Insights into the RNA Polymerase II CTD code

ROLAND SCHÜLLER

Juni 2013

AFFIDAVIT (EIDESSTATTLICHE VERSICHERUNG)

Erklärung

Hiermit erkläre ich, dass die vorliegende Arbeit mit dem Titel

„ Insights into the RNA Polymerase II CTD code "

von mir selbstständig und ohne unerlaubte Hilfsmittel angefertigt worden ist, und ich mich dabei nur der ausdrücklich bezeichneten Quellen und Hilfsmittel bedient habe.

Die Dissertation ist weder in der jetzigen noch in einer abgewandelten Form einer anderen Prüfungskommission vorgelegt worden.

Ich erkläre weiter, dass ich mich anderweitig einer anderen Doktorprüfung ohne Erfolg nicht unterzogen habe.

München, 27. Juni 2013

_____
Roland Schüller

Synopsis

Rpb1, the largest subunit of eukaryotic RNA Polymerase II (Pol II), contains a highly flexible structure at its C-terminus. This carboxyl-terminal domain (CTD) of Rpb1 is unique to eukaryotic organisms and consists of multiple tandemly repeated heptapeptides with the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$. Interestingly, the number of repeats differs from organism to organism and seems to correspond to genomic complexity, from 26 repeats in the yeast *Saccharomyces cerevisiae* to 52 repeats in the mammalian CTD (Chapman et al., 2008; Liu et al., 2010). Remarkably, five out of seven residues within the consensus sequence of the CTD can be potentially phosphorylated. In line with this, the production of monoclonal antibodies in our laboratory, against all different phosphosites within the heptad repeat confirmed the phosphorylation of Y_1, S_2, T_4, S_5 and S_7 *in vivo*. Additionally to phosphorylation, other posttranslational modifications, such as cis-trans isomerisation of the two proline residues can also take place (Egloff et al., 2008).

The potential of the CTD to be modified at each residue can create a wide range of distinct combinations which could carry information that is essential at different steps of the transcription cycle, where the modifications can be recognized as a readable code, the so-called 'CTD code'. In this respect, the CTD might serve as a dynamic platform constantly signalling between the transcription machinery and factors that interact with Pol II (Buratowski et al., 2003; Corden et al., 2007).

In this work, in order to gain new insights into the CTD code, CTD mutants were established to make the whole sequence accessible to mass spec (MS) analysis and to map phosphosites within the CTD *in vivo*. MS results showed that the CTD can be phosphorylated within all 52 repeats revealing the existence of the full repertoire of possible phosphosites within the CTD *in vivo*. Moreover, individual CTD peptides displayed many different phosphorylation patterns reflecting the great diversity of phosphorylation signatures existing in parallel within the same CTD. Data produced in this thesis showed that mono-phosphorylated CTD repeats represent the prevailing phosphorylation form *in vivo*. Additionally, dominant phosphorylation signatures in di-phosphorylated (2P) CTD repeats could be mapped along the CTD by MS analysis. Tri- and tetra-phosphorylated (3P and 4P) CTD peptides were

detected as well, but only in very low amounts. By analysing 2P-signatures in more detail it was demonstrated that different 2P-combinations predominated within distinct repeats along the CTD, suggesting that CTD phosphorylation is location dependent. Finally, known CTD-protein binding motifs could be mapped and linked to specific CTD repeats.

In conclusion, this work has established an approach for identifying high numbers of CTD-phosphosites, as well as high abundant CTD signatures along the whole CTD molecule, that contribute towards a better understanding of the 'CTD code' and open ways to yet undiscovered specific CTD-binding protein interactions.

1. Introduction

Three nuclear structurally related DNA-dependent RNA polymerases are responsible for transcribing DNA into RNA in eukaryotes (Cramer et al., 2008). RNA Polymerase I (Pol I) synthesizes most of the ribosomal RNAs (rRNAs), while RNA Polymerase III (Pol III) produces tRNAs, 5S rRNA, and other small RNAs, comprising 75% and 15% of transcripts in the cell, respectively (Grummt et al., 2003; Russell et al., 2005; Dieci et al., 2007; Werner et al., 2009).

The best studied polymerase is RNA Polymerase II (Pol II), which transcribes not only all protein-coding genes, but also a variety of small non-coding RNAs, including small nuclear and nucleolar (sn/sno) RNAs, cryptic unstable transcripts (CUTs), stable unannoted transcripts (SUTs) and XRN1-dependent unstable transcripts (XUTs) (Neil et al., 2009; Tisseur et al., 2011).

Five out of the twelve Pol II subunits are common in all three polymerases and the specific functions attributed to each polymerase are probably based on the combined action of the remaining non-identical subunits and other co-/factors (Young et al., 1991; Woychik et al., 1994; Shpakovski et al., 1995). In this respect, only the largest subunit of Pol II, Rpb1, contains a unique long and flexible carboxy-terminal domain (CTD). The CTD can be divided into three parts: (1) a flexible linker region, (2) a region consisting of tandem repeats of the consensus sequence tyrosine-serine-proline-threonine-serine-proline-serine ($Y_1S_2P_3T_4S_5P_6S_7$), and (3) a divergent C-terminal part. This unique structure is conserved from fungi to humans, although there is a variation in the number of repeats (15 repeats in amoeba; 26 repeats in budding yeast; 29 repeats in fission yeast; 52 repeats in human) as well as their deviation from the consensus sequence, reflecting to a large degree the complexity of the organism (Corden et al., 1985 and 1990; Chapman et al. 2008; Liu et al. 2010) (Figure 1). The ability of this repetitive sequence to interact with a wide range of nuclear factors is related to the dynamic plasticity of its structure and the diversity of binding surfaces generated by the multitude of posttranslational modifications it can accommodate. The association of specific posttranslational modifications of the CTD with particular events of the transcription cycle gave rise to the concept of the CTD code (Buratowksi et al., 2003). Tyrosine, threonine, and serine can all be

phosphorylated, threonine and serine can be glycosylated, and proline can undergo isomerisation (Egloff et al., 2008; Zeidan et al. 2010; Fuchs et al. 2009). In mammals, non-consensus repeats of the CTD predominantly involve substitution of S_7 of the heptad-motif by arginine or lysine, which are, therefore, excluding a phosphorylation on this position, while introducing the possibility for a broad array of additional modifications; lysine can be mono-, di-, or trimethylated, acetylated, ubiquitinylated and sumoylated, while arginine can be mono- or dimethylated (dimethylation can be asymmetric or symmetric). Sims and co-workers detected asymmetric dimethylation of arginine1810 within the CTD of Pol II in mammalian cells (Sims et al., 2011), while six out of seven lysines of the Mus musculus CTD were shown to be ubiquitinylated (Li et al., 2007). The combinatorial nature of these modifications, which is reminiscent of the histone code, is the key element of the CTD code that is generated by CTD-modifying enzymes and 'read' by CTD-associated protein factors (Buratowksi et al., 2003; Jenuwein et al., 2001).

Figure 1 Comparison of CTD sequences in budding yeast, fission yeast and human. All-consensus YSPTSPS heptad repeats are shown in red, human residues that deviate from the consensus sequence are in blue, and the numbers next to the parentheses indicate the repeat number *(Hsin and Manley, (2012), The RNA polymerase II CTD coordinates transcription and RNA processing; Genes & Development 26:2119-2137).*

The timely and precise combination of these factors leads to spatially and temporally controlled events that mediate transcription and RNA processing. A detailed

introduction to various aspects of the CTD and its interactions, CTD-modification patterns, genetic studies, as well as its role during transcription follows below.

1.1   The role of Pol II CTD in transcription and RNA processing

1.1.1 The role of CTD in transcription initiation

Initiation of transcription starts with the recruitment of gene-specific transcription factors (TFs), general transcription factors (GTFs), the Mediator complex, and Pol II. These factors form the pre-initiation complex (PIC) at the promoters of Pol II transcribed genes (Buratowski et al., 2009; Nechaev et al., 2010). The Mediator complex plays a predominant role by linking the PIC to assemblies of transcription factors bound at the upstream regulatory (activating/repressing) sequences (UAS/URS) (Svejstrup et al., 1997; Myers et al., 2000; Kornberg et al., 2005; Malik et al., 2005). The Mediator complex binds the unphosphorylated form of Pol II CTD, but when incorporated into the PIC, it strongly stimulates the CTD kinase (Kin28 in yeast, CDK7 in metazoans) of the basal transcription factor TFIIH. It has been shown, that Kin28/CDK7 and the Mediator complex subunit Srb10/CDK8 phosphorylate S_5 in vivo with Kin28/CDK7 being the predominant kinase (Phatnani et al., 2006; Hengartner et al., 1998; Dahmus et al., 1996; Bensaude et al., 1999; Palancade et al., 2003; Feaver et al., 1994; Rickert et al., 1999; Gebara et al., 1997). In turn, the serine-5 phosphorylation (S_5P) is a prerequisite for coordinating the placement of several key posttranslational modifications on chromatin like trimethylation of histone H3 at lysine 4 (H3K4me3) by Set1 and subsequent trimethylation of histone H3 at lysine 79 (H3K79me3) by Dot1 (Venters et al., 2009; Ng et al., 2003; Nakanishi et al., 2008; Wood et al., 2003). Set1 establishes two distinct chromatin zones on genes, H3K4me3 found at promoter regions and H3K4me2 located further downstream in the body of the gene (Kim and Buratowski 2009). Eventually, H3K4me2 and CTD S_5P trigger the recruitment of the histone deacetylase complexes Set3 and Rpd3C(S), leading to reduced histone acetylation levels at the 5' ends of genes, which promotes the association of Pol II and inhibits

CUT initiation at promoters (Kim and Buratowski 2009; Govind et al., 2010; Drouin et al., 2010).

The key role of S_5P, however, is the binding of the capping enzyme complex and therefore promoting transition of transcription to elongation. The CTD repeats proximal to the core Pol II are ideally located near the RNA exit tunnel to facilitate the capping reaction (Cramer et al., 2001; Ghosh et al., 2011). Although capping enzyme recognition by CTD is structurally different in mammalian and yeast (see also part 1.2.3), both organisms require S_5P for binding (Fabrega et al., 2003; Ghosh et al., 2011). In this context, specific inhibition of Kin28 has little effect on transcription of protein-coding genes, but causes a striking reduction of capping (Liu et al., 2004; Kanin et al., 2007). Interestingly, Kin28/CDK7 is also the primary kinase for CTD serine-7 phosphorylation (S_7P) (Akhtar et al., 2009; Glover-Cutter et al., 2009; Kim et al., 2009) (Figure 2). S_7P seems to be Mediator complex-dependent, but the role of this phosphorylation at promoters remains elusive and will be discussed further below (Boeing et al., 2010).

1.1.2   The role of CTD in Pol II pausing and transcription elongation

Following promoter release, transcription initiation factors are exchanged by elongation factors, playing an important role in RNA processing by moving through chromatin, and suppressing cryptic transcripts. In mammalian cells, the positive elongation factor P-TEFb kinase subunit, CDK9, phosphorylates both CTD S_2 and the DRB-sensitivity-inducing factor (DSIF), thus allowing Pol II to overcome the promoter-proximal pausing induced by the negative elongation factor (NELF) complex (Sims et al., 2004; Peterlin et al., 2006). Interestingly, about one-third of genes in both fly and human cells appear to contain a paused Pol II downstream of the transcription start site (Core et al., 2008; Nechaev et al., 2010). Pausing seems to allow rapid and coordinated transcription during development, or in response to external stimuli (Muse et al., 2007). It is unclear if promoter-proximal pausing exists in yeast, but it is known that Bur 1, the yeast homolog of CDK9, promotes elongation through phosphorylation of Spt5, the yeast homologue of DSIF (Zhou et al., 2009). Additionally, Bur1 also phosphorylates CTD S_2 downstream of the promoter and triggers the ubiquitylation of histone H2B lysine 123 (H2BK123ub) by the ubiquitin

conjugating enzyme Rad6 and Bre1 (Wood, 2003 and 2005). H2BK123ub promotes subsequent Set1 trimethylation of histone H3K4 and subsequent trimethylation of H3K79, both of which represent important marks of transcription activation (Venters et al., 2009; Ng et al., 2003; Nakanishi et al., 2008; Wood et al., 2003).

The S_5P mark is a prerequisite for the recruitment of Bur1 to the transcription complex at the promoter. It then phosphorylates S_2, priming the CTD for the recruitment of Ctk1, the major S_2 kinase which then phosphorylates S_2 further downstream in the coding region (Keogh et al., 2003; Qiu et al., 2009). Recent studies in *Drosophila* and human cells have shown that CDK12 can phosphorylate the CTD on S_2 and proposed, based on phylogenetic relationships, that the ortholog of Bur1 is CDK9, whereas CDK12 is the counterpart of Ctk1 (Guo and Stiller 2004; Bartkowiak et al., 2010) (Figure 3a). Indeed, CDK12 is required for most S_2 phosphorylation *in vivo* and is associated with elongating Pol II. Interestingly, Bur1 has been identified as an 'internal' S_7 kinase 'travelling' with Pol II and phosphorylating S_7 in later phases of the transcription cycle. Although the exact role of this modification is unclear, it is likely to be a mark that promotes elongation as genes with uniformly high levels of S_7P are transcribed at significantly higher levels (Tietjen et al., 2010).

The newly characterized CTD phosphatase Rtr1 associates with Pol II and removes S_5P marks immediately after promoter clearance (Figure 2; Figure 3a). The S_2P phosphatase Fcp1 is also recruited during elongation, but S_2P levels remain high across the transcript due to the opposing action of the S_2P kinase Ctk1 (Mosley et al., 2009; Kobor et al., 1999; Cho et al., 2001). Increasing levels of S_2P, in combination with the residual S_5P, lead to the recruitment of the Set2 methyltransferase, which di- and trimethylates H3K36, followed by the removal of acetylation from histones H3 and H4 by the histone deacetylase complex Rpd3C(S) and thus, preventing cryptic transcription initiation within open reading frames in yeast (Kizer et al., 2005; Vojnic et al., 2006; Krogan et al., 2003; Li et al., 2002; Govind et al., 2010; Carrozza et al., 2005; Keogh et al., 2005). Similarly, the splicing factor Prp40 and U2AF65 recognize the S_2P-S_5P double mark followed by the recruitment of Prp19 to activate splicing (Egloff et al., 2008; Phatnani et al., 2004; David et al., 2011). Splicing in turn triggers the binding of the yeast export factor, Yra1, to S_2P-S_5P CTD repeats (MacKellar et al., 2011). S_2P is also bound by the serine/arginine rich protein Npl3, which functions in elongation, 3'-end processing,

hnRNP formation, and mRNA export (Gilbert et al., 2001; Bucheli and Buratowski 2005; Bucheli et al., 2007). Additionally, the RNA binding factor, Ssd1, the mitotic kinase, Hrr25, and RecQ5 genome stability helicase also bind to S_2P-S_5P CTD repeats (Kanagaraj et al., 2010; Phatnani et al., 2004). However, the precise role that Ssd1 and Hrr25 play during transcription cycle remains unclear.

1.1.3 The role of CTD in 3' RNA processing and transcription termination

The CTD also plays an important role in 3' end processing of Pol II-produced transcripts and it has been shown that several 3' processing factors like CPSF, CstF and Pcf11 interact with the CTD (Shi et al., 2009; McCracken et al., 1997). Pcf11, for instance, contains an N-terminal CTD interaction domain (CID) and binds the CTD in a S_2P-dependent manner (Barilla et al., 2001; Licatalosi et al., 2002; Meinhart and Cramer 2004) (see also part 1.2.3). Indeed, S_2P is critical to the 3' end processing, as genome-wide chromatin immunoprecipitation (ChIP) experiments showed that peaks of 3' processing factors coincide with S_2P peaks, suggesting that both poly(A) sites and S_2P are a prerequisite for subsequent recruitment/assembly of the polyadenylation complex in the newly processed RNA (Kim et al., 2010; Mayer et al., 2010, 2012).

Importantly, the 3' ends of several types of Pol II transcribed RNAs, comprising snRNA and histone mRNA, are not polyadenylated, but here tight regulation of CTD phosphorylation/dephosphorylation is important to the cleavage complex recruitment. Proper snRNA 3' end formation needs the promoter and 3' box, located just downstream from the snRNA-encoding region. A multi-subunit RNA 3' end processing complex, the Integrator, is associated to the 3' box of snRNA genes. Interestingly, the RPAP2 phosphatase is recruited to snRNA genes via S_7P, close to the promoter region where S_7P is most frequent. The combined action of dephosphorylation of S_5 by RPAP2 and phosphorylation of S_2 by P-TEFb, as transcription progresses, creates a double mark consisting of S_7P on one repeat and S_2P on the following repeat which is then specifically bound by Int11, the catalytic cleavage subunit of the Integrator (Baillat et al., 2005; Egloff et al., 2010). These results support the idea that S_5P dephosphorylation is a prerequisite for 3' end formation in both, snRNA and mRNA genes (Xiang et al., 2010). In this context,

RPAP2 is also recruited to protein-coding genes but in a S_7P independent way. Two proteins, RPRD1A and RPRD1B, which interact both with RPAP2 and the CTD, could help recruit the S_5 phosphatase to mRNA genes in the absence of S_7P (Ni et al., 2011).
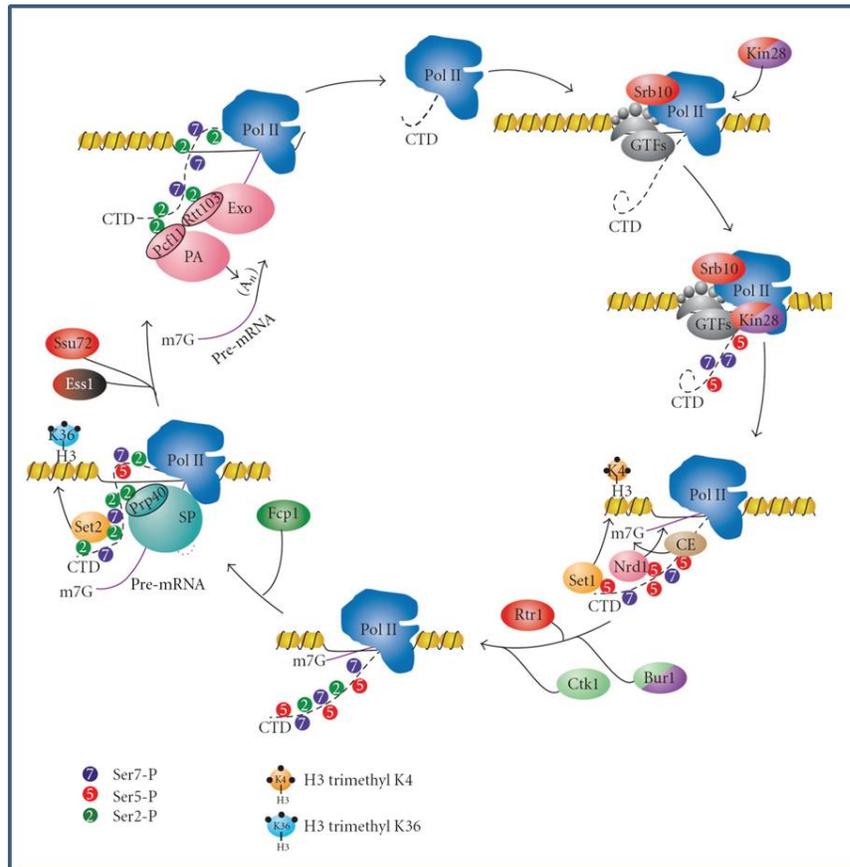
Figure 2 Transcription cycle of RNA polymerase II in yeast. Recruitment of primary RNA processing factors concomitant with dynamic modifications within both, the Pol II CTD and chromatin, along the transcription cycle are shown (*Zhang and Ansari, (2011) Emerging Views on the CTD Code; Genetics Research International; Volume 2012, Article ID 347214*).

A number of 3' cleavage factors have been shown to be critical for Pol II termination, while a functional polyadenylation signal is required for subsequent termination (Proudfoot et al., 1989; Birse et al., 1998; Dichtl et al., 2002; Ganem et al., 2003; Nedea et al., 2003; Kim et al., 2010; Zhang et al., 2012). In this line, one CTD-binding cleavage factor, Pcf11, seems to play a key role; mutated yeast Pcf11 that retains 3' cleavage activity, but is defective in CTD binding, was found to be defective in terminating Pol II mediated transcription (Sadowski et al., 2003). Two models have been suggested for Pol II termination, but most probably a combination

of the two models could best explain the mechanism. In the first model, termed the 'allosteric' model, after transcribing the poly-(A) site, Pol II undergoes conformational changes that lead to an exchange of elongation factors for termination factors. In support to this proposition, Pcf11 has been shown to directly bind the CTD of Pol II causing breakdown of the whole transcription complex (Zhang and Gilmour, 2006; Mayer et al., 2010; Kim et al., 2004).

In the second model, called the 'torpedo' model, it is proposed that cleavage of the transcript at the cleavage and polyadenylation site (CPS) creates an entry site for the 5'-3' exonuclease Rat1 (Xrn2 in mammals), which degrades the 3' RNA and triggers Pol II release by 'chasing' the complex (Kim et al., 2004; Connelly and Manley 1988; West et al., 2004). Importantly, recruitment of Rat1 seems to be indirect, probably via its partner Rtt103 which has been shown to bind S_2P CTD in a cooperative manner with Pcf11 (Lunde et al., 2010), an essential component of the cleavage factor IA (CFIA) complex that also promotes Pol II release (Zhang et al., 2005). ChIP experiments revealed that Pcf11 is located at both protein-coding and non-coding genes and Pcf11 mutations lead to transcript read-through due to inefficient cleavage, indicating that it probably plays a key role in termination, as well as processing of coding and non-coding genes (Figure 3b) (Meinhart and Cramer 2004; Licatalosi et al., 2002; Zhang et al., 2005; Sadowski et al. 2003; Kim et al., 2006; Kim et al., 2010).

