENPA for the prediction of immune therapy, which required further validations. In addition, we also screened and predicted potential drugs that targeted CENPA in cancer cells by pure computational methodology, hence, these results also required further validation with experimental evidence.

5. Conclusion

CENPA was a cell cycle biomarker in cancers with diagnostic prognostic, and therapeutic value.

References

1. R.L. Siegel, K.D. Miller, H.E. Fuchs, A. Jemal, Cancer Statistics, 2021, CA Cancer J Clin, 71 (2021) 7-33.
2. J.N. Weinstein, E.A. Collisson, G.B. Mills, et al., The Cancer Genome Atlas Pan-Cancer analysis project, Nat Genet, 45 (2013) 1113-1120.
3. W. Trust, Sharing data from large-scale biological research projects: a system of tripartite responsibility, Report of a meeting organized by the Wellcome Trust and held on 14–15 January 2003 at Fort Lauderdale, USA, Wellcome Trust London, 2003.
4. Z. Zhao, K.N. Zhang, Q. Wang, et al., Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients, Genomics, proteomics & bioinformatics, 19 (2021) 1-12.
5. T.J. Hudson, W. Anderson, A. Aretz, et al., International network of cancer genome projects, Nature, 464 (2010) 993-998.
6. D. Hanahan, R.A. Weinberg, Hallmarks of cancer: the next generation, Cell, 144 (2011) 646-674.
7. G.H. Williams, K. Stoeber, The cell cycle and cancer, The Journal of pathology, 226 (2012) 352-364.
8. B.E. Black, D.W. Cleveland, Epigenetic centromere propagation and the nature of CENP-a nucleosomes, Cell, 144 (2011) 471-479.
9. M. Schuh, C.F. Lehner, S. Heidmann, Incorporation of Drosophila CID/CENP-A and CENP-C into centromeres during early embryonic anaphase, Current biology: CB, 17 (2007) 237-243.
10. L.E. Jansen, B.E. Black, D.R. Foltz, D.W. Cleveland, Propagation of centromeric chromatin requires exit from mitosis, J Cell Biol, 176 (2007) 795-805.
11. M. Gonzalez, H. He, S. Sun, C. Li, F. Li, Cell cycle-dependent deposition of CENP-A requires the Dos1/2–Cdc20 complex, Proceedings of the National Academy of Sciences, 110 (2013) 606-611.
12. A.K. Saha, R. Contreras-Galindo, Y.S. Niknafs, et al., The role of the histone H3 variant CENPA in prostate cancer, The Journal of biological chemistry, 295 (2020) 8537-8549.
13. J. Han, R. Xie, Y. Yang, et al., 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, 10 (2021) 4076-4086.
14. Y.C. Liang, Q. Su, Y.J. Liu, H. Xiao, H.Z. Yin, 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, 191 (2021) 2117-2132.

15.   Q. Wang, J. Xu, Z. Xiong, et al., CENPA promotes clear cell renal cell carcinoma progression and metastasis via Wnt/β-catenin signaling pathway, J Transl Med, 19 (2021) 417.

16.   Y. Zhang, L. Yang, J. Shi, et al., The Oncogenic Role of CENPA in Hepatocellular Carcinoma Development: Evidence from Bioinformatic Analysis, Biomed Res Int, 2020 (2020) 3040839.

17.   A.B. Rajput, N. Hu, S. Varma, et al., Immunohistochemical Assessment of Expression of Centromere Protein-A (CENPA) in Human Invasive Breast Cancer, Cancers, 3 (2011) 4212-4227.

18.   S. Zhang, Y. Xie, T. Tian, et al., 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, 21 (2021) 410.

19.   H. Zhou, T. Bian, L. Qian, et al., Prognostic model of lung adenocarcinoma constructed by the CENPA complex genes is closely related to immune infiltration, Pathology, research and practice, 228 (2021) 153680.

20.   E. Cerami, J. Gao, U. Dogrusoz, et al., The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data, Cancer Discov, 2 (2012) 401-404.

21.   Pan-cancer analysis of whole genomes, Nature, 578 (2020) 82-93.

22.   A. Mayakonda, H.P. Koeffler, Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies, BioRxiv, (2016) 052662.

23.   C.H. Mermel, S.E. Schumacher, B. Hill, et al., GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers, Genome biology, 12 (2011) R41.

24.   Z. Tang, C. Li, B. Kang, et al., GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses, Nucleic Acids Res, 45 (2017) W98-w102.

25.   D. Szklarczyk, A.L. Gable, K.C. Nastou, et al., The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets, Nucleic Acids Res, 49 (2021) D605-d612.

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

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

28.   T.M. Malta, A. Sokolov, A.J. Gentles, et al., Machine Learning Identifies Stemness Features Associated with Oncogenic Dedifferentiation, Cell, 173 (2018) 338-354.e315.

29.   T.A. Chan, M. Yarchoan, E. Jaffee, et al., 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, 30 (2019) 44-56.

30.   S.J. Salipante, S.M. Scroggins, H.L. Hampel, E.H. Turner, C.C. Pritchard, Microsatellite instability detection by next generation sequencing, Clinical chemistry, 60 (2014) 1192-1199.

31.   B. Chen, M.S. Khodadoust, C.L. Liu, A.M. Newman, A.A. Alizadeh, Profiling Tumor Infiltrating Immune Cells with CIBERSORT, Methods in molecular biology (Clifton, N.J.), 1711 (2018) 243-259.

32.   D. Aran, Z. Hu, A.J. Butte, xCell: digitally portraying the tissue cellular heterogeneity landscape, Genome biology, 18 (2017) 220.

33.   D. Sun, J. Wang, Y. Han, et al., TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment, Nucleic Acids Res, 49 (2021) D1420-d1430.

34.   M.N. Bernstein, Z. Ni, M. Collins, et al., CHARTS: a web application for characterizing and comparing tumor subpopulations in publicly available single-cell RNA-seq data sets, BMC bioinformatics, 22 (2021) 83.

35.   H. Yuan, M. Yan, G. Zhang, et al., CancerSEA: a cancer single-cell state atlas, Nucleic Acids Res, 47 (2019) D900-d908.

36.   A.J. Tarashansky, Y. Xue, P. Li, S.R. Quake, B. Wang, Self-assembling manifolds in single-cell RNA sequencing data, eLife, 8 (2019).

37.   K.E. Yost, A.T. Satpathy, D.K. Wells, et al., Clonal replacement of tumor-specific T cells following PD-1 blockade, Nature medicine, 25 (2019) 1251-1259.

38.   C. Di Genua, S. Valletta, M. Buono, et al., C/EBPα and GATA-2 Mutations Induce Bilineage Acute Erythroid Leukemia through Transformation of a Neomorphic Neutrophil-Erythroid Progenitor, Cancer Cell, 37 (2020) 690-704.e698.

39.   C. Neftel, J. Laffy, M.G. Filbin, et al., An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma, Cell, 178 (2019) 835-849.e821.

40.   K.G. Paulson, V. Voillet, M.S. McAfee, et al., Acquired cancer resistance to combination immunotherapy from transcriptional loss of class I HLA, Nat Commun, 9 (2018) 3868.

41.   J. Fu, K. Li, W. Zhang, et al., Large-scale public data reuse to model immunotherapy response and resistance, Genome Med, 12 (2020) 21.

42.   C.J. Liu, F.F. Hu, M.X. Xia, et al., GSCALite: a web server for gene set cancer analysis, Bioinformatics, 34 (2018) 3771-3772.

43. W. Yang, J. Soares, P. Greninger, et al., Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res, 41 (2013) D955-961.

44. M.G. Rees, B. Seashore-Ludlow, J.H. Cheah, et al., Correlating chemical sensitivity and basal gene expression reveals mechanism of action, Nat Chem Biol, 12 (2016) 109-116.

45. J. Jumper, R. Evans, A. Pritzel, et al., Highly accurate protein structure prediction with AlphaFold, Nature, 596 (2021) 583-589.

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

47. A. Liberzon, C. Birger, H. Thorvaldsdóttir, et al., The Molecular Signatures Database (MSigDB) hallmark gene set collection, Cell systems, 1 (2015) 417-425.

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

49. D.N. Louis, H. Ohgaki, O.D. Wiestler, et al., The 2007 WHO classification of tumours of the central nervous system, Acta neuropathologica, 114 (2007) 97-109.

50. E.B. Claus, K.M. Walsh, J.K. Wiencke, et al., Survival and low-grade glioma: the emergence of genetic information, Neurosurg Focus, 38 (2015) E6-E6.

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

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

53. K. Kixmoeller, P.K. Allu, B.E. Black, The centromere comes into focus: from CENP-A nucleosomes to kinetochore connections with the spindle, Open biology, 10 (2020) 200051.

54. D. Schnerch, J. Yalcintepe, A. Schmidts, et al., Cell cycle control in acute myeloid leukemia, American journal of cancer research, 2 (2012) 508-528.

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

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

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