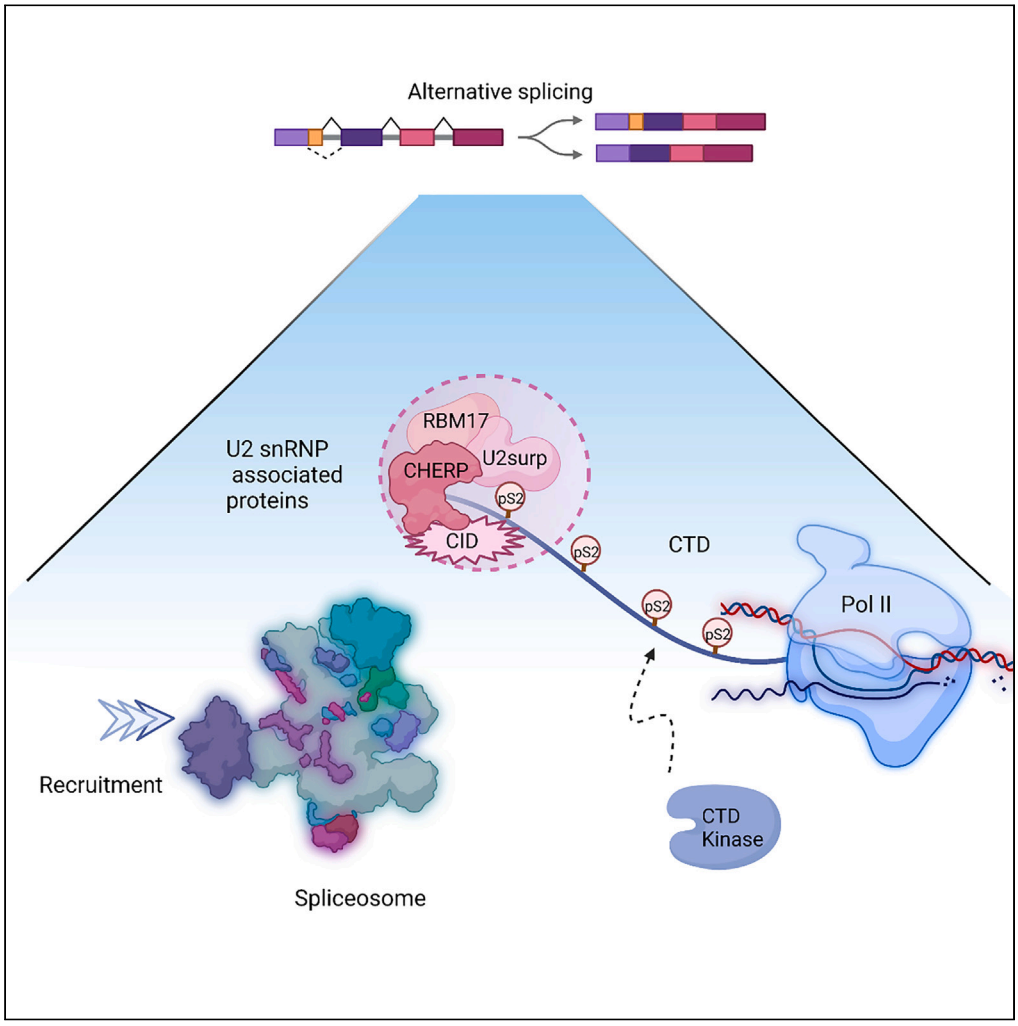
iScience

CellPress
OPEN ACCESS

Article

Distinctive interactomes of RNA polymerase II phosphorylation during different stages of transcription



Rosamaria Y. Moreno, Kyle J. Juetten, Svetlana B. Panina, ..., Edward M. Marcotte, Jennifer S. Brodbelt, Y. Jessie Zhang

jzhang@cm.utexas.edu

Highlights

We used *in vitro* reconstruction to generate the pS2/pS5 species of the CTD of Pol II

Pol II phosphorylated at S5/S2 recruits different regulators with ongoing transcription

Novel regulators specifically recruited by certain phosphoryl CTD are identified

CHERP specifically interacts with pSer2 Pol II through a CID module

Moreno et al., iScience *26*, 107581
September 15, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.107581

iScience

CellPress
OPEN ACCESS

Article

Distinctive interactomes of RNA polymerase II phosphorylation during different stages of transcription

Rosamaria Y. Moreno,[1] Kyle J. Juetten,[2] Svetlana B. Panina,[1] Jamie P. Butalewicz,[2] Brendan M. Floyd,[1] Mukesh Kumar Venkat Ramani,[1] Edward M. Marcotte,[1] Jennifer S. Brodbelt,[2] and Y. Jessie Zhang[1,3,*]

SUMMARY

During eukaryotic transcription, RNA polymerase II undergoes dynamic post-translational modifications on the C-terminal domain (CTD) of the largest subunit, generating an information-rich PTM landscape that transcriptional regulators bind. The phosphorylation of Ser5 and Ser2 of CTD heptad occurs spatio-temporally with the transcriptional stages, recruiting different transcriptional regulators to Pol II. To delineate the protein interactomes at different transcriptional stages, we reconstructed phosphorylation patterns of the CTD at Ser5 and Ser2 *in vitro*. Our results showed that distinct protein interactomes are recruited to RNA polymerase II at different stages of transcription by the phosphorylation of Ser2 and Ser5 of the CTD heptads. In particular, we characterized calcium homeostasis endoplasmic reticulum protein (CHERP) as a regulator bound by phospho-Ser2 heptad. Pol II association with CHERP recruits an accessory splicing complex whose loss results in broad changes in alternative splicing events. Our results shed light on the PTM-coded recruitment process that coordinates transcription.

INTRODUCTION

Among the three RNA polymerases in eukaryotic cells, RNA polymerase II carries the responsibility of transcribing all mRNA and a large portion of snRNA. Transcription by RNA polymerase II requires high efficiency and precision in streamlining the process from initiation to mRNA co-processing and termination.[1] The C-terminal domain (CTD) of the largest subunit of RNA polymerase II, RPB1, plays a crucial role in achieving this goal by recruiting various transcription regulatory factors to the elongating RNA polymerase II.[1,2] This unique region contains a heptad sequence (YS_2PTS_5PS) of 26–52 repeats, which undergoes phosphorylation with Ser2 and Ser5 phosphorylated at every round of transcription.[3,4] The transcription regulators specific for each stage recognize different phosphorylation patterns and recruit them over to Pol II to execute biological processes on the nascent transcripts.[5]

This model of distinct protein interactomes for each phosphorylation pattern is consistent with the observation of spatiotemporal phosphorylation on the CTD heptad according to the transcriptional stages.[6] For example, when Pol II associates with the promoter, the CTD contains no prior phosphorylation at the beginning of transcription.[7] Once Ser5 of the heptad gets phosphorylated, the added phosphates disrupt the interaction of RNA polymerase II with the Mediator complex, enabling promoter clearance.[8] RNA polymerase II with phosphorylation at Ser5 dominates at the early stage of transcription.[4] The phosphorylation of Ser2 begins after the promoter-proximal pause-release and accumulates throughout elongation until its complete removal at the end of transcription.[4] Throughout the transcription cycle, the CTD undergoes a continuous process of no phosphorylation to Ser5 phosphorylation, Ser2 phosphorylation, and then back to no phosphorylation.

Scientists have shown great interest in identifying proteins differentially recruited to Ser5 or Ser2 phosphoryl-marks on Pol II because they are highly present at each transcription stage.[9,10] However, previous efforts to elucidate distinct interactomes of Pol II have encountered challenges. RNA polymerase II is highly heterogeneous in phosphorylation, with additional phosphorylation occurring at Tyr1, Thr4, and Ser7.[11–13] Direct pulldown analysis using Pol II or phosphoryl-heptad antibodies has led to the identification of highly overlapping interactomes.[14] Conversely, short synthetic polypeptides containing a couple of heptads with phosphorylated Ser5 or Ser2 as baits result in a low signal/noise ratio due to weak interactions.[15]

A biochemical *in vitro* reconstruction strategy can potentially overcome the problem of low signal/noise and identify the distinctive interactomes. The approach involves purifying the CTD and biochemically phosphorylating it with highly specific kinases to generate different phosphorylation patterns.[16] This experimental design has several advantages over previous methods. First, the CTD's association with binding partners is mostly independent of the Pol II core subunit as it is high flexible and distanced from the core region with a 150–200 amino acid

[1]Department of Molecular Biosciences, University of Texas, Austin, TX, USA
[2]Department of Chemistry, University of Texas, Austin, TX, USA
[3]Lead contact
*Correspondence: jzhang@cm.utexas.edu
https://doi.org/10.1016/j.isci.2023.107581

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

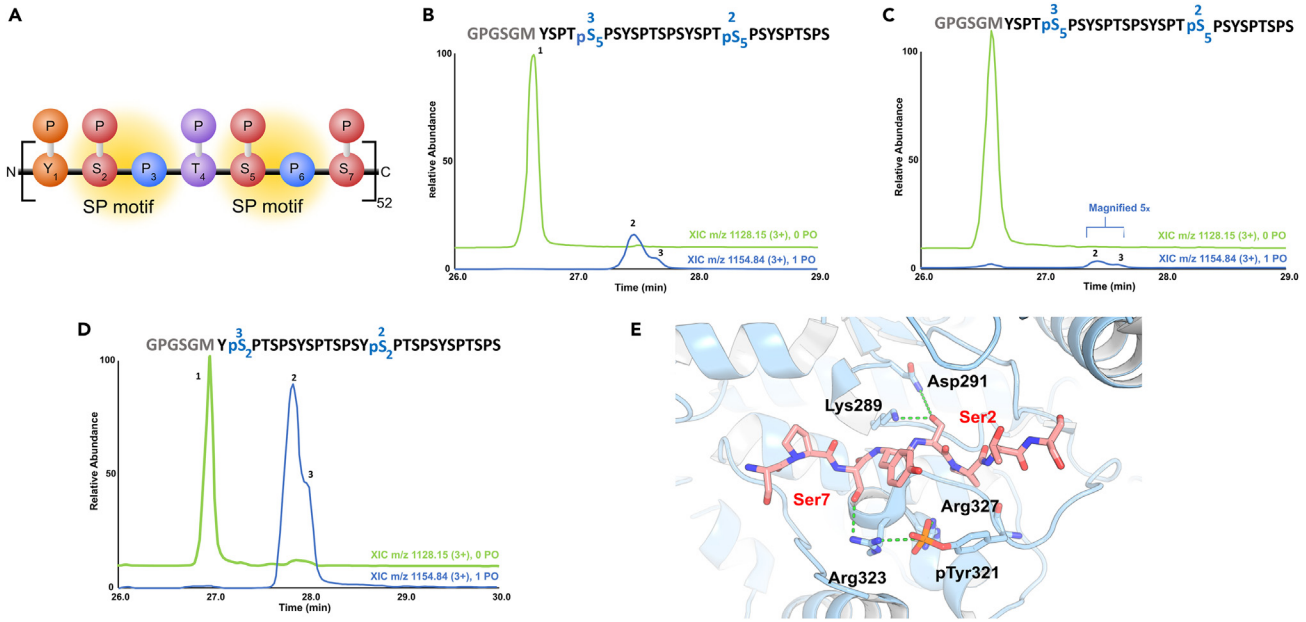


Figure 1. CTD kinases differentially phosphorylate consensus CTD

(A) Schematic of the consensus sequence and putative phosphorylation site on the CTD. Each unique residue is denoted by a distinct color, and potential phosphorylation sites are indicated with a phosphate symbol overhead. Serine-proline motifs are highlighted.

(B–D) Extracted ion chromatographic traces (XIC) for the LC-MS/MS analysis for kinase specificity toward the consensus CTD sequence. Green-colored LC traces correspond to the unphosphorylated peptide, whereas blue traces indicate the monophosphorylated species with peak numbers matching the sites of phosphorylation indicated on the sequences above the LC traces. The GPGSGM amino acid sequence is retained after 3C proteolysis of the expression tag preceding LC-MS/MS analysis. (B) TFIIH (C) P-TEFb (D) DYRK1A (E) Structural modeling of DYRK1A's interaction with a consensus CTD heptad as substrate with Ser2 subject to phosphorylation in the active site. The active site interactions of DYRK1A (light blue) with the modeled CTD peptide (salmon) are highlighted with key residues shown in stick and hydrogen bonds in dash lines.

linker.[17] This isolation of the CTD region allows us to identify proteins recruited solely based on the phosphorylation states of the CTD and avoid the co-precipitation of hundreds of proteins associated with the core subunits of Pol II. Second, using the full-length CTD increases the local concentration of epitope, thereby amplifying the signal and confidence of the interactome studies. Finally, the *in vitro* phosphorylation using CTD kinases provides us with a homogeneous phosphorylation species as bait, which avoids the interference of other co-occurring phosphorylation on Pol II.

The first attempt at using the biochemical *in vitro* construction strategy to identify Ser5 and Ser2 interactomes, using physiological Ser5 kinase (TFIIH) and Ser2 kinase (P-TEFb), reveals highly overlapping interactomes.[16] This puzzling result was due to the altered P-TEFb specificity from Ser2 to Ser5 *in vitro*.[18,19] *In vivo*, P-TEFb (CDK9), and CDK12/13, all present strong pSer2 phosphorylation but can phosphorylate both Ser5 and Ser2. To identify a kinase with a strong preference for Ser2 of the heptad *in vitro*, we used bioinformatics and structural analysis to identify a human kinase,[20] DYRK1A, and confirmed its specificity by mass spectrometry.[21] Our study using CTD treated by TFIIH and DYRK1A resulted in distinctive interactomes and confirmed several previously identified CTD-binding proteins. Notably, we discovered that CHERP (also called SCAF6), a putative RNA binding protein, selectively binds to RNA polymerase II only when phosphorylated at Ser2. We identified the domain in CHERP that recognizes the CTD and associates with Pol II upon Ser2 phosphorylation, whose loss prevents CHERP recruitment. Chromatin immunoprecipitation sequencing (ChIP-seq) revealed that CHERP co-localizes with Pol II on genes, and this interaction can be disrupted if Ser2 phosphorylation is inhibited or the CID motif on CHERP is removed. CHERP is part of an accessory splicing complex whose elimination causes extensive alternative splicing events that have been implicated in colorectal tumorigenesis.[22]

RESULTS

The kinase specificity on the serines of the CTD

To perform interactome studies to elucidate transcriptional regulators that phosphorylated Ser2 and Ser5 recruit, we needed a CTD homogenously phosphorylated at those sites. However, achieving this specificity is challenging since Ser2 and Ser5 are both parts of a Ser-Pro motif (Figure 1A). Initially, we considered physiological kinases whose activity might generate the desired phosphorylation patterns. To test the specificity of kinases *in vitro*, we generated a GST-CTD construct that contains four heptad repeats with a consensus sequence. The physiological kinase module of TFIIH, CDK7/CYCH/MAT1, generates phosphoryl-CTD heptads with only Ser5 phosphorylated (Figures 1B and S1). The specificity is consistent with our previous results using CTD constructs of varying lengths, even though longer substrate is preferred by TFIIH.[23,24] However, the physiological Ser2 kinase, P-TEFb, was reported to strongly prefer Ser5 over Ser2 as a substrate *in vitro*.[18,19] When

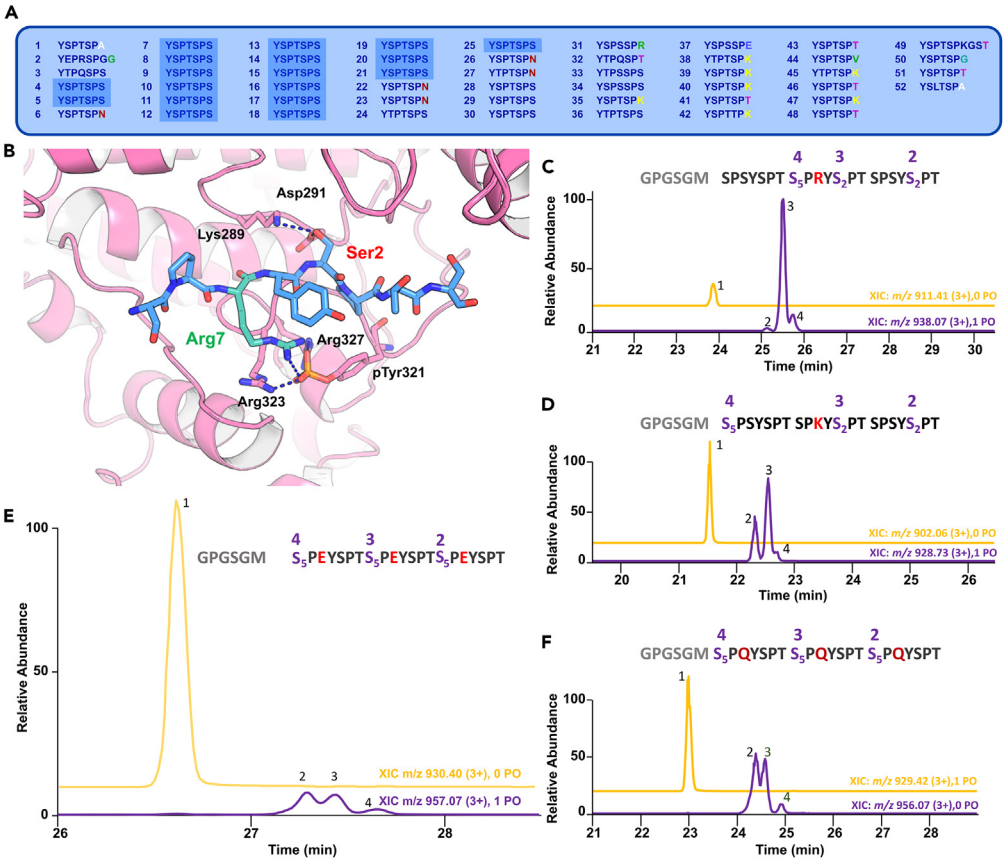


Figure 2. DYRK1A's specificity in divergent CTD heptads

(A) Diagram of the CTD sequence of human RPB1. The shaded region consists of the consensus sequence. The rest of the sequence diverges slightly from consensus, with residues different from serine in the seventh position highlighted.

(B) Structural model of DYRK1A in complex with a CTD peptide where the seventh position is occupied by arginine. DYRK1A is shown in the ribbon diagram as light pink, and the CTD peptide as sticks is shown as light blue. Hydrophilic interactions are denoted with dashed lines.

(C–F) Chromatographic traces for the LC-MS/MS analysis of DYRK1A's specificity toward CTD peptides containing three heptads. Gold-colored LC traces correspond to the unphosphorylated peptide, whereas the purple traces indicate the monophosphorylated species with peak numbers matching the sites of phosphorylation indicated on the sequences above the LC traces. (C) the middle heptad containing arginine in the seventh position. (D) the middle heptad containing lysine in the seventh position. (E) each heptad containing glutamate in the seventh position. (F) each heptad containing glutamine in the seventh position.

we used P-TEFb to treat the CTD heptad sequence, the only phosphorylation species was at Ser5 (Figures 1C and S2). Previous studies using tandem mass spectrometry confirmed that P-TEFb phosphorylates Ser2 when the CTD is primed by Tyr1 phosphorylation; otherwise, the phosphorylation occurs at Ser5.[19] Other CDKs also preferentially phosphorylate Ser5 over Ser2 with conserved recognition structural motifs.[18,19,21] The mixture of Ser5 and Ser2 phosphorylation products complicates the interpretation of the interactome.

Therefore, we sought to identify a kinase that displays exclusive activity toward Ser2 of the CTD heptad sequence. Through a thorough bioinformatic and structural search of CTD kinases, we identified DYRK1A as a candidate due to its signature motif that supports Ser2 phosphorylation.[21] Subsequent enzymatic reactions with four heptad repeats resulted in products exclusively phosphorylated at the Ser2 in the context of the consensus sequence of YSPTSPS (Figures 1D and S3). Lack of phosphorylation on the fourth repeat by DYRK1A is likely a result of higher entropy cost at the end of the flexible peptide. Likewise, detection of phosphorylation on the middle repeat is obscured as it is located further from the N or C termini of the fragmented peptides in mass spec analysis. Our structural modeling provided insight into this specificity by suggesting that the residue two positions upstream of the serine subject to phosphorylation plays a crucial role in substrate specificity (Figure 1E). If Ser2 is subject to phosphorylation, the residue two positions upstream is the 7th residue from the preceding heptad. This 7th residue (Ser7 in consensus sequence) forms hydrogen bonds with Arg323 of DYRK1A, forming salt bridges with phosphoryl-Tyr321. This favorable orientation ultimately places Ser2 in the active site to be phosphorylated.

The reduction of phosphorylation in both Ser2 and Ser5 upon DYRK1A knockdown was previously reported.[20] To investigate the possibility of alternative sites being phosphorylated with different sequence contexts, we systematically characterized DYRK1A's specificity using biochemical product profiling with high-resolution mass spectrometric characterization (Figure 2). With the 7th residue in the previous heptad

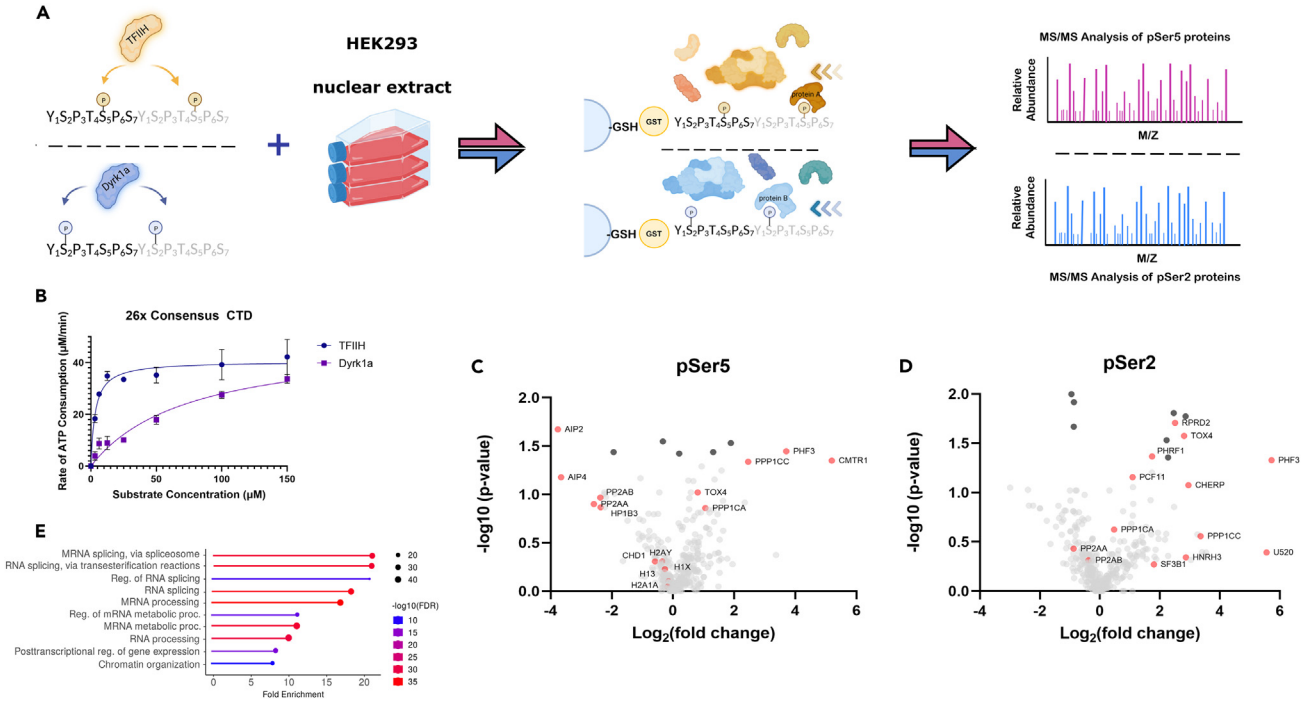


Figure 3. Proteomic study of pCTD interactomes

(A) *In vitro* phosphorylation of a GST-CTD recombinant protein containing 26x repeat consensus CTD heptads is treated with TFIIH or DYRK1A and incubated with lysate from HEK-293 cells. A no-kinase sample is treated parallelly as a control. Affinity chromatography immobilizes the GST-tagged substrate and pulls down Pol II-interacting proteins. LC-MS/MS analysis identifies proteins in each sample.

(B) Kinase activity assay of wild-type yCTD by TFIIH (dark blue) and DYRK1A (purple) fitted to the Michaelis-Menten kinetic equation. The Michaelis-Menten kinetic parameters $k_{cat}/K_m(\mu M^{-1}\ min^{-1})$ are given for each respective fit. Each measurement was conducted in triplicate, with standard deviations shown as error bars.

(C) Volcano plots comparing the pSer5 IP and the pSer2 IP (D). Both use unphosphorylated CTD IP as control. Enriched factors were determined using a p value of 0.05 and shown as dark gray dots. Factors mentioned in the text are labeled and shown as red dots.

(E) Gene ontology terms enriched for the top 100 proteins identified in the phospho-CTD interactome data for pSer2. Visualization was done with ShinyGO 0.76.

identified as key to DYRK1A (Figure 1E), we systematically replaced the preceding 7th residue and characterized the product phosphorylation sites (Figure 2). Human RNA Pol II is enriched with sequences divergent from the consensus at the 7th position of the heptad, where the most frequent replacements are positively charged residues like Lys and Arg (Figure 2A). Our binding model of DYRK1A for Ser2 phosphorylation places these positively charged residues along with Arg323 and Arg327 to stabilize the activating residue phosphoryl-Tyr321 (Figure 2B). To test this structural prediction, we used a substrate CTD with the 7th residue as Arg or Lys in the proceeding heptad; the Ser2 in the subsequent heptad gets phosphorylated preferably and effectively (Figures 2C, 2D, S4, and S5). Favorable interactions of Lys7/Arg7 with DYRK1A place the neighboring Ser2 at a highly favorable position for kinase phosphorylation (Figures 2C and 2D). In contrast, if a negatively charged residue like glutamate or phosphorylated Ser7 occupies the 7th residue, it is positioned too close to the phosphoryl-Tyr of DYRK1A. This unfavorable repelling interaction switches the mode of substrate recognition. Indeed, we previously noticed Ser5 phosphorylation as the major product when we used heptad repeats containing E at the 7th position (Figure 2E and S6).[21] Furthermore, due to space limitations, bulky residues like glutamine in this 7th position also hamper the hydrogen bond network with DYRK1A. The exclusion of Ser2 binding mode leads to Ser5 on the heptad being phosphorylated with no phosphoryl-Ser2 detected (Figures 2F and S7). Therefore, our mass spectrometric and structural analyses reveal that DYRK1A strongly prefers the phosphorylation of CTD at the Ser2 position when the 7th residue from the preceding heptad is occupied by a small polar or positively charged residue. When Ser7 is phosphorylated or occupied by bulky polar residues, DYRK1A phosphorylates Ser5 instead of Ser2. Our detailed investigation explained why DYRK1A exhibits both Ser2 and Ser5 activity in human cells, in which the 7th residue diverges widely from the consensus.[20] Most relevant to our interactome study, the analysis shows that DYRK1A produces the product with exclusive Ser2 phosphorylation if we use consensus sequence as substrate.

Interactome of differential binding of CTD domain phosphorylated at Ser5 versus Ser2

We conducted label-free proteomics analyses of pulldowns using phosphorylated CTDs that were differentially treated by kinases with well-characterized specificity (Figures 3A, and S8A). The bait used in the proteomic study was a GST-tagged 26-repeat CTD, consisting mostly of a consensus sequence, which was treated with different kinases (TFIIH kinase module and Dyrk1a). The 26-repeat CTD is a good substrate for both kinases, and the kinetic experiments revealed that this recombinant CTD was phosphorylated effectively by TFIIH and Dyrk1a with

k_{cat}/K_m as 12.3 ± 1.6 $\mu M^{-1}/min^{-1}$ and 0.63 ± 0.05 $\mu M^{-1}/min^{-1}$, respectively (Figure 3B). The activity of TFIIH is much higher on a longer repeat than on a CTD substrate with only four repeats (Figure 1B). In addition, methods such as MALDI/EMSA shows that on average more than 50% of total phosphosites on the 26X are phosphorylated. To ensure the comparability of the samples with control (unphosphorylated CTD), we divided the GST-CTD into three portions: one to be treated with TFIIH, one to be treated with Dyrk1a, and one control sample with an identical buffer with no kinase. The samples were incubated with oscillation overnight. After washing off the kinases, GST-CTD samples were incubated overnight with equal amounts of nuclear cell lysate-containing inhibitors for phosphatases and proteases (Figures 3C and 3D, Full list in Table S1). Additionally, the nuclear lysate was treated with benzonase before performing the pulldown to ensure only protein: protein interactions were observed. Finally, the samples were analyzed using label-free proteomics by comparing the abundance of pulled-down proteins in each kinase-treated sample to that in the control.

Different from the previous efforts in obtaining differential phosphorylated CTD, our study's interactomes of pSer2 and pSer5 are dramatically different, identifying several dozen proteins with differential binding patterns (Figures 3C and 3D). For pSer5 pulldown, a significant characteristic was the reduction of many proteins upon Ser5 phosphorylation compared to the control (Figure 3C). Many of these proteins were histones, histone variants, and associated accessory proteins, such as chromatin remodeling complexes (Figure 3C; and Table S1). This observation is consistent with the biological understanding that active transcription reduces chromatin density by ejecting and relocating them.[25] A top hit for depletion upon Ser5 phosphorylation is Heterochromatin protein 1-binding protein 3 (HP1B3), a component that maintains heterochromatin condensation (Figure 3C). Other top hits include AIP E3 ligase homologs, which have been implicated in RPB1 translocation and degradation.[26] Not many proteins are recruited to pSer5 CTD. The most significant hit for pSer5 binding was mRNA capping enzyme Cap1 2′O-ribose methyltransferase 1, a major function of Ser5 phosphorylation.[27,28] Furthermore, PHF3, a recently identified CTD-binding protein, was highly enriched, consistent with recent reports that it can directly bind Ser5 and/or Ser2 phosphorylated CTD[29] (Figure 3C).

The pattern of pSer2 pulldown (Figure 3D) showed a significant difference from that of pSer5 pulldown (Figure 3C), with more proteins associated with the phosphoryl mark rather than depletion. Some previously characterized pSer2-binders, such as PHF3, PCF11,[30] PHRF1,[31] and RPRD2[32] are among the hits, validating the accuracy of our pulldown (Figure 3D). Many proteins identified are spliceosome components such as U5 snRNP and splicing factors, but their direct interaction with CTD has yet to be established. A large fraction of the bound proteins are implicated in association with RNA. Ontology analysis of the top 100 enriched hits from the pSer2 pulldown revealed that RNA-binding proteins were the most enriched category (Figure 3E). Most proteins identified involve mRNA processing, translocation, or post-transcriptional modifications. This observation is consistent with the finding of pSer2-enriched Pol II during the later stage of transcriptional events after nascent mRNA appears.[33]

We observed a high representation of phosphatases in the proteomic study. PP1 is highly enriched in both pSer2 and pSer5 interactomes. We identified two PP1 catalytic subunit isoforms, alpha and gamma, and TOX high-mobility group box family member 4, which forms a stable complex with PNUTS/PP1 phosphatase complex (Figures 3C and 3D). The PNUTS/PP1 phosphatase complex is known to be involved in Pol II dephosphorylation during active transcription.[34] In contrast, PP2A is highly depleted in the pSer5 and pSer2 pulldown samples (Figures 3C and 3D). As part of the integrator complex,[35] PP2A diverts the stuck RNA polymerase II at the promoter-proximal pausing sites to abortive escape without entering productive elongation.[36] The differential recruitment of the PP1 and PP2A complexes is consistent with their biological functions. The hits identified from the proteomic study reveal distinctive interactomes recruited to Pol II by pSer5 and pSer2.

Identification of CHERP as a pSer2 binding protein

Our goal is to identify proteins that are directly recruited by specific CTD phosphoryl species. Although many proteins were found to interact with both phosphorylation forms (Table S1), we were particularly interested in those that were specific to one phosphoryl pattern. One protein that stood out was calcium homeostasis endoplasmic reticulum protein (CHERP), which was highly enriched in the pSer2 pulldown (2.9-fold enrichment) but not enriched in the pSer5 one (Figures S8B and S8C). CHERP contains a domain that resembles the CTD-binding motif found in other proteins, called the C-terminal interacting domain (CID) (Figure 4A). To corroborate the proteomics study, we examined the subcellular localization of CHERP in relation to Pol II. Immunofluorescence analysis revealed strong colocalization (correlation value of 0.77) between pSer2 Pol II and CHERP in transfected cells (Figure 4B). We also investigated whether this colocalization was dependent on pSer2 by inhibiting Ser2 phosphorylation with a small molecule compound, flavopiridol.[37] The colocalization is significantly reduced with a correlation value of 0.22, suggesting that pSer2 is required for the interaction between Pol II and CHERP (Figure 4B).

To further validate the interaction between Pol II and CHERP, we conducted co-immunoprecipitation (co-IP) experiments. The endogenously expressed CHERP interacted with total Pol II detected by an antibody against RPB1 (Figure 4C). Additionally, HA-tagged CHERP was transfected into HEK293 cells, and the association of CHERP with pSer2 Pol II was confirmed by reciprocal pulldown (Figures 4C and 4D). Notably, we observed that Thr4 phosphorylation, a post-translational modification often found together with Ser2 phosphorylation, also co-immunoprecipitated with CHERP (Figures 4C and 4D). While pThr4 and pSer2 marks are often found co-occurring during transcription, their function is not well understood.[38] These studies provide further validation of the proteomic study and confirm the association between CTD and CHERP.

CID-like domain directly interacts with the CTD domain of RPB1

To study if the direct interaction of CHERP and the CTD is through its CID-like domain (Figure 4A), we conducted an immunofluorescence experiment with a CHERP construct in which the CID domain is removed (CHERP-ΔCID) (Figures 4A and 4B). The removal of this domain

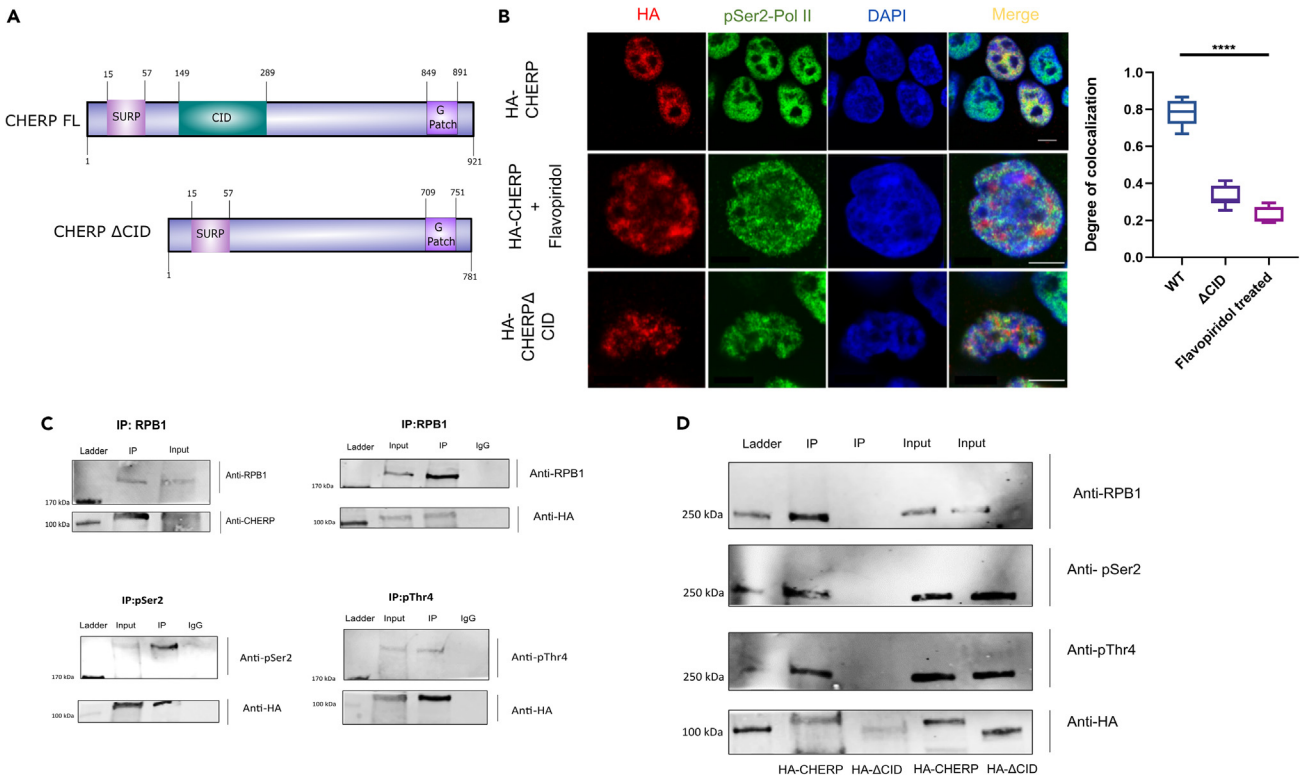


Figure 4. CHERP physically associates with phosphorylated human RNA Pol II

(A) Domain architecture of full-length CHERP and a deletion mutant that eliminates the CID domain.

(B) Representative confocal fluorescent images of HA-CHERP or HA-CHERP ΔCID(red), pSer2 Pol II (green), and DAPI (blue) in HEK293 cells. Scale bar = 5 μm. Quantification of colocalization between CHERP FL or a mutant with pSer2 under different conditions (N = 7). Box and whiskers plots with error bars representing the 10 and 90 percentiles are shown. One-way ANOVA was performed to compare groups ($p < 0.0001$). Each experiment was repeated twice with comparable results.

(C) Endogenous Pol II phosphoryl isoforms were immunoprecipitated from HEK293 cells or HEK293 with overexpression of HA-CHERP to show the interaction in the reciprocal direction. The experiment was performed three times.

(D) Anti-HA immunoprecipitation of HA-CHERP full-length protein or a deletion mutant lacking CID domain. $p \leq 0.0001$ (****).

reduced the colocalization between CHERP and pSer2 Pol II from 0.77 to 0.33 in correlation (Figure 4B). When we conducted a co-immunoprecipitation experiment using the CHERP construct lacking the CID region, we detected no Pol II or its phosphorylation forms (pSer2 and pThr4) pulled down. (Figure 4D). Both experiments indicate that the recruitment of CHERP to Pol II is dependent on the CID-like domain in cells.

To investigate the interaction of CHERP CID with Pol II CTD, we cloned and purified it to homogeneity (Figures S9A–S9C). We aimed to determine if the CHERP CID can bind to CTD heptads of different phosphorylation states and its specificity in CTD recognition. We employed Bio-layer Interferometry (BLI) to monitor binding events (both association and dissociation) by detecting the interference of light reflected from protein immobilized on a sensor (Figure 5A). When we incubated the CHERP CID domain with immobilized CTD peptides containing two and a half heptad repeats with no phosphorylation or phosphorylation at one Ser5, we observed no signal increase, indicating the lack of association (Figure 5A). In contrast, we observed a significant increase in signal when the protein was incubated with the CTD peptide phosphorylated at Ser2, which is quickly reduced when the protein is washed off (Figure 5A). We also found that the CHERP CID domain can directly interact with the Thr4-phosphorylated CTD peptide, as demonstrated by a strong binding profile (Figure 5A).

Although we observed strong signals specific for the CTD domain phosphorylated at Ser2 or Thr4, the data deviated from a simple 1:1 ratio association. Thus, BLI provides a qualitative measurement, but we need to use an alternative method to obtain a more accurate K_d. We applied fluorescence polarization to measure the binding strength between the CID-like domain of CHERP and different phosphorylation forms of the CTD (Figure 5D). We covalently attached fluorescein isothiocyanate (FITC) as a fluorescent tag to the N-terminus of the CTD peptides containing two and a half heptad repeats. Using free fluorophore as a control, we quantified the association of CHERP with CTD polypeptides phosphorylated at different sites. Our results showed the binding of the CHERP CID domain to CTD peptide phosphorylated at Ser2 with a K_d of 10 ± 2 μM and pThr4 with a K_d of 4 ± 1 μM (Figure 5D).

Although the CID-like domain did not crystalize, we used AlphaFold to predict a high-confidence structure using other CIDs (Figure 5B). To evaluate the predicted model, we identified several residues critical for the binding of CTD ligands based on the prediction (Figure S9G).

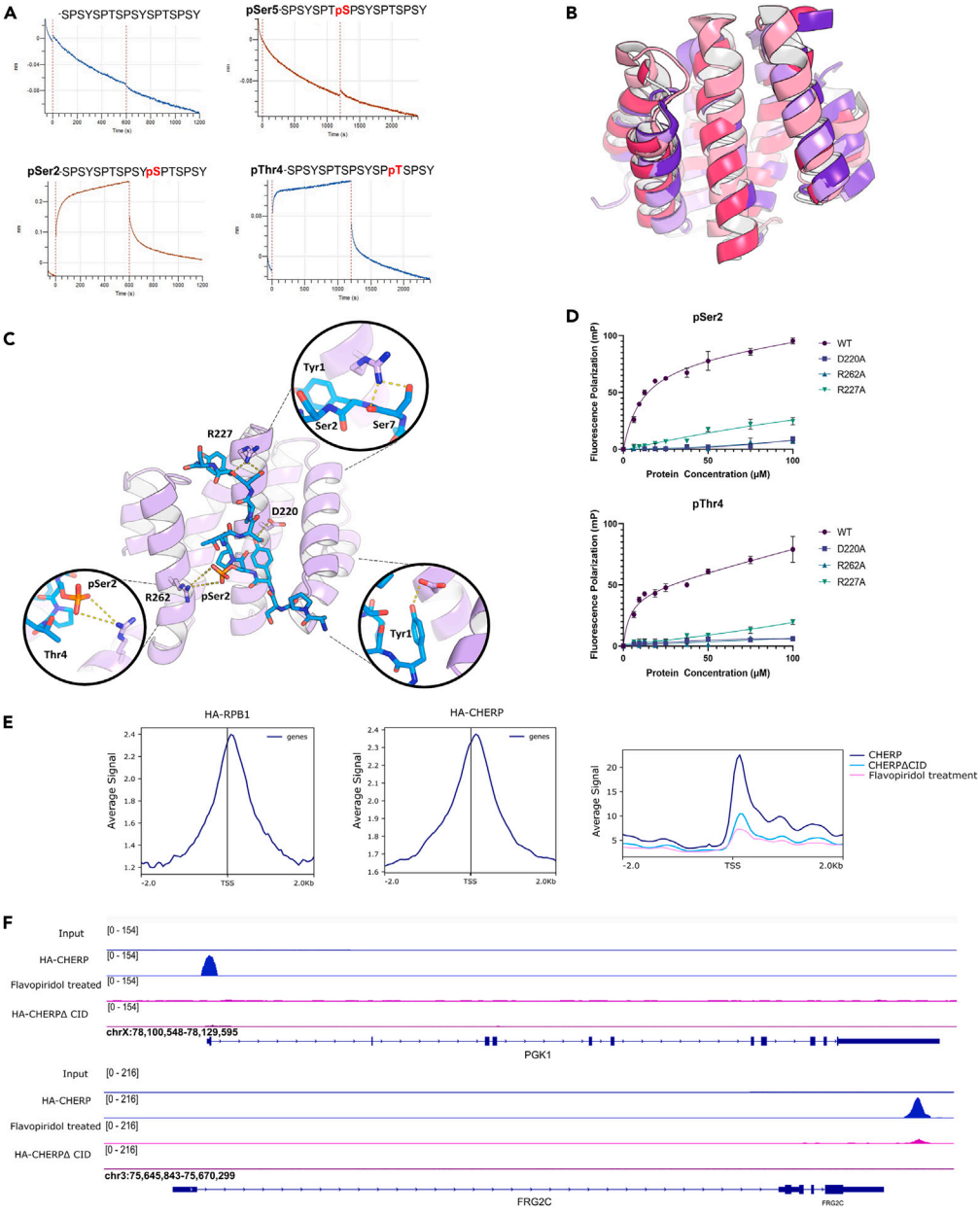


Figure 5. CHERP binds to Pol II through a CID domain

(A) Biolayer interferometry-binding assays of the interactions between the CID domain of CHERP (1 μM) with different phosphoryl isoforms of CTD peptides containing two heptad repeats. The wavelength shift (nm) generated is plotted as a function of time.

(B) Superimposition of CHERP CID in red (ALPHAFOLD structure) with SCAF4 in peach (PDB: 6XKB), RPRD1A in light purple (PDB: 4JXT), and RPRD1B in dark purple (PDB: 4Q96) (C) Modeling of the CID domain of CHERP with a pSer2 CTD ligand with emphasis on conserved interactions between CTD backbone and sidechains.

(D) Fluorescence anisotropy (FA) measurement of the binding of pS2 and pT4 FITC-labeled CTD peptides containing two repeats to the CID domain of CHERP. Experimental isotherms were fitted to a 1:1 binding model. Binding assays were performed in triplicate. Error bars indicate the standard error of the mean.

(E) Distribution of HA-RPB1 and HA-CHERP near the TSS. Profiles of HA-CHERPΔCID and HA-CHERP with flavopiridol treatment are overlaid across CHERP-bound genomic regions. Average peak coverage is shown in a bin size of 50 bp for a window 2 kb upstream/downstream of the TSS.

(F) Genome browser views of ChIP-seq signals for HA-CHERP, HA-CHERPΔCID mutant, and flavopiridol treatment samples at representative CHERP-target genes.

CHERP-CID is predicted to directly recognize the phosphorylated Ser2 in the CTD heptad through Arg262, a highly conserved residue in other CID-containing proteins that bind pSer2 marks. The salt bridge interaction between the side chain of the guanidinium group of Arg-262 of CHERP-CID and the phosphate moiety of pSer2 is critical for direct recognition (Figure 5C). Mutation in Arg262 alone abolishes any detectable binding (Figure 5D). Asp220 interacts with the hydroxyl group of the Tyr1 next to the phosphoryl-Ser2. Arg227 bridges the hydrogen bonding of the Ser7 residue in the heptad. The mutations on these residues do not alter protein folding but abolish or significantly compromise CTD binding (Figures 5D and S9D–S9F). The model also explains why phosphoryl-Thr4 can also be recognized by CHERP, as seen in the co-immunoprecipitation experiment and fluorescence anisotropy (Figures 4B and 4D). The hydroxyl group of Thr4 locates close to Ser2, facing the helix containing Arg262, and the phosphorylation on the Thr residue will be within the range for forming salt bridge interactions (Figure 5C). Thus, the structural and mutational analysis provides insight into CHERP's specificity toward CTD.

CHERP is recruited to genes via the interaction between its CID motif and pSer2 of CTD

To investigate genomic locations of CHERP association relative to RNA polymerase II, we performed a ChIP-seq analysis of CHERP and RPB1 binding. To ensure no antibody bias is introduced, we inserted full-length CHERP or RPB1 genes into an HA-tagged mammalian expression vector and transfected them into HEK293 cells (Figure S10B). Since CHERP is not expected to bind to the genome directly, we conducted ChIP-seq analysis with double crosslinking according to.[39] Both chromatin samples were pulled-down with HA antibodies and analyzed for their distribution genome-wide. We identified ~2000 peaks in each sample. Significantly, the profiles of CHERP and Pol II are highly similar, with a distribution profile peaking a little after Transcription Starting Site (TSS) (Figures 5E and 5F). Since Ser2 phosphorylation is a hallmark event upon promoter-proximal pausing release, the location of CHERP implicates its recruitment when Ser2 starts to get phosphorylated. While most of the CHERP peaks locate close to the TSS, some peaks are detected at the 3′ end of the genes (Figures 5F and S11). On the other hand, some peaks are found at the 3′ end of the genes where Thr4 phosphorylation appears (Figure S11).

To further pin down whether the location of CHERP on genes is highly dependent on the Ser2 phosphorylation on Pol II, we applied a small molecule inhibitor, flavopiridol, to lower Ser2 phosphorylation on RNA polymerase II.[37] The application of flavopiridol caused Ser2/Thr4 phosphorylation to be reduced on RNA polymerase II (Figure S10A). At 2μM, the application of flavopiridol led to a great reduction of CHERP recruitment to the genome (Figures 5E and 5F). We further tested if the CID-like domain in CHERP is crucial for the recruitment of CHERP using a construct with the CID domain omitted (CHERP-ΔCID). The ChIP-analysis reveals a weaker distribution compared to the full-length CHERP profile (Figures 5E, 5F and S10C–S10E). Thus, we found CHERP co-localizes with RNA polymerase II on genes, and its recruitment depends on the phosphorylation of Ser2 on Pol II and the presence of the CID domain of CHERP.

CHERP loss results in the alternative splicing

To investigate the role of CHERP in transcription, we performed RNA-seq analysis to measure the polyadenylated mRNA in cells where the CHERP protein is knocked down. As evaluated in the Western blot, we used a commercially available shRNA to reduce the CHERP expression to 18% (Figure S12D). We then conducted deep sequencing of the mRNA with CHERP knocked down compared to the wild type. Correlation analysis showed that both biological replicates clustered depending on the condition and strongly correlated with each other ($r > 0.99$) (Figures S12A–S12C). We found that the overall transcriptome was not significantly affected by the loss of CHERP. Analysis showed that 238 genes were differentially expressed between conditions, with 215 genes being downregulated (log2FC < −0.58, $q < 0.05$) and 23 genes upregulated after CHERP knockdown (log2FC > 0.58, $q < 0.05$) (Figure 6A). Table S2 shows a complete list of differentially expressed genes (DEGs). Whereas the number of genes affected in our study is fewer compared to another recent study that has used siRNA-mediated knockdown of CHERP in similar cell types, there was a significant overlap of differentially expressed genes.[40] We surmise that this difference in gene number is probably due to a different knockdown approach.

Deep sequencing allowed us to detect more subtle changes due to the reduction of CHERP at the level of alternatively spliced (AS) transcripts. We used rMATS software,[41] which can detect possible alterations in annotated alternative splicing of five types: skipped exon (SE), alternative 5′ and 3′ start sites (A5SS/A3SS), mutually exclusive exons (MXE), and retained intron (RI) (Figure 6B; and Table S3). We identified 2,135 AS events in 1,560 unique genes upon CHERP inhibition vs. Control (FDR <0.05, ILD, inclusion level difference, ≥10%) (Figure S12E). Based on GO analysis, alternatively, spliced transcripts were enriched for proteins related to cilium organization and cell polarity, cell cycle G2/M transition, response to endoplasmic reticulum stress, and DNA damage checkpoint (Figure 6C; and Table S3).

We analyzed alternatively spliced genes under CHERP knockdown and noticed that many genes played a role in the key cellular pathways and were involved in disease pathogenesis. Cilium organization and assembly appeared to be the most enriched biological processes among the genes that changed their splicing pattern under CHERP knockdown (Figure 6C). Primary cilia are hair-like projections that protrude from most mammalian cells and mediate various extracellular signaling pathways, including Hedgehog, Notch, Wnt, and tyrosine kinase pathways.[42] One such cilia-associated gene is a potential therapeutic target, centrosomal protein CEP164; its mutations or deficiency are related to ciliopathies (degenerative diseases affecting the kidney, retina, and brain)[43] and pancreatic cancer growth.[44] We found that CHERP knockdown leads to more frequent inclusion of exon 8 into *CEP164* mRNA (Figure 6D, *upper*; Table S4).

Furthermore, some genes with profoundly altered splicing (ILD = 100%) have been previously reported in relation to oncogenesis and cancer prognosis. For instance, a nervous system-related gene, *SNAP91* (synaptosome-associated protein 91), was found to be highly methylated in colorectal cancer tissues but not normal tissues, making *SNAP91* a potential biomarker for this cancer type[45] (Figure 6D, *lower*). Another gene, legumain (*LGMN*), encodes a cysteine endopeptidase that was demonstrated to be overexpressed in pan-cancer samples compared to normal tissues and to correlate with poor patients' prognosis and clinical stage[46] (Figure 6D, *lower*). Legumain promotes cellular

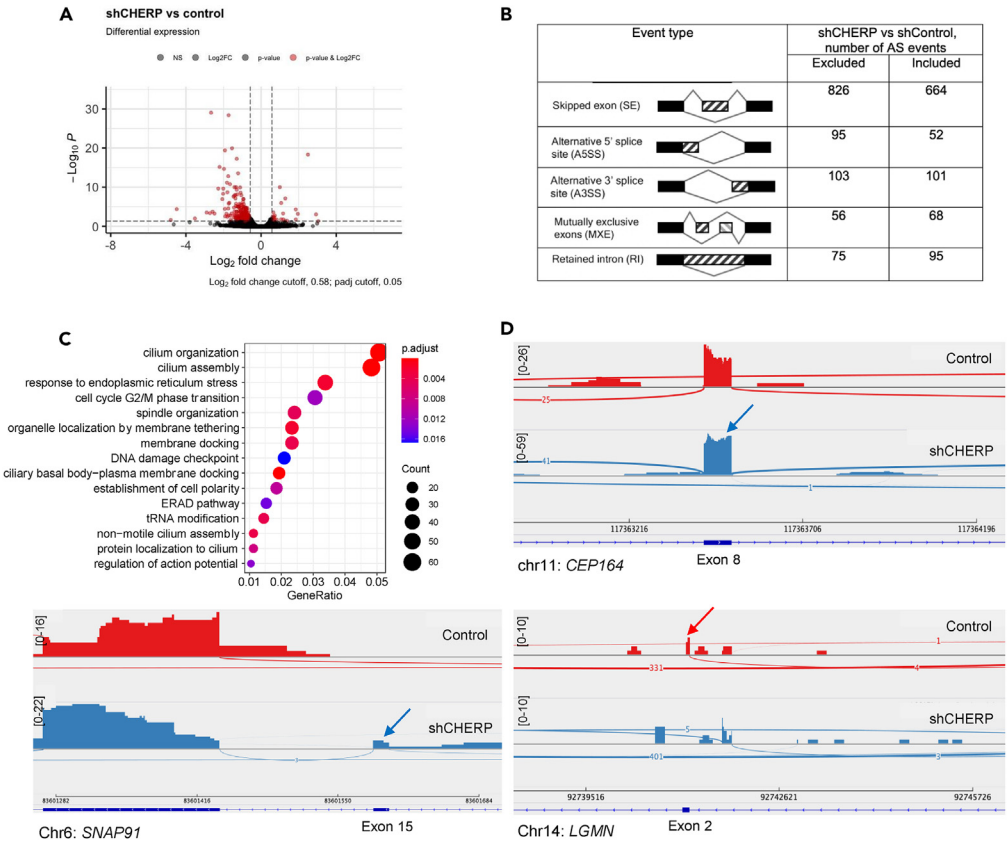


Figure 6. Whole-transcriptome effects of CHERP knockdown on gene expression and alternative splicing

(A) RNA-seq data show upregulation (red dots, *right*, log2FC > 0.58, padj <0.05) and downregulation (red dots, *left*, log2FC < −0.58, padj <0.05) in shCHERP KD compared to control. The volcano plot was built using 'Enhanced Volcano' Bioconductor package.

(B) Types and absolute numbers of annotated alternative splicing events (ASE) that were significantly different in shCHERP vs. control cells (with parameters FDR <0.05; ILD, inclusion level difference, ≥10%). In event type illustration, the constitutive exon is black, whereas alternatively, spliced exons are striped.

(C) Gene ontology of biological processes enriched among alternatively spliced transcripts.

(D) Examples of Sashimi plots of *CEP164*, *SNAP91*, and *LGMN* genes in shCHERP compared to control: read densities for shControl and shCHERP samples are shown on the y axis. Arrow indicates the inclusion of an alternative exon.

migratory and invasive activity *in vitro* and *in vivo*, inferring the significance of targeting legumain to combat tumor invasion and metastasis.[47] These examples highlight the importance of CHERP in orchestrating alternative splicing events in transcripts that play roles in cellular growth, development, and extracellular signaling.

DISCUSSION

In this study, we have undertaken a detailed investigation of distinctive interactomes associated with different phosphorylation states of RNA polymerase II during transcription. We specifically focused on pSer5 and pSer2, which represent the dominant phosphorylation states at the beginning and end of the transcription process, respectively. Given the critical role of the CTD in regulating transcription, we hypothesized that different proteins would be recruited to the transcription machinery through their interaction with the post-translational modifications of the CTD. *In vitro* phosphorylation is used to generate a hyperphosphorylated species to improve signal/noise. To test this, we accurately mimicked the phosphorylation states of the CTD, confirmed by mass spectrometry, and conducted a comprehensive proteomic investigation. Our results reveal that while some proteins can bind to both phosphorylation forms of the CTD, there are also proteins that specifically associate with either of the two different phosphorylation states. Our method differs from previous studies using phosphoryl-specific antibodies to pull down CTD-binding proteins. Since the pulldown phosphorylation RNA polymerase II is heterogeneously phosphorylated, the identified interactomes include all associated transcription regulators. In contrast, our method emphasizes more on proteins differentially recruited by different phosphoryl-CTD species. For example, CHERP identified in this study does not show any interaction with pSer5 CTD. The identified proteins are consistent with the expected role of the CTD that provides the spatiotemporal recruitment of transcription regulators at different stages of transcription.

Our study focused on the characterization of a previously understudied RNA binding protein known as CHERP for its interaction with the CTD *in vitro* and cellular contexts. Our investigations revealed that the CID-like domain in CHERP selectively recognizes the CTD when phosphorylated at Ser2 or Thr4 but not in unphosphorylated or phosphorylated at Ser5. We established the genome-wide distribution of CHERP, whose localization depends on the Ser2 phosphorylation and CID-like domain. Our findings underscore the recruitment of CHERP by Pol II in the accurate identification of the splicing sites, leading to alternative splicing.

Tumor cells frequently exploit the alternative splicing pathways to promote cell proliferation and escape apoptosis.[48] Mutations found in core or accessory splicing components have been observed in many cancer types, promoting the isoform amplification for active tumor suppression.[49,50] Abnormal splicing outcomes have been identified as biomarkers for diagnosis and a new target for treatment.[51,52] Recent research has identified CHERP as part of a stable complex with another two splicing factors, U2SURP and RBM17.[40] This complex and its associated alternative splicing events are implicated in colorectal tumorigenesis development.[22] The combination of our recent findings with previous data provides strong evidence implicating the CHERP/U2SURP/RBM17 complex in splicing events of numerous transcription factors driving tumorigenesis. Analysis of data from The Cancer Genome Atlas (TCGA) indicates that CHERP missense mutations are prevalent in 8–10% of endometrial cancer, melanoma, and cervical cancers. Interestingly, many of the known mutations in CHERP map to the CID, with the majority of them being missense mutations.[53,54] Mutations of U2SURP are also prevalent in these tumors. Since the three proteins form a stable complex and the lack of any leads to complex degradation,[40] the molecular mechanism by which the CHERP/U2SURP/RBM17 complex affect splicing outcomes is an area of ongoing investigation, with the aim of providing insight into the pathogenesis of these cancers.

We have developed the mass spectrometry method to pinpoint the exact phosphorylation sites in the CTD of RNA polymerase II. Armed with the collection of kinases and phosphatases, we can recapitulate the dynamic changes of PTM patterns during transcription and dissect the proteins recruited to Pol II at each stage. This rigorous strategy will give us the complete picture of eukaryotic transcription progression. Overall, our findings shed light on the precise molecular mechanisms that underlie the regulation of transcription and provide new insights into the complex interplay between post-translational modifications of the CTD and the recruitment of specific transcriptional regulators.

Limitations of the study

Here, we demonstrate that pSer2 and pSer5 marks recruit different transcription regulators. To narrow the scope of our study, we employed an in vitro approach that augmented the specific addition of phosphoryl marks on Ser2 and Ser5 and allowed recognition of direct binding partners of the CTD of Pol II. While the CTD in eukaryotic transcription consists of a mixture of Phospho-MARKs, there is clear dominance of pSer5 at initiation and pSer2 during elongation and termination. It is possible that some transcription factors prefer and strongly bind to a specific phosphopattern consisting of a combination of marks on the CTD that is not recapitulated in our study.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Cell lines
- METHOD DETAILS
  - shRNA transfection
  - Sequence alignment and constructs
  - Protein expression and purification
  - Differential scanning fluorimetry
  - Western blot
  - Co-immunoprecipitation
  - Immunofluorescence
  - Kinase activity assays
  - Phosphorylation sample preparation for UVPD MS/MS
  - UVPD tandem mass spectrometry and data analysis
  - Structural prediction and modeling
  - Label-free proteomics sample preparation and CTD affinity purification
  - Proteomics mass spectrometry and protein identification
  - Biolayer interferometry
  - Fluorescence polarization
  - RT-qPCR

- RNA isolation, library preparation, and RNA-Sequencing
- Analyses of RNA-seq data and alternative splicing events (ASE)
- Chromatin immunoprecipitation (ChIP) and ChIP-*Sequencing*
- Analyses of ChIP-Seq data

- QUANTIFICATION AND STATISTICAL ANALYSES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107581.

ACKNOWLEDGMENTS

This work is supported by grants from the National Institutes of Health (R01GM104896 to YJZ, R01GM125882 to YJZ and JSB, R01GM121714, and R35GM139658 to JSB, R35 GM148356 to YJZ and R35 GM122480 to EMM) and Welch Foundation (F-1515 to EMM and F-1155 to JSB), with additional support from Army Research Office grant W911NF-12-1-0390 to EMM. Immunofluorescence experiments were performed at the Center for Biomedical Research Support Microscopy and Imaging Facility at UT Austin (RRID:SCR_021756). Proteomics data was acquired in the UT Austin Center for Biomedical Research Support Biological Mass Spectrometry Facility (RRID:SCR_021728). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Robert A. Welch Foundation, Army Research Office, or the National Institutes of Health.

AUTHOR CONTRIBUTIONS

R.Y.M. and Y.J.Z. designed the experiment; K.J.J., J.P.B, and J.S.B. analyzed UVPD-MS; S.B.P. conducted bioinformatic analysis of RNA-seq; B.M.F. and E.M.M. carried out the proteomic analysis; M.K.V.R. helped with pulldown experiment. R.Y.M. conducted the rest of the experiments. The manuscript was written by R.Y.M. and Y.J.Z. with contributions from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support the inclusive, diverse, and equitable conduct of research.

Received: May 11, 2023
Revised: June 28, 2023
Accepted: August 4, 2023
Published: August 9, 2023

REFERENCES

1. Corden, J.L. (2013). RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. Chem. Rev. *113*, 8423–8455. https://doi.org/10.1021/cr400158h.
2. Jeronimo, C., Collin, P., and Robert, F. (2016). The RNA Polymerase II CTD: The Increasing Complexity of a Low-Complexity Protein Domain. J. Mol. Biol. *428*, 2607–2622. https://doi.org/10.1016/j.jmb.2016.02.006.
3. Zhang, J., and Corden, J.L. (1991). Identification of phosphorylation sites in the repetitive carboxyl-terminal domain of the mouse RNA polymerase II largest subunit. J. Biol. Chem. *266*, 2290–2296. https://doi.org/10.1016/S0021-9258(18)52242-0.
4. Komarnitsky, P., Cho, E.J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev. *14*, 2452–2460. https://doi.org/10.1101/gad.824700.
5. Yuryev, A., Patturajan, M., Litingtung, Y., Joshi, R.V., Gentile, C., Gebara, M., and Corden, J.L. (1996). The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. Proc. Natl. Acad. Sci. USA *93*, 6975–6980.
6. Buratowski, S. (2003). The CTD code. Nat. Struct. Biol. *10*, 679–680. https://doi.org/10.1038/nsb0903-679.
7. Dahmus, M.E. (1996). Reversible phosphorylation of the C-terminal domain of RNA polymerase II. J. Biol. Chem. *271*, 19009–19012. https://doi.org/10.1074/jbc.271.32.19009.
8. Kim, Y.-J., Björklund, S., Li, Y., Sayre, M.H., and Kornberg, R.D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. Cell *77*, 599–608. https://doi.org/10.1016/0092-8674(94)90221-6.
9. Phatnani, H.P., and Greenleaf, A.L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev. *20*, 2922–2936. https://doi.org/10.1101/gad.1477006.
10. Harlen, K.M., Trotta, K.L., Smith, E.E., Mosaheb, M.M., Fuchs, S.M., and Churchman, L.S. (2016). Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. Cell Rep. *15*, 2147–2158. https://doi.org/10.1016/j.celrep.2016.05.010.
11. Hsin, J.-P., Li, W., Hoque, M., Tian, B., and Manley, J.L. (2014). RNAP II CTD tyrosine 1 performs diverse functions in vertebrate cells. Elife *3*, e02112. https://doi.org/10.7554/eLife.02112.
12. Hintermair, C., Voß, K., Forné, I., Heidemann, M., Flatley, A., Kremmer, E., Imhof, A., and Eick, D. (2016). Specific threonine-4 phosphorylation and function of RNA polymerase II CTD during M phase progression. Sci. Rep. *6*, 27401. https://doi.org/10.1038/srep27401.
13. Egloff, S. (2012). Role of Ser7 phosphorylation of the CTD during transcription of snRNA genes. RNA Biol. *9*, 1033–1038. https://doi.org/10.4161/rna.21166.
14. Harlen, K.M., and Churchman, L.S. (2017). The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. Nat. Rev. Mol. Cell Biol. *18*, 263–273. https://doi.org/10.1038/nrm.2017.10.
15. Phatnani, H.P., Jones, J.C., and Greenleaf, A.L. (2004). Expanding the functional repertoire of CTD kinase I and RNA polymerase II: novel

phosphoCTD-associating proteins in the yeast proteome. Biochemistry *43*, 15702–15719. https://doi.org/10.1021/bi048364h.

16. Ebmeier, C.C., Erickson, B., Allen, B.L., Allen, M.A., Kim, H., Fong, N., Jacobsen, J.R., Liang, K., Shilatifard, A., Dowell, R.D., et al. (2017). Human TFIIH Kinase CDK7 Regulates Transcription-Associated Chromatin Modifications. Cell Rep. *20*, 1173–1186. https://doi.org/10.1016/j.celrep.2017.07.021.
17. Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001). Structural Basis of Transcription: RNA Polymerase II at 2.8 Ångstrom Resolution. Science *292*, 1863–1876. https://doi.org/10.1126/science.1059493.
18. Czudnochowski, N., Bösken, C.A., and Geyer, M. (2012). Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. Nat. Commun. *3*, 842. https://doi.org/10.1038/ncomms1846.
19. Mayfield, J.E., Irani, S., Escobar, E.E., Zhang, Z., Burkholder, N.T., Robinson, M.R., Mehaffey, M.R., Sipe, S.N., Yang, W., Prescott, N.A., et al. (2019). Tyr1 phosphorylation promotes phosphorylation of Ser2 on the C-terminal domain of eukaryotic RNA polymerase II by P-TEFb. Elife *8*, e48725. https://doi.org/10.7554/eLife.48725.
20. Di Vona, C., Bezdan, D., Islam, A.B.M.M.K., Salichs, E., López-Bigas, N., Ossowski, S., and de la Luna, S. (2015). Chromatin-wide profiling of DYRK1A reveals a role as a gene-specific RNA polymerase II CTD kinase. Mol. Cell *57*, 506–520. https://doi.org/10.1016/j.molcel.2014.12.026.
21. Ramani, M.K.V., Escobar, E.E., Irani, S., Mayfield, J.E., Moreno, R.Y., Butalewicz, J.P., Cotham, V.C., Wu, H., Tadros, M., Brodbelt, J.S., and Zhang, Y.J. (2020). Structural Motifs for CTD Kinase Specificity on RNA Polymerase II during Eukaryotic Transcription. ACS Chem. Biol. *15*, 2259–2272. https://doi.org/10.1021/acschembio.0c00474.
22. Wang, Q., Wang, Y., Liu, Y., Zhang, C., Luo, Y., Guo, R., Zhan, Z., Wei, N., Xie, Z., Shen, L., et al. (2019). U2-related proteins CHERP and SR140 contribute to colorectal tumorigenesis via alternative splicing regulation. Int. J. Cancer *145*, 2728–2739. https://doi.org/10.1002/ijc.32331.
23. Escobar, E.E., Venkat Ramani, M.K., Zhang, Y., and Brodbelt, J.S. (2021). Evaluating Spatiotemporal Dynamics of Phosphorylation of RNA Polymerase II Carboxy-Terminal Domain by Ultraviolet Photodissociation Mass Spectrometry. J. Am. Chem. Soc. *143*, 8488–8498. https://doi.org/10.1021/jacs.1c03321.
24. Mayfield, J.E., Robinson, M.R., Cotham, V.C., Irani, S., Matthews, W.L., Ram, A., Gilmour, D.S., Cannon, J.R., Zhang, Y.J., and Brodbelt, J.S. (2017). Mapping the Phosphorylation Pattern of Drosophila melanogaster RNA Polymerase II Carboxyl-Terminal Domain Using Ultraviolet Photodissociation Mass Spectrometry. ACS Chem. Biol. *12*, 153–162. https://doi.org/10.1021/acschembio.6b00729.
25. Zhou, C.Y., Johnson, S.L., Gamarra, N.I., and Narlikar, G.J. (2016). Mechanisms of ATP-Dependent Chromatin Remodeling Motors. Annu. Rev. Biophys. *45*, 153–181. https://doi.org/10.1146/annurev-biophys-051013-022819.
26. Karakasili, E., Burkert-Kautzsch, C., Kieser, A., and Sträßer, K. (2014). Degradation of DNA damage-independently stalled RNA polymerase II is independent of the E3 ligase Elc1. Nucleic Acids Res. *42*, 10503–10515. https://doi.org/10.1093/nar/gku731.
27. Schneider, S., Pei, Y., Shuman, S., and Schwer, B. (2010). Separable functions of the fission yeast Spt5 carboxyl-terminal domain (CTD) in capping enzyme binding and transcription elongation overlap with those of the RNA polymerase II CTD. Mol. Cell Biol. *30*, 2353–2364. https://doi.org/10.1128/MCB.00116-10.
28. Ho, C.K., and Shuman, S. (1999). Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. Mol. Cell *3*, 405–411.
29. Appel, L.M., Franke, V., Bruno, M., Grishkovskaya, I., Kasiliauskaite, A., Kaufmann, T., Schoeberl, U.E., Puchinger, M.G., Kostrhon, S., Ebenwaldner, C., et al. (2021). PHF3 regulates neuronal gene expression through the Pol II CTD reader domain SPOC. Nat. Commun. *12*, 6078. https://doi.org/10.1038/s41467-021-26360-2.
30. Meinhart, A., and Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. Nature *430*, 223–226. https://doi.org/10.1038/nature02679.
31. Lee, J.-Y., Fan, C.-C., Chou, N.-L., Lin, H.-W., and Chang, M.-S. (2020). PHRF1 promotes migration and invasion by modulating ZEB1 expression. PLoS One *15*, e0236876. https://doi.org/10.1371/journal.pone.0236876.
32. Ni, Z., Xu, C., Guo, X., Hunter, G.O., Kuznetsova, O.V., Tempel, W., Marcon, E., Zhong, G., Guo, H., Kuo, W.-H.W., et al. (2014). RPRD1A and RPRD1B are human RNA polymerase II C-terminal domain scaffolds for Ser5 dephosphorylation. Nat. Struct. Mol. Biol. *21*, 686–695. https://doi.org/10.1038/nsmb.2853.
33. Gu, B., Eick, D., and Bensaude, O. (2013). CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II *in vivo*. Nucleic Acids Res. *41*, 1591–1603. https://doi.org/10.1093/nar/gks1327.
34. Cortazar, M.A., Sheridan, R.M., Erickson, B., Fong, N., Glover-Cutter, K., Brannan, K., and Bentley, D.L. (2019). Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism. Molecular Cell *76*, 896–908.e894. https://doi.org/10.1016/j.molcel.2019.09.031.
35. Vervoort, S.J., Welsh, S.A., Devlin, J.R., Barbieri, E., Knight, D.A., Offley, S., Bjelosevic, S., Costacurta, M., Todorovski, I., Kearney, C.J., et al. (2021). The PP2A-Integrator-CDK9 axis fine-tunes transcription and can be targeted therapeutically in cancer. Cell *184*, 3143–3162.e32. https://doi.org/10.1016/j.cell.2021.04.022.
36. Stein, C.B., Field, A.R., Mimoso, C.A., Zhao, C., Huang, K.-L., Wagner, E.J., and Adelman, K. (2022). Integrator endonuclease drives promoter-proximal termination at all RNA polymerase II-transcribed loci. Mol. Cell *82*, 4232–4245.e11. https://doi.org/10.1016/j.molcel.2022.10.004.
37. Chao, S.-H., and Price, D.H. (2001). Flavopiridol Inactivates P-TEFb and Blocks Most RNA Polymerase II Transcription in Vivo. J. Biol. Chem. *276*, 31793–31799. https://doi.org/10.1074/jbc.M102306200.
38. Jasnovidova, O., Klumpler, T., Kubicek, K., Kalynych, S., Plevka, P., and Stefl, R. (2017). Structure and dynamics of the RNAPII CTDsome with Rtt103. Proc. Natl. Acad. Sci. USA *114*, 11133–11138. https://doi.org/10.1073/pnas.1712450114.
39. Tian, B., Yang, J., and Brasier, A.R. (2012). Two-Step Cross-linking for Analysis of Protein–Chromatin Interactions. In Transcriptional Regulation: Methods and Protocols, A. Vancura, ed. (Springer New York), pp. 105–120. https://doi.org/10.1007/978-1-61779-376-9_7.
40. De Maio, A., Yalamanchili, H.K., Adamski, C.J., Gennarino, V.A., Liu, Z., Qin, J., Jung, S.Y., Richman, R., Orr, H., and Zoghbi, H.Y. (2018). RBM17 Interacts with U2SURP and CHERP to Regulate Expression and Splicing of RNA-Processing Proteins. Cell Rep. *25*, 726–736.e7. https://doi.org/10.1016/j.celrep.2018.09.041.
41. Shen, S., Park, J.W., Lu, Z.X., Lin, L., Henry, M.D., Wu, Y.N., Zhou, Q., and Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. Proc. Natl. Acad. Sci. USA *111*, E5593–E5601. https://doi.org/10.1073/pnas.1419161111.
42. Higgins, M., Obaidi, I., and McMorrow, T. (2019). Primary cilia and their role in cancer. Oncol. Lett. *17*, 3041–3047. https://doi.org/10.3892/ol.2019.9942.
43. Chaki, M., Airik, R., Ghosh, A.K., Giles, R.H., Chen, R., Slaats, G.G., Wang, H., Hurd, T.W., Zhou, W., Cluckey, A., et al. (2012). Exome capture reveals ZNF423 and CEP164 mutations, linking renal ciliopathies to DNA damage response signaling. Cell *150*, 533–548. https://doi.org/10.1016/j.cell.2012.06.028.
44. Kobayashi, T., Tanaka, K., Mashima, Y., Shoda, A., Tokuda, M., and Itoh, H. (2020). CEP164 Deficiency Causes Hyperproliferation of Pancreatic Cancer Cells. Front. Cell Dev. Biol. *8*, 587691. https://doi.org/10.3389/fcell.2020.587691.
45. Rademakers, G., Massen, M., Koch, A., Draht, M.X., Buekers, N., Wouters, K.A.D., Vaes, N., De Meyer, T., Carvalho, B., Meijer, G.A., et al. (2021). Identification of DNA methylation markers for early detection of CRC indicates a role for nervous system-related genes in CRC. Clin. Epigenetics *13*, 80. https://doi.org/10.1186/s13148-021-01067-9.
46. Zhen, Y., Chunlei, G., Wenzhi, S., Shuangtao, Z., Na, L., Rongrong, W., Xiaohe, L., Haiying, N., Dehong, L., Shan, J., et al. (2015). Clinicopathologic significance of legumain overexpression in cancer: a systematic review and meta-analysis. Sci. Rep. *5*, 16599. https://doi.org/10.1038/srep16599.
47. Liu, C., Sun, C., Huang, H., Janda, K., and Edgington, T. (2003). Overexpression of legumain in tumors is significant for invasion/metastasis and a candidate enzymatic target for prodrug therapy. Cancer Res. *63*, 2957–2964.
48. Kahles, A., Lehmann, K.-V., Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C.; Cancer Genome Atlas Research Network, and Rätsch, G. (2018). Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. Cancer Cell *34*, 211–224.e6. https://doi.org/10.1016/j.ccell.2018.07.001.
49. Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous Mutations Frequently Act as Driver Mutations in Human Cancers. Cell *156*, 1324–1335. https://doi.org/10.1016/j.cell.2014.01.051.

50. Jung, H., Lee, D., Lee, J., Park, D., Kim, Y.J., Park, W.-Y., Hong, D., Park, P.J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat. Genet. *47*, 1242–1248. https://doi.org/10.1038/ng.3414.
51. Kim, Y.-J., and Kim, H.-S. (2012). Alternative Splicing and Its Impact as a Cancer Diagnostic Marker. Genomics Inform. *10*, 74–80. https://doi.org/10.5808/GI.2012.10.2.74.
52. Bonnal, S., Vigevani, L., and Valcárcel, J. (2012). The spliceosome as a target of novel antitumour drugs. Nat. Rev. Drug Discov. *11*, 847–859. https://doi.org/10.1038/nrd3823.
53. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404. https://doi.org/10.1158/2159-8290.Cd-12-0095.
54. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1. https://doi.org/10.1126/scisignal.2004088.
55. Stauffer, W., Sheng, H., and Lim, H.N. (2018). EzColocalization: An ImageJ plugin for visualizing and measuring colocalization in cells and organisms. Sci. Rep. *8*, 15764. https://doi.org/10.1038/s41598-018-33592-8.
56. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.
57. Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat. Protoc. *7*, 1728–1740. https://doi.org/10.1038/nprot.2012.101.
58. Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. *44*, W160–W165. https://doi.org/10.1093/nar/gkw257.
59. Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360. https://doi.org/10.1038/nmeth.3317.
60. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.
61. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.
62. Klein, D.R., Holden, D.D., and Brodbelt, J.S. (2016). Shotgun Analysis of Rough-Type Lipopolysaccharides Using Ultraviolet Photodissociation Mass Spectrometry. Anal. Chem. *88*, 1044–1051. https://doi.org/10.1021/acs.analchem.5b04218.
63. Shivakumar, D., Williams, J., Wu, Y., Damm, W., Shelley, J., and Sherman, W. (2010). Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. J. Chem. Theor. Comput. *6*, 1509–1519. https://doi.org/10.1021/ct900587b.
64. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.
65. Floyd, B.M., Drew, K., and Marcotte, E.M. (2021). Systematic Identification of Protein Phosphorylation-Mediated Interactions. J. Proteome Res. *20*, 1359–1370. https://doi.org/10.1021/acs.jproteome.0c00750.
66. Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. *46*, W537–W544. https://doi.org/10.1093/nar/gky379.
67. Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS A J. Integr. Biol. *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-beta tubulin (rabbit polyclonal)	Abcam	Cat#ab6046, RRID: AB_2210370
Anti-pSer2 (Clone 3E10) (rat monoclonal)	Millipore	Cat#04-1571, RRID:AB_11212363
Anti-pSer5 (Clone 3E8) (rat monoclonal)	Millipore	Cat#04-1572, RRID:AB_10615822
Anti-POLR2C (rabbit monoclonal)	Millipore	Cat#ab182150
Anti-pThr4 (rat monoclonal)	Active Motif	Cat# 61361, RRID: AB_2750848
Anti-CHERP (rabbit polyclonal)	Thermofischer Scientific	Cat# A304-621A; RRID:AB_2620816
Anti-HA (rabbit monoclonal)	Cell Signaling	Cat#3724S
Goat anti-rat IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	Thermofischer Scientific	cat#:A-11006; RRID:AB_2534074
Goat anti-rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 568	Thermofischer Scientific	Cat# A-11011; RRID:AB_143157
Chemicals, peptides, and recombinant proteins		
CHERP CID synthetic gene	Biomatik	N/A
S7K, S7R, S7E, S7Q, 4X WT CTD	IDT	N/A
CHERP cDNA	DNASU	HsCD00879118
RPB1 cDNA	Addgene	Cat #75284
TFIIH (Cdk7/CyclinH/MAT1 (CAK complex))	Millipore	Cat #14-476
P-TEFb (Cdk9/Cyclin T1)	Millipore	Cat #14-685
DYRK1A	Addgene	Cat #79690
Biotinylated CTD peptides	Biomatik	N/A
FITC-CTD peptides	Biomatik	N/A
Flavopiridol	Selleck Chemicals	Cat# S1230
Ribonuclease A	VWR lifesciences	CAS# 9001-99-4
Proteinase K	Ambion	Cat #2542
Glycogen	Thermofischer scientific	Cat#R0561
16% Formaldehyde solution (w/v), Methanol-free	Thermo scientific	Ref # 28908
Critical commerical assays		
NEBNext Ultra II DNA Library Prep kit for Illumina	NEB	E7645S
NEBNext Multiplex Oligos for Illumina (index primers set 1)	NEB	E7335S
Kinase Glo Luminscent Kinase Assay	Promega	V6711
Deposited data		
ChIP-seq	GEO	GSE226908
RNA-seq	GEO	GSE221328
Proteomics	PRIDE	PXD039903
Gels	Mendeley	DOI: https://data.mendeley.com/datasets/9mb9vgsw3n/draft?a=08fc1869-91af-49f8-960a-f22c413fde53

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental models: Cell lines		
HEK293T	ATCC	RRID:CVCL_0045
HEK293	ATCC	RRID:CVCL_0063
Oligonucleotides		
qPCR primers	IDT	Table S4
Software and algorithms		
Image J	NIH	https://imagej.nih.gov/ij/download.html
EzColocalization	Stauffer et al.[55]	https://github.com/DrHanLim/EzColocalization
Bowtie2	Langmead et al.[56]	https://github.com/BenLangmead/bowtie2
MACS2	Feng et al.[57]	https://pypi.org/project/MACS2/
Rstudio	R Core	https://www.r-project.org/
IGV	Broad Institute	https://software.broadinstitute.org/software/igv/
Deep-tools	Ramirez et al.[58]	https://deeptools.readthedocs.io/en/develop/index.html
rMATS turbo	Shen et al.[41]	https://github.com/Xinglab/rmats-turbo
HISAT2	Kim et al.[59]	http://daehwankimlab.github.io/hisat2/
DESeq2	Love et al.[60]	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
TrimGalore!	Babahram Bioinformatics	https://github.com/FelixKrueger/TrimGalore
featureCounts	Liao et al.[61]	https://subread.sourceforge.net/
ProSight Lite	Proteomics Center of Excellence Northwestern University	http://prosightlite.northwestern.edu/
Xtract algorithm	ThermoFischer Scientific	https://www.thermofisher.com/order/catalog/product/IQLAAEGABSFANOMBAQ
Proteome Discoverer	ThermoFischer Scientific	https://www.thermofisher.com/us/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/multi-omics-data-analysis/proteome-discoverer-software.html
Other		
Dynabeads Protein A	ThermoFischer Scientific	Cat# 10001D
AMPure XP beads	Beckman Coulter	Cat# A63881
Vivaspin	Sartorius	Cat#: VS2002
Ni-NTA	Qiagen	Cat#: 30210
DirectZol RNA Miniprep kit	Zymo Research	Cat#: R2050

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to the Lead Contact, Dr. Yan Jessie Zhang (jzhang@cm.utexas.edu)

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The RNA-seq data produced by this study is available under GEO: GSE221328. The mass spectrometry proteomics data generated in this study has been deposited in the ProteomeXchange Consortium under PRIDE:PXD039903. The ChIP-seq data produced by this

study are available under GEO: GSE226908. Any unprocessed western blots were deposited under Mendeley Data: 10.17632/9mb9vgsw3n.

- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cell lines

Human embryonic kidney cells (HEK293) were purchased from ATCC (Manassas, VA, USA). Cells were routinely cultured in Dulbecco's modified Eagle's media (Sigma-Aldrich, St. Louis, MO, USA, product number #D6429), supplemented with 10% Opti-Gold fetal bovine serum (GenDEPOT, Katy, TX, USA) at 37°C in humidified atmosphere with 5% CO_2. HyClone penicillin and streptomycin mix (Cytiva, Marlborough, MA, USA) were added to the media to reach a final concentration of 1%. Cell lines have not been authenticated by STR genotyping. Cell lines were tested for mycoplasma contamination prior to use.

METHOD DETAILS

shRNA transfection

HEK293 cells were infected at a multiplicity of infection 1 using MISSION shRNA lentiviral particles (Sigma, clone: TRCN0000053624) against CHERP. Hexadimethrine bromide was added to the cells at a final concentration of 8 μg/ml. Transduced cells were selected with puromycin at a concentration of 1 μg/ml for 7 days. Parallelly, the control cells were transfected with MISSION non-mammalian shRNA negative control plasmid (Sigma, Cat: SHC002) using Fugene (Promega, Madison, WI, USA) with a DNA to Fugene ratio of 1:3 for the same duration of time.

Sequence alignment and constructs

The sequences of CID-containing proteins were obtained from NCBI (RPRD2-Q5VT52, RPRD1A-Q96P16, RPRD1B-Q9NQG5, Scaf4-O95104, Scaf8-Q9UPN6, CHERP-Q8IWX8). The sequences were aligned in Jalview using ClustalO, and visualization of the alignment was done with ESPript 3.

CTD constructs were cloned using ligation-independent cloning with varying lengths of CTD heptad gene block inserted. The CHERP CID domain (encoding residues 105-328) was ordered as a synthetic gene and was subcloned into a pET28a (Novogene, Sacramento, CA, USA) derivative vector encoding a 6xHis-tag followed by a GST-tag and a 3C protease site. The DYRK1A kinase domain (127-485) was obtained from Addgene. The full-length CHERP cDNA (clone: HsCD00879118) encoding residues 1-916 were cloned into a mammalian expression vector containing a CMV promoter and an N-terminal HA tag.

Protein expression and purification

For protein expression, BL21 (DE3) cells expressing CHERP, DYRK1A, or GST-CTD substrates were grown in one-liter cultures at 37°C in Luria-Bertani (LB) broth (Thermo Scientific, Waltham, MA, USA) containing 50μg/ml kanamycin. Once the cultures reached an OD 600 value of 0.6-0.8, the protein expression was induced with 0.25mM isopropyl-β-D-thiogalactopyranoside (IPTG), and the cultures were grown an additional 16h at 18°C. The cells were pelleted and resuspended in lysis buffer (50mM Tris-HCl pH 8.0, 500mM NaCl, 15mM imidazole, 10% glycerol, 0.1% Triton X-100, and 10mM 2-mercaptoethanol (BME)) and sonicated at 90 A for 2.5min of 1 s on/5 s off cycles on ice. The lysate was cleared by centrifugation at 15000 rpm for 45 min at 4°C. The supernatant was loaded over 3ml of Ni-NTA beads (Qiagen, Germany) equilibrated in lysis buffer, then washed through with wash buffer containing 50mM Tris-HCl pH 8.0, 500mM NaCl, 30mM imidazole, and 10mM BME. The recombinant protein was eluted with buffer containing 50mM Tris-HCl pH 8.0, 500mM NaCl,300mM imidazole, and 10mM BME. Protein fractions were pooled and dialyzed overnight at 4°C in a 10.0 kDa dialysis membrane (Thermo Scientific) against dialysis buffer (50mM Tris HCl pH 7.5,100mM NaCl, and 10mM BME). The protein was polished using gel filtration chromatography and loaded onto a Superdex 75 size exclusion column (GE) in gel filtration buffer. Peak fractions were analyzed by SDS-PAGE before the fractions were pooled, concentrated, and flash-frozen at -80°C.

Differential scanning fluorimetry

Purified recombinant CHERP CID domain at a final concentration of 5μM was incubated with 10X SYPRO Orange (Molecular Probes) in a 96-well low-profile PCR plate (ABgene, Thermo Scientific), and fluorescence was captured in a LightCycler 480 (Roche). Protein melting curves were carried out with a temperature acquisition mode using a total of 10 acquisitions per 1°C in each cycle from 20°C to 95°C. The melting temperature was derived using the Boltzmann equation.

Western blot

Cells were lysed in RIPA lysis buffer (50 mM Tris-Cl pH 8.0, 150 mM NaCl, NP-40, 0.5% sodium deoxycholate, 0.1% SDS) and 1× protease inhibitor cocktail (Roche, Indianapolis, IN, USA). Protein concentrations were quantified with the Bradford protein assay. Briefly, 25 μg of protein extracts were loaded and separated by SDS-PAGE gels. Blotting was performed with standard protocols using a PVDF membrane

(Bio-Rad, Hercules, CA, USA). Membranes were blocked for 1 h in blocking buffer (5% BSA in PBST) and probed with primary antibodies at 1:1,000 dilution at 4 °C overnight. After three washes with PBST, the membranes were incubated with diluted goat anti-rabbit secondary IR-Dye 680RD antibody at 1:10,000 (LI-COR, Lincoln, NE, USA) for 1 h at room temperature. After washing, membranes were visualized on LI-COR Odyssey CLx image reader. All antibodies used for immunoblotting are listed in the supplementary section.

Co-immunoprecipitation

Cellular extracts were prepared by incubating cells with lysis buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.5% NP-40, 1 mM PMSF, and 1× protease inhibitor) for 30 min on ice. Supernatants were collected by centrifugation at 12,000 x g for 20 min at 4°C. For immunoprecipitation, Dynabeads Protein G (20 µl, Invitrogen) was incubated with 3 µg of antibody or 3 µg of control IgG overnight at 4°C with rotation. Subsequently, 250 µg of protein was incubated with the antibody-bound beads for an additional 2 hours and washed three times with lysis buffer. The precipitated proteins were eluted from the beads with 2× SDS loading buffer and boiled for 5 min, followed by western blot analyses. At least three independent replicates of each IP experiment were performed.

Immunofluorescence

In brief, HEK293 cells were transfected with HA-CHERP or HA-CHERPΔCID using Fugene (1:3 plasmid to reagent ratio) to overexpress the protein of interest. Cells were washed with PBS and fixed in 1% formaldehyde for 15 min at room temperature. For pSer2 inhibition, cells were treated with 2µM flavopiridol (Selleckchem, Houston, TX, USA) for 3 hours, then fixed with formaldehyde. Cells were permeabilized with 0.2% Triton X-100 to allow antibody labeling. Subsequently, the samples were blocked with 2% BSA for 30 min and incubated with primary antibody for 1 h at room temperature. After washing with PBS, the cells were stained with secondary antibody (Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody Alexa Fluor 488 or Goat anti-Rat IgG (H+L) Cross-Adsorbed Secondary Antibody Alexa Fluor 568, Thermo Scientific) for 1 h at room temperature. Cells were counterstained with DAPI for nuclear visualization, and coverslips were mounted with antifade fluorescent mounting media (Abcam, cat#: ab104135). Standard fluorescence images were captured using a confocal microscope (Zeiss LSM 710). Confocal images were acquired with the Plan-Apo 63x oil immersion lens and analyzed using the Zen/ImageJ program. Quantification of colocalization was done on ImageJ using the EzColocalization plugin.[55] Coefficients were calculated using the Pearson correlation threshold. Box and whiskers plots were generated from N=7 cells of at least two biological replicates.

Kinase activity assays

The DYRK1A and TFIIH kinetic activity assay were performed in a 25µl reaction volume containing 0-100µM of the substrate (GST-yCTD 26x) in a reaction buffer containing 50mM Tris at pH 8.0 and 20mM $MgCl_2$. The reaction was initiated by adding 1µM of kinase and incubating at 30°C for 10 minutes before being quenched with 25µl of water and 50µl of room temperature Kinase-Glo Detection Reagent (Promega). The mixtures were incubated at room temperature for 10 minutes with the reagent before obtaining luminescence readings in a Tecan plate reader 200. The readings were translated to ATP concentration using an ATP standard curve determined with Kinase-Glo Detection Reagent. Kinetic data were obtained in a triplicate fashion and fitted to the Michaelis-Menten kinetic equation to obtain respective kinetic parameters k_{cat} (min^{-1}) and K_m (µM) in GraphPad Prism 9.

Phosphorylation sample preparation for UVPD MS/MS

Kinase reactions were performed in a buffer containing 2mM ATP, 50mM Tris pH 8.0, and 10mM $MgCl_2$ and supplemented with 1 mg/ml of CTD substrate for 15h. Reactions were initiated by adding either 0.2µM CDK7/CycH/MAT1 (Proqinase, Cat:#0366-0360-4) or PTEFb or 0.6µM DYRK1A. The reaction time was optimized so that no further phosphorylation occurred on the substrate. Reactions were quenched with the addition of 10mM EDTA. All samples were digested with 3C-protease to cleave the tag at a molar ratio of 100:1 protein/protease.

UVPD tandem mass spectrometry and data analysis

Peptides were desalted with Pierce C_{18} spin columns according to the manufacturer's instructions and eluted with water/acetonitrile (30:70, v/v). The solvent was evaporated, and the peptides were reconstituted in water/acetonitrile/formic acid (98:2:0.1, v/v/v) before liquid chromatography analysis. Peptides were separated using a Dionex Ultimate 3000 nano liquid chromatography system (Thermo Scientific) plumbed for direct injection onto a 20 cm C_{18} (1.8 µm, 300 A pore size, 75 µm ID) Picofrit analytical column (New Objective, Woburn, MA). Mobile phases A and were comprised of HPLC-grade water and acetonitrile, respectively, each containing 0.1% formic acid. Separations were carried out with a flow rate of 0.300 µl/min using a linear gradient from 2 to 55% B over 40 min.

Eluted peptides were analyzed with an Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Scientific, San Jose, CA, USA) equipped with a Coherent ExciStar XS excimer laser operating at 193 nm as described previously.[62] In brief, UVPD was implemented by adding a fused-silica window at the back end of the dual linear ion trap to allow entrance of the laser beam into the high-pressure trap of the instrument. A periscope and appropriate mirrors were used to guide the laser beam through the high-pressure ion trap, and a pulse generator and customized instrument code were used to trigger the laser during the time of ion trapping. UVPD mass spectra were acquired using two pulses at 3mJ. All spectra were acquired using resolution settings of 60 and 30K (at *m/z* 200) for MS1 and MS/MS events, respectively.

MS/MS spectra were deconvoluted using the Xtract algorithm with a signal-to-noise threshold of 3. Fragments were matched to the nine ion types (*a, a+1, b, c, x, x+1, y, y–1, z*) observed from UVPD of peptides using ProSight Lite with a 10-ppm error tolerance. Phosphate

localization was performed by adding the mass of a phosphoryl group (+79.97 Da) at each of the possible Tyr or Ser residues to identify fragment ions that were phosphorylated.

Structural prediction and modeling

The substrate-bound configuration of DYRK1A (PDB code: 2WO6) was used initially to model CTD binding. The model containing the CTD substrate was optimized with Maestro (Schrodinger, LLC), which utilizes a simple minimization routine based on the OPLS_2005 Forcefield.[63] Virtual mutation of the substrate ligand to the CTD sequence was done in PyMOL and fit the likeliest rotamer configuration. PyMOL was used to prepare all graphical illustrations for protein structures.

The superimposition of the CID of CHERP with other CID proteins was done by using the PDB file generated by Alphafold.[64] The CID domain region of CHERP has a very high confidence score (pLDDT > 90) as dictated by the Alphafold algorithm. The Alphafold structure was then superimposed with the complex structure of pSer2 CTD peptide and the CID of RPRD1B (PDB: 4Q96) to visualize the predicted binding pocket of the CID of CHERP and the CTD.

Label-free proteomics sample preparation and CTD affinity purification

0.6μM DYRK1A and 0.2 μM CDK7/CycH/MAT1 were used to phosphorylate 1mg/ml of the 26x yeast GST-CTD substrate in a 100μl reaction for 15 hours. Glutathione Agarose beads were washed in Buffer C (20mM Tris pH 8.0, 150mM NaCl, 10mM BME) thrice, and the treated GST-CTD samples were added to the beads and incubated overnight. 200 million HEK293 cells were grown, collected, and the cell pellet was resuspended in Buffer A (10mM HEPES pH 7.4, 100mM NaCl, 300 mM Sucrose, 3mM $MgCl_2$, 0.5% Triton X-100, 1:100 Protein, and Phosphatase Inhibitor). Cells were then vortexed, incubated on ice for 15 minutes, and centrifuged at 15,000 x g for 10 minutes at 4°C. The supernatant is discarded, and the cell pellet is resuspended in buffer B (10mM Tris pH 8.0, 150mM NaCl, 1:100 PPI) supplemented with 1:1000 benzonase to remove protein:DNA interactions. This mixture was incubated at room temperature for 1 hour and centrifuged at 15,000 x g for 10 minutes. The supernatant was collected as the nuclear fraction. After overnight incubation, the GST-CTD bound beads were washed twice with buffer C and once with buffer B. The nuclear fraction was added to the substrate-bound beads and incubated at 4°C overnight. Then the beads were centrifuged at 4,000 x g for 2 minutes at 4°C. The beads were washed twice with low salt buffer (20mM Tris pH 8.0, 150mM NaCl, 10% glycerol, 0.1% Triton X-100, and 1:100 PPI) for 5 minutes per wash and thrice with high salt buffer (20mM Tris pH 8.0, 500mM NaCl, 10% glycerol, 0.1% Triton X-100, and 1:100 PPI). To the beads, 100μl of elution buffer was added and spun at 4°C for 2 hours. Then the beads were centrifuged at 4,000 x g for 2 minutes at 4°C, and the supernatant was collected for the pulldown.

Pulldown samples were exchanged into 5 mM Tris-HCl using 3 kDa Amicon filters. Samples were then denatured in 2,2,2-trifluoroethanol (TFE) and 5 mM tris(2-carboxyethyl)phosphine (TCEP) at 55°C for 45 min. Proteins were alkylated in the dark with 5.5 mM iodoacetamide, and the remaining iodoacetamide was quenched with 100 mM dithiothreitol (DTT). MS-grade Trypsin was then added to the solution at an enzyme: protein ratio of 1:50, and the digestion reaction was incubated at 37°C for 4 h. Trypsin was quenched by adding 10% formic acid, and the volume was reduced to 500 μL in a vacuum centrifuge. Samples were then filtered using a 10 kDa Amicon filter and desalted using Pierce C18 tips (Thermo Scientific). The samples were resuspended in 95% water, 5% acetonitrile, and 0.1% formic acid prior to MS.

Proteomics mass spectrometry and protein identification

Peptides were separated on a 75 μM × 25 cm Acclaim PepMap100 C-18 column (Thermo Scientific) using a 5–50% acetonitrile + 0.1% formic acid gradient over 120 min and analyzed online by nanoelectrospray-ionization tandem MS on a Thermo Scientific Fusion Tribrid Orbitrap mass spectrometer, using a data-dependent acquisition strategy and analyzing two biological replicates per sample. Full precursor ion scans (MS1) were collected at high resolution (120,000). MS2 scans were acquired in the ion trap in rapid scan mode using the Top Speed acquisition method and fragmenting by collision-induced dissociation. Dynamic exclusion was activated with a 60 s exclusion time for ions selected more than once.

Proteins were identified with Proteome Discoverer 2.3 (Thermo Scientific), searching against the UniProt human reference proteome. Methionine oxidation [+15.995 Da], N-terminal acetylation [+42.011 Da], N-terminal methionine loss [−131.04 Da], and N-terminal methionine loss with the addition of acetylation [−89.03 Da] were all included as variable modifications. Peptides and proteins were identified using a 1% false discovery rate.

To score changes in protein abundance, a z-score was estimated between the unmodified control and the kinase-treated sample for each protein, as in.[65] To generate volcano plots, datasets from both replicates were log_2 transformed, missing values were imputed, and data were quantile normalized. Enriched proteins were defined using a p-value of 0.05. P-values in volcano plot analyses were calculated using a two-tailed, two-sample t-test.

Biolayer interferometry

Biotinylated CTD peptides (Biotin-SPSYSPTSPSYSPTSPSY, pSer5: Biotin-SPSYSPTpSPSYSPTSPSY, pSer2: Biotin-SPSYSPTSPSYpSPTSPSY, and pThr4: Biotin-SPSYSPTSPSYSPpTSPSY) were immobilized onto streptavidin sensor tips (ForteBio) using an Octet RED96e (ForteBio). The sensor tip was dipped into 100nM of CHERP CID to measure association. Then, subsequently dipped into a well containing only buffer composed of (50mM Tris pH 8.0, 150mM NaCl, 0.05% Tween-20, and 1mg/ml BSA) to measure the dissociation phase.

Fluorescence polarization

CTD peptides with double repeats were labeled with fluorescein isothiocyanate (FITC) and purchased from Biomatik. Protein and peptide concentrations were determined according to their absorbance at 280nm. Fluorescence polarization values were collected on a Tecan F200 plate reader in buffer (50mM Tris pH 8.0, 300mM NaCl, 0.005% Tween-20, and 10mM BME) at room temperature. Samples were excited with vertically polarized light at 485 nm and at an emission wavelength of 535nm. CHERP-CID protein was titrated into a reaction mixture containing buffer supplemented with 10nM of FITC-peptide. Measurements were taken in triplicates and fitted to the cubic equation applying a 1:1 binding mode to obtain Kd values using GraphPad Prism v9.

RT-qPCR

Total RNA was harvested from HEK293 cells using DirectZol RNA Miniprep kit (Zymo Research, Irvine, CA, USA, product number #R2050). cDNA was generated using AzuraQuant cDNA synthesis kit (Azura Genomics) using the manufacturer's instructions. qPCR was done using the AzuraQuant Green Fast qPCR Mix Lo-Rox (Azura Genomics) in a ViiA-7 Real-Time PCR system (Applied Biosystems). All qPCR experiments were conducted in biological triplicates, error bars represent mean ± standard error mean. Relative gene expression was assessed using the ΔΔCt method normalized to *ACTB* expression. Student's t-test was used to compare groups. The specificity of amplification was controlled with PCR product melting curves. All primers used in this study can be found in the supplementary section in Table S4.

RNA isolation, library preparation, and RNA-Sequencing

Total RNA was isolated from HEK293 cells (at least ~10^6 cells per sample) using DirectZol RNA Miniprep kit (Zymo Research). Then, mRNA was isolated from total RNA using the Poly(A) Purist-MAG kit (Thermo Scientific). Briefly, a starting amount of 1.3-1.7 ng RNA per sample was combined in a 96-well plate with washed and resuspended Oligo (dT) MagBeads. The mixture was heated in a thermocycler set to 70C for 5 - minutes, followed by incubation at room temperature with gentle vortexing for 30 minutes. Subsequently, the beads were captured and washed, and bound poly(A) RNA was eluted in water. mRNA quality was assessed on the Agilent Bioanalyzer using the Agilent RNA 6000 Pico kit (Agilent Technologies, Santa Clara, CA, USA). Libraries were prepared at the University of Texas Genomic Sequencing and Analysis Facility (GSAF) according to the manufacturer's instructions for the NEBNext Ultra II Direction RNA kit (NEB, Ipswich, MA, USA, product number #E7760). The Strandedness of the library was preserved using the dUTP method. The resulting libraries tagged with unique dual indices were checked for size and quality using the Agilent High Sensitivity DNA Kit (Agilent). Library concentrations were measured using the KAPA SYBR Fast qPCR kit and loaded for sequencing on the NovaSeq 6000 (Illumina, San Diego, CA, USA) instrument (paired-end 2X150, 100 cycles). A minimum number of reads was set to 40x10^6 per sample.

Analyses of RNA-seq data and alternative splicing events (ASE)

The quality of raw reads was assessed using FastQC read quality reports (https://usegalaxy.org).[66] Adapter Illumina sequences (–illumina) were trimmed off by TrimGalore! v.0.6.7 with default parameters. Next, reads were aligned to a human reference genome, GRCh38 version, using HISAT2 fast aligner v.2.2.1 with default parameters, except Reverse (RF) –rna-strandedness. Gencode v38 gtf file was used as annotation gtf.[59] Lastly, mapped fragments were quantified by featureCounts v.2.0.1 in Galaxy.[61] Differential expression was analyzed using DESeq2 v.1.30.1 in R; genes with adjusted p-value < 0.05 were considered as differentially expressed.[60] RNA-seq data were deposited in GEO under the accession number GSE221328. rMATS turbo v.4.1.2 in the command line was employed for the detection of alternatively spliced events upon SCAF6 loss vs. control (with parameters FDR < 0.05; ILD, inclusion level difference, ≥ 10%).[41] As input files for rMATS, we used alignment .bam files from HISAT2 mapper and gencode v38 annotation gtf. Enrichment analysis of gene clusters was performed using Bioconductor R package 'clusterProfiler' v.3.18.1.[67]

Chromatin immunoprecipitation (ChIP) and ChIP-*Sequencing*

Briefly, cells were double crosslinked with 2mM DSG for 15 min followed by secondary fixation with 1% formaldehyde for 10 min at room temperature. Crosslinking was quenched with 0.125 M glycine for 5 min. Cells were successively lysed in lysis buffer LB1 (50 mM HEPES-KOH, pH 7.5, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100, 1× PI), LB2 (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 1× PI) and LB3 (10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% *N*-lauroylsarcosine, 1× PI). Chromatin was sonicated to an average size of ~200–500 bp using UCD-200 Biorupter (30s on and 30 s off for 30 min). A total of 5 μg of HA antibody that was pre-mixed in a 50ul volume of Dynabeads protein A (Invitrogen) was added to each sonicated chromatin sample and incubated overnight at 4°C. The chromatin-bound beads were washed two times with low salt buffer (0.1% Na Deoxycholate, 1% Triton X-100, 1mM EDTA, 50mM HEPES pH 7.5, 150mM NaCl), once with high salt wash buffer (0.1% Na Deoxycholate, 1% Triton X-100, 1mM EDTA, 50mM HEPES pH 7.5, 500mM NaCl), once with LiCl wash buffer (250mM LiCl, 0.5% NP-40, 0.5% Na-Deoxycholate, 1mM EDTA, 10mM Tris-Cl pH 8.0) and twice in TE buffer. The chromatin was reverse crosslinked overnight at 65°C with shaking at 750rpm. After DNA extraction using phenol-chloroform, the DNA was resuspended in 10mM Tris-HCl pH 8.0. The purified DNA was subjected to qPCR to confirm target region enrichment before moving on to deep sequencing library preparation. For sequencing, the extracted DNA was used to construct the ChIP-seq library using the NEBNext Ultra II DNA Library Prep Kit, followed by sequencing with an Illumina NovaSeq 6000 system.

Analyses of ChIP-Seq data

After an initial assessment of read quality, CHERP (HA-tag) ChIP-seq data were mapped onto human reference genome hg38 with Bowtie2 v. 2.5.0 aligner for paired-end reads using default parameters.[56] After alignment, MACS2 v.2.2.7.1 in Galaxy (parameters: –broad; –broad-cutoff of q<0.1; others -default) was used to call peaks for IP-samples against input.[57] Coverage tracks in .bigwig format were generated from bed-graph files of scores and viewed in IGV v.2.4.16 software. TSS profiling was done using plotProfile on matrices generated with 50-bp bins using the computeMatrix function from the Deep-tools v.2.2.3.[58] The reproducibility of data was assessed by Pearson correlation analysis using the plotCorrelation function.[58] ChIP-seq data were deposited in GEO under the accession number GSE226908.

QUANTIFICATION AND STATISTICAL ANALYSES

Statistical analyses were performed using RStudio v.4.0.5 and GraphPad Prism v9. Two-tailed, independent sample *t*-test was used to compare the two groups. $p < 0.05$ was considered significant. For colocalization analysis, one-way ANOVA was performed to determine p-values. Correlations were assessed using two-tailed Pearson *r* coefficients. Protein bands were quantified and compared using ImageJ software.