In yeast, Pol II transcript processing is exerted through two different gene class-specific pathways (Figure 3b). The majority of small mRNAs (<550 bp), CUTs, snRNA, and snoRNAs are processed through the Nrd1-Nab3 pathway, while longer mRNAs are processed in a polyadenylation-dependent process (Lykke-Andersen et al., 2007; Arigo et al., 2006; Thiebaut et al., 2006; Egloff et al., 2008; Richard and Manley 2009; Buratowski et al., 2005; Kim et al., 2006; Birse et al., 1998; Gudipati et al., 2008; Steinmetz et al., 2001). The pathway selection is dependent on the CTD phosphorylation state, with Nrd1 preferentially binding to S_5P and its recruitment is additionally enhanced through histone H3me3K4 (Vasiljeva et al., 2008; Terzi et al., 2011). The helicase Sen1 (senataxin in humans), which exists in a complex with Nrd1 and Nab3, associates with the exosome complex, linking transcription termination to 3' exonuclease activity that can 'trim' snoRNA ends, or completely degrade cryptic transcripts. Remaining in yeast, interestingly, while transcription proceeds, phosphorylation of S_2P CTD blocks the use of the Sen1/Nrd1/Nab3

termination pathway, providing a mechanism by which S_2 phosphorylation could enhance downstream elongation (Gudipati et al., 2008). Senataxin, the Sen1 homolog in higher eukaryatoes, has not been implicated to date in termination and it therefore still remains to be proven whether a similar early termination pathway exists in higher animals too. No functional homologs of Nrd1 and Nab3 have been identified yet, and there is no evidence for the existence of a Nrd1-like complex in human cells. SCAF8, which shares sequence similarity with Nrd1, specifically binds to CTD via its CID domain, however, its function is unknown and does not appear to involve termination (see also part 1.2.3) (Yuryev et al., 1996; Patturajan et al., 1998; Becker et al., 2008).

The second pathway, responsible for the processing of most mRNA transcripts, includes the cleavage and polyadenylation factor (CPF) complex, cleavage factor IA and IB (CFIA and CFIB) complexes, and the exosome (Richard and Manley 2009; Kim et al., 2006; Birse et al., 1998) (Figure 3b). Importantly, the majority of termination and 3' processing factors involved in this pathway tend to preferentially bind to S_2P or S_2P/S_5P enriched CTD including: Npl3, Rtt103, Rna14, Rna15, Ydh1, Yhh1, Pta1, and Pcf11. Binding of Rna15 to nascent RNA triggers endonucleolytic cleavage followed by polyadenylation by the polyadenylate polymerase (Pap1). Subsequently, polyadenylation–binding proteins (PAB) protect the mature transcript from exonucleolytic degradation (Birse et al., 1998; Minvielle-Sebastia et al., 1994).

In both pathways, the CTD is hypophosphorylated by the combined action of two essential phosphatases at the end of transcription, Ssu72 and Fcp1. Ssu72 which is primarily localized at the 3' end of genes, is the main S_5P phosphatase and its activity is enhanced by the prolyl isomerase Ess1 (Pin1 in humans) and by interacting with Pta1 (Figure 2, Figure 3a) (Nedea et al., 2003; Krishnamurthy 2004 and 2009; Ghazy et al., 2009; Singh et al., 2009). In contrast to Ssu72, Fcp1 is found across the entire transcribed region and mainly dephosphorylates S_2P CTD (Kobor et al., 1999; Cho et al., 2001; Archambault et al., 1997; Kong et al., 2005; Hausmann et al., 2004; Ghosh et al., 2008). Recent data showed that Ssu72 may be the phosphatase that removes S_7P at both 5' and 3' ends of genes (Zhang et al., 2012). Global dephosphorylation of the CTD promotes the release of Pol II from DNA, which can then bind to promoters for a new round of transcription (Steinmetz and Brow 2003; Cho et al., 1999; Dichtl et al., 2002). Interestingly, it has been suggested that transcription termination and subsequent dephosphorylation of the CTD is coupled to

transcription re-initiation via gene looping, during which the promoter and terminator regions come in close proximity, allowing Pol II to form a new PIC more rapidly (O'Sullivan et al., 2004; Singh et al., 2009). In line with this, Ssu72 and TFIIB have been shown to be essential in gene looping (Ansari and Hampsey 2005; Singh et al., 2007).
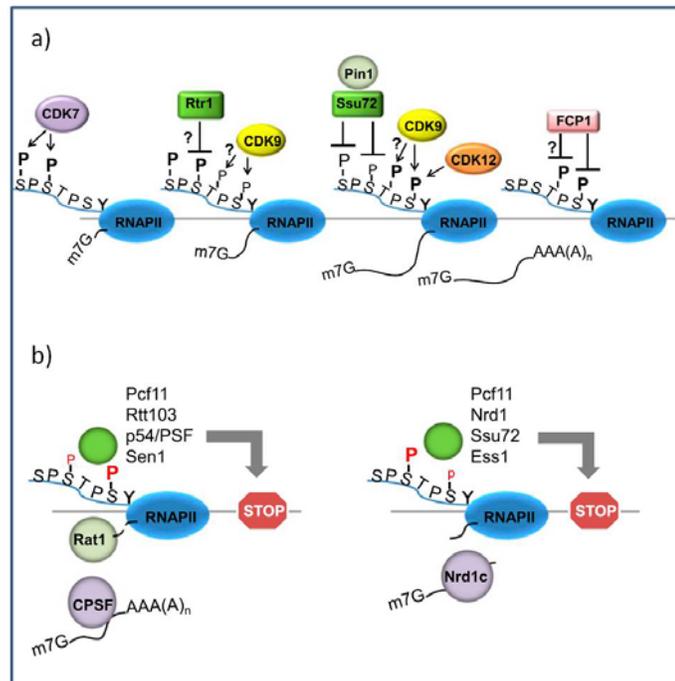
Figure 3
a) Dynamic modifications of the CTD during transcription cycle in mammals. Dynamic phosphorylation pattern of CTD due to the recruitment of CTD kinases, CTD phosphatases, prolyl isomerase Pin1 at different stages (initiation, elongation and termination) within the transcription cycle is shown.
b) Different pathways for transcription termination for protein-coding and noncoding genes. Left: poly(A)-dependent termination: RNA is cleaved by 3' end processing factors at the polyadenylation site. The CTD with S_2P is involved in recruiting factors like Pcf11, Rtt103, p54/PSF, and Sen1, to facilitate termination of long polyadenylated transcripts. Right: Nrd1-dependent termination: The Nrd1 complex (Nrd1-Nab3-Sen1) interacts via Nrd1 with S_5P CTD which is present at the 3' ends of short genes, such as snoRNAs and CUTs *(Hsin and Manley, (2012), The RNA polymerase II CTD coordinates transcription and RNA processing; Genes & Development 26:2119-2137).*

In summary, the phosphorylation and dephosphorylation of the CTD is a complex and highly controlled mechanism, which is clearly involved in every stage of transcription, from initiation, to elongation, to termination, and possibly re-initiation. Phosphorylated residues may be individually conceived as marks, but in a broader sense, waves of phosphorylation and simultaneous waves of dephosphorylation of

specific residues create combinatorial platforms that highly coordinate every step of the transcription cycle.

1.2 CTD- a closer look

The largest subunit of eukaryotic RNA Polymerase II, Rpb1, consists of a unique structure at its C-terminal domain, the CTD, with tandem repeats of the heptapeptide sequence Y_1-S_2-P_3-T_4-S_5-P_6-S_7 (Allison et al., 1985; Corden et al., 1985). The length of the domain is a direct link to the genetic complexity of the organism where it is encountered. The CTD is dispensable for Pol II activity *in vitro*, (West and Corden 1995; Bartolomei et al., 1988), however, deletion of the entire CTD in mice, drosophila and yeast is lethal (Egloff et al., 2008). Moreover, it is well proven that the CTD plays a direct and major role in coupling transcription with co-transcriptional nuclear processes, such as chromatin modification and RNA processing (see also part 1.1) (Egloff et al., 2008). The CTD has been also implicated in a variety of transcription-extrinsic processes like mRNA export and stress response. The process of mRNP export is controlled by the protein Sus1. This key player in mRNA export directly interacts with S_5P and S_2P/S_5P CTD, Ub8 subunit of the SAGA complex, Yra1 subunit of the TREX1 complex, and Sac3 subunit of the TREX2 complex at the nuclear pore (Stewart et al., 2010; Pascual-Garcia et al., 2008; Jani et al., 2009). Interestingly, in response to DNA damage, the ubiquitin ligase Rsp5 binds the CTD and ubiquitylates Pol II (Huibregtse et al., 1997; Beaudenon et al., 1999). Similarly, UV-induced DNA damage in mammalian fibroblasts leads to hyperphosphorylation of the CTD by P-TEFb, which then promotes Pol II ubiquitylation and subsequent degradation (Heine et al., 2008). Additionally, S_5P CTD can also recruit the Asr1 ubiquitin ligase, promoting ejection of the Rpb4/7 heterodimer from the core polymerase that may provide a mechanism for stopping polymerases engaged in abortive or cryptic transcription (Daulny et al., 2008).

Recent studies have identified new posttranslational modifications of the CTD repeats, new CTD-binding factors and there are new insights into the relation between the different sites within the CTD based on genetic studies and ChIP data mainly performed in yeast. Some of these new exciting data linked to Pol II CTD will be discussed below.

1.2.1   Posttranslational modifications within the CTD of Pol II

Serine-5 and Serine-2 phosphorylation: As discussed above, dynamic phosphorylations of the three serine residues are the best-characterized CTD modifications and in particular the exchange between S_2P and S_5P patterns play a pivotal role in the mediation of transcription and RNA processing. In ChIP experiments it has been shown that S_5P marks are highly abundant at TSSs (transcription start sites) and strongly decline in the body of active genes (Kim et al., 2010; Mayer et al., 2010; Tietjen et al., 2010; Bataille et al., 2012; Koch et al., 2011; Brookes et al., 2012) (Figure 4). The two main functions of this mark described to date are the recruitment of the capping machinery during transcription initiation and the interaction with Nrd1 that plays an important role in the 3' end formation and early termination of non-polyadenylated transcripts (Cho et al., 1997; McCracken et al., 1997; Gudipati et al., 2008; Vasiljeva et al., 2008). In contrast to S_5P, S_2P is absent at TSS, progressively increases within the body of active genes while its peak is found in proximity to the poly(A) site (Kim et al., 2010; Mayer et al., 2010; Tietjen et al., 2010; Bataille et al., 2012; Koch et al., 2010; Brookes et al., 2012) (Figure 4). Interestingly, in *S.cerevisiae*, S_2 phosphorylation is controlled by two different kinases, Bur1, which directly binds to S_5P CTD, and Ctk1. Most S_2P sites on elongating Pol II seem to be catalyzed by Ctk1 (Buratowski et al., 2009; Qiu et al., 2009). Similarly, *S. pombe* has two S_2 kinases equivalent to Ctk1 and Bur1, CDK9 and Lsk1 (Viladevall et al., 2009). In metazoans, CDK9, the kinase subunit of P-TEFb, phosphorylates both elongation factor Spt5 and CTD S_2 (Bres et al., 2008). Due to the dual functionality of CDK9, it was thought that P-TEFb combines the activities of both yeast Bur1/CDK9 and Ctk1/Lsk1 homologues, respectively. However, recent studies in Drosophila and human cells have discovered two additional S_2 CTD kinases, CDK12 and CDK13 (Blazek et al., 2011; Bartkowiak et al., 2010; Bartkowiak and Greenleaf 2011). ChIP data showed that CDK12 contributes the majority of S_2P sites on elongating Pol II and that its abundance at the 5' end of genes is rather low (Bartkowiak et al., 2010; Bartkowiak and Greenleaf 2011). In addition, CDK12 can promote the expression of a subset of human genes, including the DNA damage response genes. The function of CDK13 during the transcription cycle remains elusive (Kohoutek et al., 2012). A recent study has

revealed the existence of yet another S_2 kinase, bromodomain protein Brd4, which is an atypical CTD S_2 kinase that can phosphorylate S_2 *in vitro* and *in vivo* (Devaiah et al., 2012). Brd4 is able to recruit P-TEFb but can also activate transcription of a subset of genes independent of P-TEFb (Devaiah et al., 2012; Rahman et al., 2011). The identification of these new metazoan S_2 kinases emphasizes that the maintenance of distinct homologues in yeast species between Bur1- and Ctk1-type kinases has also been preserved in higher organisms. Apart from their role in gene transcription, CTD S_2 and S_5 phosphorylation has been implicated in other processes. For instance, S_2P has been shown to be essential for additional cellular pathways, like sexual differentiation in *S.pombe* (Coudreuse et al., 2010), whereas the mitotic phosphatase Cdc14 has been shown to remove S_2P and S_5P and thereby repressing transcription during mitosis (Clemente-Blanco et al., 2011) (see also part 1.2.2.2 and Table 2).

Serine-7 phosphorylation: Next to S_2P and S_5P, other CTD posttranslational modifications, such as S_7P, have been discovered more recently and fulfil gene class-specific tasks. The requirement for S_7 phosphorylation was the first example of a specifically modified form of Pol II involved in expressing a particular type of gene and therefore strengthening the idea of a gene-specific CTD code (Chapman et al., 2007; Egloff et al., 2007). S_7P is required for expression of snRNA genes in mammalian cells and mutations of this residue leads to a marked defect in transcription of human snRNA genes and 3' processing of the transcripts (Egloff et al., 2007). S_7P specifically recruits the RPAP2 S_5P phosphatase and the RNA 3' processing integrator complex to snRNA genes, ensuring proper transcription and processing of transcripts (Egloff et al., 2012). Surprisingly, CDK7/Kin28, the kinase responsible for S_5 phosphorylation turned out to be critical for S_7 phosphorylation in yeast and humans (Boeing et al., 2010; Akhtar et al., 2009; Glover-Cutter et al., 2009; Kim et al., 2009). Consequently, S_7P ChIP-profiles at the beginning of snRNA and protein-coding genes generally resemble those of S_5P, and knockdown of CDK7 dramatically decreases both S_5 and S_7 phosphorylation (Akhtar et al., 2009). However, a very important finding is that, in contrast to S_5P, the S_7P levels remain high toward the 3' end of coding and non-coding genes, suggesting that CDK7 is not the only S_7- specific kinase. Indeed, it has been shown that the inactivation of Bur1 kinase reduces the levels of S_7P within coding regions and CDK9, the Bur1 homolog

in humans, is capable of phosphorylating S_7 *in vitro* (Tietjen et al., 2010; Glover-Cutter et al., 2009). Interestingly, Ssu72 has recently been shown to also remove phosphates from S_7P (Zhang et al., 2012) underlying the close connection between the S_5P and S_7P marks that share a common kinase, as well as a phosphatase.

Modulation of residue-specific serine phosphorylation pattern during transcription

Distinct patterns of CTD phosphorylation are detected between non-coding and protein-coding genes. S_5P marks the initiation site and is also detected on paused genes. High levels of S_7P along transcribed regions suggest an important function of this mark in transcription elongation (Figure 4). Indeed, high levels of S_7P are present on highly transcribed genes (Kim et al., 2010; Tietjen et al., 2010). S_2P levels are lower on non-coding genes whereas S_7P is equivalent to, or higher, on non-coding genes than on protein-coding genes (Kim et al., 2010; Tietjen et al., 2010). The reason for this could be the short length of non-coding genes, since S_2P generally occurs later in the transcription cycle. Accordingly, S_2P is important for elongation and in activating splicing and 3' end processing in protein-coding genes. Because snRNA genes are intronless, the requirement for a high level of S_2P might be bypassed. S_7P specifically recruits the RPAP2 S_5P phosphatase and the RNA 3' end processing integrator complex to snRNA genes for accurate transcription and processing of transcripts (Egloff et al., 2012).

In yeast, short genes also exhibit a lower level of S_2P for non-coding snoRNAs, compared to protein-coding genes. In respect of this, termination factors are recruited to protein-coding genes at S_2P sites, whereas in snoRNAs the termination factor Nrd1 specifically binds to S_5P CTD. Additionally, S_7P levels can be found at high levels on non-coding genes (Kim et al., 2010; Tietjen et al., 2010; Vasiljeva et al., 2008). In yeast, it is not known, whether this could also be a positive signal for the recruitment of gene-specific factors to non-coding genes. In summary, a low level of S_2P and an abundance of S_7P at non-coding genes could stand for a CTD gene-type specific signal.

Threonine-4 phosphorylation: T_4P was the fourth identified posttranslational modification in the CTD heptad repeat (Hsin et al., 2011; Hintermair et al., 2012). In ChIP analysis, T_4P signals are very weak or absent at the TSS, remain low in the

gene body, but strongly rise downstream of the poly(A) site (Hintermair et al., 2012) (Figure 4). Interestingly, in mammalian cells, T_4P peaks at about 300 bp downstream of S_2P, suggesting that the increase of S_2P might be a prerequisite for the subsequent phosphorylation of T_4 (Hintermair et al., 2012). Likewise, T_4P is tightly associated with S_2P in co-IP experiments and no T_4P can be detected in a serine-2/alanine (S_2/A) mutant. Moreover, the S_2/A mutant promotes a global defect in RNA elongation, while few genes become activated, and show an enrichment of Pol II within the gene body (Hintermair et al., 2012). Additionally, the elongation defect was concomitant with a local accumulation of polymerases immediately downstream of the initiation site in mutant cells (Hintermair et al., 2012). In contrast, no enrichment of T_4P at 3' regions of genes can be detected in yeast, which correlates with the finding that T_4P would block the binding of the termination factor Pcf11 (Mayer et al., 2012; Meinhart et al., 2004). In a recent study in chicken cells, it has been shown that T_4P is crucial for processing, but not transcription, of the intron-less replication-activated histone genes, whereas expression of other protein-coding genes or non-coding RNA genes remains unaffected by T_4 mutation (Hsin et al., 2011). This observation reflects another example of a gene-specific role of the Pol II CTD. In human cells, Plk3 can phosphorylate T_4 CTD under physiological conditions, as well as under stress conditions, leading to the contribution of a new class of CTD-specific kinases (Hintermair et al., 2012). Accordingly, inhibition of CDK9, the known S_2P kinase, by DRB and flavopiridol also leads to diminished T_4P levels, suggesting that T_4 phosphorylation is also CDK9-dependent (Hsin et al., 2011). However, this could be an inhibitory effect caused by the lack of S_2P, which has been shown to be a prerequisite for priming T_4 phosphorylation. Finally, to date, no specific T_4 CTD kinase has been identified in yeast and no T_4 CTD phosphatase has been described in any species.

Tyrosine-1 phosphorylation: Recently, the CTD code has been expanded by another CTD posttranslational modification, tyrosine-1 phosphorylation (Y_1P), which plays an important role in the regulation of transcription termination (Baskaran et al., 1993; Mayer et al., 2012). Y_1 is phosphorylated in yeast and ChIP data showed that this modification can be found at all active genes. Importantly, Y_1P levels drop before reaching the poly(A) site, whereas S_2P levels still remain high (Figure 4). This led to the discovery that Y_1P has a key function in suppressing termination during

elongation by blocking the recruitment of Rtt103 and Pcf11 (Mayer et al., 2012). Y_1P blocks CTD binding to the conserved CTD-interacting domain (CID) of termination factors *in vitro*, whereas within the gene body, the CTD binding of elongation factor Spt6 through its CTD-binding domain is maintained and accompanied by high Y_1P levels.

This fundamental role of Y_1P in gene expression might also explain the lethal phenotype of the tyrosine-1/phenylalanine (Y_1/F) mutant in yeast and human. Nevertheless, an *in vivo* kinase screen including the CTD kinases Kin28, Srb10, Bur1 and Ctk1 has not led to the discovery of the responsible Y_1P kinase in yeast, suggesting that Y_1 phosphorylation of the yeast CTD depends on a kinase other than the known CTD kinases (Mayer et al., 2012). Concurrently, Y_1 phosphorylation in human is performed by c-Abl, a kinase that lacks a yeast homolog (Baskaran et al. 1999).
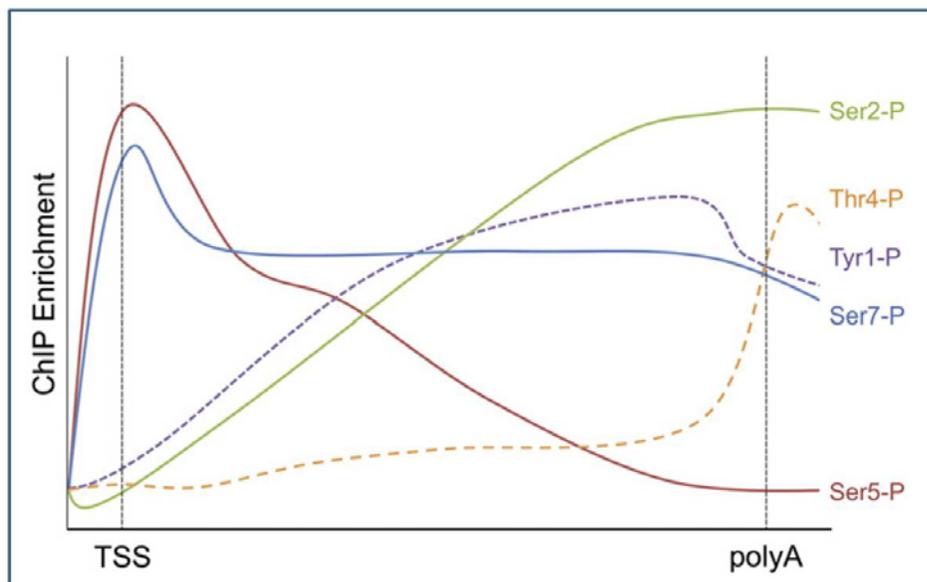
Figure 4 Average profile of CTD phosphorylation profiles along genes from ChIP experiments. Schematic graph of genome-wide distribution for all CTD phosphorylation marks *(Heidemann et al., (2012), Dynamic phosphorylation patterns of RNA Polymerase II CTD during transcription; Biochim Biophys Acta. 2013 Jan; 1829(1):55-62).*

Phosphospecific CTD antibodies: Hallmarks and considerations

Monoclonal antibodies against CTD Y_1P, S_2P, T_4P, S_5P and S_7P have also been established successfully in our lab and are a powerful tool for the study and identification of new phosphoresidues in ChIP experiments, as well as *in vivo* cell

studies. However, it is very important to be aware of limitations in the epitope recognition capacities of these antibodies, due to epitope masking. Furthermore, the epitope specificity of antibodies is also an issue, as it can change at high antigen, as well as high antibody concentration. Additionally, the signal strength of antibodies in western blotting or ChIP analysis reflects the number of accessible CTD-marks and not the overall number of existing modifications, which are physically present in the CTD under investigation. Accordingly, the absence of a signal can be explained either with its physical absence or with the masking of the targeted epitope by other modifications. Consequently, in order to get a more detailed insight into how other adjacent modifications can influence or inhibit epitope recognition, our CTD antibodies were tested in enzyme-linked immunosorbent assays (ELISA) using a panel of di-heptad CTD peptides with various combinations of modifications. With this analysis we obtained a comprehensive overview of inhibitory modifications that interfere with binding of specific antibodies to the CTD (Figure 5).

The results of this investigation demonstrate that all CTD-specific antibodies underlie specific restrictions in recognition of their respective epitope (Hintermair et al., 2012; Heidemann et al., 2012). For example, Y_1P in the same repeat together with S_2P influences the recognition by the S_2P specific antibody 3E10 (Figure 5). In another case, phosphorylation of S_2 or S_5 next to T_4P inhibits the epitope recognition of the T_4P specific antibody 6D7 (Hintermair et al., 2012; Heidemann et al., 2012). Consequently, the T_4P specific antibody cannot distinguish between *de-novo* phosphorylation of a T_4 residue and the unmasking of pre-existing T_4P marks, which leaves us for the interpretation of the strong increase of T_4P in the 3' region of genes either or both options (Figure 4) (Hintermair et al., 2012). Nevertheless, T_4P-linked structural changes occur in the CTD downstream of the poly(A) site. Additionally, the CTD-specific antibodies were used to purify and define different fractions of the hyperphosphorylated form of Pol II (IIO form) biochemically. IP-experiments showed that three different populations of Pol IIO regarding their CTD marks exist. A population associated preferentially with (i) S_5P marks, (ii) S_5P and S_2P marks, and (iii) S_2P and T_4P marks. The S_7P mark can be found in all three populations (Hintermair et al., 2012). These data indicate that T_4P is strictly associated with the S_2P mark and that S_5P/S_7P and S_2P marks are associated with different populations of Pol IIO in human cells. Importantly, recent ChIP studies in yeast revealed that the phosphoserine marks are placed and removed as a function of the distance from

transcription start site (TSS) and termination site, respectively, with no significant detectable difference between genes (Mayer et al., 2010; Bataille et al., 2012). In other words, the CTD cycle is very similar at all genes including the fact that short genes will have higher levels of S_5P and lower levels of S_2P than long genes when they reach the termination site. In line with this, short genes, such as snoRNAs and genes coding for small proteins tend to use an alternative mechanism for termination compared to most class II genes (Kim et al., 2006; Lykke-Andersen and Jensen, 2007) (Figure 3b).

αTyr1-P (3D12)	$T_4S_5P_6S_7Y_1S_2P_3T_4S_5$
αSer2-P (3E10)	$S_5P_6S_7Y_1S_2P_3T_4S_5$
αThr4-P (6D7)	$Y_1S_2P_3T_4S_5P_6S_7Y_1$
αSer5-P (3E8)	$Y_1S_2P_3T_4S_5P_6S_7Y_1$
αSer7-P (4E12)	$P_3T_4S_5P_6S_7Y_1S_2P_3$
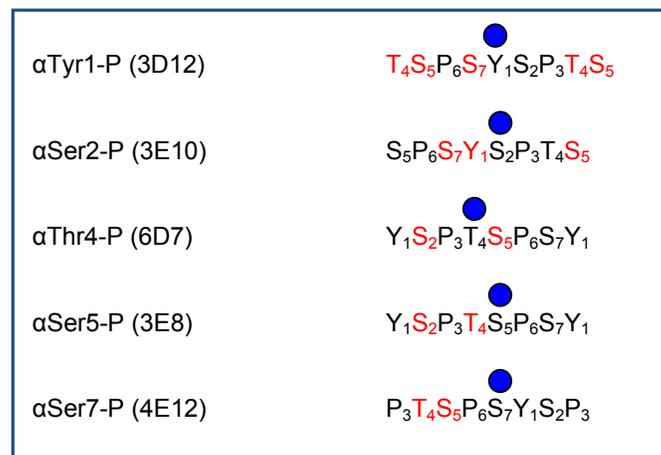
Figure 5 Overview of the characteristics in epitope recognition of monoclonal phospho-specific-CTD antibodies established in our lab. Blue circle indicates phospho-specific epitope of each antibody. Red amino acid residues reveal full or partial inhibition of antibody binding.

Arginine methylation: Interestingly, in addition to 21 consensus repeats, the mammalian CTD consists of 31 non-consensus repeats that are mainly found in the distal part of Pol II CTD. Recently, for the first time, a specific role in CTD function within non-consensus repeats has been discovered. In human cells, the CTD of Pol II is methylated at arginine1810 (R1810) of CTD repeat 31 by the methyltransferase CARM1 *in vitro* and *in vivo* (Sims et al., 2011). This specific modification is linked to the regulation of snRNA and snoRNA expression since substitution of R1810 to alanine, as well as genetic knockout of CARM1, lead to a specific up-regulation of these classes of RNA species (Sims et al., 2011). Contrary to the inhibitory effect on snRNA gene expression of the S_7 mutation, expression of snRNA and snoRNA were up-regulated when R1810 was mutated to alanine, indicating a repressive rather

than activating function of this mark (Egloff et al., 2007; Sims et al., 2011). Thus, methylation of R1810 by CARM1 controls the expression of a subclass of RNAs, further expanding the gene-specific functions associated with the CTD. Interestingly, R1810 methylation is inhibited by S_5P and S_2P marks *in vitro*, suggesting that the methylation is placed before early initiation. However, this novel CTD mark can be detected within the actively transcribed Pol IIO form *in vivo* showing that arginine methylation is maintained within Pol II CTD during transcription (Sims et al., 2011). Additionally, the Tudor domain of TDRD 3 specifically binds to dimethylated R1810 (Sims et al., 2011). However no specific role of this novel interaction has been yet found. Although the mechanism of how this new CTD mark of R1810 interferes with the expression of short transcripts is unknown, this and maybe other modifications within the non-consensus repeats of the distal part of mammalian CTD may play a key role in recruiting the transcription machinery to certain gene loci or keep it away. In addition to R1810, eight lysines are located within the distal part of the CTD and are potential residues for acetylation, methylation, sumoylation and ubiquitylation. Mass spectrometry might be a powerful tool for mapping new additional posttranslational modifications within non-consensus repeats of the distal part of Pol II CTD.

Proline isomerisation: Proline-3 (P_3) and proline-6 (P_6) are totally conserved in all 52 repeats of mammalian CTD and are surrounded by phosphorylation sites on each side. Prolines can be in either *cis* or *trans* orientation, resulting in four possible configurations of each repeat further expanding the complexity of the CTD code (Egloff et al., 2008). The peptidyl proline isomerase Ess1 in yeast and Pin1 in mammals can isomerize the prolines at position 3 and 6 of each CTD repeat (Egloff et al., 2008). CTD-protein binding studies revealed that the polyadenylation/termination factor Pcf11 binds exclusively to repeats with S_2P and prolines in the *trans* configuration whereas the Ssu72 S_5P phosphatase recognizes repeats with S_5P and the downstream proline in the *cis* configuration (Noble et al., 2005; Werner-Allen et al., 2011) (see also part 1.2.3). These findings show that the isomerisation status of prolines interferes with the CTD phosphorylation pattern directly and *vice versa* suggesting a further regulatory mechanism in regulating recruitment and binding of CTD-interacting factors. Furthermore, Ess1 and Pin1 can activate Ssu72 to promote S_5P dephosphorylation (Noble et al., 2005; Werner-Allen

et al., 2011; Xiang et al., 2010). In addition, Pin1 is also involved in the hyperphosphorylation of the CTD during the mitotic (M) phase performed by Cdc2/cyclin B (Xu et al., 2003).

Serine and threonine glycosylation: Serine and threonine residues within the CTD can be glycosylated by the addition of a monosaccharide N-acetylglucosamine (O-GlcNAc) to their hydroxyl groups (Kelly et al., 1993). Importantly, phosphorylation and O-GlcNAcylation of single CTD residues are mutually exclusive proposing a role for CTD glycosylation in inhibiting CTD phosphorylation (Comer et al., 2001). In line with this, a recent study showed that dynamic glycosylation of CTD S_5 and S_7 mediated through O-GlcNAc transferase (OGT) and O-GlcNAc aminidase (OGA) exists during the assembly of the pre-initiation complex. A reduction in the transcription and Pol II occupancy at several B-cell promoters could be observed by the knockdown of OGT (Ranuncolo et al., 2012). These data suggest that the glycosylated form of Pol II is recruited to the promoter and that OGA acts at this stage to selectively remove the O-GlcNAc group before phosphorylation occurs. However, so far, no clear evidence has been demonstrated that glycosylation within the CTD plays an important role in gene expression.

1.2.2   Genetic analysis of the CTD of Pol II

The CTD is dispensable for polymerase activity *in vitro*, but deletion of the entire CTD in mice or yeast is lethal. In mammals, CTD with only 31 repeats interferes with cell viability (Meininghaus et al., 2000) whereas mice homozygous for a CTD containing 39 repeats show a high degree of neolethality (Litingtung et al., 1999). In yeast, but not in mammals, cells expressing Rpb1 with CTDs consisting of only ~ 50% of the original numbers of heptads are viable. In line with this, the CTD of budding yeast contains 26 repeats, but only eight heptads are required for cell viability and 13 are needed for wild-type-like growth (West and Corden, 1995). Genetic studies in mammalian cells revealed that a CTD composed of 55 consensus repeats can fulfil all essential functions for proliferation (Chapman et al., 2005). Moreover, a mutant consisting of only non-consensus repeats showed a severe growth defect compared to mutants of similar length, containing consensus repeats

(Chapman et al., 2005). Interestingly, these repeats differ from the consensus sequence mainly at S_7 suggesting an important role for this position within the CTD. These data also imply that the highly conserved composition of the mammalian CTD consensus and non-consensus repeats is probably not entirely essential for life but rather 'optimized' for efficient function or other, as yet unknown, purposes that could come along with a survival advantage e.g. in response to cellular stress. In addition, non-consensus repeats may play an important role in the expression of specific genes as shown for R1810 in regulating the expression of snRNA and snoRNA genes (Sims et al., 2011). Furthermore, it has been shown that the effect on cell viability and growth appears to be dependent on the number of repeats. In other words, the greater the number of consensus repeats, the greater the rate of proliferation and cell survival (Litingtung et al., 1999: Meininghaus et al., 2000: Chapman et al., 2005). The number of repeats that comprise the Pol II CTD in different organisms may reflect the requirement for complex pre-mRNA processing events and transcriptional control. In this line, by both increasing the number of repeats and diverging their sequence, a greater number and diversity of factors can bind the CTD. Chapman et al. could show that both the last repeat 52 that contains a unique site for the binding of the CTD tyrosine kinases Abl1 and Abl2 as well as repeats 1-3 serve to regulate the stability of Pol II by preventing its degradation to the CTD-less RNA Pol II form (IIb form). Interestingly, all other repeats could be deleted without inducing degradation (Chapman et al., 2004 and 2005). In more detail, mutagenesis of CTD repeat 52 showed a requirement for acidic amino acids at its C-terminus independent of their specific sequence. In addition, repeats 1-3 can not be replaced by consensus repeats suggesting that these repeats may serve as a spacer between the Linker region and the CTD rather than being a binding site for a specific factor (Chapman et al., 2004 and 2005).

A CTD-less RNA polymerase II can stimulate capping but the presence of the CTD increases the efficiency of this reaction fourfold (Mortillaro et al., 1996). However, a CTD-less RNA polymerase II is not able to initiate on the endogenous chromatin template (West and Corden, 1995; Bartolomei et al., 1988; Meininghaus et al., 2000). In a different study it has been demonstrated that the CTD independently stimulates each of the three major pre-mRNA processing events *in vivo* (Fong and Bentley, 2001). Interestingly, there is a difference between the amino- and carboxy-terminal halves of the CTD in the ability to stimulate different processing steps. While

the proximal part of the CTD can support capping without efficient splicing or 3' processing, the distal part of the CTD supports all of the three pre-mRNA processing steps (Fong and Bentley, 2001). Moreover, the CTD carboxyl-terminus (repeat 27-52) was sufficient for 3' processing and splicing concluding that although the CTD is a highly repetitive structure, there seems to be functional specialization of different segments within it (Fong and Bentley, 2001). In different studies, mutants where the positions S_2, T_4, S_5 or S_7 of the CTD had been replaced by alanine were transfected into mammalian cells and the viability of these Pol II CTD mutants was measured over a period of 4 days. As a result, mutants containing replacements of S_2/A, T_4/A, and S_5/A in 48 out of 52 CTD repeats revealed a strong growth defect with a dramatically reduced cell count after 4 days. Additionally, the T_4/S and S_7/A mutants showed an attenuated phenotype with almost constant cell numbers (Chapman et al., 2007; Hintermair et al., 2012).

Most genetic studies of CTD have been performed in yeast and the next chapter will focus on different aspects of yeast CTD mainly focussing on deciphering the CTD code in both budding yeast and fission yeast.

1.2.2.1 Genetic analysis of the budding yeast CTD

In a recent genome-wide ChIP study in budding yeast new insights in the complex interplay between the CTD-modifying enzymes have been gained. In a *kin28* mutant strain the levels of both S_5P and S_7P are strongly reduced at the 5' end of genes as expected, however, a dramatic increase of these phospho marks could be detected throughout the ORF compared to wild-type cells (Bataille et al., 2012). Consequently, the distribution but not the overall level of S_5P and S_7P is affected in the absence of Kin28. This result can be explained by the fact that Bur1 is a potent S_5 and S_7 kinase but its activity is repressed in the presence of a functional Kin28 complex. Another interesting finding is that Kin28 and Bur1 seem to have opposing roles during early transcription as the accumulation of Pol II at the 5' end of genes in a *kin28* mutant can be rescued by simultaneous knockout of Bur1. Bataille et al. also showed that the depletion of CTD phosphatase Ssu72 led to a similar pattern in both S_5P and S_7P marks extending further to the end of the gene, implying that Ssu72 dephosphorylates S_5 and S_7 prior to termination. In line with this, new data also revealed that the Ess1 isomerase can specifically stimulate the dephosphorylation of

both S_5 and S_7 at the 3' end of genes. Ess1 catalyzes the *cis/trans* inter-conversion of the peptidyl-proline bond between S_5-P_6 and the *cis*-isomer form is known to be the preferred substrate of Ssu72. Therefore Ess1 plays a crucial role in CTD dephosphorylation by Ssu72 (Bataille et al., 2012).

Bataille et al. also suggests that distinct variants of the same phosphoserine can be found within the CTD. They observed that S_5P is removed in two waves, first by Rtr1 and later by Ssu72. The same is true for S_2P by removing one part of it prior to termination whereas the rest is dephosphorylated after termination. These different subclasses of the same phosphoserine may arise because of *trans* versus *cis* conformations of the S-P bonds or due to differential phosphorylation of neighbouring residues. Another explanation would be that the degenerate repeats located within the distal part of the CTD might be functionally distinct from consensus repeats located mainly in the proximal part of the CTD with respect to targeting by the different CTD modifying enzymes.

Genetic studies in budding yeast have shown that Y_1, S_2, and S_5 are essential for CTD function (West and Corden, 1995; Pei et al., 2001) and that A-insertions between adjacent repeats are lethal whereas individual residues inserted between pairs of heptapeptides are well tolerated (Liu, 2010; Stiller et al., 2004). These findings propose that the minimal function unit of CTD lies within a di-heptad. Based on this, Stiller and colleagues discovered the irreducible unit of CTD function in budding yeast performing genetic analyses of CTD mutants (Table 1). The two essential sequence motifs defining the functional unit are paired tyrosines placed 7 amino acids apart (Y_1-Y_8) as well as three potential phosphoserines in a 2-5-9 orientation with respect to the Y_8 residue to a given di-heptapeptide (Liu, 2008 and 2010). Since these two essential elements are somewhat independent the final functional unit consists of the sequence Y_1-S_2-P_3-X_4-S_5-P_6-X_7-Y_8 that is either linked to a proximal S_2-P_3-X-S_5-P_6-X or to a distal S_2-P_3. In this line, yeast mutants are viable with repeats containing only a minimal sequence of these two essential elements (Y_1-S_2-P_3-T_4-S_5-P_6-S_7-Y_1-S_2-P_3-T_4) or by replacing the right-hand S_5-P_6-S_7 residues by alanines (mutant '252' and 'AR'; Table 1). Additionally, 9 or more tandem copies of this 11-mer sequence unit were sufficient for wild-type growth (Liu, 2008 and 2010). Since this mutant contains a row of non-overlapping individual minimal functional units it seems like that the overall sequence required for most or all CTD functions is not based on tandemly repeated heptads but is rather defined within

repeated units of three consecutive S-P pairs interspersed with Y residues that are spaced at a heptad interval. In a different approach it has been shown that additional distance between essential units results in a progressive decline in CTD efficiency. A very slow growth phenotype could be detected in mutants with five A insertions between every diheptad and complete lethality was obtained when units were separated by seven A residues (mutant '5A' and '7A'; Table 1) (Liu et al., 2010). Importantly, the lethal phenotype could be rescued by replacing alanine with proline in position 3 and 6 suggesting that the quality of the inserted sequence rather than the physical distance between functional units is important for CTD function (mutant 'AP'; Table 1). The long stretch of alanine residues tends to form stably secondary structure like α-helices and by replacing alanine with proline disrupts this structure leading to a more structurally unordered sequence around each given functional unit. In this line, a CTD mutant where seven alanine insertions are placed between every tri-heptad instead of every di-heptad grows vigorously explained by the fact that the amount of normal, structurally unordered sequence around the essential sequences has been increased (mutant 'A7'; Table 1) (Liu et al., 2010). This leads to the conclusion that placing ordered structures directly next to essential CTD units negatively influences the interactions between these units and binding partners. In this respect, CTDs with only two heptads between each 7 alanine stretch are not recognized as substrate for any tested CTD kinase (human Cdk7 and Cdk9, yeast Ctk1), whereas CTDs with three heptads between each 7 alanine stretch showed efficient substrate specificity for all three tested kinases. In an experiment focussing on the question what is most important for optimal CTD function: the total overall length of the CTD, the absolute number of essential units present, or the spacing of the essential elements along the CTD? It turned out that independent of the sequence repeated, the length variant in each strain set revealing the highest growth rate always contained the CTD with closest to normal length, rather than the CTD with a WT-equivalent number of essential functional units. These genetic studies in budding yeast show that on the one hand CTD repeats are functionally redundant but on the other hand overall length of the CTD is most important and also strongly conserved within species. One idea based on this observed CTD characteristics is that a certain length is required to establish an optimal 'loading platform' for CTD- and phospho CTD-associating proteins (PCAPs) (Liu, 2008 and 2010). In more detail, binding of protein factors needed for key functions determines the minimum

number of repeats and is responsible for the strong purifying selection on CTD towards overlapping tandem functional units. Next, length beyond the minimum gains extra space to bind proteins involved in additional or accessory functions and finally non-consensus repeats provide landing platforms for proteins that play a role in more taxon-specific functions. In conclusion, the need for maintenance of a dynamic microenvironment around each functional CTD unit in combination with an optimized macroenvironment for overall binding across the full CTD length probably led to the highly conserved evolution of CTD across many species (Liu et al., 2010).

Table 1 Overview of CTD mutants used in genetic studies in budding yeast.

CTD Mutant	Repeated Sequence	Cell phenotype at around WT CTD length
AL	ASPTSPS YSPTSPS	Lethal
YAA	YAATSPS YSPTSPS	Lethal
YATA	YAPTAPS YSPTSPS	Lethal
YAP	YAPTSPS YSPTSPS	Lethal
APS	YSPTAPS YSPTSPS	Viable ++++
AR	YSPTSPS YSPTAAA	Viable ++++
252	YSPTSPS YSPS	Viable +++++
AA	AA YSPTSPS YSPTSPS	Viable ++++
5A	AAAA YSPTSPS YSPTSPS	Slow growth +
7A	AAAAAAAYSPTSPS YSPTSPS	Lethal
A7	AAAAAAAYSPTSPS YSPTSPS YSPTSPS	Viable +++
AP	AAPAAPA YSPTSPS YSPTSPS	Viable ++++

The '+' marks for each viable CTD mutant indicate the relative vigor of the yeast cells bearing the mutants, compared to WT cells (WT is labeled as five pluses). Adapted from: *(Liu and Greenleaf, (2010), Genetic Organization, Length Conservation, and Evolution of RNA Polymerase II Carboxyl-Terminal Domain; Mol. Biol. Evol. 27(11):2628-2641. 2010).*

1.2.2.2 Genetic analysis of the fission yeast CTD

In Shuman's lab, recent genetic studies in fission yeast shed light into the key rules that govern the CTD code in this organism by manipulating the composition and structure of the Rpb1 CTD. The fission yeast is an ideal model system for CTD studies as the native heptad repeat array is relatively homogeneous consisting of 29 repeats of which 24 follow the consensus sequence (Figure 1). First they investigated the importance of all individual amino acids within the canonical repeat by introducing alanine in lieu of Y_1, S_2, P_3, T_4, S_5, P_6, and S_7 of every heptad of the

Rpb1 CTD array (Table 2). The key results from this approach were that Y_1, P_3, S_5, and P_6 are essential for viability whereas S_2, T_4, and S_7 are not (Table 2) (Schwer and Shuman, 2011). Interestingly, the S_2/A mutant grew well at 30°C revealing that S_2 phosphorylation is not essential and this observation contrasts clearly with the situation in budding yeast, where the same mutant was lethal (Table 2). In line with this, a similar result was obtained by studying a different mutant where Y_1 was replaced by phenylalanine along the whole CTD. This Y_1/F mutant was viable, though cold sensitive in fission yeast, whereas the analogous mutant in budding yeast is lethal. The Y_1/F phenotype in fission yeast suggests that the Y_1 hydroxyl group and therefore any tyrosine phosphorylation are not essential for the growth but, instead, the phenyl ring of Y_1 is indispensable at this position within the CTD. The lethal phenotype of both, the S_5/A mutant and the S_5/E mutant occurs in both yeast strains and the negative outcome of replacing S_5 with glutamate indicates that a state simulating constitutive S_5 phosphorylation is detrimental across species (Table 2) (Schwer and Shuman, 2011). Additionally, the finding that replacing S_5 with threonine is lethal, too, might be explained by the fact that the extra methyl group of threonine is directly deleterious due to steric hindrance with CTD binding proteins in fission yeast. The dominant role of S_5 as the sole serine phosphorylation source for the Pol II CTD code regarding vegetative growth was underlined by the viability of the double mutant S_2/A-S_7/A at 30°C in fission yeast (Table 2). Another interesting finding of the Shuman group was the requirement of S_2 for transcription during sexual differentiation and that this specific function of this CTD residue could be bypassed by subtracting S_7. They found out that the S_7P signal was higher in the S_2/A mutant compared to wildtype even though the total Rpb1 signal was higher in WT than in the S_2/A background. Based on this, Shuman and co-workers proposed that an imbalance in the CTD phosphorylation array and not the absence of a particular phospho-CTD residue reflects a CTD-associated pathology in this specific case and thereby adding a new aspect in how to read a CTD code (Schwer and Shuman, 2011).

On top of this, Shuman's lab could show in a very exciting experiment that the lethality of S_5/A is rescued by the fusion of the Mce1 capping protein in-frame to the mutant S_5/A cassette. The key finding of this experiment is that the essentiality of the S_5P mark reflects its singular requirement for capping enzyme recruitment, which can be bypassed by fusing the capping enzyme to the S_5/A CTD. In a similar

approach the Mce1 coding region was fused to mutant Y_1/A, P_3/A, P_6/A, and S_2/A-S_5/A CTD cassettes showing that the need for P_6 as an essential letter is withdrawed when the capping enzyme is directly tethered to the CTD. Additionally, S_2 does not turn into an essential factor regarding viability in the absence of S_5. Finally, the requirement for Y_1 and P_3 could not be bypassed by Mce1 fusion suggesting that Y_1 and P_3 might be used as recognition sites for CTD binding factors other than the capping enzymes (Schwer and Shuman, 2011).

Following the pioneer work performed in the labs of Stiller and Greenleaf focussing on the discovery of the minimal essential CTD unit in budding yeast Shuman's lab set up similar approaches for getting first insights in the grammar and punctuation of the CTD code in fission yeast. First, they confirmed that the essential CTD information lies within a di-heptad repeat in fission yeast, too. In genetic studies using CTD mutants where blocks of alanine insertions replaced the native amino acid residues at the distal end of each di-heptad unit they found out that the essential S_5-P_6 dipeptide need not to be located in consecutive heptad repeats. Additionally, the deca-peptide unit Y_1-S_2-P_3-T_4-S_5-P_6-S_7-Y_1-S_2-P_3 is the minimal coding unit of the CTD and the spacing between adjacent units is flexible (Schwer and Shuman, 2012). Moreover, Y_1 must be present in consecutive heptads and proper spacing between consecutive tyrosines is important for CTD function, a common feature of Y_1 in both, budding yeast and fission yeast. Interestingly, knockout of Ssu72 in fission yeast is not lethal, in contrast to its essentiality in budding yeast suggesting that a phosphatase other than Ssu72 is mainly responsible for dephosphorylating S_5P and S_7P in fission yeast.

Shuman's lab investigated the effect of CTD mutations on CTD phosphorylation *in vivo*. One result of this study is that independent pathways exist for S_2 and S_5 phosphorylation in fission yeast. Furthermore, S_2P and S_5P marks can be placed in T_4/A and S_7/A cells as well as in Y_1/F cells implying that neither S_2 nor S_5 phosphorylation requires concomitant phosphorylations of Y_1, T_4, or S_7. On the contrary, the P_6/G mutant directly influences the Rpb1 S_5P mark strengthening the idea that P_6 plays a specific role in the recognition of S_5 residues by cyclin–dependent kinases *in vivo* (Schwer and Shuman, 2012).

In summary, genetic studies of the Rpb1 CTD in both, budding yeast and fission yeast share important common features defining the minimal functional unit in respect of the need of an adjacent pair of tyrosines spaced 7 amino acids apart and

the essentiality of three consecutive serine residues in a 2-5-2 configuration. These results correlate with findings based on crystal structures of proteins complexed with the CTD showing e.g. that paired tyrosines anchor binding of the CTD to several associated protein partners (see also part 1.2.3). In addition, CTD-protein binding studies emphasize the importance of contiguous and properly spaced phosphoserines and that specific protein-CTD contacts although highly variable, tend to occur over relatively short stretches within the CTD (see also part 1.2.3). The last chapter of the introduction part will discuss the structural features of the CTD and will summarize the basic rules that govern the CTD-specific recognition of several CTD binding proteins.

Table 2 Overview of CTD mutants used in genetic studies in fission yeast.

rpb1	CTD	Growth (30°C)	Mating
"WT"	$(YSPTSPS)_{14}$	++	++
Y1A	$(ASPTSPS)_{14}$	lethal	
Y1F	$(FSPTSPS)_{14}$	++	++
Y1L	$(LSPTSPS)_{14}$	lethal	
S2A	$(YAPTSPS)_{14}$	++	sterile
S2T	$(YTPTSPS)_{14}$	++	++
S2E	$(YEPTSPS)_{14}$	lethal	
P3A	$(YSATSPS)_{14}$	lethal	
P3G	$(YSGTSPS)_{14}$	lethal	
T4A	$(YSPASPS)_{14}$	++	++
S5A	$(YSPTAPS)_{14}$	lethal	
S5T	$(YSPTTPS)_{14}$	lethal	
S5E	$(YSPTEPS)_{14}$	lethal	
P6A	$(YSPTSAS)_{14}$	lethal	
P6G	$(YSPTSGS)_{14}$	+	
S7A	$(YSPTSPA)_{14}$	++	++
T4A+S7A	$(YSPASPA)_{14}$	++	++
Y1F+S7A	$(FSPASPA)_{14}$	++	++
S2A+S7A	$(YSPTSPA)_{14}$	++	++

A summary of the mutational effects on growth and mating in fission yeast is shown. The alleles are named according to the amino acid substitutions introduced into all 14 consensus heptads appended to the 'rump' that connects the CTD to the body of Pol II *(Schwer and Shuman, (2011), Deciphering the RNA Polymerase II CTD Code in Fission Yeast; Molecular Cell 43, 311-318, July 22, 2011).*

1.2.3   Structural view of the CTD of Pol II

The CTD of yeast spans up to 650 Å in an extended conformation and is located near the RNA exit channel of Rpb1. Consequently, both the location and the flexible nature of the CTD support the binding of many protein factors in close proximity of

the nascent transcript (Phatnani et al., 2006; Munoz et al., 2010). Multiple phosphorylations of the CTD create not only new recognition sites, but also lead to an overall structural change of the CTD. The CTD structure becomes more extended upon phosphorylation due to charge repulsion between phosphate groups (Zhang and Corden 1991: Cagas and Corden 1995; Cramer et al., 2001; Noble et al., 2005). In line with this, phosphorylations and other modifications not only change the chemical structure of the CTD but also increase or restrict the conformational variability on the domain and effect recognition by other factors. For example, the phosphate group of phosphoserine contains a double negative charge and can form multiple hydrogen bonds and salt bridges. In contrast, a methyl group, linked to arginine or lysine, deletes the possibility for the formation of H-bonds but promotes hydrophobic interactions instead. Importantly, phosphorylation and glycosylation occurs on the same hydroxyl groups of serine and threonine residues, causing the modifications to be mutually exclusive. Glycosylation inserts a relatively large sugar rest to the peptide chain serving as a steric block to prevent aberrant phosphorylation (Kelly et al., 1993; Comer et al., 2001). In the same line, ubiquitination and sumoylation would lead to even more drastic changes to the CTD structure. The addition of carbohydrate, ubiquitin, and SUMO likely inhibit access to the neighbouring amino acids and might prevent or control the dynamic exchange of other posttranslational modifications and binding factors. However, no detailed information about the functional relevance of these modifications within the CTD is available yet. Another important feature of the CTD code is the *cis/trans* isomerisation of the S_2-P_3 and S_5-P_6 peptide bonds (Lu et al., 2007; Shaw 2002 and 2007). The majority of peptide bonds in the *cis* conformation appears in surface-accessible bend, coil, or turn motifs (Lu et al., 2007). Switches between the *cis* and *trans* isomers induce large structural changes, leading to sharp turns into the backbone that destroys previous interactions and also creates new epitopes for recognition. Due to its flexibility, the CTD has not been detected in the crystal structures of Pol II. Nuclear magnetic resonance (NMR) spectroscopy data, however, imply that the free CTD is largely flexible, although it also contains some residual structure and shows a tendency to form ß-turns at two SPXX motifs (S_2-P_3-T_4-S_5 and S_5-P6-S_7-Y_1) (Suzuki et al., 1989). Importantly, available structures of bound CTD peptides demonstrate how complex and diverse the recognition of basically the same peptide sequence may be implying that no simple rules of the CTD code exist.

In other words, the specificity of binding depends on different modification isoforms, modification patterns, length of bound CTD peptides, and *cis* or *trans* conformations of the phosphoserine-proline peptidyl-prolyl bonds. A combination of steric constrains, intramolecular and intermolecular hydrogen bonds, van der Waals forces, and electrostatic and stacking interactions contribute to the specific recognition and binding of protein factors to the CTD. Based on this, each CTD-binding protein has its own minimal requirement for the functional unit of a CTD peptide, which can be as long as three repeats (Cgt1; Figure 6c), or as short as four residues, similar to Scp1 (Figure 6b) (Fabrega et al., 2003; Zhang et al., 2006).

CIDs (CTD-interacting domains) are the best studied family of the CTD binding domains and can be found in Rtt103, SCAF8, Pcf11, and Nrd1 (Figure 6a) (Meinhart and Cramer 2004; Becker et al., 2008; Lunde et al., 2010; Kim et al., 2004; Patturajan et al., 1998; Barilla et al., 2001; Sadowski et al., 2003; Steinmetz et al., 1998; Vasiljeva et al., 2008). This domain contains eight α-helices and binds from 8 to 11 residues of the CTD. The CTD forms a classical ß-turn conformation which is positioned in the binding groove of the CID. The ß-turn consists of S_{2b}-P_{3b}-T_{4b}-S_{5b} and is always stabilized by three intramolecular H-bonds independent from the phosphorylation pattern (Figure 6a) (Becker et al., 2008). Contacts of the CTD-CID interaction are made by H-bonds between the CID and $P_{6a,b}$, S_{7a}, Y_{1b}, and S_{5b}. In addition, the side chain hydroxyl group of Y_{1b} forms an H-bond with a conserved aspartate of CID (Figure 6a). Interestingly, both the Rtt103 and SCAF8 CID binds the S_2P CTD with a higher affinity than the Pcf11 CID due to the presence of a conserved arginine that creates a salt bridge interaction with the phosphate group of S_2P (Lunde et al., 2010; Noble et al., 2005). In all CTD-CID complexes the S-P peptidyl-prolyl bonds are in *trans* conformation with the exception of Nrd1 that binds to S_5P CTD favouring a *cis* conformation of the $S_{5a}P$-P_{6a} peptidyl-prolyl bond (Figure 6a) (Kubicek et al., 2012). A common feature shared by all CTD-CID interactions involves the Y_1 residue. Its hydroxyl group forms an H-bond with a conserved aspartate of the CID and its aromatic ring is tightly placed in the hydrophobic pocket of the CID. In line with this, recently it has been demonstrated that the phosphorylation of Y_1 impairs the binding to all three yeast CID-containing proteins, Nrd1, Pcf11, and Rtt103 (Mayer et al., 2012).

Both, fission yeast CTD phosphatase Fcp1 and human small CTD phosphatase Scp1 belong to the family of Mg^{2+} dependent S-P/T-P-specific phosphatases (Zhang et al., 2006; Ghosh et al., 2008). Scp1 contains a FCPH (FCP homology) domain and binds the CTD in a CID-CTD similar way. In more detail, the peptide forms a ß-turn-like structure spanning residues S_{2b}-P_{3b}-T_{4b}-$S_{5b}P$ and one intramolecular H-bond is formed between the hydroxyl groups of S_{2b} and T_{4b} (Figure 6b). Additionally, P_{3b} is placed in a hydrophobic pocket while the S_{2b} and T_{4b} backbone carbonyls develop H-bonds with arginine178 (Figure 6b) (Zhang et al., 2006).

Fcp1 instead needs the minimal CTD peptide stretch of S_{5a}-P_{6a}-S_{7a}-Y_{1b}-$S_{2b}P$-P_{3b}-T_{4b} for binding and in contrast with Scp1, Y_{1b} and P_{3b} residues flanking $S_{2b}P$ are important for the phosphatase activity (Ghosh et al., 2008; Hausmann et al., 2004). These differences in positioning the CTD peptide within the structure of Scp1 and Fcp1 respectively, might explain that Scp1 preferentially dephosphorylates S_5, whereas Fcp1 favours S_2P residues as a substrate (Hausmann et al., 2002; Yeo et al., 2003).

Interestingly, mouse capping enzyme Mce1, and *Candida albicans* guanyltransferase Cgt1 use distinct CTD binding interfaces to read the same pattern of modification (Fabrega et al., 2003; Ghosh et al., 2011). Both enzymes bind the CTD peptides that contain S_5P in an extended ß-like conformation. However, the structure of Cgt1 covers almost three CTD heptads with S_5P in each whereas Mce1 interacts with a short doubly phosphorylated $S_{5a}P$-P_{6a}-S_{7a}-Y_{1b}-$S_{2b}P$-P_{3b} peptide (compare Figure 6c and 6d) (Fabrega et al., 2003; Ghosh et al., 2011). In the Cgt1-CTD interaction, the two terminal phosphoserines are anchored in two positively charged pockets whereas the central $S_{5b}P$ is not recognized. In more detail, interactions take place with $Y_{1b,c}$, P_{3b}, P_{6a}, and the terminal $S_{5a,c}P$ side chains whereas the middle heptad forms an exposed loop and may serves as a binding platform for other CTD-protein interactions (Figure 6c). In other words, the Cgt1 protein is associated with two distinct functional units, with the intervening CTD segment looped out suggesting that efficient CTD-Cgt1 interaction requires contact with more than one minimum CTD unit and enough flexibility between them to support cooperative binding. Similar to the Cgt1 structure, the most important residues that are involved in Mce1-CTD binding are Y_{1b} and $S_{5a}P$. While S_5P is placed in a positively charged pocket and forms several H-bonds, Y_{1b} is located in a

hydrophobic pocket forming an H-bond via its side chain hydroxyl group (Figure 6d) (Fabrega et al., 2003; Ghosh et al., 2011).
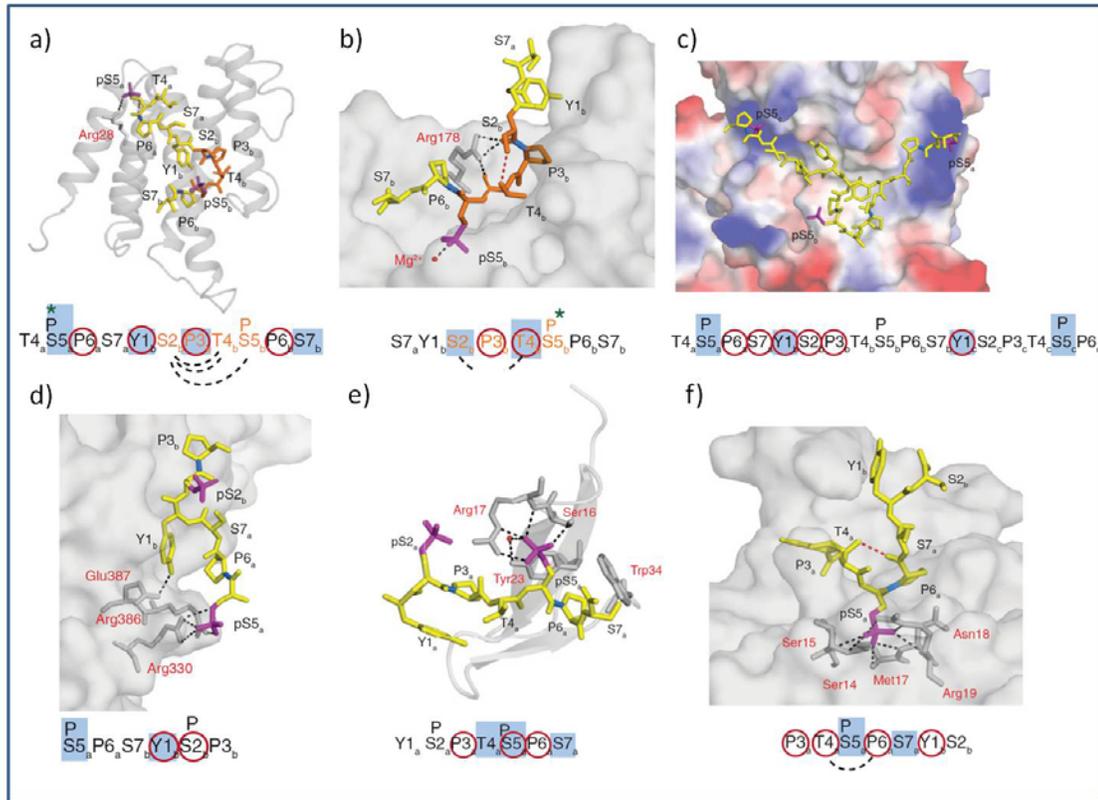

Figure 6 Overview of crystal structures between CTD-interacting proteins and CTD peptides.
a) Nrd1 CID-CTD complex; b) Scp1-CTD complex; c) Cgt1-CTD complex; d) Mce1-CTD complex; e) Pin1 WW-CTD complex; f) Ssu72-CTD complex; The CTD residues forming the ß-turn conformation are highlighted in orange, the phosphate group of serine is shown in magenta, the serine-proline peptide bonds are highlighted in blue, and the dashed lines indicate H-bonds. CTD peptide sequence below: blue boxes indicate residues involved in the intermolecular H-bonds, dashes lines indicate residues forming the intramolecular H-bonds, green asterisks indicate a direct recognition of the phosphorylated serine, red circles indicate other types of electrostatic interactions contributing to the binding. Alphabetical subscripts indicate the sequential number of the heptads. *(Jasnovidova and Stefl, (2012), The CTD code of RNA Polymerase II: a structural view; WIREs RNA 2013, 4:1-16. Doi: 10.1002).*

The human peptidyl-prolyl *cis-trans* isomerase PIN1 binds the CTD via its WW domain forming a compact triple-stranded anti parallel ß-sheet (Verdecia et al., 2000). Its WW domain belongs to the class IV group that specifically recognizes compact S-P motifs within peptide sequences. The binding interface includes one canonical heptad repeat of the CTD, which is phosphorylated at positions $S_{2a}P$ and $S_{5a}P$. Main contacts in the Pin1-CTD complex originate from P_{3a}, $S_{5a}P$, and P_{6a} and

the specificity towards S_5P recognition can be explained by the generation of several H-bonds between the phosphate group and aginine17 and serine16 of the WW domain (Figure 6e) (Verdecia et al., 2000).

Recent crystal structures of human and fruit fly Ssu72 revealed that this CTD-phosphatase requires the *cis* conformation of the S_5P-P_6 for placing the S_5P residue into the catalytic side of Ssu72 (Werner-Allen et al., 2011; Xiang et al., 2010). Upon *cis*-configuration, the phosphate group is attached to the active side via multiple H-bonds and additional interactions include intramolecular H-bonds between T_{4a} and P_{6a}, H-bonds of $S_{5a}P$ and S_{7a} backbone amides and Ssu72 residues. Additionally, van der Waals forces, electrostatic and stacking interactions take part in the binding of P_{3a}, T_{4a}, P_{6a} and Y_{1b} residues (Figure 6f) (Werner-Allen et al., 2011; Xiang et al., 2010).

In summary, extensive structural studies have shown an enormous diversity of interactions within CTD-protein complexes, however, contacts to CTD Y_1 and phosphorylated S_5 side chains predominate in most of these complexes which is consistent with the fact that these two positions within the CTD heptad array are the least degenerate in Nature. In the future, it will be important to combine the structural and dynamic data of the CTD interactions which may then lead to the deciphering of how the CTD code is written, read, and erased.

Aim of present study:

Mammalian RNA polymerase II (Pol II) largest subunit Rpb1 contains a unique and highly repetitive domain at the carboxy-terminus (CTD) with the consensus heptad Y_1-S_2-P_3-T_4-S_5-P_6-S_7. As discussed earlier (see Introduction) each single residue of the consensus sequence can potentially be modified. Whereas phosphorylation of serine-2, serine-5 and serine-7 have already been extensively characterised, the role of tyrosine-1 and threonine-4 phosphorylation are poorly understood. Several CTD modification specific antibodies have been raised in our laboratory, providing valuable information while mapping these modifications *in vivo* by ChIP analyses. However, in spite of the availability of all these CTD-specific antibodies, many questions about the modification pattern of the CTD *in vivo* still remain elusive and a number of questions remain unanswered. For example, how extensively are

individual repeats modified? Are there distinct phosphorylation marks excluding each other? Are there distinct 'signatures' marking the proximal or the distal part of the CTD? Do these 'signatures' form platforms for specific recruitment of CTD-binding proteins? And finally, is there a difference in the degree of phosphorylation among the different CTD heptad repeats?

So far, several CTD-phosphosites have already been published (e.g. http://www.phosphosite.org) and accordingly, the first part of this work was to set up a protocol for mapping phosphosites in order to confirm existing patterns. Of note here, is that the distal part of the wild-type (WT) CTD of Pol II is accessible to mass spectrometry analysis (MS) due to lysine (K) and arginine (R) residues at the seventh position of the non-consensus heptad repeats. However, the main part of WT CTD-sequence is inaccessible. The main reason for this is that the mass spectrometer is most efficient at obtaining sequence information from peptides that are up to 30 residues long, rather than from longer peptides or whole proteins.

Therefore, in a second step, CTD mutants were established to make the whole sequence accessible to MS and to map phosphosites within the entire CTD *in vivo*. MS analysis of four CTD mutants showed that it was difficult to 'read' longer peptides with more than 28 amino acids in length and thus, the sequence coverage of the proximal part of the CTD was still incomplete. Consequently, a second round of CTD mutants was created, where the proximal part of the CTD can be fragmented into only di- and tri-heptads that are readily assessable by MS.

This work has established a workflow for identifying abundant CTD signatures that define a so-called 'CTD-code' and can finally be linked to specific CTD-binding protein interactions.

2. Results

The establishment of a reliable protocol for mapping phosphosites within the CTD of Pol II using tandem mass spectrometry (MS/MS) was attained in this work. Additionally, development of a standardized procedure summarized in the workflow scheme in Table 3 was attempted and the initial results appear very promising in opening new ways to the evaluation of phosphorylation patterns within the CTD of Pol II in *vivo.*

Table 3 The principle of my workflow.

Lysis of Raji cells and preparation of
whole cell extracts

↓

Immunoprecipitation of Rpb1 by using CTD- or Rpb1-specific monoclonal antibodies

↓

SDS-PAGE of immunoprecipitates, coomassie staining
and excision of Rpb1 corresponding bands

↓

In-gel trypsinization of purified Rpb1 proteins
(hydrolysis only at carbonyl group contributed by Arg/Lys)

↓

Tandem mass spectrometry (MS/MS) of CTD-peptides and
subsequent database matching with obtained tandem-MS-spectra data
using Sequest computer software algorithms

↓

Data analysis using bioinformatics approaches

2.1 Mapping phosphosites of WT CTD peptides of Raji cells

In order to map posttranslational modifications of proteins via MS the protein of interest has to be digested into peptides using a sequence-specific protease, such as trypsin. The aim of producing optimally up to 30 amino acid long peptidic

fragments is imposed by the detection limit of the mass spectrometer, which can provide sequence information up to this size. Here, trypsin is used for the digestion of Rpb1 proteins into peptides. Trypsin is an aggressive and stable protease, which very specifically cleaves proteins on the carboxy-terminal side of R and K residues. For the case of the Rpb1, this creates peptides both in the preferred mass range for sequencing and with a basic residue at the carboxyl terminus of the peptide. Such peptides are 'information-rich' and provide easily interpretable peptide-fragmentation spectra.

2.1.1 Closer look to the mammalian WT CTD sequence

Mammalian WT CTD of Rpb1 of Pol II consists of a long stretch of consensus heptad repeats with the sequence YSPTSPS in its proximal part, whereas the last 26 repeats contain only three consensus repeats (see Figure 7). Among the non-canonical repeats, position 7 is the most variant residue. Importantly, eight lysines and one arginine can be found at this position which makes the distal part of mammalian CTD accessible to MS analysis.

Figure 7 Human CTD sequence. <u>Left:</u> Scheme of the 52 repeats of human CTD. Amino acids that diverge from the consensus sequence are marked in blue. Repeat number is shown in front of each corresponding sequence. <u>Right:</u> Predicted fragmented CTD peptides after trypsin digestion. Repeats 39 and 40 as well as repeats 46-47 and 48-49 can not be separated in the MS analysis due to identical masses.

The first step was to set up a protocol that allows the efficient mapping of phosphosites within the distal part of the WT CTD. The analysis of the WT CTD served as an original 'internal control' of this protocol, by comparing and confirming

the outcome of the spectra with published data of phosphorylated CTD peptides originating from whole phosphoproteomic approaches.

2.1.2 Purification of the WT Rpb1 protein

A key point for the purification of the Rpb1 protein was to use phospho-CTD specific antibodies that purify the hyperphosphorylated Pol IIO form very efficiently. Importantly, our lab has established monoclonal antibodies against all different phosphorylated epitopes within the consensus sequence of CTD-heptad repeats. These antibodies were all tested individually as well as in all different combinations to find the CTD-antibodies with the highest IP-efficiency. The combination of CTD α-S_2-P/α-S_5-P (view also Figure 5, Introduction) in a 1:1 ratio turned out to be the best choice resulting in the purification of approximately 80-90% of the Pol IIO form using whole Raji cell lysates (Figure 8). Whole cell lysate extracts of Raji cells growing under normal cell culture conditions were used for the IP reaction and the western blot analysis using a specific Rpb1 antibody (Pol3.3) for detection was performed via the ECL-detection system applying two different exposure time points (view also 4.4.5.3 in Material and Methods).
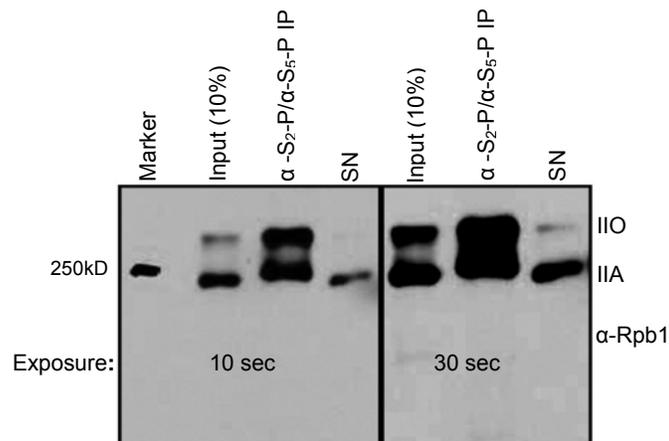
Figure 8 ECL-Western blot analysis of WT Raji Rpb1 purification. <u>Left:</u> Exposure time of membrane for 10 seconds. <u>Right:</u> Exposure time of membrane for 30 seconds. Rpb1 (IIO and IIA form) was purified with α-S2P/ αS5P in a 1:1 ratio and detected on the ECL blot via α-Rpb1. SN=Supernatant fraction of IP.

On the membrane with the longer exposure duration (30 seconds) saturation of the signal of the Rpb1 antibody is obtained and about 10% of the Pol IIO form can be found left in the supernatant fraction of the IP reaction. Interestingly, although the IP

antibodies used (α-S$_2$-P/α-S$_5$-P) are specifically recognising the Pol IIO form they also purify a large amount of the hypophosphorylated Pol IIA form when using high amounts of cells in the IP-reaction (Figure 8).

Next step was to isolate enough Rpb1 protein for subsequent MS analysis. The detection of the protein of interest after Coomassie gel staining served as a quantitative assessment of the amount of isolated protein. Consequently, 80 Mio cells were needed to visualize both the Pol IIO and Pol IIA form on the Coomassie gel (~ 50 ng of protein) following IP and SDS-PAGE (Figure 9).
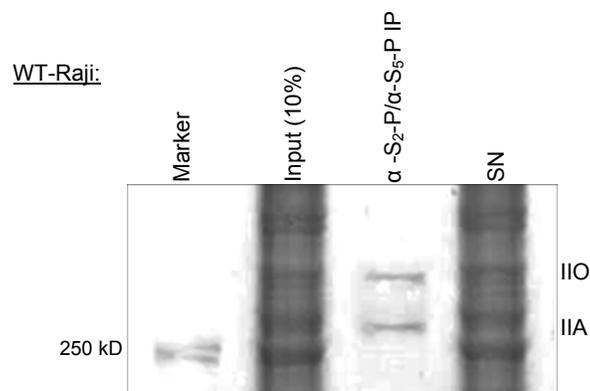

Figure 9 Coomassie gel after WT Raji Rpb1 purification. 80 Mio cells were used for the IP and Rpb1 (IIO and IIA form) was purified with α-S2P/ αS5P in a 1:1 ratio and detected on the SDS-PAGE gel after Coomassie staining. Both forms of Rpb1 (IIO and IIA) can be detected in the IP lane (Lane 3). SN=supernatant fraction of IP.

Both Rpb1 forms (IIO+IIA) were analysed via MS/MS analysis and the results are summarized in Figure 10. When using 80 Mio cells as starting material the MS output was rather low, detecting only five phosphorylated CTD peptides. Due to this weak outcome the protein amount was increased using 300 Mio cells, instead of 80 Mio cells, for every MS run. An additional phosphopeptide purification step using TiO$_2$ beads was included. The TiO$_2$ enrichment strategy is based on the general affinity of phosphorylated peptides towards metal oxides and reduces the underrepresentation of phosphorylated peptides in a high 'background' of non-phosphorylated peptides. These improvements led to a 10-fold increase in the MS outcome and not yet described to date, new phosphorylated CTD peptides, that could be found in several positions within the WT CTD of Pol II (Figure 10).

Figure 10 MS analysis of WT CTD of Pol II. <u>Left:</u> Detected phospho-CTD peptides using 80 Mio cells in one run. <u>Right:</u> Detected phospho-CTD peptides using 300 Mio cells and including a phosphopeptide purification step. Repeat numbers and detection counts are indicated for corresponding phospho-CTD peptide (shown in brackets). Amino residues diverging from the consensus sequence are marked in blue and amino acid residues that are found to be phosphorylated are marked in red.

Preliminary results of the MS analysis of WT CTD showed that 101 phosphosites could be mapped within 184 heptad repeats, which is a rather low average phosphorylation frequency of 0.55 phosphosites / CTD-heptad. Most abundant phosphorylated residues within the CTD were T_4 and S_5 with 26 and 32 phosphorylation counts, respectively. As depicted in figure 10, most frequent phosphorylated CTD peptides contained three phosphorylation sites and the preferred distribution of these triple-phosphosites was one phosphosite per heptad repeat.

2.2   Mapping phosphosites of mutated CTDs of Pol II

To get insights into phosphorylation patterns of the entire CTD of Pol II *in vivo*, CTD mutants were established for optimal mapping of phosphosites via MS. Mutated CTD sequences were synthesized by *Geneart* in Regensburg, subcloned into a final Rpb1

expression vector and transfected into Raji cells. Recombinant Rpb1 was then purified with CTD-specific antibodies and phosphosites of mutated CTD sequences were finally mapped in a MS/MS approach.

2.2.1 Designing CTD mutants to obtain complete sequence coverage for subsequent MS analysis

In order to obtain MS data including the complete CTD, K and R residues were introduced at position 7 of heptad repeats mainly within the proximal part of the CTD. K and R residues are a prerequisite for subsequent peptide fragmentation using trypsin for protein digestion. To receive full sequence coverage, it is essential that all fragmented peptides have different masses based on their length and amino acid composition. Consequently, additional amino acid residues next to K and R were strategically inserted into the CTD sequence. All in all, 9 different CTD mutants have been designed according to different numbers and lengths of fragmented peptides and different numbers of CTD mutations. The length of the different CTD peptides ranged from mono-heptads (7 residues) to hexa-heptads (42 residues). Subsequently, CTD mutants that comprise longer CTD peptides are less mutated and *vica versa*. Figure 11 shows one example of how the CTD can be mutated for achieving CTD peptides comprising the whole CTD sequence that can be then analysed MS. In this CTD mutant (M-8K4R), eight lysines and four arginines were inserted at the seventh position of CTD heptad repeats via amino acid substitutions leading to the fragmentation of 20 CTD peptides each containing a unique mass. In the M-8K4R, the longer CTD peptides cover four heptad repeats (28 residues) with the longest peptide covering 34 residues (repeat 50-52), whereas the shortest CTD peptides can be found with repeat 39 and 40 including only 7 residues. The majority of mutations were introduced into the CTD via amino acid substitutions, however, in CTD peptide 19-21 an alanine was placed into the CTD in an additive manner expanding the overall CTD length. Importantly, while most mutations were placed in the proximal part of the CTD, two lysines were replaced by arginine at position seven of repeat 40 and repeat 47 in the distal part of the CTD in order to avoid peptides with identical sequences and, therefore, to obtain different masses for CTD peptide 39 and 40, as well as for CTD peptides 46-47 and 48-49. For the same reason, further amino acid substitutions (alanine in peptide 13-15; serine in peptide 16-18;

leucine in repeat 19-21), next to arginine and lysine, were created by using aliphatic non-polar amino acids (Figure 11).

Figure 11 CTD sequence and predicted CTD peptides after trypsin digestion of CTD mutant M-8K4R. <u>Left:</u> Scheme of the 52 repeats of CTD mutant M-8K4R. Amino acids that diverge from the consensus sequence are marked in blue and mutated residues are marked in red. Repeat number is shown in front of each corresponding sequence. <u>Right:</u> Predicted CTD peptides after trypsin digestion. CTD peptides cover whole CTD sequence (all repeats) and contain unique masses.

2.2.2 Establishing cell lines expressing Pol II CTD mutants

All 9 CTD mutants were successfully cloned into the final expression vector RX4-267 (Meininghaus et al., 2000). This vector contains the full length, haemagglutinin-(HA)-tagged, mouse Rpb1 gene comprising 28 exons (Figure 12). The CTD is encoded by exon 28 and can be exchanged for any given CTD sequence using flanking restriction sites on both ends (AgeI and NotI) as insertion points. The Rpb1 gene contains a point mutation (N793D) conferring α-amanitin resistance (Bartolomei and Corden, 1987). Consequently, in the presence of α-amanitin, the endogenous Pol II is effectively, chemically 'knocked-out', thereby allowing the properties of the mutant RNA polymerases to be examined *in vivo*. Since the Rpb1-expression vector uses the replication origin of the Epstein-Barr virus (EBV-oriP), it is episomally maintained instead of being integrated into the genome, when using human cells expressing the EBV-nuclear antigen 1 (EBNA1). Episomes replicate like extra chromosomes thereby offering a great advantage of avoiding position effects within the genome, as well as allowing cell lines to be produced as a 'batch-culture'.
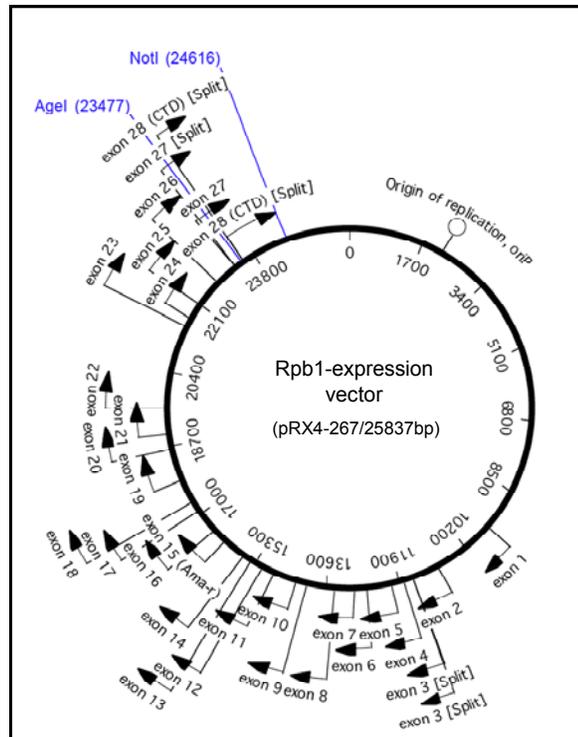
Figure 12 Scheme of the final Rpb1 expression vector. This vector encodes the whole mouse Rpb1 gene containing 28 exons. CTD is encoded by the last exon and can be easily exchanged using the restriction sites AgeI and NotI on both ends, respectively. For more detailed informations view text above.

The Rpb1-expression vector containing the mutated CTD was transfected into Raji cells via electroporation and was positively selected due to its neomycin resistance. After 2-4 weeks under neomycin selection, cell viability of 80-90% was obtained and α-amanitin was added thereby knocking out the endogenous Pol II within the next 24 hours.
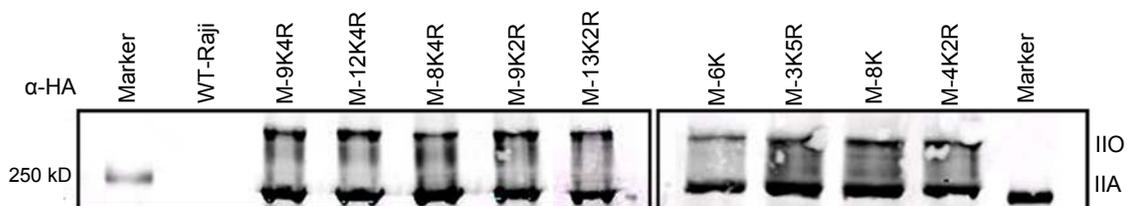
Figure 13 Western blot analysis showing stable expression of all 9 CTD mutants. Cell lysates of CTD mutants were produced after 2 weeks under α-amanitin treatment and stably expressed recombinant Rpb1 containing hyperphosphorylated and hypophosphorylated mutated CTDs, respectively. IIO- and IIA-forms were detected via α-HA. WT Raji lysate (second lane) was used as a negative control obtaining no signal with α-HA.

Viability of WT Raji cells dramatically decreases after 5-7 days under α-amanitin

selection. All 9 CTD mutants showed full viability and stable expression of their recombinant polymerases over months under treatment with α-amanitin using normal cell culture conditions and therefore all CTD mutants were included in the final MS analysis (Figure 13).

2.2.3 Purification of Rpb1 CTD mutants

Compared to WT Rpb1 immunoprecipitation from Raji cells and subsequent band isolation, for the recombinant Rpb1 proteins there was an additional option, due to an HA-tag inserted in their sequence. Therefore, in this case, both purification schemes were adopted, either with the combination of the α-S_2P/α-S_5P-IP, or with an α-HA antibody (3F10) and their efficiencies were compared by western blot analysis (Figure 14).
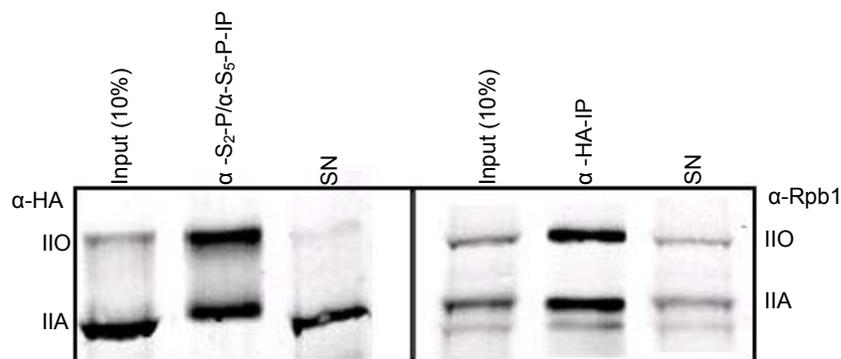
Figure 14 Odyssey-Western blot analysis of CTD mutant M-3K5R purification. Left: IP purification of IIO and IIA form using α-S_2P/α-S_5P. Detection antibody: α-HA; Right: IP purification of IIO and IIA form using α-HA. Detection antibody: α-Rpb1; SN=Supernatant fraction of corresponding IP.

As it can be seen in Figure 14, in both α-S_2P/α-S_5P and α-HA immunoprecipitates recovery of the hyperphosphorylated Pol IIO, as well as the hypophosphorylated Pol IIA form can be observed. The α-S_2P/α-S_5P IP appears to enrich CTD mutant M-3K5R as efficiently as it has been shown for WT cells (see Figure 8), purifying 80-90% of the hyperphosphorylated Pol IIO form and the results were consistent for all 9 CTD mutants. The α-HA IP seems to be less efficient in the recovery of the hyperphosphorylated form, apparently due to the fact that the amounts of the overexpressed unmodified recombinant protein are more abundant in the cell and in this case there is no phospho-selection of the precipitates.

For subsequent MS analysis, 450 Mio cells were used for one experimental round for

each CTD mutant comprising IP purification either with α-S$_2$P/α-S$_5$P or α-HA and subsequent band isolation after SDS-PAGE and Coomassie staining (Figure 15). Interestingly, after Coomassie staining, a third RNA Pol II form, termed IIO low, located between the IIO and IIA form could be detected in the α-HA immunoprecipitates. The reason why this intermediate Rpb1-form was not observed in the western blot analysis might be due to several causes, such as the different amount of cells used in western blot and coomassie stained gel as well as the arbitrary resolution of the gel bands in each case. For western blot analysis lower cell amounts are preferred in order to avoid saturated signals in the antibody detection reaction, whereas in the final purification destined for subsequent MS analysis more material is used to obtain a more informative MS data outcome. These results concerning the differences between the two IP procedures and what they recover will be discussed in more detail later (see Discussion).
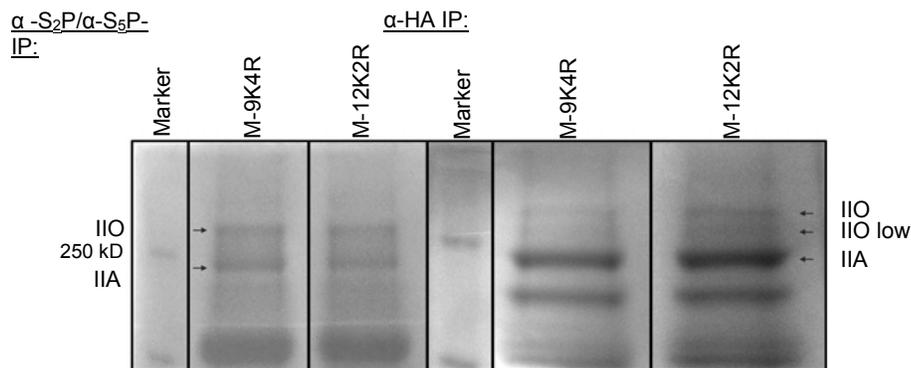
Figure 15 Coomassie gel after purification of recombinant Rpb1 of M-9K4R and M-12K2R. Left: Coomassie staining of IP-purified IIO and IIA form of M-9K4R and M-12K2R using α-S$_2$P/α-S$_5$P. Right: Coomassie staining of IP-purified IIO, IIO low and IIA form of M-9K4R and M-12K2R using α-HA. Purified recombinant Rpb1 extracted from 450 Mio cells were loaded in each well.

Following Coomassie staining, the IIO form and IIA form purified from the α-S$_2$P/α-S$_5$P-IP as well as the IIO form, IIO low form and the IIA form purified from the α-HA-IP were excised from the gels and further processed as described before (see Material and Methods) for final MS analysis. Massive MS data from the 9 different CTD mutants were collected, performing multiple replicate rounds of each mutant and the final results were implemented into bioinformatics evaluations.

2.3 Mass spec results of Pol II CTD mutants

All data obtained for the same CTD mutant independent from the origin of the probe (IP-approach; different Pol II forms) were collected as one data set for the corresponding mutant. The raw MS data were further processed using the computer software Sequest. Sequest is a MS/MS database search program originally developed in 1993 in the Yates lab at the University of Washington. It correlates MS/MS spectra of peptides against peptide sequences from a sequence database. Classical Sequest applies a two-stage scoring method for each search. The first stage applies the preliminary score to filter through all candidate peptides in the sequence database. The best scoring candidate peptides are then re-scored using the cross correlation algorithm. The sensitivity of the cross correlation algorithm is enhanced by the correction factor that is applied in its calculation (copied from: http://proteomicsresource.washington.edu/sequest.php).

Sequest data files of each CTD mutant were then used for further more detailed bioinformatics analysis addressing the degree and patterns of modifications in specific heptad repeats of each individual mutant.

2.3.1 CTD sequence coverage by MS/MS analysis

The MS outcome clearly showed that CTD peptides consisting of di- or tri-heptads exhibited by far the highest total count numbers in the MS detection. A massive drop in data acquisition could be observed when analysing very short CTD peptides (i.e., one heptad repeat). Similar weak results were obtained when increasing the length of CTD peptides beyond three heptad repeats. Consequently, 100% CTD sequence coverage could be obtained in three CTD mutants (M-13K2R, M-9K4R and M-12K2R), for which protein digestion led to the fragmentation of only di- and tri-heptads. Furthermore, in these three mutants, all potential phosphosites within the CTD were found to be phosphorylated (Figure 16). Interestingly, all CTD repeats were also found in the unphosphorylated state except repeat 52.
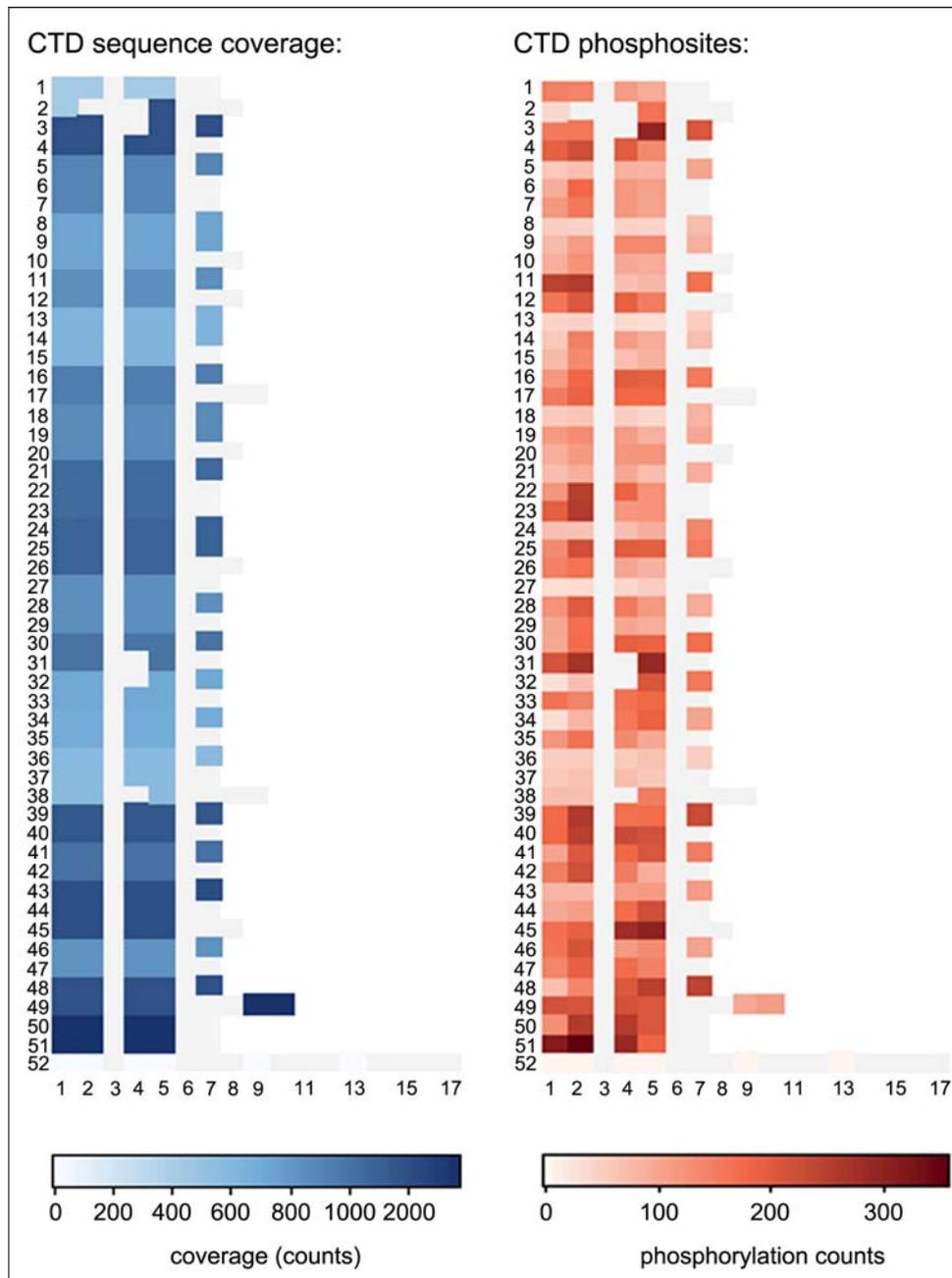
Figure 16 CTD sequence coverage and CTD phosphosites of M-12K2R All 52 repeats of M-12K2R are shown with repeat number in front of corresponding repeat. Each box represents a residue with corresponding position within the repeat written at the bottom. Repeat 2 lies on two different peptides due to an arginine at position 4. Left: Blue color code refers to total coverage count of each repeat. Right: Red color code refers to total phosphorylation counts of each phospho-residue within the CTD.

The CTD phosphosite graph of M-12K2R shows that next to the well described phosphorylation residues S_2P and S_5P, also the remaining three Y_1, T_4 and S_7 are frequently highly phosphorylated. In more detail, a tendency towards equal distribution of phosphorylation counts among phosphosites could be observed in

most of the CTD repeats. In addition, phosphorylated residues in heptad repeats diverging from the consensus CTD sequence and mainly located in the distal part of the CTD were found to be strongly phosphorylated as well.

2.3.2 Different phosphorylation patterns in adjacent heptad repeats

Analysing adjacent CTD repeats one interesting observation was that in some cases neighbouring repeats were phosphorylated on the same residue reflecting synchronised phospho-heptad repeats (Figure 17A+D). However, the majority of adjacent heptad repeats displayed different phosphorylation patterns (Figure 17B+C).
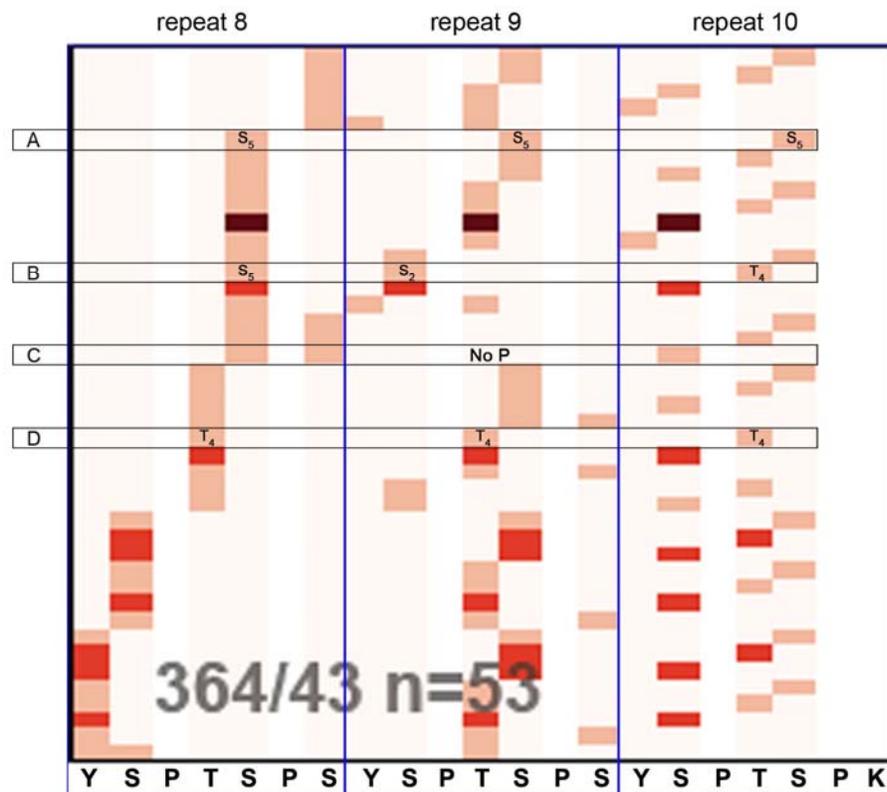
Figure 17 Comparison of phosphorylation patterns detected within the adjacent repeats 8-10 of M-9K4R. Each row corresponds to one specific amino acid annotated at the bottom, whereas red boxes indicate phosphosites. Blue squares divide the detected CTD peptide in the three adjacent heptad repeats 8-10. **A+D:** Identical phosphorylation patterns can be found between adjacent repeats. **B+C:** Different phosphorylation patterns can be found between adjacent repeats. **C:** Central CTD heptad (repeat 9) is in the unphosphorylated state, while the flanking heptad repeat on each side is phosphorylated. 364 = all possible 3P-combinations of repeat 8-10; 43 = detected 3P-combinations of repeat 8-10; 53 = total counts of detected 3P-peptides comprising repeats 8-10. Dark red: 3P-combinations found three times; red: 3P-combinations found twice; light red: 3P-combinations found once.

Additionally, the MS analysis detected unphosphorylated CTD heptads flanked by phosphorylated neighbour CTD repeats at both ends as well (Figure 17C). All 14 potential phosphorylation-sites participated in the formation of specific patterns.

2.3.3 Phosphorylation frequencies within mono-, di- and tri-heptads

Interestingly, the highest phosphorylation frequency found within CTD peptides detected in the mass spectra, were four phosphosites (4P) in parallel and this appeared to be independent from the length of the peptide.
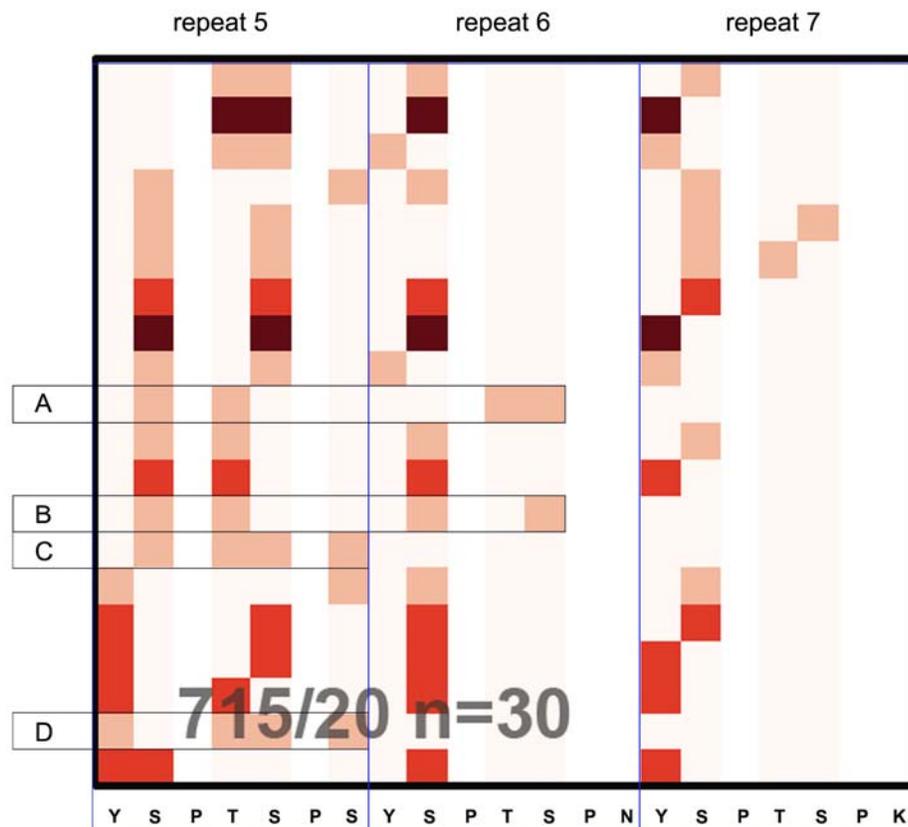
Figure 18 Phosphorylation frequencies within mono-, di- and tri-heptads shown for adjacent CTD repeats 5-7 of M-12K2R. Each row corresponds to one specific amino acid annotated at the bottom, whereas red boxes indicate phosphosites. Blue squares divide the detected CTD peptide in the three adjacent heptad repeats 5-7. **A+B:** Four phosphosites in parallel can be found within di-heptads. **C+D:** Four phosphosites in parallel can be found within mono-heptads. Majority of detected 4P-combinations are found within tri-heptads (16 out of 20). 715 = all possible 4P-combinations of repeat 5-7; 20 = detected 4P-combinations of repeat 5-7; 30 = total counts of detected 4P-peptides comprising repeats 5-7. Dark red: 4P-combinations found three times; red: 4P-combinations found twice; light red: 4P-combinations found once.

In other words, mono-heptad repeats (Figure 18C+D) showed the same highest degree of phosphorylation as that observed for di- and tri- heptad repeats, although the longer CTD peptides tended to carry more 4P-combinations than shorter ones. More specifically, to date, 4P-combinations have been detected either in mono-, or in di- or in tri-heptads in 4 out of 9 CTD mutants. All 13 potential phosphorylation-sites contributed to the formation of specific 4P-signatures (Figure 18).

However, the average counts of detected 3P- and 4P-peptides were rather low in the MS analysis. A much better outcome was achieved when analysing mono- and double-phosphorylated (2P/3P) di- and -tri-heptads. In more detail, in the CTD mutants, M-13K2R, M-9K4R and M-12K2R, from which CTD peptides were detected that covered their whole CTD sequence, on average, 80% of possible 2P-combinations have been found within di-heptads and 63% of possible 2P-signatures within tri-heptads. Moreover, all possible mono-phosphorylated-forms of all CTD peptides originating from these three mutants were found except for the last CTD repeat.

2.3.4 Phosphorylation patterns within the consensus heptad sequence

A profound and to date unclarified question in the field of CTD research addresses the actual phospho-combinations occuring within the highly conserved CTD-consensus sequence consisting of the residues $Y_1S_2P_3T_4S_5P_6S_7$ *in vivo*. Importantly, the total count numbers (absolute abundance) of consensus heptads containing only one phosphosite (1P) are dominant compared to the higher P-levels (2P, 3P and 4P), with total counts ranging between 3000 and 5000, taking into account the data obtained with all 9 CTD mutants. 2P-heptad consensus repeats were found between 100 and 500 times, whereas detection counts of 3P- and 4P-levels were very low (Figure 19A).

In these analyses, at the mono-phospho level (1P) all 5 potential phosphosites were found to be phosphorylated in rather equal amounts comparing the relative abundance (ranging from 15-20%) among all 1P-signatures (Figure 19B). Furthermore, all possible 2P-combinations (10 out of 10) were mapped within the CTD-consensus sequence. Among the 2P-combinations, the two most abundant

ones were S_2-P/S_7-P (more than 20%) and Y_1-P/S_7-P (13%) whereas the combination Y_1-P/S_2-P displayed the lowest relative abundance (3%) among all ten possible 2P-signatures (Figure 19B). Additionally, all possible 3P-combinations (10 out of 10) with their relative abundance ranging from 5-17% could also be identified, with the two most relative abundant 3P-combinations being S_2-P/S_5-P/S_7-P (17%) and Y_1-P/S_2-P/S_7-P (17%). On the contrary, at the 4P-level of consensus heptads, only 2 out of 5 possible combinations could be observed in very low equal counts with S_2-P/T_4-P/S_5-P/S_7-P and Y_1-P/T_4-P/S_5-P/S_7-P (relative abundance 50% each). In this analysis all consensus heptad repeats of all mutants were included independent from their location within the CTD, as well as from variations in sequence of adjacent repeats (Figure19).
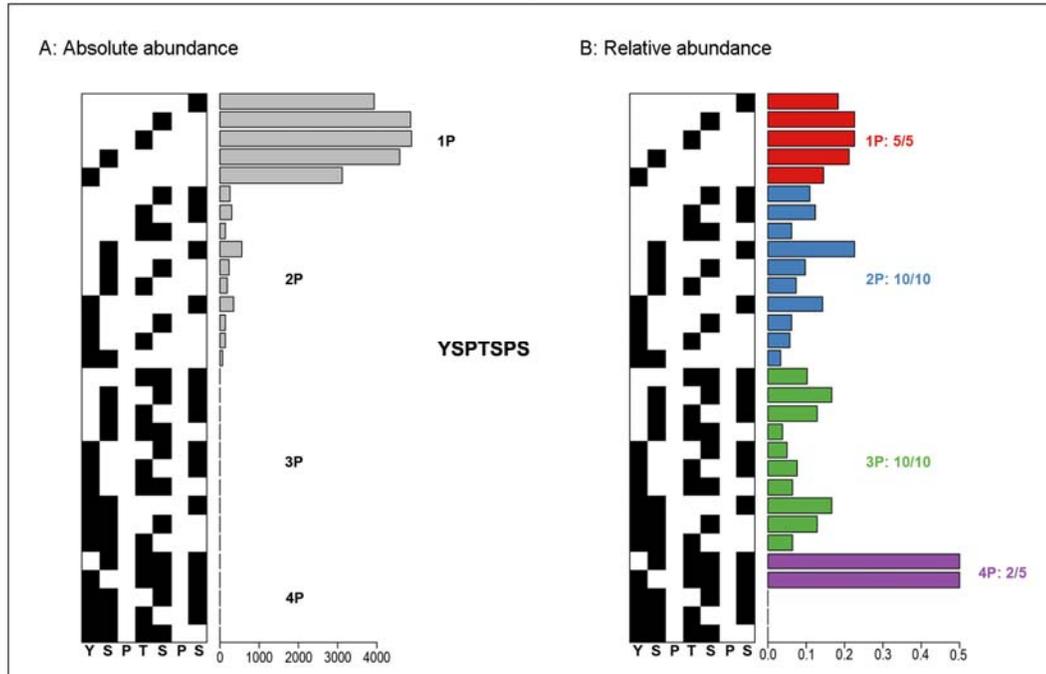
Figure 19 Phosphorylation patterns of the CTD consensus sequence. A: Absolute abundance (total counts form 0-4000 counts) of all detected 1P- , 2P-, 3P- or 4P-combinations found within consensus heptads. **B:** Relative abundance (0-50%) among all detected 1P- (5 out of 5/red), 2P-(10 out of 10/blue), 3P-(10 out of 10/green) or 4P- (2 out of 5/violet) signatures within consensus heptads. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites.

2.3.5 Phosphorylation patterns within di-consensus heptad repeats

Next to mono-consensus repeats, di-consensus repeats were also scanned for phosphorylation patterns extracting data form all 9 CTD mutants (Figure 20). Similar

to mono-heptad repeats, in di-heptad consensus repeats all possible 1P- (10/10) and 2P-combinations (45/45) could be detected. However, in contrast to 100% saturation of 3P-combinations within mono heptad repeats, in di-heptad consensus sequences only 83 out of 120 3P-combinations could be found (69%). Additionally, 16 different 4P-di-consensus heptads were mapped in this approach (16 out of 210, 7,6%). Interestingly, when performing the same analysis deleting position 14 (S_7 in second repeat) of the di-consensus heptad mapping phosphosites within the sequence YSPTSPSYSPTSP, higher percentages of different combinations of detected 3P- and 4P-levels were obtained (**3P:** 81 out of 84, 96%; **4P:** 40 out of 126, 31%) (Figure 20).
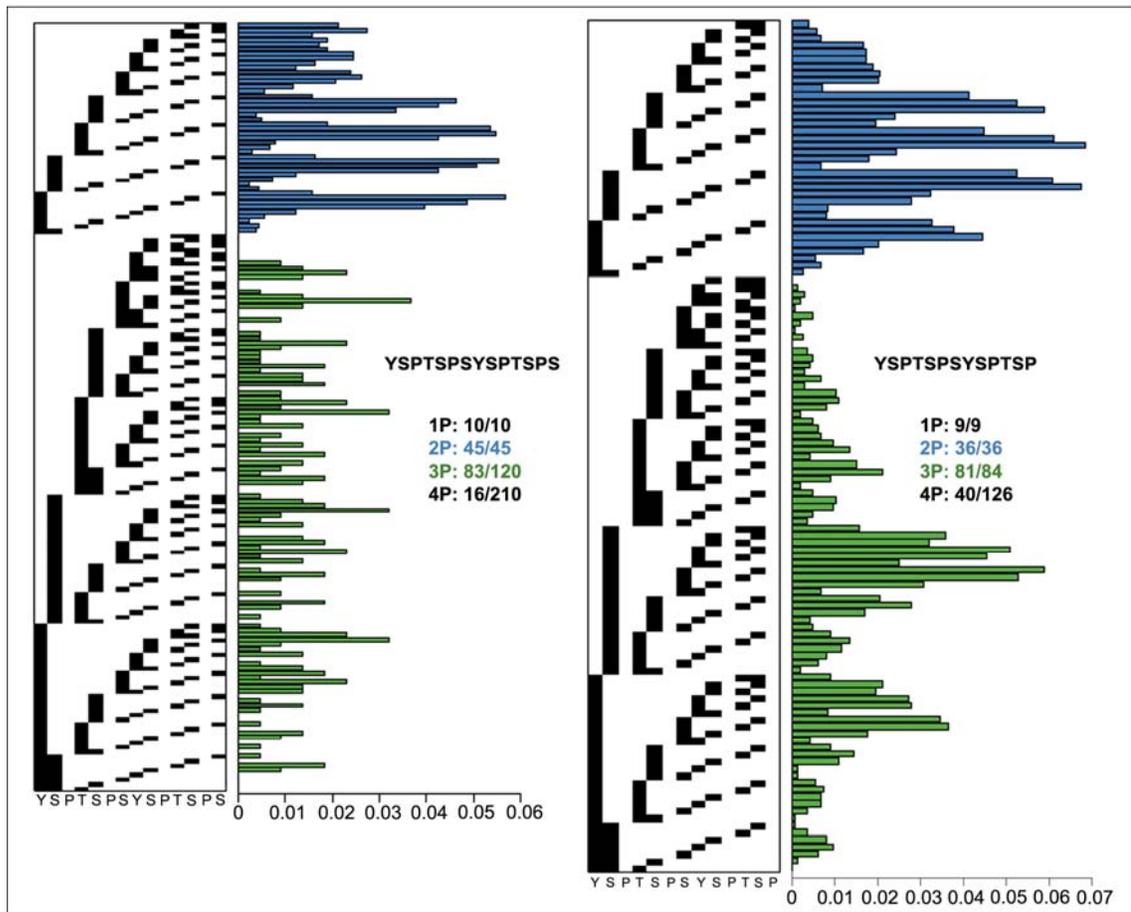
Figure 20 Phosphorylation patterns of the CTD consensus sequence containing 14 (left) and 13 residues (right). Left: Relative abundance (0-6%) among all detected 2P- (45 out of 45/blue) and 3P- signatures (83 out of 120/green) within CTD di-consensus heptad sequences. Right: Relative abundance (0-7%) among all detected 2P- (36 out of 36/blue) and 3P- signatures (81 out of 84/green) within CTD di-consensus heptad sequences missing residue 14. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites. Numbers of detected 1P- and 4P-combinations for both sequences are additionally indicated in a text box.

The reason for higher 3P-and 4P-levels in the consensus sequence, comprising 13 residues instead of 14 residues, might be explained by the fact that all CTD peptides with consensus sequences containing a lysine or arginine in position 14 were included in this second analysis.

Consequently, higher data input led to higher percentages of 3P-and 4P-combinations in the 13 consensus sequence stretch compared to the di-consensus heptads. In line with this, more dominant 3P-patterns could be observed within the YSPTSPSYSPTSP sequence. The three most abundant 3P-combinations included S_{2a}-P/S_{7a}-P/S_{5b}-P, S_{2a}-P/S_{7a}-P/T_{4b}-P and S_{2a}-P/Y_{1b}-P/S_{5b}-P (Figure 20).

2.3.6 Mapping phosphosites within the minimal functional unit of CTD

It has been shown that in yeast the minimal functional unit of CTD requires three consecutive serine residues in a row in the configuration 2-5-2, as well as paired tyrosines spaced 7 amino acids apart (Y_1-Y_8). These requirements lead to two possible sequences comprising the minimal functional unit of CTD, either **Y**SP**T**SP**S**Y**SP** or **SP**T**SP**S**Y**SP**TSPS**Y**. These two newly defined CTD consensus sequences were analysed in order to obtain data on highly abundant phospho-patterns within these sequences in mammalian cells (Liu et al., 2008 and 2010: Schwer et al., 2012; see also chapter 1.2.2 of Introduction) (Figure 21).

The three most abundant 2P-combinations found within the short functional unit of CTD (YSPTSPSYSP) were S_{5a}-P/S_{2b}-P, T_{4a}-P/S_{2b}-P and S_{2a}-P/S_{2b}-P. Within the long version of the minimal functional unit of CTD (SPTSPSYSPTSPSY) 5 predominant 2P-combinations could be observed with S_{5a}-P/S_{5b}-P, T_{4a}-P/S_{5b}-P, T_{4a}-P/T_{4b}-P, S_{2a}-P/S_{5b}-P and S_{2a}-P/T_{4b}-P. Interestingly, although the same three residues (S_2, T_4 and S_5) were found in all of these high abundant 2P-combinations there was no overlap of these combinations between the two differently defined functional units. Accordingly, no interference between the most abundant 3P-combinations of the short functional unit (S_{2a}-P/Y_{1b}-P/S_{2b}-P, S_{2a}-P/S_{7a}-P/S_{2b}-P, S_{2a}-P/S_{5a}-P/S_{2b}-P and Y_{1a}-P/S_{7a}-P/S_{2b}-P) and the long functional unit (S_{7a}-P/S_{2b}-P/S_{5b}-P, S_{5a}-P/S_{2b}-P/Y_{1c}-P and T_{4a}-P/S_{2b}-P/Y_{1c}-P) could also be found (Figure 21).
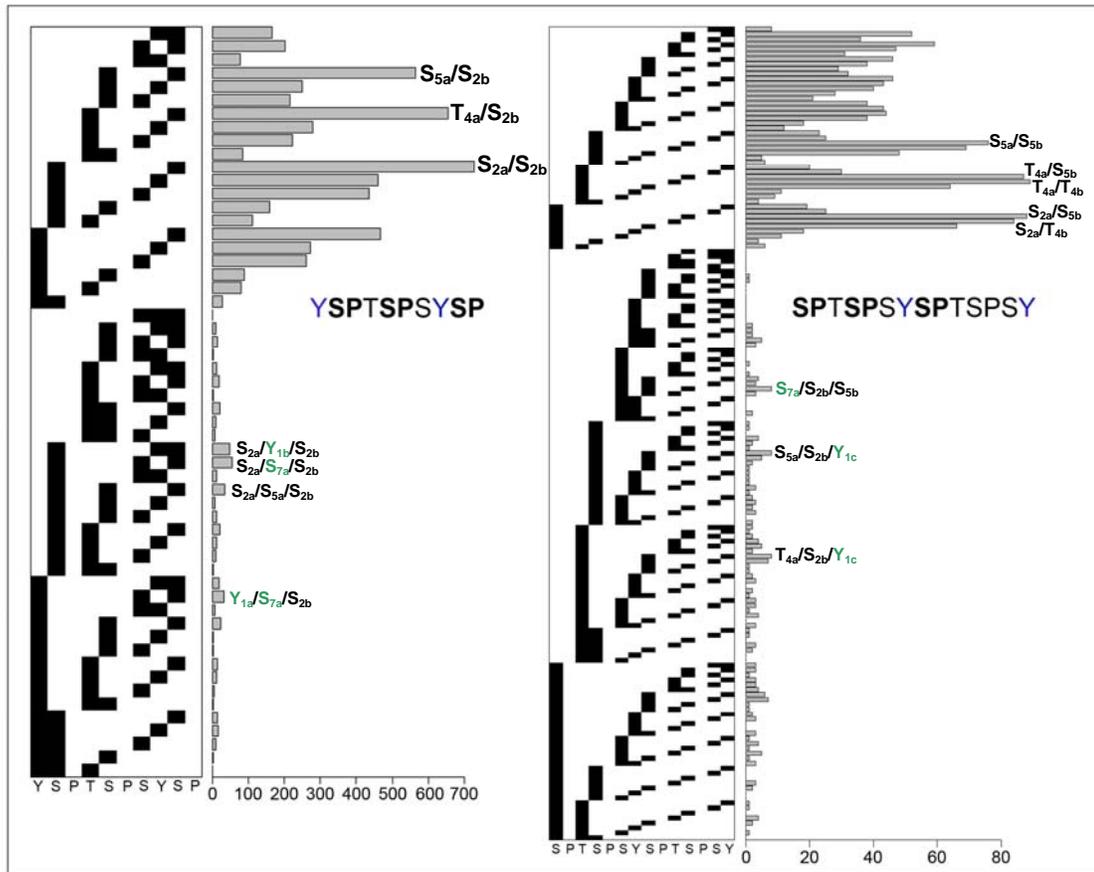
Figure 21   2P- and 3P- combinations of defined functional units within CTD.
Left: Total counts (0-700) of all 2P- and 3P- combinations within the defined sequence YSPTSPSYSP. Right: Total counts (0-80) of all 2P- and 3P- combinations within the defined sequence SPTSPSYSPTSPSY. Most abundant phospho-signatures within 2P- and 3P-levels are indicated. Green letters within 3P-combinations mark phospho-residues absent in the dominant 2P-combinations. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites.

However, most of the 3P-combinations contained two CTD residues, Y_1 and S_7, which were absent in all dominant 2P-combinations within the two functional units. Interestingly, in the short functional unit, two out of three of the most dominant 2P-combinations could   be again detected within some of the corresponding most abundant 3P-combinations (**S_{2a}-P/S_{2b}-P** integrated in **S_{2a}-P**/Y_{1b}-P/**S_{2b}-P**, **S_{2a}-P**/S_{7a}-P/**S_{2b}-P**, **S_{2a}-P**/S_{5a}-P/**S_{2b}-P** and **S_{5a}-P/S_{2b}-P** integrated in S_{2a}-P/**S_{5a}-P/S_{2b}-P),** whereas in the long functional unit, none of the most frequent 2P-combinations were recovered in any of the related most abundant 3P-combinations (Figure 21).

2.3.7 Dominant phosphorylation signatures in non-consensus repeats within the distal part of CTD.

Since the distal part of the CTD was less mutated compared to the proximal part containing most of the consensus sequences, many WT-repeats can be found in the last 26 repeats of all 9 CTD mutants. These non-consensus repeats consisting of WT-sequences were scanned for dominant phosphorylation signatures within 2P- and 3P-combinations.
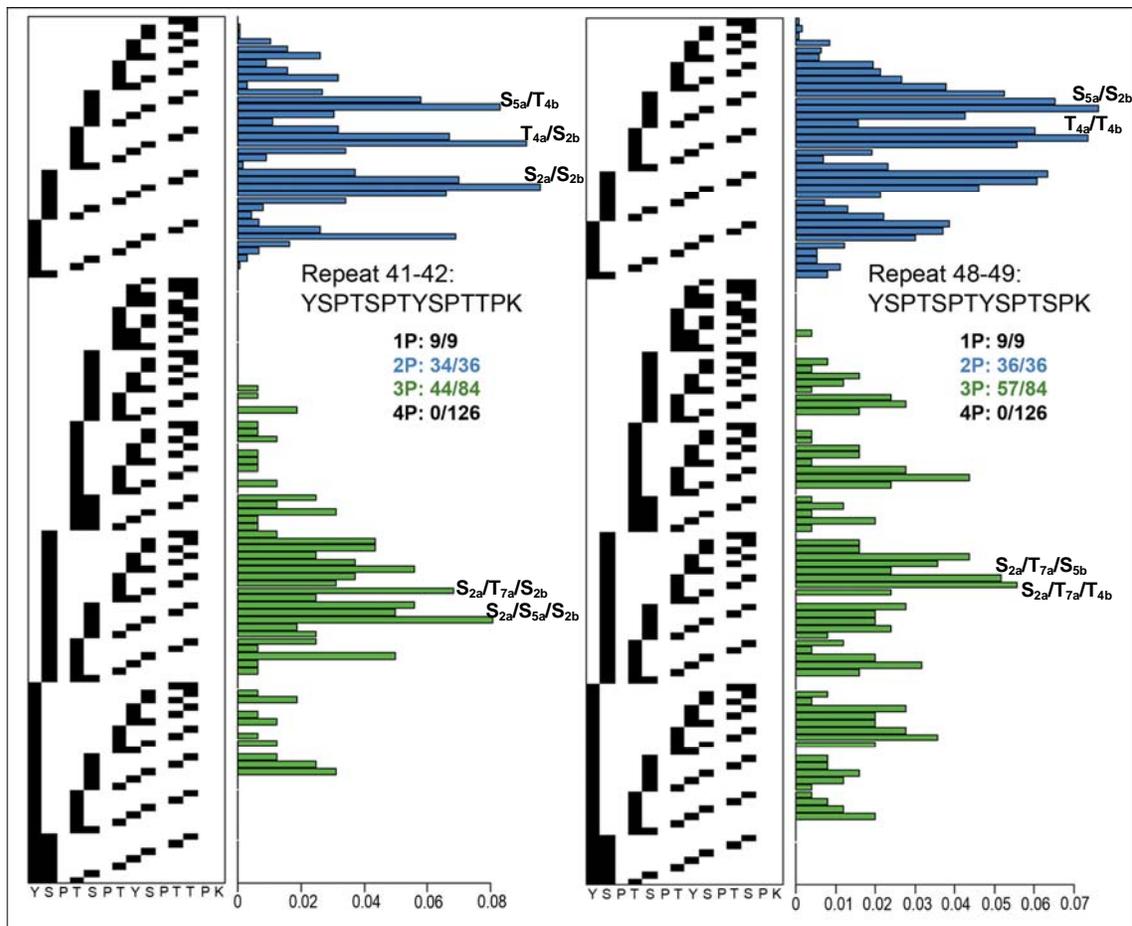
Figure 22 2P- and 3P- combinations of non-consensus repeats 41-42 and 48-49. Left: Relative abundance (0-8%) among all detected 2P- (34 out of 36/blue) and 3P-signatures (44 out of 84/green) of repeat 41-42 (YSPTSPTYSPTTPK) is shown. Right: Relative abundance (0-7%) among all detected 2P- (36 out of 36/blue) and 3P- signatures (57 out of 84/green) of repeat 48-49 (YSPTSPTYSPTSPK) is shown. Most abundant phospho-signatures within 2P- and 3P-levels are indicated. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites. Portions of detected 1P-, 2P-, 3P- and 4P-combinations are shown as text.

In Figure 22 two different di-heptad repeats with defined location within the CTD (repeat 41-42 and repeat 48-49) are shown displaying all mapped 2P- and 3P-combinations found within these sequences.

In repeat 41-42 (YSPTSPTYSPTTPK) the three most abundant 2P-combinations were S_{5a}-P/T_{4b}-P, T_{4a}-P/S_{2b}-P and S_{2a}-P/S_{2b}-P whereas two predominant signatures regarding 2P-levels could be detected in repeat 48-49 (YSPSPTYSPTSPK) including S_{5a}-P/S_{2b}-P and T_{4a}-P/T_{4b}-P. Although these two sequences only differ in one position, residue number 12, and their most frequent 2P-combinations shared the same residues including S_2, T_4 and S_5, no overlap among the most dominant 2P-levels of repeat 41-42 and repeat 48-49 could be found. The two most abundant 3P-signatures of repeat 41-42 and repeat 48-49 were S_{2a}-P/T_{7a}-P/S_{2b}-P, S_{2a}-P/S_{5a}-P/S_{2b}-P and S_{2a}-P/T_{7a}-P/S_{5b}-P, S_{2a}-P/T_{7a}-P/T_{4b}-P, respectively. When comparing the most abundant 2P-and 3P-signatures, only one phospho-combination could be detected within both forms (S_{2a}-P/S_{2b}-P integrated in **S_{2a}-P**/T_{7a}-P/**S_{2b}-P** and **S_{2a}-P**/S_{5a}-P/**S_{2b}-P** of repeat 41-42). Although no crossover between the most frequent 3P-signatures of repeat 41-42 and repeat 48-49 could be observed as well, the non-consensus residue T_7 (instead of S_7) was found in three out of the 4 most abundant 3P-combinations (S_{2a}-P/**T_{7a}-P**/S_{2b}-P in repeat 41-42; S_{2a}-P/**T_{7a}-P**/S_{5b}-P and S_{2a}-P/**T_{7a}-P/**T_{4b}-P in repeat 48-49).

2.3.8 Phosphorylation patterns within the CTD are location dependent

In a further approach, comparison of phosphorylation patterns of an identical sequence located within different repeats along the CTD was addressed. In more detail, the consensus sequence YSPTSPS was analysed regarding detectable 2P-and 3P-combinations facing different repeat numbers. In Figure 23, consensus heptad repeats 11 and 13 show very similar phosphorylation patterns. Both repeats are predominantly phosphorylated at S_2/S_7 and Y_1/S_7 concerning 2P-levels. However, repeat 9 and repeat 25 of the same sequence showed different dominant 2P-signatures, indicating that phosphorylation of CTD is location dependent.
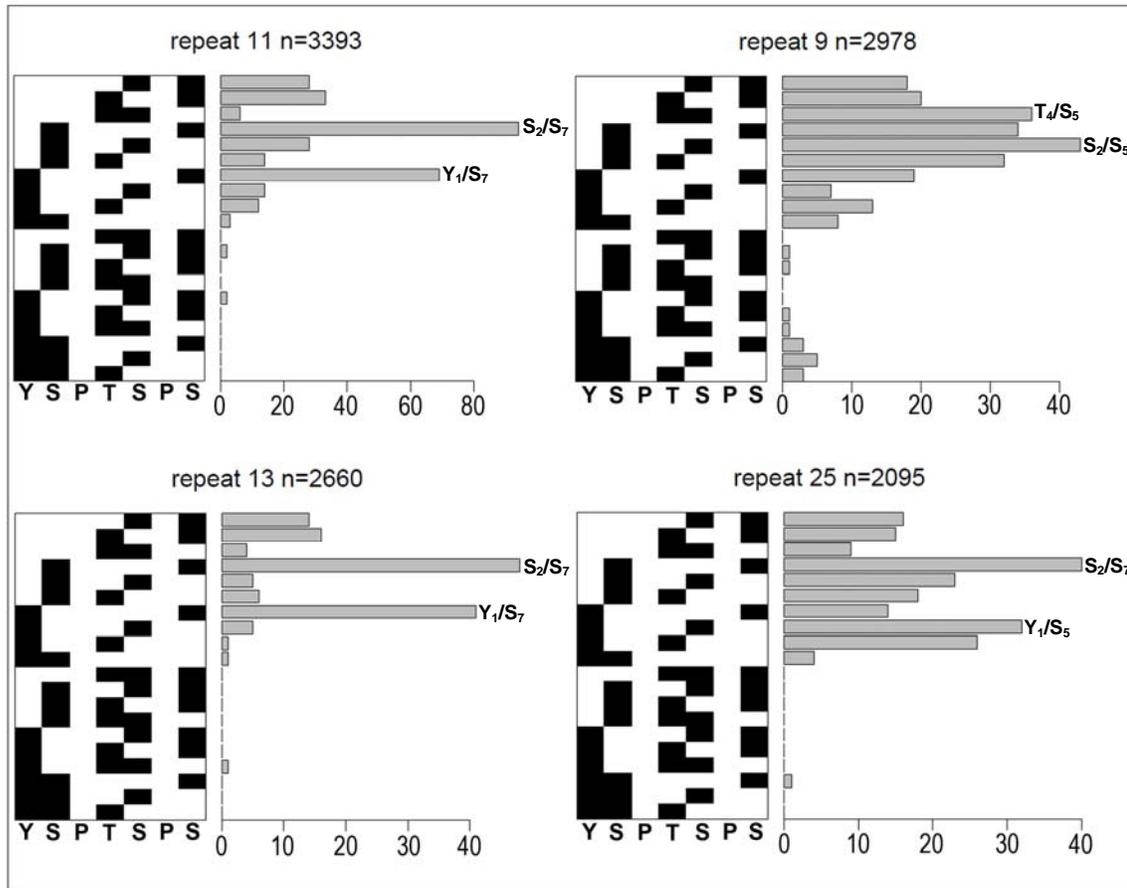
Figure 23   Total counts of 2P- and 3P- combinations of consensus heptad repeats 9, 11, 13 and 25. Repeat number and corresponding total counts (n) are shown on top of each graph. The two most abundant 2P-signatures of each repeat are indicated. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites. All found 2P-and 3P-combinations are counted.
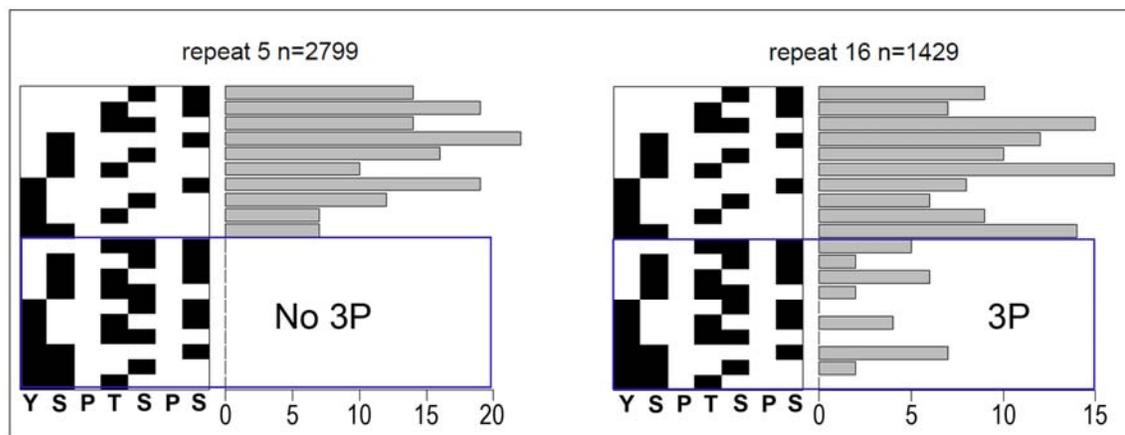
Figure 24   Total counts of 2P- and 3P- combinations of consensus heptad repeats 5 and 16. Repeat number and corresponding total counts (n) are shown on top of each graph. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites. All found 2P-and 3P-combinations are counted. Blue square marks 3P-combination part of the graph of both repeats.

Although repeat 25 revealed the same most abundant 2P-combination (S_2-P/S_7-P) compared to repeat 11 and 13 other high frequent 2P-signatures like Y_1-P/S_5-P could be observed in this repeat that was only detectable at very low abundance within the other three repeats. Moreover, in repeat 9, other most abundant 2P-signatures were mapped comprising T_4-P/S_5-P and S_2-P/S_5-P, which seem to play only a minor role in the repeats 11, 13 and 25.

Comparing consensus repeat 5 and 16 reflected another example of location dependency of phosphorylation patterns within the CTD (Figure 24). While the frequency of detected 2P-signatures was rather similar among these two repeats, big differences could be observed at the 3P-levels between these two locations. No 3P-combinations could be found in repeat 5, whereas several 3P-signatures were mapped in repeat 16.
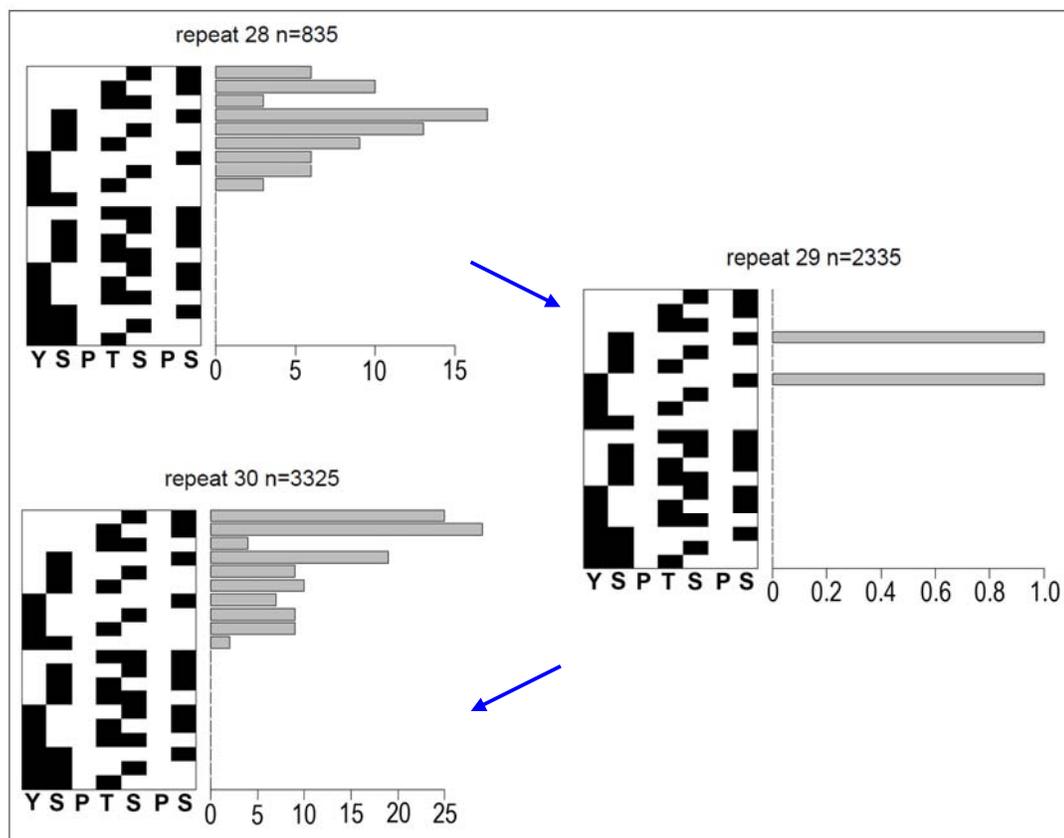
Figure 25 Total counts of 2P- and 3P- combinations of neighbouring consensus heptad repeats 28, 29 and 30. Repeat number and corresponding total counts (n) are shown on top of each graph. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites. All found 2P- and 3P-combinations are counted. Blue arrows indicate the right order of repeat 28, 29 and 30.

Additionally, differences in phosphorylation patterns could also be found within neighbouring consensus heptad repeats within the CTD. By comparing adjacent repeats 28, 29 and 30, it became apparent that repeat 28 and 30 shared a very common pattern at the 2P-levels, comprising all possible 2P-combinations, whereas repeat 29, located in between the two, displayed a very unique pattern including only two 2P-combinations in total (Figure 25).

2.3.9 Scanning for known CTD-binding motifs of CTD-interacting proteins

Ghosh et al. revealed the crystal structure capturing the interaction of the human capping enzyme Mce1 with the CTD. It has been shown that both the N-terminal nucleotidyl transferase domains (NTD) of the two Mce1 protomers interact with a 6 amino acid long CTD segment that is connected via T_4 and phosphorylated at residues S_5 and S_2. The CTD binding motif for this interaction is, therefore, $\mathbf{S_{5a}}$-$\mathbf{P}P_{6a}S_{7a}Y_{1b}\mathbf{S_{2b}}$-$\mathbf{P}P_{3b}T_{4b}\mathbf{S_{5b}}$-$\mathbf{P}P_{6b}S_{7b}Y_{1c}\mathbf{S_{2c}}$-$\mathbf{P}P_{3c}$, where a, b, c correspond to the residues originating from the three consecutive heptad repeats (Ghosh et al., 2011). Our MS data confirmed the existence of the Mce1 binding motif *in vivo* (Figure 26). All in all, 30 out of 126 possible 4P-combinations could be detected within the defined sequence SPSYSPTSPSYSP among which the 4P-combination S_{7a}-P/S_{2b}-P/S_{5b}-P/S_{2c}-P showed the highest frequency (8%). Additionally, the Mce1 CTD-binding motif could be mapped to repeat 9 (repeat number refers to the central heptad repeat), whereas all other repeats containing the defined sequence lack the specific 4P–combination essential for binding of Mce1 to the CTD (Figure 26).

Interestingly, when comparing different repeats within the CTD comprising the Mce1-CTD binding sequence SPSYSPTSPSYSP a tendency towards lower frequent phosphorylated repeats within the distal part of the CTD could be observed. In more detail, several 4P-combinations could be mapped within the repeats 12 and 14, whereas repeats 25 and 30, located more distal within the CTD, did not carry any 4P-signatures. The comparative study of phosphorylation patterns within the Mce1-binding sequence of different repeats along the CTD revealed again that CTD phosphorylation signatures tend to be location dependent (Figure 27).
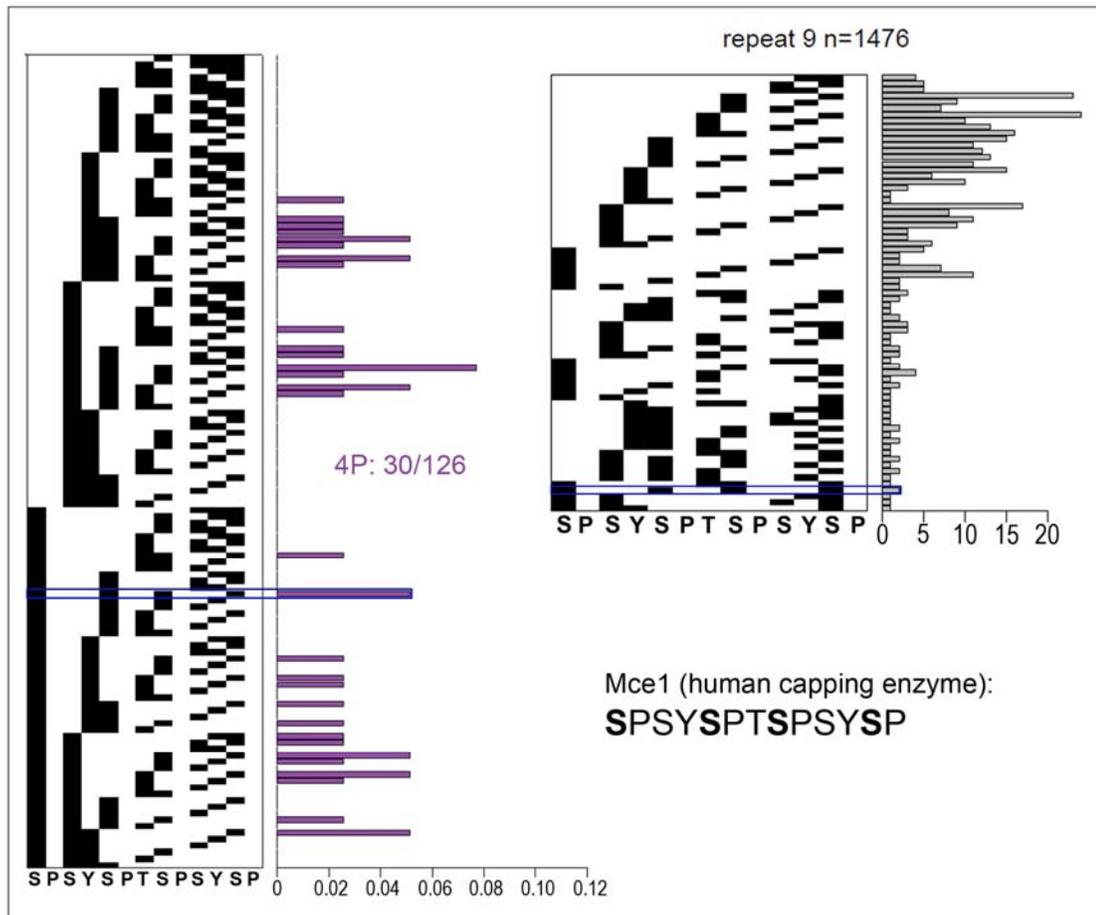
Figure 26   Mapping of the Mce1 CTD-binding motif within the CTD. <u>Left:</u> Relative abundance (0-12%) among all detected 4P-combinations (30 out of 126/purple) of the defined Mce1 sequence binding motif SPSYSPTSPSYSP is shown. <u>Right:</u> Total counts of all detected 2P-, 3P- and 4P-combinations of the defined Mce1 sequence binding motif SPSYSPTSPSYSP in repeat 9 (repeat number refers to central heptad repeat) are shown. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites. Blue squares in both graphs mark the detected specific 4P-CTD binding motif of human capping enzyme Mce1.

The most common CTD-binding domain is the CTD interacting domain (CID) and crystal structures of CTD peptides and the CID-domains of Rtt103, Pcf11 (both yeast), SCAF8 and SCP1 (both human) have been obtained (Lunde et al., 2010; Meinhart and Cramer, 2004; Becker et al., 2008; Zhang et al., 2006). The CTD-binding motif of the CID domain comprises 8 residues of the sequence PSYSPTSPS. Rtt103 and Pcf11 bind the S_5-P form, whereas SCAF8 is recruited to the S_2-P form of PSYSPTSPS. In contrast, SCP1 binds the double-phosphorylated PSY**S_2-P**PT**S_5-P**PS CTD sequence. The CTD binding motifs of all four proteins could be found in the MS analysis (Figure 28).
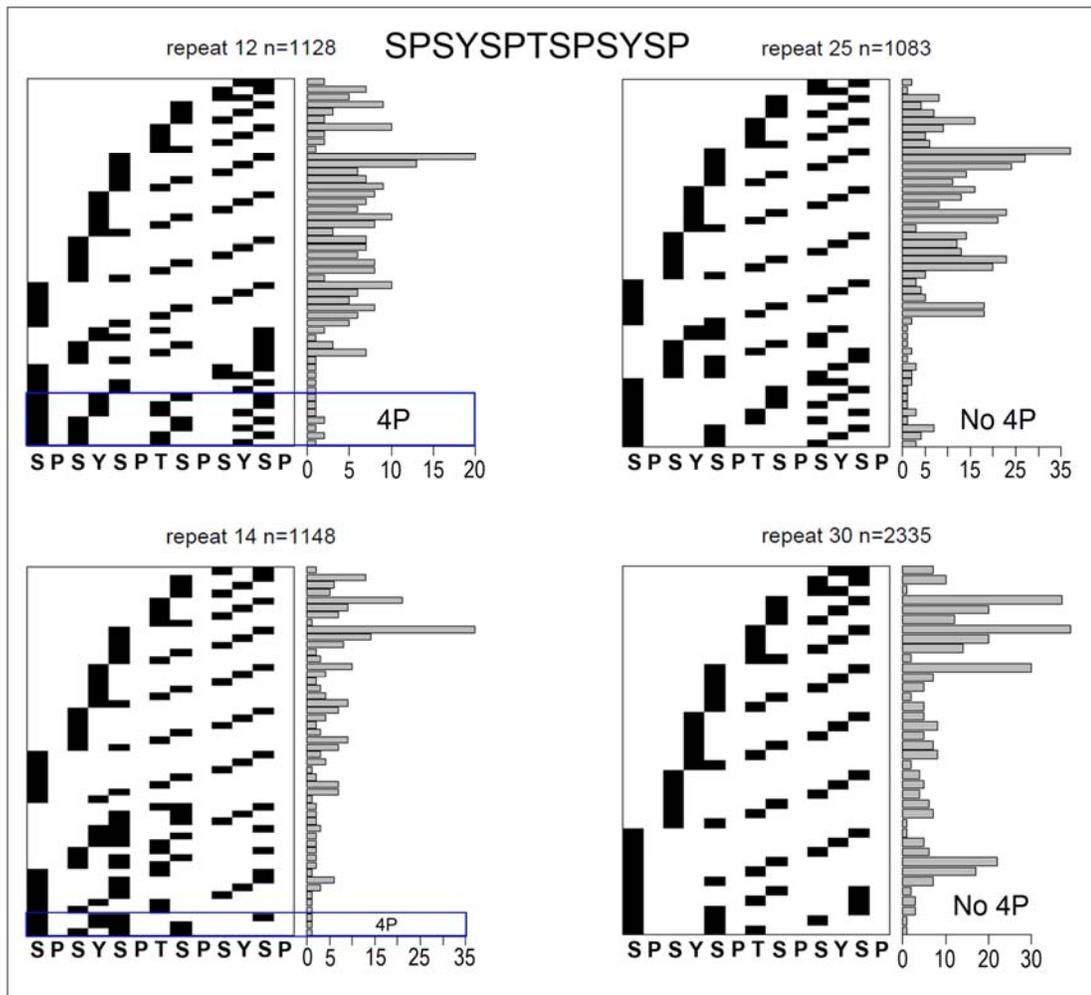
Figure 27 Total counts of 2P-, 3P- and 4P-combinations of CTD repeats 12, 14, 25 and 30 containing the Mce1-binding sequence. Repeat number (refers to central heptad repeat) and corresponding total counts (n) are shown on top of each graph. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites. All found 2P-, 3P- and 4P-combinations are counted. Blue square marks 4P-combination part of repeat 12 and 14.

In general, all possible 1P- and 2P-combinations were mapped within the CTD sequence PSYSPTSPS. Additionally, 11 out of 20 possible 3P-combinations could be detected as well. Furthermore, the S_2-P binding motif of Rtt103 and Pcf11, as well as the S_5-P binding motif of SCAF8 could be mapped in all repeats containing the defined sequence motif PSYSPTSPS. In contrast, the 2P-binding motif S_2-P/S_5-P of SCP1 was found in only 8 out of 15 possible CTD repeats indicating that CTD-SCP1 binding might be restricted to specific CTD repeats (examples are shown in Figure 28).
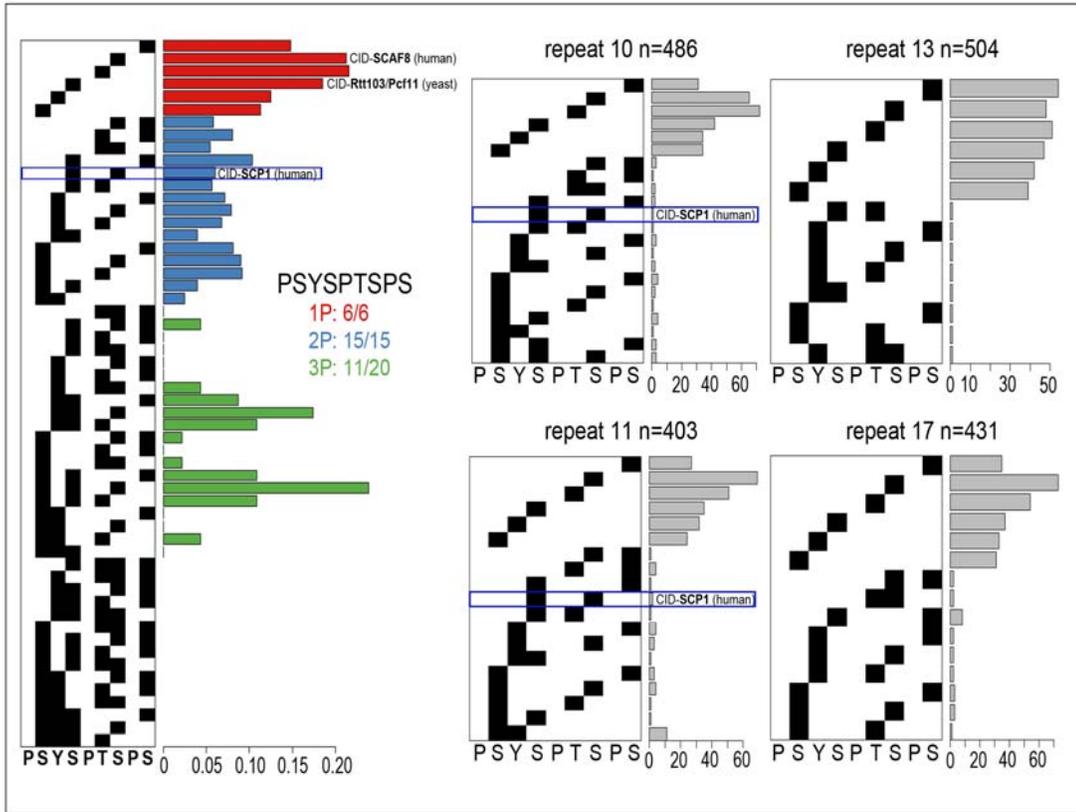
Figure 28 Mapping of the CID CTD binding motifs of Rtt103, Pcf11, SCAF8 and SCP1. Left: Relative abundance (0-20%) among all detected 1P- (6 out of 6/red), 2P- (15 out of 15/blue) and 3P- signatures (11 out of 20/green) of the defined CID sequence binding motif PSYSPTSPS is shown. Specific binding motifs of Rtt103, Pcf11, SCAF8 and SCP1 are indicated. Blue square marks SCP1 binding motif. Right: Repeat number (refers to the consensus heptad repeat) and corresponding total counts (n) are shown on top of each graph. Each row corresponds to one specific amino acid annotated at the bottom, whereas black boxes indicate phosphosites. All found 1P-, 2P- and 3P-combinations are counted. Blue squares mark SCP1-binding motif found in repeat 10 and 11.

3. Discussion

3.1 Phosphopeptide analysis by mass spectrometry (MS)

In order to obtain a successful phosphopeptide detection and phosphosite localization by MS one has to overcome several barriers. The primary goal is to detect the phosphopeptide of interest in a typically enormous pool of non-modified peptides. Generally, phosphopeptides are present at sub-stoichiometric amounts when compared to the unmodified peptides of the protein of interest. Additionally, the ionization efficiency of phosphopeptides in positive ion mode MS is lower than in unphosphorylated peptides due to the negative phosphate group (Boersema et al., 2009). Consequently, to overcome the underrepresentation of phosphorylated targets in a high background of non-phosphorylated species, several enrichment strategies have been established. Enrichment is based on two main features that distinguish phosphorylated from non-phosphorylated peptides: the negative charge of the phosphate group due it its low pK_a in combination with its steric structure. In this work, metal oxide affinity chromatography, in particular titanium dioxide (TiO_2), was used for phosphopeptide purification taking in advantage the strong affinity of negatively charged phosphate groups towards metal ions (Larsen et al., 2005). However, enrichment of phosphopeptides using TiO_2 can be influenced by the presence of acidic and basic residues and is known to preferentially purify low phosphorylated peptides. In contrast, immobilized metal affinity chromatography (IMAC) enriches multiphosphorylated peptides better than TiO_2 (Thingholm et al., 2009; Bodenmiller et al., 2007). Consequently, the combined and consecutive use of IMAC and TiO_2 should lead to the separation of mono-phosphorylated (TiO_2) from multi-phosphorylated peptides (IMAC), thereby decreasing sample complexity (Thingholm et al., 2008). In more detail, acidic buffers mainly elute mono-phosphorylated peptides from IMAC material, whereas subsequent basic elution recovers multiply phosphorylated peptides that are normally hard to detect. Thus, this two step purification procedure should enable separation of non-, mono- and multi-phosphorylated peptides in distinct fractions. However, in this work, a single purification step using only TiO_2 turned out to be the most efficient method in yielding both low and high phosphorylated CTD peptides. Nevertheless, mono-

phosphorylated peptides were the dominant detected form as shown by the MS data and synthetic peptides carrying different numbers of phosphate groups will be tested in the near future in order to exclude a technical bias towards low phosphorylated CTD peptides leading to falsified MS results. Additionally, other phosphopeptide purification methods will also be tested or optimized, like the capacity and binding efficiency of TiO_2 beads, e.g. by repeated incubation of the same sample with fresh TiO_2 beads have rarely been investigated yet. An interesting approach comes from Mamone and co-workers who have recently developed an efficient method for the separation of mono- and multi-phosphorylated peptides called hydroxyapatite (HAP) affinity chromatography (Mamone et al., 2010). A strong interaction of HAP with phosphate and calcium ions takes place and due to the higher affinity of multi-phosphorylated peptides to the HAP surfaces, stepwise elution with a phosphate buffer gradient allows the selective isolation of mono- and multi-phosphorylated peptides. