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]

[1] Department of Molecular Biosciences and [2] Chemistry, University of Texas, Austin, Texas

* Corresponding should be addressed to Y. Jessie Zhang (jzhang@cm.utexas.edu)

Keywords: Transcription, phosphorylation, proteomics, RNA-binding protein, alternative splicing

Abstract

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 binding transcriptional regulators. The phosphorylation of Ser5 and Ser2 of CTD heptad occurs spatialtemporally 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 recruits 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 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 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 local concentration of epitope, thereby amplifying the signal and confidence of the interactome studies. Finally, the *in vitro* phosphorylation using CTD kinases provides us a homogenous 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). To identify a kinase with a strong preference for Ser2 of the heptad, we used bioinformatics and structural analysis to identify a human kinase (20), DYRK1A, and confirmed its specificity by mass spectrometry (21). Our study using the 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 it is 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 a part of an accessory splicing complex whose elimination causes extensive alternative splicing events, which have been implicated in colorectal tumorigenesis.

Results

The kinase specificity on the serines of the CTD

To perform interactome studies to elucidate transcriptional regulators that are recruited by phosphorylated Ser2 and Ser5, 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 (**Figure 1B and S1**). The specificity is consistent with our previous results using CTD constructs of varying lengths (22,23). However, the physiological Ser2 kinase, P-TEFb, was reported to strongly prefer Ser5 over Ser2 as a substrate *in vitro* (18,19). When we used P-TEFb to treat the CTD heptad sequence, the only phosphorylation species was at Ser5 (**Figure 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 are 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 towards 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 (**Figure 1D and S3**). 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 ofDYRK1A, which in turn, forms 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 DYRK1Aknockdown was previously reported (20). To investigate the possibility of alternative sites being phosphorylated with different sequence context, we conducted a systematic characterization of DYRK1Aspecificity using biochemical product profiling with high-resolution mass spectrometric characterization (**Figure 2**). With the 7th residue in the previous heptad 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 7th residue as Arg or Lys in the proceeding heptad; the Ser2 in the subsequent heptad gets phosphorylated preferably and effectively (**Figures 2C and 2D and Figures S4 and S5**). Favorable interactions of Lys7/Arg7 with DYRK1A places the neighboring Ser2 at a highly favorable position for kinase phosphorylation (**Figure 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, chunky 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 residue or a 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 (**Figure 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). 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**). 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 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 Supplementary Table S1**). 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, the interactomes of pSer2 and pSer5 are dramatically different in our study, 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 the accessory proteins associated with them, such as chromatin remodeling complexes (**Figure 3C, Supplementary Table S1**). This observation is consistent with the biological understanding that active transcription reduces chromatin density by ejecting and relocating them (24). 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 (25). 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, which is a major function of Ser5 phosphorylation (26,27). 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 (28) (**Figure 3C**).

The pattern of pSer2 pulldown (**Figure 3D**) showed a significant difference from that of pSer5 pulldown (**Figure 3C**), with more proteins found in association with the phosphoryl mark rather than depletion. Some previously characterized pSer2-binders, such as PHF3, PCF11 (29), PHRF1 (30), and RPRD2 (31) are among the hits, validating the accuracy of our pulldown (**Figure 3D**). Many of the 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 is 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 are involved in 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 (32).

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 (**Figure 3C and 3D**). The PNUTS/PP1 phosphatase complex is known to be involved in Pol II dephosphorylation during active transcription (33). In contrast, PP2A is highly depleted in the pSer5 and pSer2 pulldown samples (**Figures 3C and 3D**). As part of the integrator complex (34), PP2A diverts the stuck RNA polymerase II at the promoter proximal pausing sites to abortive escape without entering productive elongation (35). 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 novel proteins that are directly recruited by a specific CTD phosphospecies. Although many proteins were found to interact with both phosphorylation forms (**Supplementary**

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 (**Figure S8B-C**). 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 proteomics study, we examined the sub-cellular localization of CHERP in relation to Pol II. Immunofluorescence analysis revealed strong co-localization (correlation value of 0.77) between pSer2 Pol II and CHERP in transfected cells (**Figure 4B).** We also investigated whether this co-localization was dependent on pSer2 by inhibiting Ser2 phosphorylation with a small molecule compound flavopiridol (36). The co-localization 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 (**Figure 4C and 4D**). Notably, we observed that Thr4 phosphorylation, a post-translational modification often found together with Ser2 phosphorylation, also co-immunoprecipitated with CHERP (**Figure 4C and 4D**). While pThr4 and pSer2 marks are often found co-occurring during transcription their function is not well understood (37). 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-∆CID) (**Figure 4A and 4B**). The removal of this domain reduced the co-localization between CHERP and pSer2 Pol II from 0.77 to 0.33 in correlation (**Figure 4B**). When we conducted co-immunoprecipitation experiment using the CHERP construct lacking 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 (**Figure S9A-C**). 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 CHERP CID domain can directly interact with 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 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 didn't 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**). 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 don't alter protein folding but abolish or significantly compromise CTD binding (**Figure 5D and S9D-F**). The model also explains why phosphoryl-Thr4 can also be recognized by CHERP, as seen in the co-immunoprecipitation experiment and fluorescence anisotropy (**Figure 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 an insight into CHERP's specificity towards 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 ChIP-seq analysis of CHERP and RPB1 binding. We inserted full-length CHERP or RPB1 genes into a 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 cross-linking according to (38). Both chromatin samples were pulled-down with HA antibody and analyzed for its distribution genome-wide. We identified ~2000 peaks in each sample. The profiles of CHERP and Pol II are highly similar with a distribution profile peaking a little after Transcription Starting Site (TSS) (**Figure 5D and 5E**).

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 (36). The application of flavopiridol caused Ser2 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 (**Figure 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 (**Figure 5E and 5F and S10C-E**). 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 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 western blot, we used a commercially available shRNA to reduce the CHERP expression to 18% (**Figure S11D**). 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 ($r > 0.99$) (**Figure S11A-C)**. We found that 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**). Supplementary 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 type, there was a significant overlap of differentially expressed genes (39). 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 (40), 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**, **Supplementary 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 S11E**). 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, Supplementary 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 (41). One such cilia-associated gene is a potential therapeutic target, centrosomal protein CEP164; its mutations or deficiency are related to ciliopathies (degenerative diseases affecting kidney, retina, and brain) (42) and pancreatic cancer growth (43). We found that CHERP knockdown leads to more frequent inclusion of exon 8 into *CEP164* mRNA (**Figure 6D**, *upper*, **Supplementary 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 (44) (**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 (45) (**Figure 6D,** *lower*). Legumain promotes cellular migratory and invasive activity *in vitro* and *in vivo* inferring significance of targeting legumain to combat tumor invasion and metastasis (46). 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. 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. 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 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. We observed that loss of CHERP function leads to alternative splicing events. 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 (47). Mutations found in core or accessory splicing components have been observed in many cancer types, promoting the isoforms amplification for active tumor suppression (48,49). Abnormal splicing outcomes have been identified as novel biomarkers for diagnosis and new target for treatment (50,51). Recent research has identified to CHERP part of a stable complex with another two splicing factors, U2SURP and RBM17 (39). This complex and its associated alternative splicing events are implicated in the colorectal tumorigenesis development (52). The combination of our recent findings with previous data provides strong evidence implicating 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. 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 (39), the molecular mechanism by which 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.

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.

Data Availability

The mass spectrometry proteomics data generated in this study have been deposited in the ProteomeXchange Consortium via the PRIDE partner repository under accession code

PXD039903. The RNA-seq data used in this study are available in GEO under accession code GSE221328. The ChIP-seq data replicates used in this study is available in GEO under accession code GSE226908. Any unprocessed images were deposited on Mendeley at doi:https://data.mendeley.com/datasets/9mb9vgsw3n/draft?a=08fc1869-91af-49f8-960a-f22c413fde53

Author Contributions

RYM and YJZ designed the experiment, KJJ, JPB, and JSB analyzed UVPD-MS, SBP conducted bioinformatic analysis of RNA-seq, BMF and EMM carried out proteomic analysis, MKVR helped with pulldown experiment. RYM conducted the rest of the experiments. The manuscript was written by RYM and YJZ with contribution from all authors.

Competing Interests

The authors declare no competing interests.

Methods

Cell culture. 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), was added to the media to reach a final concentration of 1%.

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, Wadison, 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.

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 IRDye 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 (53). 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 1mg/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(54). 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 phospho 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 (55). 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 (56). The CID domain region of CHERP has a very high confidence score (pLDDDT > 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 was resuspended in buffer B (10mM Tris pH 8.0, 150mM NaCl, 1:100 PPI) supplemented with 1:1000 benzonase. 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 (57). To generate volcano plots, datasets from both replicates were $\log_2$ transformed, missing values were imputed, and data was 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 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 $\Delta\Delta$Ct method normalized to *ACTB* expression. Student's t-test was used to compare groups. Specificity of amplification was controlled with PCR product melting curves. All primers used in this study can be found in the supplementary section as Table 4.

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 manufacturer's instructions for the NEBNext Ultra II Direction RNA kit (NEB, Ipswich, MA, USA, product number #E7760). 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). Minimum number of reads was set to 40x10^6 per sample.

Analyses of RNA-seq data and alternative splicing events (ASE). Quality of raw reads was assessed using FastQC read quality reports (https://usegalaxy.org) (58). Adapter Illumina sequences (--illumina) were trimmed off by TrimGalore! v.0.6.7 with default parameters. Next, reads were aligned to 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 (60). Differential expression was analyzed using DESeq2 v.1.30.1 in R; genes with adjusted *p*-value < 0.05 were considered as differentially expressed (61). RNA-seq data was deposited in GEO under the accession number GSE221328. rMATS turbo v.4.1.2 in command line was employed for detection of alternatively spliced events upon SCAF6 loss vs. control (with parameters FDR < 0.05; ILD, inclusion level difference, ≥ 10%) (40). 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 (62).

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 initial assessment of read quality, CHERP (HA-tag) ChIP-seq data was mapped onto human reference genome hg38 with Bowtie2 v. 2.5.0 aligner for paired-end reads using default parameters (63). 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 (64). Coverage tracks in .bigwig format were generated from bedgraph 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 (65). Reproducibility of data was assessed by pearson correlation analysis using the plotCorrelation function (65). ChIP-seq data was deposited in GEO under the accession number GSE226908.

Statistical analyses. Statistical analyses were performed using RStudio v.4.0.5 and GraphPad Prism v9. Two-tailed, independent sample *t*-test was used for comparing the two groups. $p < 0.05$ was considered as significant. Correlations were assessed using two-tailed Pearson *r* coefficients. Protein bands were quantified and compared using ImageJ software.

References

1. Corden, J.L. (2013) RNA polymerase II C-terminal domain: Tethering transcription to transcript and template. *Chem Rev*, **113**, 8423-8455.
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.
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. *Journal of Biological Chemistry*, **266**, 2290-2296.
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.
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 U S A*, **93**, 6975-6980.
6. Buratowski, S. (2003) The CTD code. *Nat Struct Biol*, **10**, 679-680.
7. Dahmus, M.E. (1996) Reversible phosphorylation of the C-terminal domain of RNA polymerase II. *J Biol Chem*, **271**, 19009-19012.
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.
9. Phatnani, H.P. and Greenleaf, A.L. (2006) Phosphorylation and functions of the RNA polymerase II CTD. *Genes Dev*, **20**, 2922-2936.
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.
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.
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. *Scientific Reports*, **6**, 27401.
13. Egloff, S. (2012) Role of Ser7 phosphorylation of the CTD during transcription of snRNA genes. *RNA Biology*, **9**, 1033-1038.
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.
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.
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.
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.
18. Czudnochowski, N., Bosken, 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.
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**.

20. Di Vona, C., Bezdan, D., Islam, A.B., Salichs, E., Lopez-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.
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. *et al.* (2020) Structural Motifs for CTD Kinase Specificity on RNA Polymerase II during Eukaryotic Transcription. *ACS Chem Biol*, **15**, 2259-2272.
22. 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*.
23. 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.
24. Zhou, C.Y., Johnson, S.L., Gamarra, N.I. and Narlikar, G.J. (2016) Mechanisms of ATP-Dependent Chromatin Remodeling Motors. *Annual Review of Biophysics*, **45**, 153-181.
25. Karakasili, E., Burkert-Kautzsch, C., Kieser, A. and Strasser, K. (2014) Degradation of DNA damage-independently stalled RNA polymerase II is independent of the E3 ligase Elc1. *Nucleic Acids Res*, **42**, 10503-10515.
26. 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.
27. 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.
28. 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.
29. Meinhart, A. and Cramer, P. (2004) Recognition of RNA polymerase II carboxy-terminal domain by 3′-RNA-processing factors. *Nature*, **430**, 223-226.
30. 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.
31. 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. *Nature structural & molecular biology*, **21**, 686-695.
32. 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.
33. 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 &#x201c;Sitting Duck Torpedo&#x201d; Mechanism. *Molecular Cell*, **76**, 896-908.e894.
34. 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 e3132.
35. 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. *Molecular Cell*, **82**, 4232-4245.e4211.
36. Chao, S.-H. and Price, D.H. (2001) Flavopiridol Inactivates P-TEFb and Blocks Most RNA Polymerase II Transcription in Vivo *. *Journal of Biological Chemistry*, **276**, 31793-31799.

37. Jasnovidova, O., Klumpler, T., Kubicek, K., Kalynych, S., Plevka, P. and Stefl, R. (2017) Structure and dynamics of the RNAPII CTDsome with Rtt103. *Proceedings of the National Academy of Sciences*, **114**, 11133-11138.
38. Tian, B., Yang, J. and Brasier, A.R. (2012) In Vancura, A. (ed.), *Transcriptional Regulation: Methods and Protocols*. Springer New York, New York, NY, pp. 105-120.
39. 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 e727.
40. 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 U S A*, **111**, E5593-5601.
41. Higgins, M., Obaidi, I. and McMorrow, T. (2019) Primary cilia and their role in cancer. *Oncol Lett*, **17**, 3041-3047.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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.
47. Kahles, A., Lehmann, K.-V., Toussaint, N.C., Hüser, M., Stark, S.G., Sachsenberg, T., Stegle, O., Kohlbacher, O., Sander, C., Caesar-Johnson, S.J. *et al.* (2018) Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell*, **34**, 211-224.e216.
48. 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.
49. 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. *Nature Genetics*, **47**, 1242-1248.
50. Kim, Y.-J. and Kim, H.-S. (2012) Alternative Splicing and Its Impact as a Cancer Diagnostic Marker. *Genomics Inform*, **10**, 74-80.
51. Bonnal, S., Vigevani, L. and Valcárcel, J. (2012) The spliceosome as a target of novel antitumour drugs. *Nature Reviews Drug Discovery*, **11**, 847-859.
52. 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. *International Journal of Cancer*, **145**, 2728-2739.
53. Stauffer, W., Sheng, H. and Lim, H.N. (2018) EzColocalization: An ImageJ plugin for visualizing and measuring colocalization in cells and organisms. *Scientific Reports*, **8**, 15764.
54. Klein, D.R., Holden, D.D. and Brodbelt, J.S. (2016) Shotgun Analysis of Rough-Type Lipopolysaccharides Using Ultraviolet Photodissociation Mass Spectrometry. *Analytical Chemistry*, **88**, 1044-1051.

55. 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. *Journal of Chemical Theory and Computation*, **6**, 1509-1519.
56. 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.
57. Floyd, B.M., Drew, K. and Marcotte, E.M. (2021) Systematic Identification of Protein Phosphorylation-Mediated Interactions. *Journal of Proteome Research*, **20**, 1359-1370.
58. 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 Research*, **46**, W537-W544.
59. Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357-360.
60. 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.
61. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.
62. 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 Journal of Integrative Biology*, **16**, 284-287.
63. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357-359.
64. Feng, J., Liu, T., Qin, B., Zhang, Y. and Liu, X.S. (2012) Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, **7**, 1728-1740.
65. 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 Research*, **44**, W160-W165.

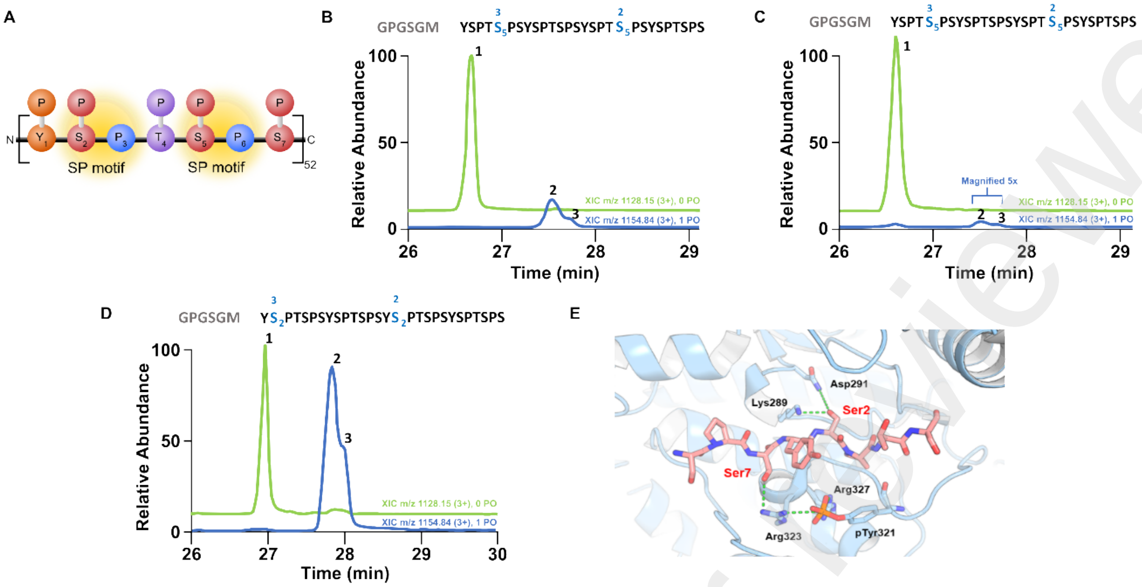


Figure 1: CTD kinases differentially phosphorylate consensus CTD. Schematic of the consensus sequence and putative phosphorylation site on the CTD. Each unique residue is denoted by a distinct color and potential phosphorylation sites are indicated with a phosphate symbol overhead. Serine-proline motifs are highlighted. (B-D) Extracted ion chromatographic traces (XIC) for the LC–MS/MS analysis for kinase specificity towards 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) is highlighted with key residues shown in stick and hydrogen bonds in dash lines.

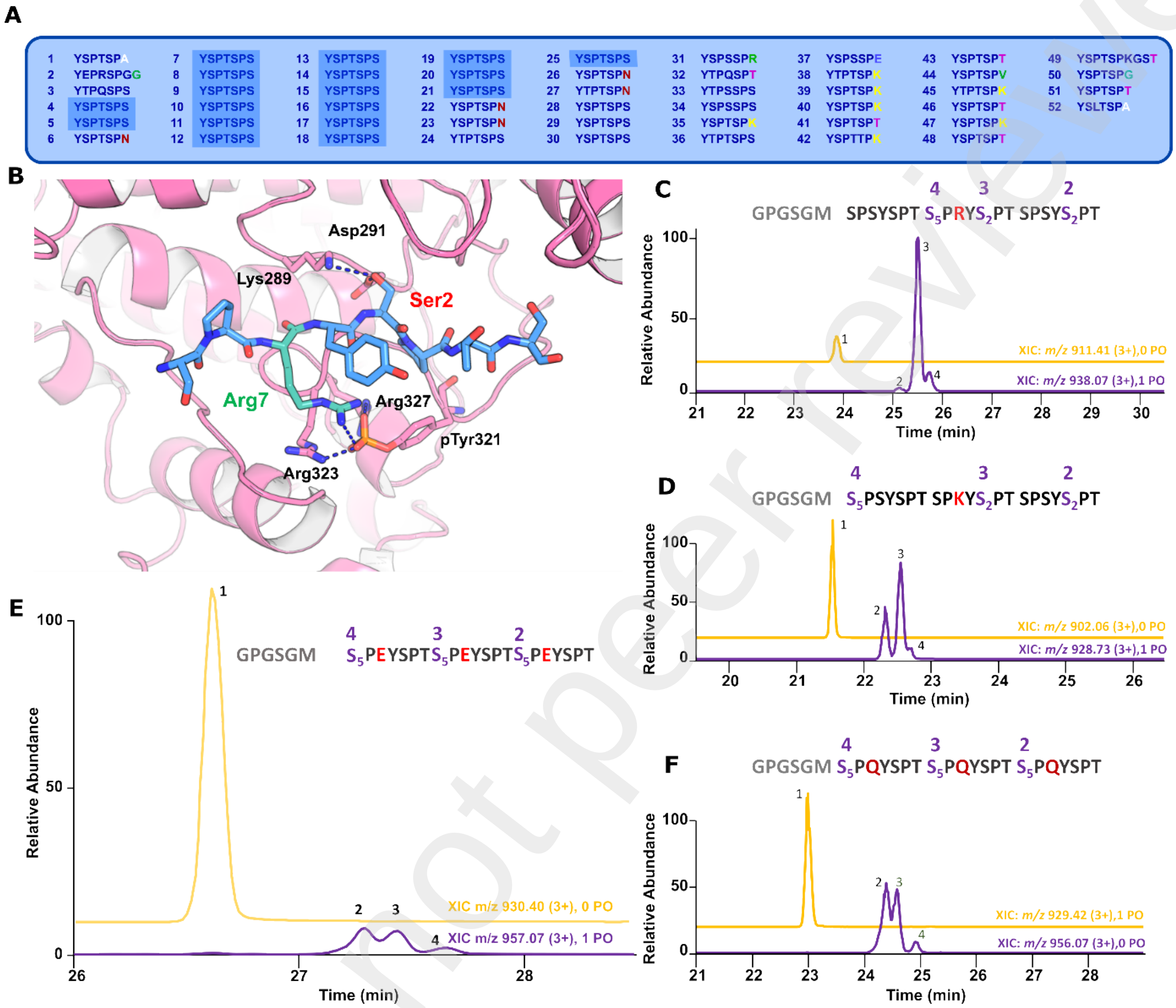


Figure 2. DYRK1A's specificity in divergent CTD heptads. (A) Diagram of the CTD sequence of human RPB1. The shaded region consists of the consensus sequence. The rest of the sequence diverges slightly from consensus with residues different from serine in the seventh position highlighted. (B) Structural model of DYRK1A in complex with a CTD peptide where the seventh position is occupied by arginine. DYRK1A is shown in ribbon diagram as light pink and the CTD peptide as sticks 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 towards 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.

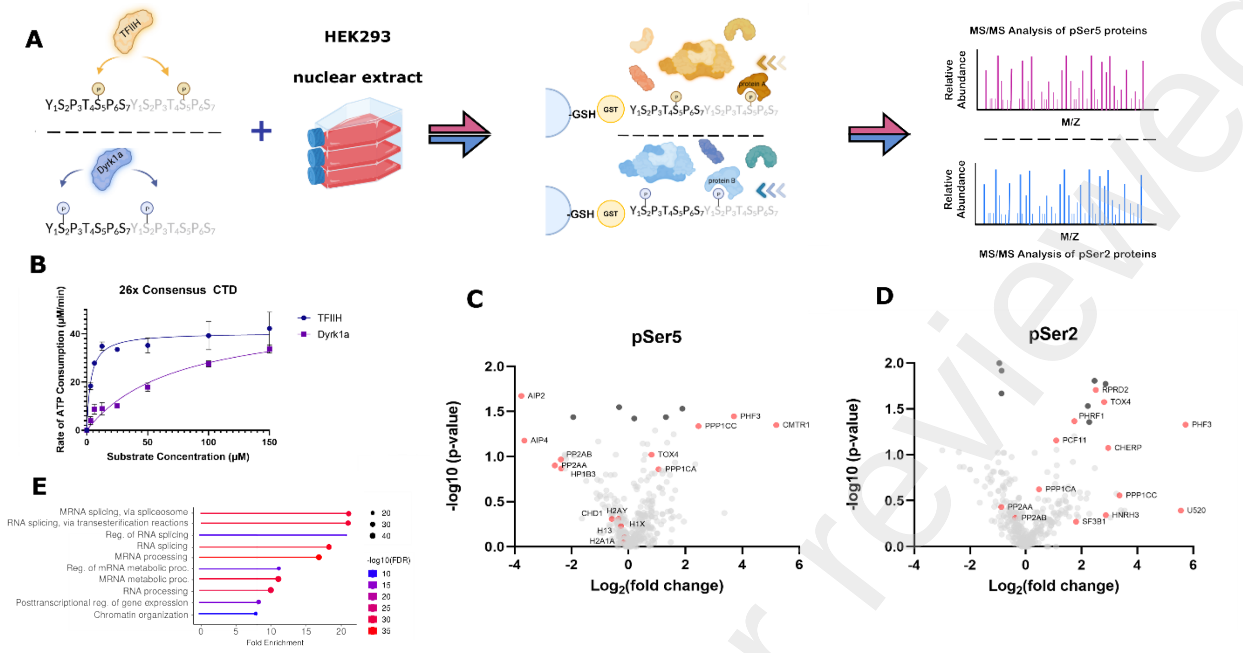


Figure 3. Proteomic study of pCTD interactomes. *In vitro* phosphorylation of a GST-CTD recombinant protein containing 26x repeat consensus CTD heptads is treated with TFIIH or DYRK1A and incubated with lysate from HEK-293 cells. A no-kinase sample is treated parallelly as control. Affinity chromatography 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(µM^{-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 grey 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.

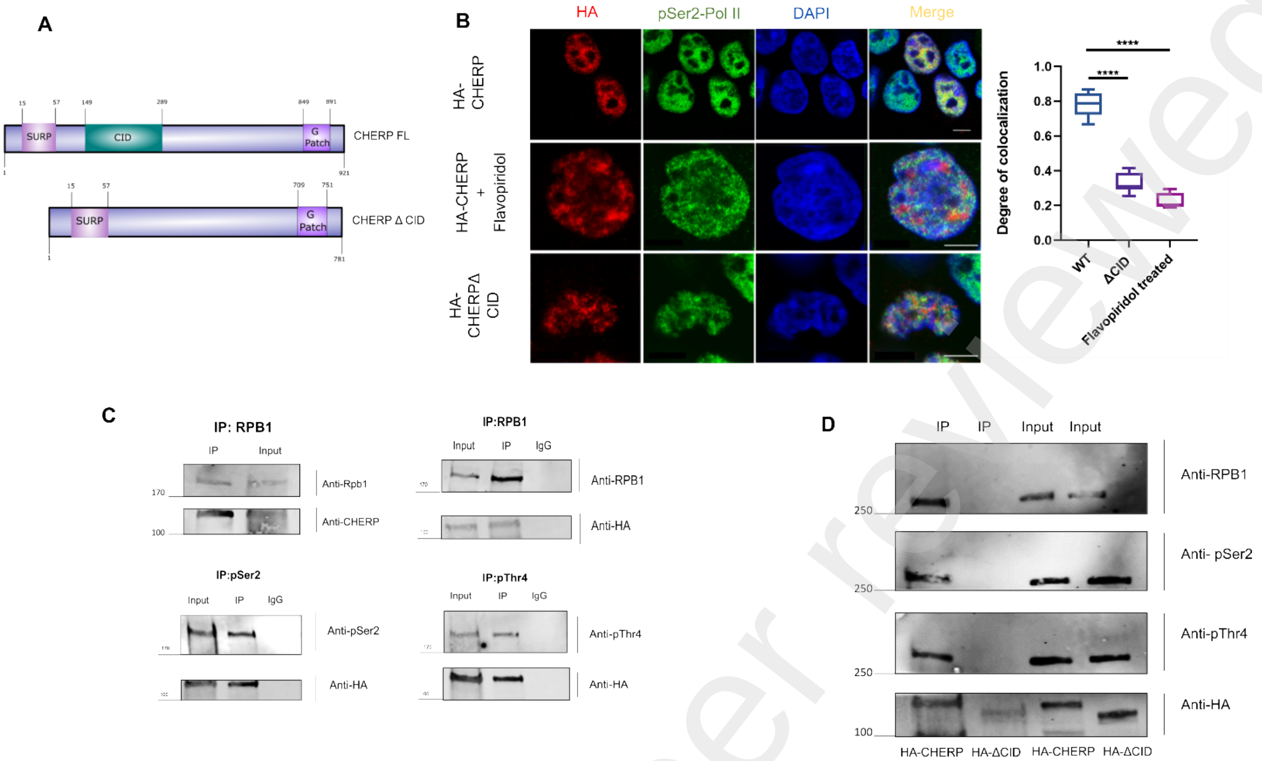


Figure 4. CHERP physically associates with phosphorylated human RNA Pol II. (A) Domain architecture of full-length CHERP and a deletion mutant that eliminates the CID domain. (B) Representative confocal fluorescent images of HA-CHERP or HA-CHERP ΔCID(red), pSer2 Pol II (green) and DAPI (blue) in HEK293 cells. Scale bar = 5 μm. Quantification of colocalization between CHERP FL or mutant with pSer2 under different conditions (N = 7). Box and whiskers plot 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 phosphoisoforms 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.

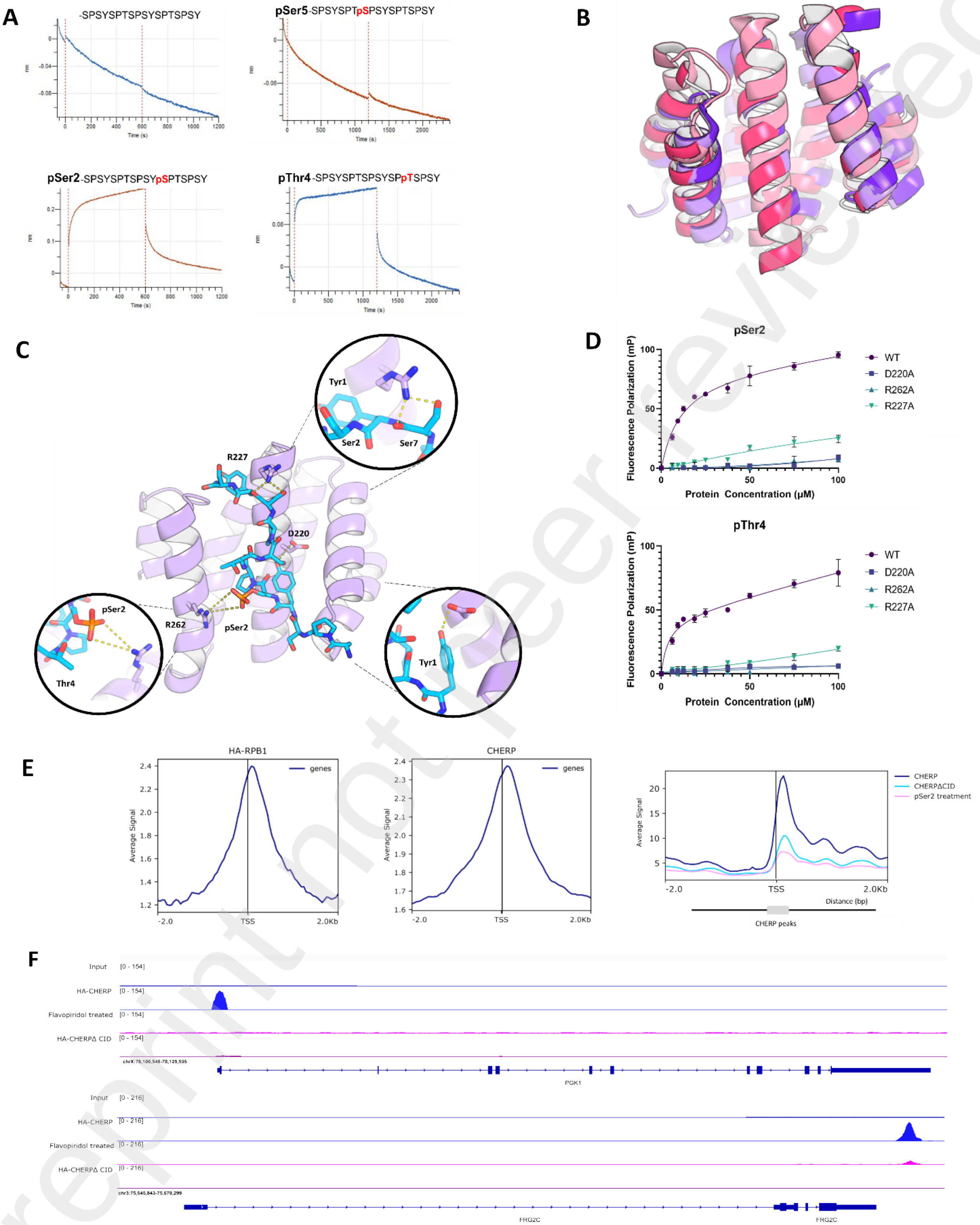
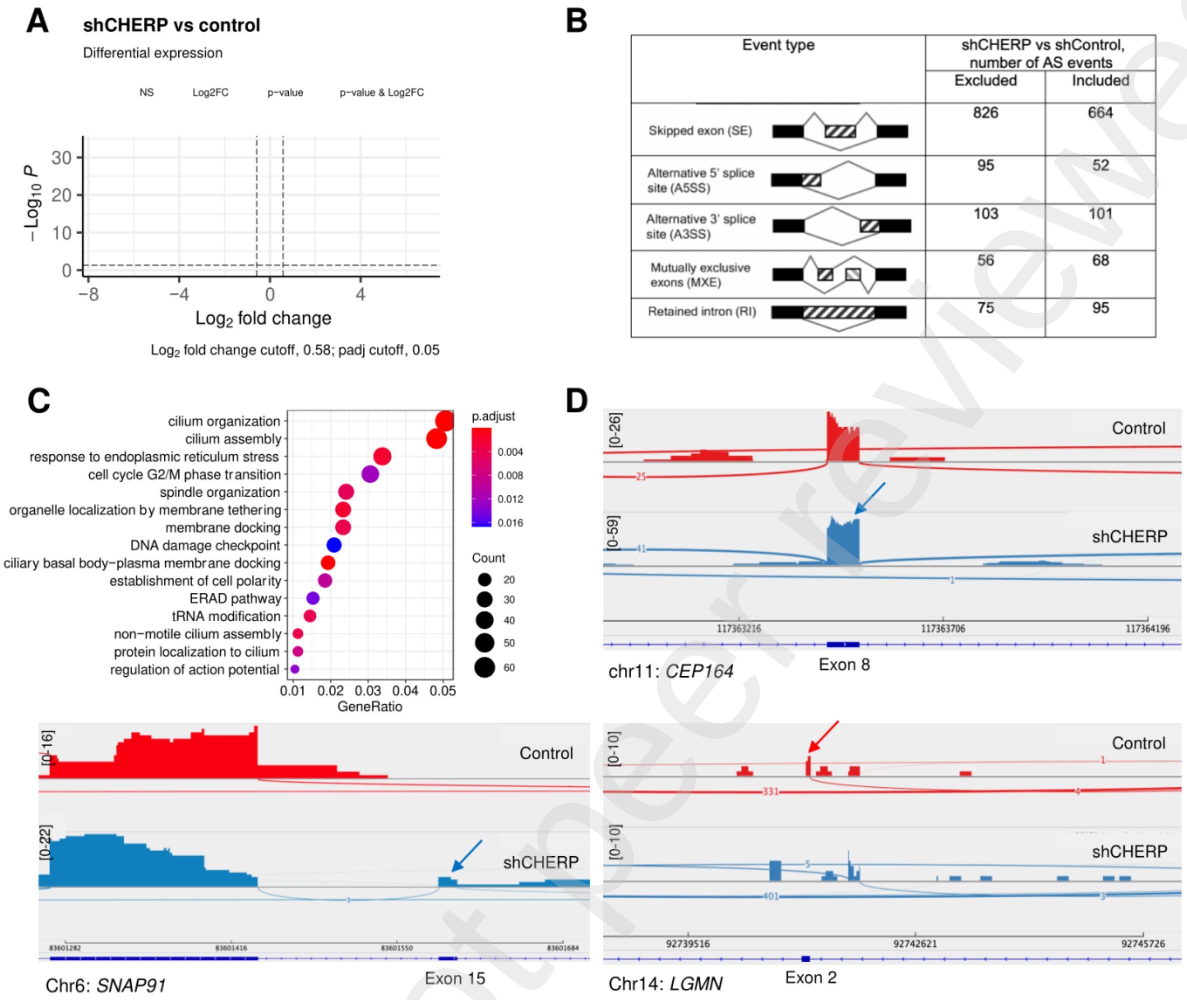


Figure 5. CHERP binds to Pol II through a CID domain. (A) Biolayer interferometry-binding assays of the interactions between the CID domain of CHERP (1 μM) with different phosphoisoforms of

CTD peptides containing two heptad repeats. 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 one: one binding model. Binding assays were performed in triplicate. Error bars indicate 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 overlayed across CHERP-bound genomic regions. Average peak coverage is shown in a bin size of 50 bp for a window 2 kb upstream/downstream from 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.



Event type	shCHERP vs shControl, number of AS events	
	Excluded	Included
Skipped exon (SE)	826	664
Alternative 5' splice site (A5SS)	95	52
Alternative 3' splice site (A3SS)	103	101
Mutually exclusive exons (MXE)	56	68
Retained intron (RI)	75	95

Figure 6. Whole-transcriptome effects of CHERP knockdown on gene expression and alternative splicing. (A) RNA-seq data shows 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. 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, 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 inclusion of alternative exon.

KEY RESOURCES TABLE

Key resources table

REAGENT or	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
Anti-HA (rabbit monoclonal)	Cell Signaling	Cat#3724S
Goat anti-rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 488	Thermofischer Scientific	Cat # A-11008
Goat anti-rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 568	Thermofischer Scientific	Catalog # A-11011
Chemicals, peptides, and recombinant proteins		
CHERP CID synthetic gene	Biomatik	
S7K, S7R, S7E, S7Q, 4X WT CTD	IDT	
CHERP cDNA	DNASU	HsCD00879118
RPB1 cDNA	Addgene	Cat #75284
TFIIH (Cdk7/CyclinH/MAT 1 (CAK complex))	Millipore	Cat #14-476
P-TEFb (Cdk9/Cyclin T1)	Millipore	Cat #14-685
DYRK1A	Addgene	Cat #79690

Biotinylated CTD peptides	Biomatik	
FITC-CTD peptides	Biomatik	
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
Commerical assay		
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:
Experimental models: Cell lines		
HEK293T	ATCC	
HEK293	ATCC	
Oligonucleotides		
qPCR primers	IDT	Table S4
Software and algorithms		
Image J	NIH	https://imagej.nih.gov/ij/download.html
EzColocalization	Stauffer, et al. (1)	https://github.com/DrHanLim/EzColocalization
Bowtie2	Langmead, et al. (2)	https://github.com/BenLangmead/bowtie2
MACS2	Feng, et al.(3)	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.(4)	https://deeptools.readthedocs.io/en/develop/index.html

rMATS turbo	Shen, et al. (5)	https://github.com/Xinglab/rmats-turbo
HISAT2	Kim, et al. (6)	http://daehwankimlab.github.io/hisat2/
DESeq2	Love, et al. (7)	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
TrimGalore!	Babahram Bioinformatics	https://github.com/FelixKrueger/TrimGalore
featureCounts	Liao, et al. (8)	https://subread.sourceforge.net/
Proteome Discoverer	ThermoFischer Scientific	
Xtract algorithm	ThermoFischer Scientific	
ProSight Lite	Proteomics Center of Excellence Northwestern University	http://prosightlite.northwestern.edu/
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

References

1. Stauffer, W., Sheng, H. and Lim, H.N. (2018) EzColocalization: An ImageJ plugin for visualizing and measuring colocalization in cells and organisms. *Scientific Reports*, **8**, 15764.
2. Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357-359.
3. Feng, J., Liu, T., Qin, B., Zhang, Y. and Liu, X.S. (2012) Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, **7**, 1728-1740.
4. 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 Research*, **44**, W160-W165.
5. 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 U S A*, **111**, E5593-5601.
6. Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, **12**, 357-360.
7. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 550.
8. 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.

Supplementary Information

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]

[1] Department of Molecular Biosciences and [2] Chemistry, University of Texas, Austin, Texas

[*] Corresponding should be addressed to Y. Jessie Zhang (jzhang@cm.utexas.edu)

Table of Contents:

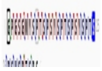
Antibodies

For western blot analysis, phospho-specific antibodies, pThr4 (Active Motif, cat: 61361 , 1:800 dilution) , pSer5 (Sigma, SKU: SAB4200638-100UL, 1:1000 dilution for WB and IF), pSer2 (Sigma, SKU: MABE953 , 1:1000 dilution for WB and IF). The RPB1 antibody is from Abcam (Cat: ab76123, 1:1000 dilution). The HA antibody used for western blot and immunofluorescence is from Cell signaling (cat: C29F4, 1:1000 dilution for WB and 1:800 for IF). The CHERP antibody is from Bethyl Laboratories (cat: A304-620A, 1:1000 dilution). Secondary antibodies for Western blotting were obtained from Licor (IRDye series). Secondary antibodies for double immunofluorescence are the Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 488 (cat: A-11008) at 4 μg/ml and the Goat anti-Rabbit IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor™ 568 (cat: A-11011) at 2 μg/ml.

Differential Scanning Fluorimetry

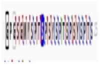
Purified recombinant CHERP CID domain at a final concentration of 5μM was incubated with 10X SYPRO Orange (Molecular Probes) in a 96-well low-profile PCR plate (ABgene, Thermo Scientific) and fluorescence was captured in a LightCycler 480 (Roche). Protein melting curves were carried out with a temperature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

2



Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
A_{+12}	1091.483	1091.478	-4.6
A_{+13}	1178.515	1178.506	-8.1
A_{+15}	1428.611	1428.609	-1.4
A_{+16}	1525.663	1525.653	-6.7
A_{+19}	1810.796	1810.802	3.3
A_{+9}	806.351	806.346	-5.4
A_{10}	906.391	906.394	3.3
A_{12}	1090.475	1090.475	-0.5
A_{14}	1340.571	1340.572	1.1
A_{16}	1524.655	1524.649	-4.1
A_{18}	1712.735	1712.744	4.8
A_{22}	2146.915	2146.931	7.4
A_{29}	2946.194	2946.209	5.1
A_{30}	3043.247	3043.228	-6.2
A_{32}	3231.327	3231.305	-6.9
A_4	270.133	270.132	-1.8
A_7	621.258	621.257	-1.4
A_8	708.290	708.290	-0.1
B_{10}	934.385	934.385	-0.3
B_{11}	1021.417	1021.418	0.1
B_{15}	1455.598	1455.597	-0.8
B_{16}	1552.650	1552.639	-7.5
B_{17}	1653.698	1653.699	0.4
B_{18}	1740.730	1740.730	0.2
B_{19}	1837.783	1837.775	-4.3
B_2	154.074	154.074	-2.1
B_{21}	2087.878	2087.879	0.4
B_{22}	2174.910	2174.916	2.5
B_{25}	2540.009	2540.005	-1.7
B_{27}	2724.094	2724.113	7.1
B_{29}	2974.189	2974.186	-1.0
B_3	211.096	211.095	-1.8
B_4	298.128	298.127	-2.0
B_6	486.190	486.190	-0.3
B_7	649.253	649.252	-0.9
B_8	736.285	736.285	-0.4
B_9	833.338	833.337	-0.5
C_{10}	951.412	951.413	0.8
C_{12}	1135.497	1135.488	-7.9
C_{13}	1222.529	1222.528	-0.6
C_{14}	1385.592	1385.590	-1.7
C_{16}	1569.677	1569.682	3.1
C_{19}	1854.809	1854.813	1.9
C_2	171.101	171.100	-0.7
C_{20}	1941.841	1941.844	1.4
C_3	228.122	228.122	-0.4
C_4	315.154	315.154	-1.3
C_5	372.175	372.175	-0.7
C_6	503.216	503.216	-0.3
C_7	666.279	666.280	1.7
C_9	850.364	850.367	3.6
X_{+25}	2655.066	2655.085	7.1
X_{+33}	3431.383	3431.404	6.1
X_{+6}	601.247	601.248	1.4
X_4	416.154	416.154	-1.0
Y_{10}	1088.406	1088.407	0.5
Y_{11}	1189.454	1189.451	-2.2
Y_{12}	1286.507	1286.507	0.3
Y_{13}	1373.539	1373.538	-0.5
Y_{16}	1720.687	1720.688	0.6
Y_{17}	1807.719	1807.717	-1.2
Y_{19}	2005.819	2005.818	-0.5
Y_2	202.095	202.095	-1.8
Y_{20}	2092.851	2092.852	0.5
Y_{-23}	2438.992	2438.992	0.0
Y_{-24}	2526.024	2526.038	5.7
Y_{-26}	2724.124	2724.113	-4.0
Y_{-27}	2811.156	2811.163	2.2
Y_{29}	3106.268	3106.293	8.0
Y_3	289.127	289.127	-1.7
Y_4	390.175	390.174	-1.7
Y_5	487.228	487.227	-1.5
Y_6	574.260	574.259	-1.5
Y_7	737.323	737.323	0.0
Y_{-8}	823.347	823.355	8.9
Y_9	921.408	921.407	-0.8
Z_{24}	2511.013	2511.001	-4.6
Z_4	374.156	374.156	-1.1

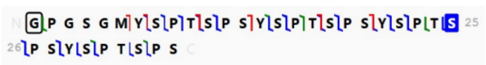
3

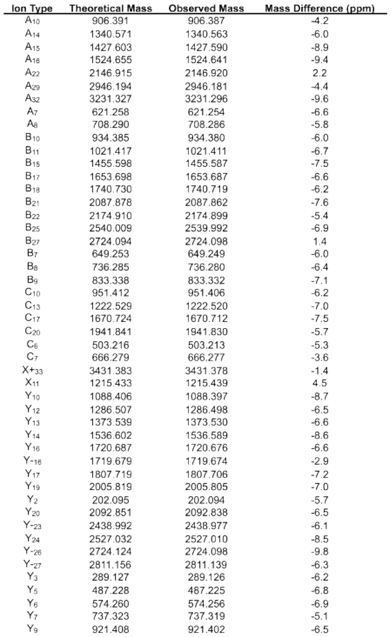


Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
A_{+10}	907.398	907.396	-2.9
A_{+12}	1171.449	1171.445	-3.7
A_{+15}	1508.577	1508.565	-7.6
A_{+16}	1605.630	1605.624	-3.5
A_{+19}	1890.762	1890.748	-7.4
A_{+26}	2610.075	2610.057	-6.7
A_{+9}	806.351	806.346	-6.1
A_{10}	906.391	906.393	2.8
A_{16}	1604.622	1604.621	-0.6
A_{18}	1792.702	1792.696	-3.4
A_{22}	2226.882	2226.876	-2.8
A_{29}	2946.194	2946.210	5.4
A_{30}	3043.247	3043.246	-0.3
A_{32}	3231.327	3231.310	-5.1
A_7	621.258	621.257	-1.6
A_8	708.290	708.290	-0.3
B_{10}	934.385	934.385	-0.2
B_{11}	1101.384	1101.385	0.8
B_{14}	1448.532	1448.530	-1.6
B_{15}	1535.564	1535.561	-2.1
B_{17}	1733.664	1733.664	-0.4
B_{18}	1820.696	1820.696	-0.4
B_{21}	2167.845	2167.844	-0.4
B_{22}	2254.877	2254.876	-0.2
B_{24}	2452.977	2452.977	0.0
B_{25}	2540.009	2540.007	-1.0
B_{26}	2637.062	2637.067	2.0
B_{27}	2724.094	2724.112	6.7
B_{29}	2974.189	2974.185	-1.5
B_3	211.096	211.095	-1.8
B_{32}	3259.322	3259.330	2.5
B_4	298.128	298.127	-2.0
B_7	649.253	649.253	-0.8
B_8	736.285	736.285	-0.6
B_9	833.338	833.337	-1.4
C_{10}	951.412	951.413	1.5
C_{12}	1215.463	1215.455	-6.6
C_{13}	1302.495	1302.493	-1.5
C_{17}	1750.691	1750.693	1.3
C_{20}	2021.808	2021.806	-0.9
C_{23}	2368.956	2368.957	0.4
C_{27}	2741.120	2741.120	-0.2
C_3	228.122	228.122	-0.4
C_4	315.154	315.154	-1.3
C_5	372.175	372.175	-1.0
C_6	503.216	503.216	-0.3
C_7	666.279	666.280	1.5
C_9	850.364	850.368	4.7
X_{+6}	601.247	601.248	1.3
X_{17}	1753.732	1753.724	-4.3
X_7	763.302	763.301	-2.6
Y_{10}	1008.440	1008.440	0.1
Y_{11}	1109.488	1109.488	-0.2
Y_{12}	1206.540	1206.540	-0.8
Y_{13}	1293.572	1293.572	-0.7
Y_{14}	1456.636	1456.633	-2.1
Y_{15}	1543.668	1543.671	2.1
Y_{16}	1640.721	1640.719	-1.1
Y_{17}	1727.753	1727.752	-0.4
Y_{-17}	1726.745	1726.746	0.6
Y_{19}	1925.853	1925.852	-0.7
Y_2	202.095	202.095	-1.8
Y_{20}	2012.885	2012.886	0.6
Y_{21}	2175.948	2175.936	-5.8
Y_{-23}	2359.025	2359.019	-2.8
Y_{-24}	2526.024	2526.027	1.3
Y_{-26}	2724.124	2724.112	-4.4
Y_{-27}	2811.156	2811.163	2.6
Y_{-29}	3105.260	3105.271	3.6
Y_3	289.127	289.127	-1.7
Y_4	390.175	390.174	-1.7
Y_5	487.228	487.227	-1.7
Y_6	574.260	574.259	-1.7
Y_7	737.323	737.323	0.2
Y_9	921.408	921.408	-0.5
Z_{15}	1527.649	1527.652	2.0
Z_{24}	2511.013	2511.005	-3.3
Z_4	374.156	374.156	-1.3

Figure S1: Lists of fragment ions for mono-phosphorylated peptides (*m/z* 1154.84) analyzed in Figure 1B by UVPD-MS. In each case , the 3+ charge state was selected, and UVPD was performed using 2 pulses (1.5 mJ per pulse). The identified site of phosphorylation is shaded in blue in the sequence map above each table. Fragment ions are named as T_n where T is the type of ion (A = a, B = b, C = c, X = x, Y = y, Z = z, for which A,B, and C originate from the N-terminus of the protein and X,Y and Z originate from the C-terminus of the protein), the subscript indicates the number of amino acids contained in the fragment ion, and a plus or minus sign in the subscript designates whether the fragment ion contains one extra hydrogen atom or lacks one hydrogen atom.

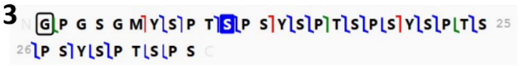
2

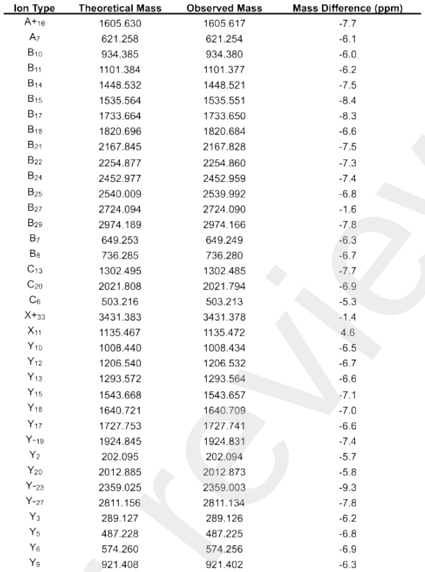




Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
A_{10}	906.391	906.387	-4.2
A_{14}	1340.571	1340.563	-6.0
A_{15}	1427.603	1427.590	-8.9
A_{16}	1524.655	1524.641	-9.4
A_{22}	2146.915	2146.920	2.2
A_{29}	2946.194	2946.181	-4.4
A_{32}	3231.327	3231.296	-9.6
A_{7}	621.258	621.254	-6.6
A_{8}	708.290	708.286	-5.8
B_{10}	934.385	934.380	-6.0
B_{11}	1021.417	1021.411	-6.7
B_{15}	1455.598	1455.587	-7.5
B_{17}	1653.698	1653.687	-6.6
B_{18}	1740.730	1740.719	-6.2
B_{21}	2087.878	2087.862	-7.6
B_{22}	2174.910	2174.899	-5.4
B_{25}	2540.009	2539.992	-6.9
B_{27}	2724.094	2724.098	1.4
B_{7}	649.253	649.249	-6.0
B_{8}	736.285	736.280	-6.4
B_{9}	833.338	833.332	-7.1
C_{10}	951.412	951.406	-6.2
C_{13}	1222.529	1222.520	-7.0
C_{17}	1670.724	1670.712	-7.5
C_{20}	1941.841	1941.830	-5.7
C_{6}	503.216	503.213	-5.3
C_{7}	666.279	666.277	-3.6
X_{+33}	3431.383	3431.378	-1.4
X_{11}	1215.433	1215.439	4.5
Y_{10}	1088.406	1088.397	-8.7
Y_{12}	1286.507	1286.498	-6.5
Y_{13}	1373.539	1373.530	-6.6
Y_{14}	1536.602	1536.589	-8.6
Y_{16}	1720.687	1720.676	-6.6
Y_{-16}	1719.679	1719.674	-2.9
Y_{17}	1807.719	1807.706	-7.2
Y_{19}	2005.819	2005.805	-7.0
Y_{2}	202.095	202.094	-5.7
Y_{20}	2092.851	2092.838	-6.5
Y_{-23}	2438.992	2438.977	-6.1
Y_{24}	2527.032	2527.010	-8.5
Y_{-26}	2724.124	2724.098	-9.8
Y_{-27}	2811.156	2811.139	-6.3
Y_{3}	289.127	289.126	-6.2
Y_{5}	487.228	487.225	-6.8
Y_{6}	574.260	574.256	-6.9
Y_{7}	737.323	737.319	-5.1
Y_{9}	921.408	921.402	-6.5

3

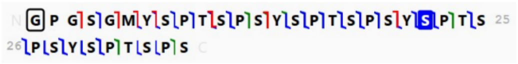




Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
A_{+16}	1605.630	1605.617	-7.7
A_{7}	621.258	621.254	-6.1
B_{10}	934.385	934.380	-6.0
B_{11}	1101.384	1101.377	-6.2
B_{14}	1448.532	1448.521	-7.5
B_{15}	1535.564	1535.551	-8.4
B_{17}	1733.664	1733.650	-8.3
B_{18}	1820.696	1820.684	-6.6
B_{21}	2167.845	2167.828	-7.5
B_{22}	2254.877	2254.860	-7.3
B_{24}	2452.977	2452.959	-7.4
B_{25}	2540.009	2539.992	-6.8
B_{27}	2724.094	2724.090	-1.6
B_{29}	2974.189	2974.166	-7.8
B_{7}	649.253	649.249	-6.3
B_{8}	736.285	736.280	-6.7
C_{13}	1302.495	1302.485	-7.7
C_{20}	2021.808	2021.794	-6.9
C_{6}	503.216	503.213	-5.3
X_{+33}	3431.383	3431.378	-1.4
X_{11}	1135.467	1135.472	4.6
Y_{10}	1008.440	1008.434	-6.5
Y_{12}	1206.540	1206.532	-6.7
Y_{13}	1293.572	1293.564	-6.6
Y_{15}	1543.668	1543.657	-7.1
Y_{16}	1640.721	1640.709	-7.0
Y_{17}	1727.753	1727.741	-6.6
Y_{-19}	1924.845	1924.831	-7.4
Y_{2}	202.095	202.094	-5.7
Y_{20}	2012.885	2012.873	-5.8
Y_{-23}	2359.025	2359.003	-9.3
Y_{-27}	2811.156	2811.134	-7.8
Y_{3}	289.127	289.126	-6.2
Y_{5}	487.228	487.225	-6.8
Y_{6}	574.260	574.256	-6.9
Y_{9}	921.408	921.402	-6.3

Figure S2: Lists of fragment ions for mono-phosphorylated peptides (*m/z* 1154.84) analyzed in Figure 1C by UVPD-MS. In each case , the 3+ charge state was selected, and UVPD was performed using 2 pulses (1.5 mJ per pulse). The identified site of phosphorylation is shaded in blue in the sequence map above each table. Fragment ions are named as T_n where T is the type of ion (A = a, B = b, C = c, X = x, Y = y, Z = z, for which A,B, and C originate from the N-terminus of the protein and X,Y and Z originate from the C-terminus of the protein), the subscript indicates the number of amino acids contained in the fragment ion, and a plus or minus sign in the subscript designates whether the fragment ion contains one extra hydrogen atom or lacks one hydrogen atom.

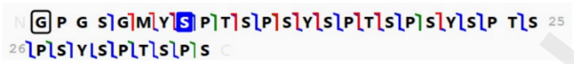
2



Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
A_{+10}	907.398	907.401	3.1
A_{+15}	1428.611	1428.614	2.4
A_{+18}	1713.743	1713.746	1.5
A_{+19}	1810.796	1810.804	4.4
A_{+22}	2227.890	2227.892	1.0
A_{+23}	2324.942	2324.935	-3.1
A_{+33}	3329.387	3329.407	5.9
A_{+9}	806.351	806.347	-4.9
A_{12}	1090.475	1090.475	-0.7
A_{13}	1177.507	1177.503	-4.0
A_{14}	1340.571	1340.575	3.0
A_{16}	1524.655	1524.656	0.4
A_{17}	1625.703	1625.715	7.0
A_{18}	1712.735	1712.750	8.9
A_{19}	1809.788	1809.801	7.3
A_{29}	2946.194	2946.214	6.7
A_{30}	3043.247	3043.241	-2.2
A_{8}	708.290	708.291	0.6
B_{10}	934.385	934.386	0.1
B_{11}	1021.417	1021.419	1.0
B_{15}	1455.598	1455.600	1.3
B_{16}	1552.650	1552.653	1.8
B_{17}	1653.698	1653.701	1.7
B_{18}	1740.730	1740.733	1.4
B_{19}	1837.783	1837.772	-6.0
B_{21}	2087.878	2087.879	0.4
B_{22}	2254.877	2254.881	1.8
B_{24}	2452.977	2452.984	2.8
B_{25}	2540.009	2540.013	1.4
B_{27}	2724.094	2724.120	9.4
B_{29}	2974.189	2974.193	1.4
B_{4}	298.128	298.128	-0.7
B_{7}	649.253	649.253	-0.1
B_{8}	736.285	736.285	0.0
B_{9}	833.338	833.335	-3.1
C_{10}	951.412	951.413	1.6
C_{13}	1222.529	1222.531	1.9
C_{20}	1941.841	1941.844	1.5
C_{3}	228.122	228.122	1.0
C_{5}	372.175	372.176	0.9
C_{6}	503.216	503.216	0.3
C_{7}	666.279	666.281	2.1
X_{+6}	601.247	601.249	2.9
X_{+9}	948.395	948.398	2.8

Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
Y_{10}	1008.440	1008.442	2.0
Y_{12}	1206.540	1206.542	1.5
Y_{13}	1373.539	1373.541	1.2
Y_{14}	1536.602	1536.597	-3.2
Y_{16}	1720.687	1720.690	2.0
Y_{17}	1807.719	1807.716	-1.6
Y_{-19}	2004.812	2004.801	-5.2
Y_{2}	202.095	202.095	-0.8
Y_{20}	2092.851	2092.855	1.5
Y_{23}	2440.000	2440.003	1.3
Y_{-24}	2526.024	2526.047	9.3
Y_{-26}	2724.124	2724.120	-1.7
Y_{-27}	2811.156	2811.169	4.7
Y_{-28}	2974.220	2974.193	-8.8
Y_{3}	289.127	289.127	-1.0
Y_{5}	487.228	487.228	-0.7
Y_{6}	574.260	574.260	-0.6
Y_{7}	737.323	737.324	0.7
Y_{8}	824.355	824.353	-2.8
Y_{9}	921.408	921.408	0.1
Z_{24}	2511.013	2511.012	-0.2

3

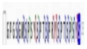


Name	Theoretical Mass	Observed Mass	Mass Difference (ppm)
A_{+12}	1171.449	1171.445	-3.5
A_{+26}	2610.075	2610.063	-4.7
A_{+30}	3044.255	3044.260	1.7
A_{+33}	3329.387	3329.407	5.9
A_{+8}	789.264	789.260	-5.5
A_{+9}	886.317	886.314	-4.0
A_{16}	1604.622	1604.614	-4.7
A_{18}	1792.702	1792.698	-1.7
A_{19}	1889.754	1889.741	-7.0
A_{22}	2226.882	2226.881	-0.1
A_{29}	2946.194	2946.213	6.4
A_{32}	3231.327	3231.318	-2.7
A_{7}	621.258	621.257	-1.4
B_{10}	1014.352	1014.355	3.1
B_{11}	1101.384	1101.385	1.1
B_{13}	1285.469	1285.459	-7.5
B_{14}	1448.532	1448.533	0.9
B_{15}	1535.564	1535.565	0.3
B_{16}	1632.617	1632.621	2.6
B_{17}	1733.664	1733.666	0.9
B_{18}	1820.696	1820.696	0.0
B_{21}	2167.845	2167.844	-0.1
B_{22}	2254.877	2254.879	1.2
B_{24}	2452.977	2452.982	2.1
B_{25}	2540.009	2540.012	1.0
B_{27}	2724.094	2724.113	7.0
B_{29}	2974.189	2974.193	1.2
B_{31}	3172.290	3172.295	1.7
B_{4}	298.128	298.127	-1.0
B_{7}	649.253	649.253	-0.6
B_{8}	816.251	816.252	0.3
C_{10}	1031.378	1031.382	3.6
C_{12}	1215.463	1215.453	-7.8
C_{13}	1302.495	1302.496	0.8
C_{14}	1465.558	1465.563	3.5
C_{16}	1649.643	1649.655	7.3
C_{17}	1750.691	1750.698	4.0
C_{20}	2021.808	2021.807	-0.1
C_{5}	372.175	372.176	0.6
C_{6}	503.216	503.216	-0.1
C_{7}	666.279	666.281	2.3
X_{+6}	601.247	601.249	3.1
Y_{10}	1008.440	1008.441	0.9
Y_{12}	1206.540	1206.541	0.4

Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
Y_{13}	1293.572	1293.575	1.6
Y_{14}	1456.636	1456.643	4.6
Y_{16}	1640.721	1640.722	0.6
Y_{17}	1727.753	1727.755	1.2
Y_{18}	1828.800	1828.811	5.8
Y_{19}	1925.853	1925.852	-0.4
Y_{2}	202.095	202.095	-1.3
Y_{20}	2012.885	2012.887	0.7
Y_{21}	2175.948	2175.950	0.7
Y_{23}	2360.033	2360.035	0.7
Y_{-23}	2359.025	2359.021	-2.0
Y_{-27}	2811.156	2811.170	5.0
Y_{-28}	2974.220	2974.193	-9.0
Y_{3}	289.127	289.127	-1.3
Y_{4}	390.175	390.175	-1.4
Y_{5}	487.228	487.227	-0.9
Y_{6}	574.260	574.259	-1.0
Y_{8}	824.355	824.353	-2.6
Y_{9}	921.408	921.408	-0.3

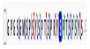
Figure S3: Lists of fragment ions for mono-phosphorylated peptides (*m/z* 1154.84) analyzed in Figure 1C by UVPD-MS. In each case , the 3+ charge state was selected, and UVPD was performed using 2 pulses (1.5 mJ per pulse). The CTD substrate contains four consensus heptads. The identified site of phosphorylation is shaded in blue in the sequence map above each table. Fragment ions are named as T_n where T is the type of ion (A = a, B = b, C = c, X = x, Y = y, Z = z, for which A,B, and C originate from the N-terminus of the protein and X,Y and Z originate from the C-terminus of the protein), the subscript indicates the number of amino acids contained in the fragment ion, and a plus or minus sign in the subscript designates whether the fragment ion contains one extra hydrogen atom or lacks one hydrogen atom.

2



Ion Type	Observed mass	Theoretical mass	Mass difference (ppm)
A+10	893.383	893.383	0.29
A+11	980.415	980.414	-0.29
A+18	1768.796	1768.798	0.75
A+21	2053.929	2053.920	-4.53
A+22	2150.982	2150.987	2.31
A+8	643.287	643.288	1.38
A17	1680.757	1680.763	3.63
A18	1767.789	1767.794	3.27
A20	1965.889	1965.894	2.69
A24	2400.069	2400.076	2.71
A25	2567.068	2567.083	5.93
B13	1205.502	1205.513	9.00
B14	1292.534	1292.539	3.91
B16	1545.688	1545.694	3.79
B17	1708.752	1708.754	1.22
B18	1795.784	1795.789	2.77
B20	1993.884	1993.889	2.28
B21	2080.916	2080.923	3.14
B23	2265.001	2265.007	2.91
B24	2428.064	2428.072	3.32
B25	2595.062	2595.067	1.62
B6	486.190	486.191	2.32
C16	1562.714	1562.718	2.49
C17	1725.778	1725.780	1.52
C24	2445.090	2445.086	-1.64
Y10	1102.422	1102.425	2.86
Y13	1518.639	1518.639	-0.38
Y14	1605.671	1605.675	2.41
Y-16	1802.764	1802.757	-4.04
Y17	1890.804	1890.807	1.60
Y-20	2236.944	2236.951	3.26
Y21	2324.984	2324.990	2.66
Y23	2513.046	2513.053	3.06
Y3	383.109	383.110	1.83
Y6	730.257	730.259	2.32
Y7	817.290	817.291	2.27
Y9	1015.390	1015.392	2.26
Z17	1874.785	1874.787	1.02

3



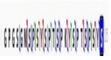
Ion Type	Observed mass	Theoretical mass	Mass difference (ppm)
A+11	980.415	980.414	-1.14
A+18	1848.763	1848.763	-0.13
A+19	1945.816	1945.824	4.56
A+21	2133.895	2133.894	-0.70
A+22	2230.948	2230.947	-0.60
A10	892.375	892.379	4.13
A14	1264.539	1264.538	-0.81
A17	1680.757	1680.758	0.95
A24	2480.036	2480.054	7.29
A25	2567.068	2567.073	2.28
B13	1205.502	1205.513	8.74
B14	1292.534	1292.538	2.66
B16	1545.688	1545.690	1.36
B20	2073.850	2073.854	1.85
B21	2160.882	2160.887	2.18
B23	2344.967	2344.979	4.91
B24	2508.030	2508.037	2.51
C13	1222.529	1222.530	0.78
C16	1562.714	1562.716	1.10
C24	2525.057	2525.065	3.15
C9	774.333	774.335	2.67
X+16	1830.759	1830.776	9.63
X+9	962.411	962.408	-2.99
Y10	1102.422	1102.425	2.36
Y13	1518.639	1518.638	-0.78
Y-13	1517.631	1517.631	-0.41
Y-13	1517.631	1517.631	-0.41
Y14	1605.671	1605.675	2.17
Y-16	1802.764	1802.756	-4.53
Y17	1890.804	1890.805	0.88
Y18	2053.867	2053.866	-0.25
Y-20	2236.944	2236.949	2.27
Y21	2324.984	2324.990	2.79
Y-23	2512.038	2512.052	5.62
Y3	303.143	303.144	1.57
Y6	650.291	650.292	1.78
Y7	737.323	737.325	1.91
Z19	2124.880	2124.875	-2.59

4

Ion Type	Observed mass	Theoretical mass	Mass difference (ppm)
A+11	980.415	980.413	-1.84
A+12	1077.468	1077.473	5.50
A+16	1598.667	1598.670	1.70
A+18	1848.763	1848.760	-1.25
A+21	2133.895	2133.901	2.57
A10	892.375	892.378	3.59
A17	1760.723	1760.725	1.08
A22	2229.940	2229.958	8.03
A25	2567.068	2567.073	2.05
B10	920.370	920.374	4.10
B11	1007.402	1007.402	0.43
B17	1788.718	1788.719	0.49
B20	2073.850	2073.858	3.71
B21	2160.882	2160.887	2.07
B23	2344.967	2344.976	3.75
B24	2508.030	2508.038	2.88
C16	1642.681	1642.685	2.35
C20	2090.877	2090.880	1.49
C24	2525.057	2525.079	8.79
C9	774.333	774.335	2.88
X+16	1830.759	1830.768	5.28
X+9	962.411	962.412	1.50
Y10	1022.456	1022.453	-2.85
Y13	1438.673	1438.671	-1.21
Y-13	1437.665	1437.666	0.50
Y14	1605.671	1605.673	1.31
Y16	1803.772	1803.769	-1.67
Y17	1890.804	1890.808	2.39
Y-20	2236.944	2236.947	1.18
Y21	2324.984	2324.983	-0.46
Y-23	2512.038	2512.035	-1.08
Y3	303.143	303.144	2.17
Y3	303.143	303.144	2.17
Y6	650.291	650.292	1.83
Y7	737.323	737.326	3.45
Z19	2124.880	2124.871	-4.28

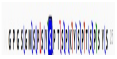
Figure S4: Lists of fragment ions for mono-phosphorylated peptides (m/z 938.07) analyzed in Figure 2C by UVPD-MS. In each case , the 3+ charge state was selected, and UVPD was performed using 2 pulses (1.5 mJ per pulse). The CTD substrate contains three consensus heptads where the 7th position is occupied by arginine instead of serine in the middle heptad. The identified site of phosphorylation is shaded in blue in the sequence map above each table. Fragment ions are named as T_n where T is the type of ion (A = a, B = b, C = c, X = x, Y = y, Z = z, for which A,B, and C originate from the N-terminus of the protein and X,Y and Z originate from the C-terminus of the protein), the subscript indicates the number of amino acids contained in the fragment ion, and a plus or minus sign in the subscript designates whether the fragment ion contains one extra hydrogen atom or lacks one hydrogen atom.

2



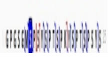
Ion type	Theoretical mass	Observed mass	Mass Difference (ppm)
A+12	1077.468	1077.468	0.503002
A+26	2637.122	2637.142	7.678434
A+8	643.2874	643.2873	-0.06068
A10	892.3749	892.374	-1.02535
A17	1652.75	1652.751	0.345485
A18	1739.782	1739.776	-3.5924
A22	2121.968	2121.967	-0.32281
A24	2372.063	2372.066	1.294232
B13	1205.502	1205.503	0.188303
B14	1292.534	1292.535	0.33268
B16	1517.682	1517.68	-1.21172
B18	1767.777	1767.778	0.256254
B20	1965.878	1965.878	0.236536
B21	2052.91	2052.909	-0.36193
B23	2236.995	2236.999	1.889589
B24	2400.058	2400.059	0.42749
B25	2567.056	2567.058	0.828186
B6	486.1897	486.189	-1.33898
C10	937.3961	937.3974	1.408156
C16	1534.708	1534.708	-0.42288
C20	1982.904	1982.908	1.732812
C23	2254.021	2254.021	-0.12023
C6	503.216	503.2146	-2.69069
C9	774.3328	774.3324	-0.50495
Y11	1265.485	1265.484	-0.91664
Y13	1490.633	1490.63	-1.78045
Y-13	1489.625	1489.622	-2.18039
Y14	1577.665	1577.664	-0.38982
Y16	1775.766	1775.764	-0.82162
Y17	1862.798	1862.796	-0.68553
Y20	2209.946	2209.946	0.151135
Y-20	2208.938	2208.937	-0.60163
Y21	2296.978	2296.977	-0.36004
Y23	2485.04	2485.038	-0.75934
Y3	383.1094	383.1089	-1.33904
Y6	730.2575	730.2573	-0.30948
Y7	817.2895	817.2894	-0.14683
Y9	1015.39	1015.39	-0.38901

3



Ion type	Theoretical mass	Observed mass	Mass Difference (ppm)
A+16	1570.661	1570.66	-0.56157
A+18	1820.757	1820.753	-1.87836
A+21	2105.889	2105.891	0.709899
A10	892.3749	892.3739	-1.06233
A16	1569.653	1569.659	3.651124
A17	1732.717	1732.722	3.163241
A19	1916.802	1916.79	-6.1733
A22	2201.934	2201.953	8.75276
A25	2539.061	2539.062	0.238671
B14	1372.501	1372.501	0.008015
B17	1760.712	1760.713	0.57136
B18	1847.744	1847.745	0.498987
B21	2132.876	2132.876	-0.08627
B6	486.1897	486.1892	-1.02429
C13	1302.495	1302.494	-0.73014
C16	1614.675	1614.674	-0.29232
C19	1961.823	1961.826	1.814639
Y12	1313.614	1313.612	-1.75394
Y13	1410.667	1410.666	-0.55151
Y-13	1409.659	1409.657	-1.42656
Y14	1497.699	1497.7	0.524137
Y-16	1694.791	1694.799	4.25895
Y17	1862.798	1862.796	-0.64527
Y18	2025.861	2025.858	-1.49763
Y20	2209.946	2209.944	-0.74391
Y-20	2208.938	2208.936	-0.94659
Y21	2296.978	2296.976	-0.67654
Y23	2485.04	2485.035	-2.01486
Y3	303.143	303.1428	-0.72903
Y6	650.2912	650.2912	0.032293
Y7	737.3232	737.3231	-0.07053
Y9	935.4236	935.424	0.450063
Z19	2096.874	2096.853	-9.90713

4



Ion type	Theoretical mass	Observed mass	Mass Difference (ppm)
A+18	1820.757	1820.757	0.097744
A10	972.3412	972.3377	-3.63247
A17	1732.717	1732.718	0.630801
A17	1732.717	1732.718	0.630801
B10	1000.336	1000.333	-3.02398
B11	1087.368	1087.369	0.823088
B13	1285.469	1285.457	-8.97105
B17	1760.712	1760.712	0.111319
B18	1847.744	1847.743	-0.25003
B20	2045.844	2045.843	-0.33336
B21	2132.876	2132.876	-0.28506
B24	2480.024	2480.023	-0.41209
B25	2567.056	2567.057	0.248923
B6	486.1897	486.1896	-0.0761
C16	1614.675	1614.674	-0.30594
C9	854.2991	854.2985	-0.7035
Y11	1185.519	1185.519	0.295229
Y13	1410.667	1410.666	-0.50756
Y-13	1409.659	1409.655	-2.4041
Y14	1497.699	1497.699	0.223676
Y-16	1694.791	1694.797	3.604002
Y17	1782.831	1782.832	0.379733
Y-20	2128.971	2128.969	-1.01785
Y21	2296.978	2296.974	-1.51068
Y-22	2427.01	2427.01	-0.20188
Y3	303.143	303.1427	-1.04901
Y6	650.2912	650.2907	-0.7043
Y7	737.3232	737.3227	-0.60353
Y9	935.4236	935.4249	1.376916
Z19	2016.908	2016.894	-7.07618

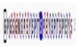
Figure S5: Lists of fragment ions for mono-phosphorylated peptides (*m/z* 928.73) analyzed in Figure 2D by UVPD-MS. In each case , the 3+ charge state was selected, and UVPD was performed using 2 pulses (1.5 mJ per pulse).The CTD substrate contains three consensus heptads where the 7th position is occupied by lysine instead of serine in each heptad. The identified site of phosphorylation is shaded in blue in the sequence map above each table. Fragment ions are named as T_n where T is the type of ion (A = a, B = b, C = c, X = x, Y = y, Z = z, for which A,B, and C originate from the N-terminus of the protein and X,Y and Z originate from the C-terminus of the protein), the subscript indicates the number of amino acids contained in the fragment ion, and a plus or minus sign in the subscript designates whether the fragment ion contains one extra hydrogen atom or lacks one hydrogen atom.

2

Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
$A+_{12}$	1119.478	1119.484	5.6
$A+_{14}$	1307.558	1307.561	2.6
$A+_{15}$	1404.611	1404.618	5.0
$A+_{22}$	2245.900	2245.911	5.1
$A+_{6}$	459.203	459.204	4.0
$A+_{8}$	643.287	643.291	4.9
$A+_{9}$	772.330	772.333	3.4
A_{10}	934.385	934.391	6.4
A_{11}	1021.417	1021.425	7.5
A_{13}	1219.518	1219.515	-2.1
A_{15}	1403.603	1403.610	5.5
A_{17}	1695.709	1695.717	4.9
A_{18}	1782.741	1782.753	6.8
A_{23}	2373.935	2373.955	8.7
A_{25}	2624.030	2624.040	3.9
A_{26}	2721.083	2721.105	8.3
A_{4}	270.133	270.134	3.7
A_{7}	545.227	545.230	5.9
B_{10}	962.380	962.385	4.9
B_{11}	1049.412	1049.419	5.8
B_{13}	1247.513	1247.519	5.3
B_{14}	1334.545	1334.552	5.1
B_{15}	1431.598	1431.606	6.1
B_{16}	1560.640	1560.649	5.9
B_{17}	1723.704	1723.714	5.8
B_{18}	1810.736	1810.747	6.2
B_{20}	2008.836	2008.835	-0.4
B_{23}	2401.930	2401.949	8.1
B_{24}	2564.993	2565.007	5.2
B_{25}	2652.025	2652.039	5.2
B_{26}	2749.078	2749.101	8.4
B_{4}	298.128	298.129	3.7
B_{6}	486.190	486.192	4.4
B_{7}	573.222	573.224	4.7
B_{8}	670.274	670.276	2.5
B_{9}	799.317	799.322	6.1
C_{13}	1264.539	1264.546	5.7
C_{16}	1577.667	1577.677	6.3
C_{17}	1740.730	1740.742	6.8
C_{2}	171.101	171.101	5.1
C_{23}	2418.956	2418.967	4.5
C_{3}	228.122	228.123	4.9
C_{4}	315.154	315.155	4.1
C_{5}	372.175	372.177	5.2

Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
C_{8}	687.301	687.305	6.8
C_{9}	816.343	816.349	6.6
$X+_{12}$	1463.526	1463.528	1.3
$X+_{15}$	1748.658	1748.665	3.8
$X+_{17}$	1932.743	1932.751	4.3
$X+_{8}$	987.335	987.337	2.2
X_{18}	2094.798	2094.806	3.7
X_{7}	885.279	885.284	4.8
Y_{10}	1144.433	1144.438	5.1
Y_{11}	1307.496	1307.494	-1.8
Y_{12}	1436.538	1436.536	-2.0
Y_{13}	1533.591	1533.598	4.1
Y_{14}	1620.623	1620.633	6.0
Y_{16}	1818.724	1818.734	5.6
$Y-_{16}$	1817.716	1817.726	5.7
Y_{17}	1905.756	1905.771	7.8
Y_{2}	216.111	216.112	4.6
$Y-_{20}$	2293.907	2293.916	4.3
Y_{21}	2381.946	2381.959	5.4
$Y-_{22}$	2511.979	2511.978	-0.5
$Y-_{23}$	2569.001	2569.014	5.1
Y_{3}	303.143	303.144	3.8
Y_{4}	466.206	466.208	3.7
Y_{6}	692.302	692.305	5.0
Y_{7}	859.300	859.305	5.7
Y_{9}	1057.401	1057.406	5.4
Z_{12}	1420.520	1420.532	8.7
Z_{3}	287.124	287.126	4.3
Z_{5}	579.230	579.233	4.9

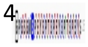
3



Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
$A+_{12}$	1119.478	1119.483	4.5
$A+_{16}$	1613.619	1613.626	3.7
$A+_{17}$	1776.683	1776.699	9.2
$A+_{18}$	1863.715	1863.723	4.2
$A+_{19}$	1960.768	1960.776	4.4
$A+_{21}$	2148.847	2148.853	2.7
$A+_{22}$	2245.900	2245.911	5.0
$A+_{25}$	2625.038	2625.048	3.8
$A+_{8}$	643.287	643.291	4.9
$A+_{9}$	772.330	772.333	3.3
A_{10}	934.385	934.391	6.0
A_{11}	1021.417	1021.426	8.4
A_{13}	1219.518	1219.517	-0.8
A_{14}	1386.516	1386.515	-1.1
A_{16}	1612.612	1612.626	9.0
A_{23}	2373.935	2373.954	8.0
A_{26}	2721.083	2721.103	7.5
A_{4}	270.133	270.134	4.1
A_{7}	545.227	545.230	5.9
A_{9}	771.322	771.327	6.6
B_{10}	962.380	962.385	5.2
B_{11}	1049.412	1049.419	6.1
B_{13}	1247.513	1247.520	5.4
B_{16}	1640.607	1640.618	6.9
B_{17}	1803.670	1803.682	6.9
B_{18}	1890.702	1890.716	7.7
B_{20}	2088.802	2088.816	6.4
B_{21}	2175.834	2175.847	6.0
B_{23}	2401.930	2401.949	7.9
B_{24}	2564.993	2565.007	5.4
B_{25}	2652.025	2652.039	5.3
B_{3}	211.096	211.097	4.4
B_{4}	298.128	298.129	4.0
B_{6}	486.190	486.192	4.6
B_{7}	573.222	573.225	5.1
B_{8}	670.274	670.278	5.3
B_{9}	799.317	799.322	6.6
C_{13}	1264.539	1264.550	8.9
C_{15}	1528.590	1528.594	2.4
C_{16}	1657.633	1657.644	6.7
C_{20}	2105.829	2105.840	5.3
C_{3}	228.122	228.123	4.9
C_{4}	315.154	315.155	3.8
C_{5}	372.175	372.178	5.5

Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
C_{6}	503.216	503.219	5.0
C_{8}	687.301	687.306	7.3
C_{9}	816.343	816.349	6.9
$X+_{11}$	1254.517	1254.529	9.5
$X+_{12}$	1383.559	1383.563	2.8
$X+_{13}$	1480.612	1480.623	7.5
$X+_{15}$	1748.658	1748.670	6.8
$X+_{9}$	1004.421	1004.420	-1.2
X_{12}	1382.551	1382.559	5.3
X_{15}	1747.650	1747.648	-1.5
X_{18}	2094.798	2094.809	5.2
X_{19}	2223.841	2223.849	3.5
X_{8}	906.361	906.366	6.1
Y_{10}	1064.466	1064.473	5.9
Y_{11}	1227.530	1227.537	5.7
Y_{13}	1453.625	1453.632	5.0
$Y-_{13}$	1452.617	1452.621	3.0
Y_{14}	1620.623	1620.634	6.7
Y_{15}	1721.671	1721.678	3.8
Y_{16}	1818.724	1818.734	5.9
$Y-_{16}$	1817.716	1817.726	5.8
Y_{17}	1905.756	1905.771	7.9
$Y-_{18}$	2067.811	2067.823	5.4
Y_{2}	216.111	216.112	4.6
$Y-_{20}$	2293.907	2293.917	4.6
Y_{21}	2381.946	2381.960	5.5
$Y-_{23}$	2569.001	2569.010	3.6
$Y-_{25}$	2713.054	2713.067	4.8
$Y-_{25}$	2713.054	2713.047	-2.5
Y_{3}	303.143	303.144	3.8
Y_{4}	466.206	466.208	4.2
Y_{6}	692.302	692.306	5.5
Y_{7}	779.334	779.339	6.2
Y_{8}	880.381	880.388	7.2
Y_{9}	977.434	977.439	5.2
Z_{3}	287.124	287.126	4.3

4



Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
$A+_{10}$	1015.360	1015.359	-0.2
$A+_{12}$	1199.444	1199.448	3.0
$A+_{14}$	1387.524	1387.525	0.8
$A+_{16}$	1613.619	1613.625	3.5
$A+_{18}$	1863.715	1863.720	2.8
$A+_{19}$	1960.768	1960.777	4.9
$A+_{25}$	2625.038	2625.052	5.3
$A+_{8}$	723.254	723.257	5.1
$A+_{9}$	852.296	852.300	3.9
A_{10}	1014.352	1014.354	2.2
A_{11}	1101.384	1101.389	4.3
A_{16}	1612.612	1612.616	2.8
A_{17}	1775.675	1775.683	4.2
A_{22}	2244.892	2244.893	0.3
A_{23}	2373.935	2373.953	7.8
A_{26}	2721.083	2721.099	6.0
A_{4}	270.133	270.134	3.7
A_{7}	625.193	625.197	5.9
B_{10}	1042.347	1042.353	5.8
B_{11}	1129.379	1129.385	5.3
B_{12}	1226.431	1226.436	3.8
B_{13}	1327.479	1327.480	0.8
B_{14}	1414.511	1414.518	4.9
B_{15}	1511.564	1511.573	6.1
B_{16}	1640.607	1640.616	5.6
B_{17}	1803.670	1803.681	6.1
B_{18}	1890.702	1890.714	6.6
B_{20}	2088.802	2088.817	7.1
B_{21}	2175.834	2175.846	5.1
B_{23}	2401.930	2401.945	6.5
B_{24}	2564.993	2565.004	4.2
B_{25}	2652.025	2652.037	4.6
B_{3}	211.096	211.097	3.9
B_{4}	298.128	298.129	3.3
B_{6}	486.190	486.192	4.2
B_{7}	653.188	653.191	4.7
C_{10}	1059.373	1059.378	5.0
C_{12}	1243.458	1243.459	1.1
C_{13}	1344.505	1344.512	4.6
C_{15}	1528.590	1528.597	4.7
C_{16}	1657.633	1657.642	5.7
C_{20}	2105.829	2105.836	3.5
C_{3}	228.122	228.123	4.5
C_{4}	315.154	315.155	4.1

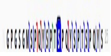
Ion Type	Theoretical Mass	Observed Mass	Mass Difference (ppm)
C_{8}	767.267	767.270	4.3
C_{9}	896.310	896.316	6.5
$X+_{12}$	1383.559	1383.567	5.3
$X+_{18}$	2015.840	2015.853	6.5
X_{15}	1667.684	1667.685	0.7
X_{16}	1764.737	1764.738	0.5
X_{19}	2143.875	2143.890	7.2
Y_{10}	1064.466	1064.472	5.2
Y_{11}	1227.530	1227.536	4.9
Y_{13}	1453.625	1453.632	4.7
Y_{14}	1540.657	1540.666	5.8
Y_{16}	1738.757	1738.767	5.6
$Y-_{16}$	1737.750	1737.758	5.1
Y_{17}	1825.789	1825.801	6.2
Y_{19}	2117.895	2117.908	6.0
Y_{2}	216.111	216.112	4.1
$Y-_{20}$	2213.940	2213.948	3.6
$Y-_{21}$	2380.939	2380.950	4.7
$Y-_{22}$	2511.979	2511.981	0.8
$Y-_{23}$	2569.001	2569.005	1.7
Y_{25}	2714.062	2714.083	7.9
Y_{3}	303.143	303.144	3.5
Y_{4}	466.206	466.208	3.7
Y_{5}	595.249	595.251	3.3
Y_{6}	692.302	692.305	4.9
Y_{7}	779.334	779.338	5.1
Y_{9}	977.434	977.439	4.7
Z_{3}	287.124	287.125	3.6
Z_{5}	579.230	579.233	4.0

Figure S6: Lists of fragment ions for mono-phosphorylated peptides (*m/z* 957.03) analyzed in Figure 2E by UVPD-MS. In each case , the 3+ charge state was selected, and UVPD was performed using 2 pulses (1.5 mJ per pulse). The CTD substrate contains three consensus heptads where the 7th position is occupied by glutamate instead of serine in the middle heptad. The identified site of phosphorylation is shaded in blue in the sequence map above each table. Fragment ions are named as T_n where T is the type of ion (A = a, B = b, C = c, X = x, Y = y, Z = z, for which A,B, and C originate from the N-terminus of the protein and X,Y and Z originate from the C-terminus of the protein), the subscript indicates the number of amino acids contained in the fragment ion, and a plus or minus sign in the subscript designates whether the fragment ion contains one extra hydrogen atom or lacks one hydrogen atom.

2

Ion type	Theoretical mass	Observed mass	Mass Difference (ppm)
A+11	1021.441	1021.444	2.91
A+12	1118.494	1118.492	-1.77
A10	933.4014	933.4022	0.80
B10	961.3964	961.3988	2.54
B11	1048.428	1048.432	3.60
B13	1246.529	1246.533	3.47
B14	1333.561	1333.565	3.07
B16	1558.672	1558.681	5.84
B17	1721.736	1721.74	2.64
B18	1808.768	1808.773	2.94
B20	2006.868	2006.874	2.94
B6	486.1897	486.1906	2.02
B9	798.333	798.3353	2.89
C10	978.4227	978.427	4.48
C16	1575.699	1575.703	2.94
C23	2416.004	2416.009	2.21
C9	815.3593	815.3622	3.51
Y10	1143.449	1143.452	2.87
Y11	1306.512	1306.523	8.46
Y13	1531.623	1531.629	3.86
Y-13	1530.615	1530.619	2.42
Y14	1618.655	1618.665	5.74
Y16	1816.756	1816.762	3.27
Y17	1903.788	1903.795	3.95
Y18	2066.851	2066.855	1.83
Y21	2378.994	2379.004	3.93
Y4	466.2064	466.2077	2.83
Y5	594.2649	594.266	1.73
Y6	691.3177	691.3194	2.46
Y7	858.3161	858.318	2.30
Y8	959.3637	959.3639	0.16
Y9	1056.417	1056.42	2.96
Z5	578.2462	578.2476	2.41

3



Ion type	Theoretical mass	Observed mass	Mass Difference (ppm)
A+11	1021.441	1021.432	-9.48
A+12	1118.494	1118.488	-5.85
A10	933.4014	933.4029	1.57
A9	770.3381	770.3371	-1.29
B10	961.3964	961.3988	2.53
B11	1048.428	1048.432	3.70
B13	1246.529	1246.533	3.03
B14	1413.527	1413.533	3.94
B16	1638.639	1638.645	3.90
B16	1638.639	1638.645	3.96
B17	1801.702	1801.706	2.37
B18	1888.734	1888.747	6.87
B20	2086.834	2086.841	3.23
B21	2173.866	2173.872	2.46
B24	2562.041	2562.049	3.15
B6	486.1897	486.1906	1.98
C16	1655.665	1655.669	2.79
C8	687.3008	687.3017	1.35
C9	815.3593	815.362	3.22
X+9	1003.437	1003.434	-3.67
X11	1252.525	1252.521	-3.35
Y10	1063.482	1063.486	3.52
Y11	1226.546	1226.551	4.53
Y13	1451.657	1451.661	2.90
Y13	1451.657	1451.664	5.12
Y-13	1450.649	1450.644	-3.67
Y16	1816.756	1816.76	2.60
Y17	1903.788	1903.795	3.95
Y18	2066.851	2066.847	-1.73
Y20	2291.962	2291.968	2.55
Y21	2378.994	2379.002	3.39
Y4	466.2064	466.2074	2.15
Y6	691.3177	691.3192	2.11
Y7	778.3497	778.3521	3.04
Y8	879.3974	879.4007	3.70
Y9	976.4502	976.4525	2.39

4

Ion type	Theoretical mass	Observed mass	Mass Difference (ppm)
A+6	459.2026	459.2037	2.40
B10	1041.363	1041.366	2.80
B11	1128.395	1128.397	2.07
B13	1326.495	1326.496	0.48
B14	1413.527	1413.53	2.01
B14	1413.527	1413.531	2.91
B16	1638.639	1638.644	3.16
B17	1801.702	1801.706	2.47
B18	1888.734	1888.739	2.92
B20	2086.834	2086.835	0.32
B21	2173.866	2173.872	2.65
B23	2398.978	2398.992	5.92
B24	2562.041	2562.049	3.04
B6	486.1897	486.1906	2.00
B7	653.188	653.1893	1.88
B9	878.2994	878.2965	-3.21
C16	1655.665	1655.67	2.99
C9	895.3257	895.3239	-1.91
X+11	1253.533	1253.541	7.03
X+8	906.3845	906.3809	-3.94
Y10	1063.482	1063.486	3.47
Y13	1451.657	1451.661	3.14
Y14	1538.689	1538.695	3.84
Y16	1736.789	1736.795	3.16
Y-16	1735.782	1735.787	3.16
Y17	1823.821	1823.826	2.79
Y-20	2210.988	2210.992	1.88
Y21	2378.994	2379.002	3.00
Y4	466.2064	466.2073	1.90
Y6	691.3177	691.3192	2.17
Y7	778.3497	778.3522	3.18
Y9	976.4502	976.4526	2.49
Z16	1720.771	1720.781	6.31

Figure S7: Lists of fragment ions for mono-phosphorylated peptides (*m/z* 956.07) analyzed in Figure 2F by UVPD-MS. In each case , the 3+ charge state was selected, and UVPD was performed using 2 pulses (1.5 mJ per pulse). The CTD substrate contains three consensus heptads where the 7th position is occupied by glutamine instead of serine in every heptad. The identified site of phosphorylation is shaded in blue in the sequence map above each table. Fragment ions are named as T_n where T is the type of ion (A = a, B = b, C = c, X = x, Y = y, Z = z, for which A,B, and C originate from the N-terminus of the protein and X,Y and Z originate from the C-terminus of the protein), the subscript indicates the number of amino acids contained in the fragment ion, and a plus or minus sign in the subscript designates whether the fragment ion contains one extra hydrogen atom or lacks one hydrogen atom.

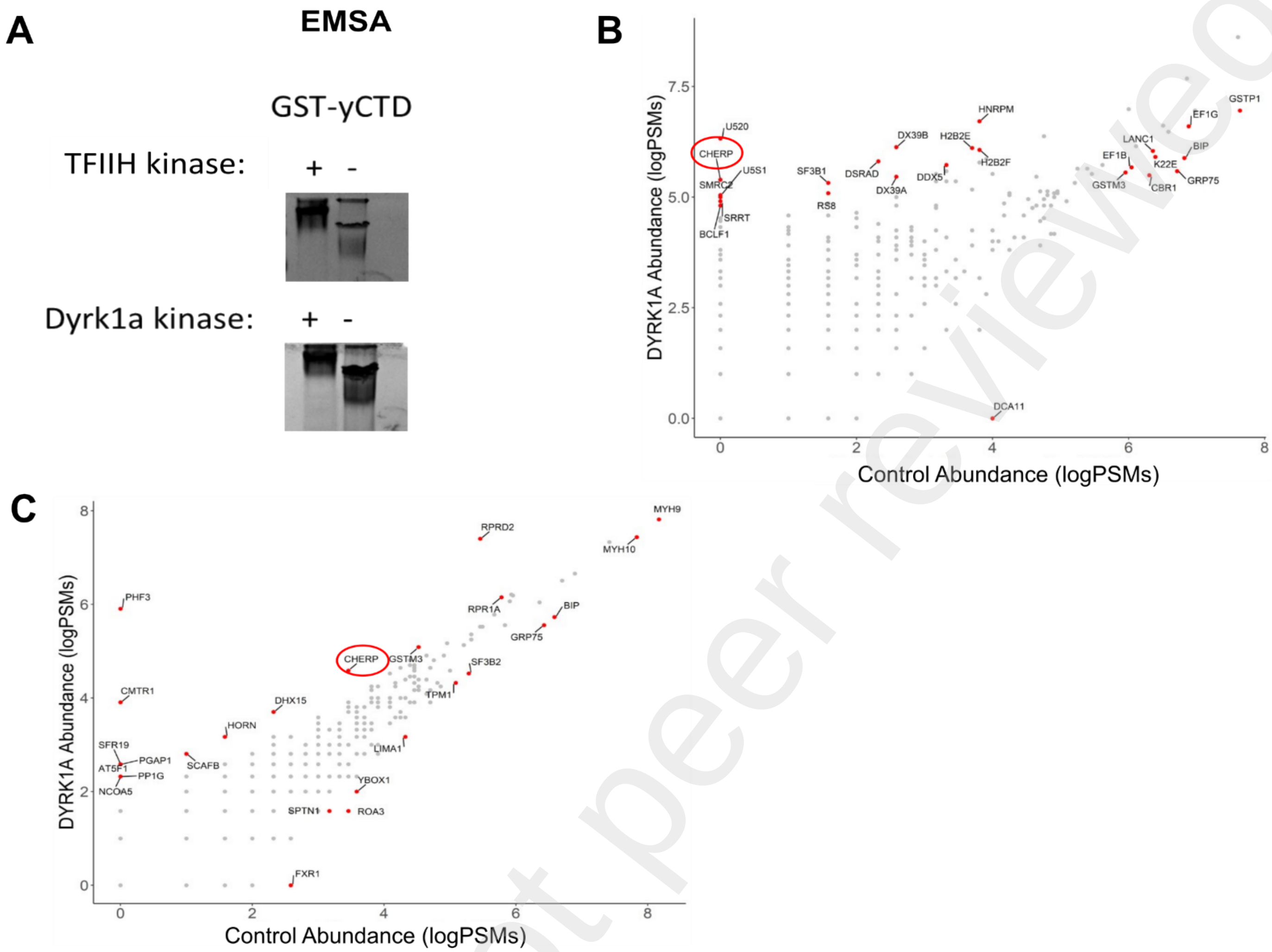


Figure S8. Identification of CHERP in Dyrk1a-treated pulldown sample. (A) SDS-PAGE EMSA of 26 repeat yeast CTD and treated with TFIIH (top, right band) and Dyrk1a (bottom, left band) paired with control reactions in the absence of either kinase (right bands) before initiating pulldown experiment with CTD as the bait protein. (B-C) Spectral counts of proteins identified in the dyrk1a sample compared to the control sample (unphosphorylated) for the first replicate (top graph) and second replicate (bottom graph) with CHERP highlighted as being differentially enriched in the Dyrk1a-treated sample.

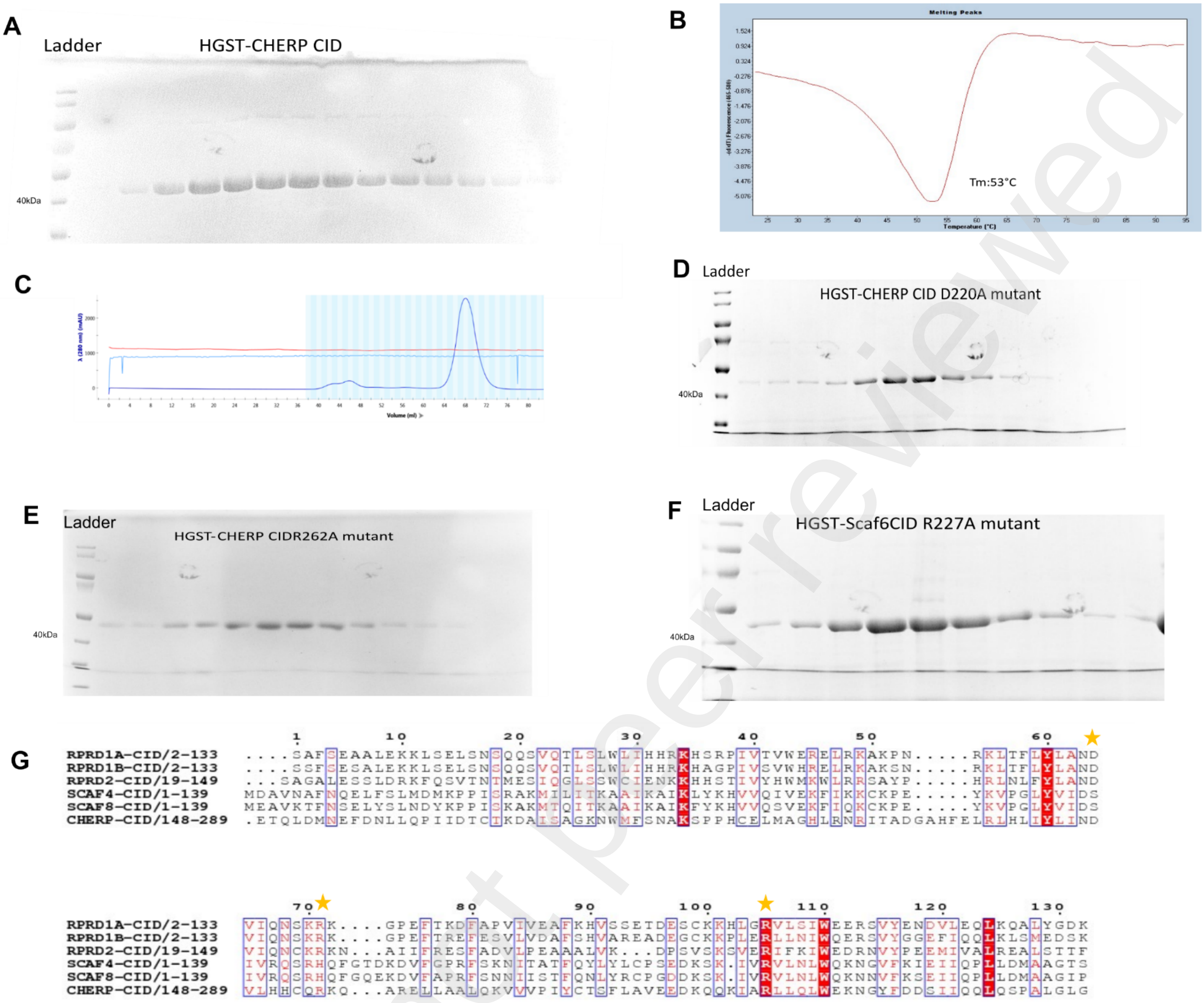


Figure S9. Protein purification and characterization of CHERP. (A) Coomassie-stained gel of individual fractions from gel filtration chromatography of purified CHERP CID domain. (B) Differential scanning fluorometry plot showing the melting temperature of the CID domain of CHERP. (C) Size exclusion chromatography profile of purified recombinant CID domain of CHERP. (D-F) Coomassie-stained gel of fractions from size-exclusion chromatography for various CID domain mutants of CHERP. (G) Multiple sequence alignment of CID domains from RPRD1A (Q96P16), RPRD1B (Q9NQG5), RPRD2 (Q5VT52), SCAF4 (O95104) , SCAF8 (Q9UPN6), and CHERP(Q8IWX8). Conserved residues are boxed and highlighted in red. Star symbols denote residues that were mutated.

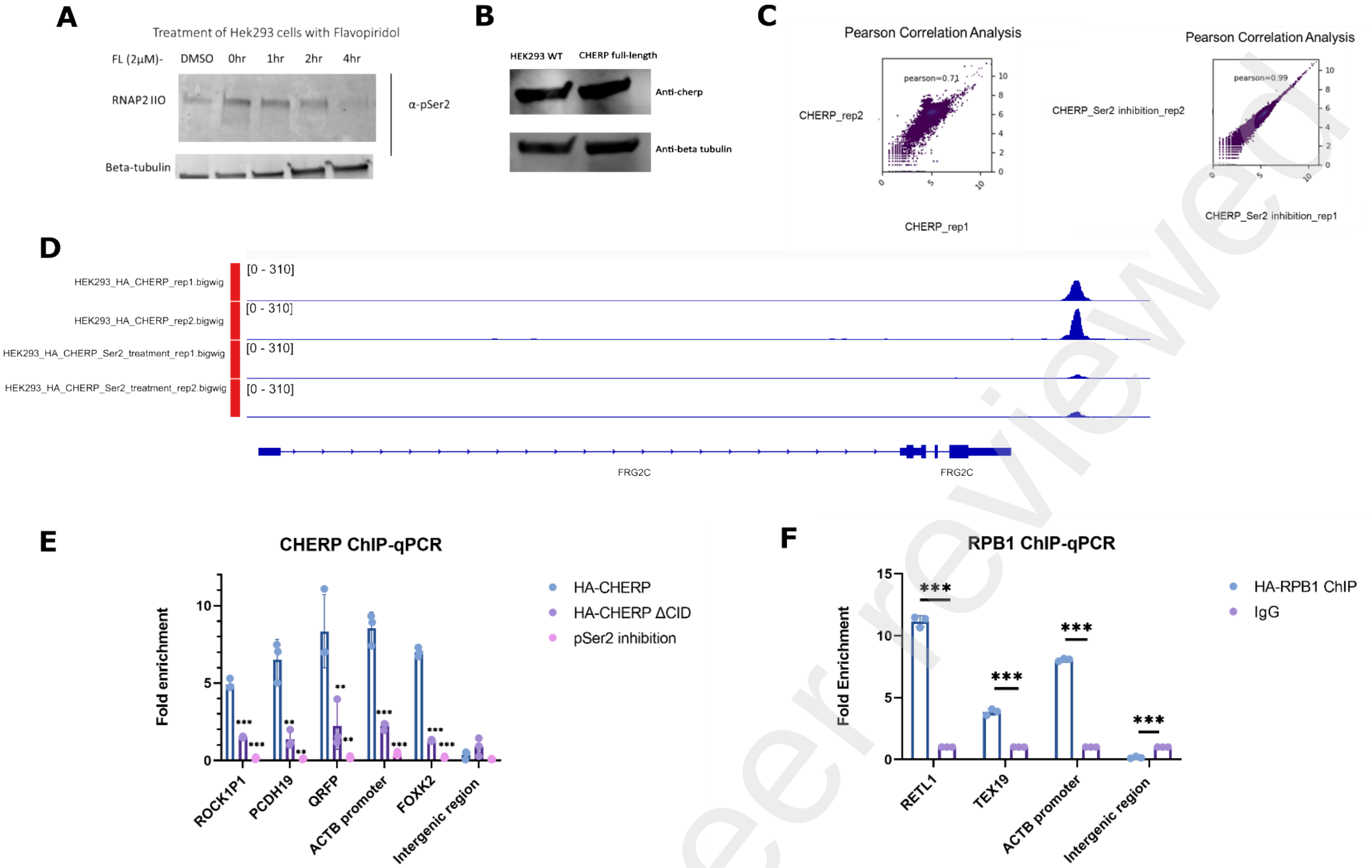


Figure S10. Quality Control of ChIP-seq data. (A) Time course of HEK293 cells treated with flavopiridol and cell lysate was probed for levels of pSer2 Pol II (B) ectopic overexpression of full-length CHERP in HEK293 cells compared to wild-type expression of CHERP used in ChIP assay (C) Scatter plots showing the pearson correlation between ChIP-seq replicate datasets of HA-CHERP-bound regions and between pSer2 inhibition replicates. The genome was divided into bins of 15 kb and the number of mapped reads in the individual bins was calculated. (D) IGV track example of CHERP replicates and biological replicates of CHERP with pSer2 inhibition at a selected genomic site. (E) ChIP-qPCR analysis of peaks at promoter sites of selected genes from three biological replicates for HA-CHERP WT, CHERP ΔCID, HA-CHERP pSer2 inhibition, and IgG control samples. (F) ChIP-qPCR analysis of HA-RPB1 biological replicates. Fold enrichment was calculated by comparing the positive locus sequence in ChIP DNA over the negative IgG sample. For each data point, *n* = 3,

error bars indicate standard deviation of three biological replicates. $^{**}p \le 0.001$, $^{***}p \le 0.0001$.

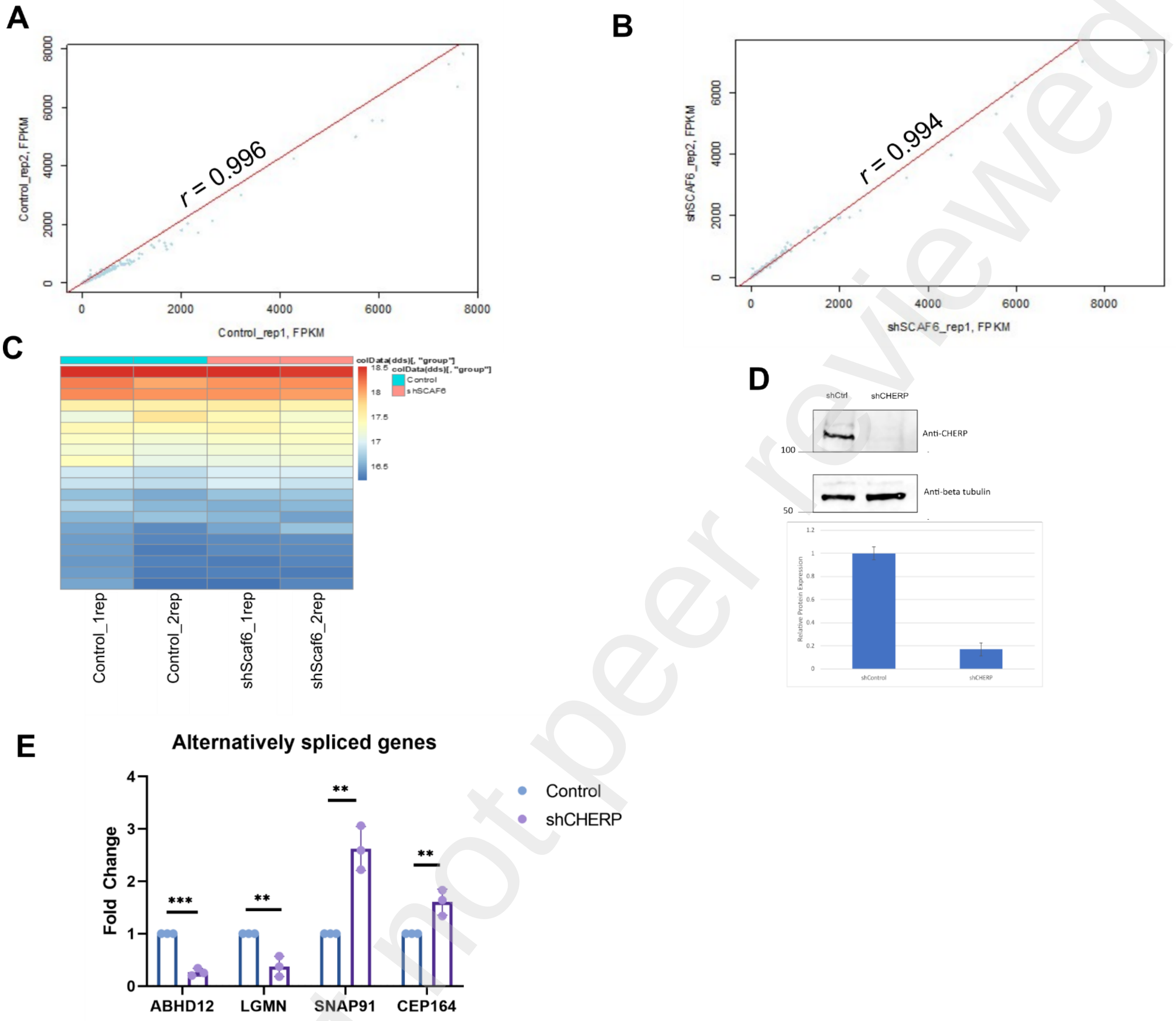


Figure S11. Quality control of RNA-seq data. (A) Control and (B) shSCAF6 (shCHERP) HEK293 RNA-Seq data shows strong between-replicate correlation. (C) Biological replicates of shControl and shSCAF6 samples cluster according to the condition. The heatmap was built using Bioconductor package ‘DESeq2’ on rlog-normalized counts. (D) Western blot analysis of CHERP knockdown efficiency by shRNA where representative blot of shCHERP vs control (MISSION non-mammalian shRNA control plasmid) shown (50 μg each). Quantification of western blot was done with three biological replicates and is shown below. (E) Relative transcript expression of various CHERP-controlled isoforms normalized by total mRNA of each target gene.

For each data point, *n* = 3, error bars indicate standard deviation of three biological replicates. *p* ≤ 0.001 (**),****p* ≤ 0.0001.

Supplementary Table 4

QPCR primers	
Gene	**Sequence**
LGMN	FP: AGT GGC ACA ATC TTG GCT CA RP:ACCATTCTGCACCTTGGAGT
ABHD12	FP:TCTTTGCCTTGGGCGTTCTTC RP:GCACTCCACGTTTTTGACTGG
CEP164	FP:AGTGTCCACAGCTCAAGTGA RP:ACATCCTTCTTCTCCTCTGG
SNAP91	FP:GAGAGGATTCTTTGGCTGC RP:AACAGTTGTAGTAGTGGAGGC
LGMN normalization	FP: TGGAAGATTCGGACGTGGAAG RP: ATACTGCATGACGTGGCTGG
ABHD12 normalization	FP: GACCATTGGAGTCTGGCACA RP: GGAAGCCAAGGCATCCTCAT
CEP164 normalization	FP: TGGGGGAGCGGGAGAAATA RP: ACCTCTCCTCCTTAGCCCAAT
SNAP91 normalization	FP: TGGAGACGCTTGAACAGCAT RP: AGAGGGAGCACCAGATCCTT
FOXK2	FP: ACCACAGGGAGGTCAAAGGTA RP: TGGTCTCCCCTCTCCTCCTTT
RTEL1	FP: TGGAAAACCCCAAGTGTGGC RP: ACGGAAACGTGGAAACCAAGG
TEX19	FP: TTCCCTCAGTTTCCCTCAAG RP: AGGGAACCTGAGGGAAGCT
ACTB promoter	FP: GTGCAATCAAAGTCCTCGG RP: CAAGATGAGATTGGCATGGC
PCDH19	FP: TTTGACAAGTCTTTGTACTT RP: CACCTTTCTAATGGAACCCC
ROCK1P1	FP: TTGCGCCTTTTCCAAGGCA RP: GAACCGCAAGGAACCTTCC
QRFP	FP: GTTGAAGTCCTCGTTGTCTTG RP: CCTACCTGTGGATGAAGTT
Intergenic region	FP: TGGTGGCTAGGAGCTACCAT RP: GACAATAAACCACCATGCAG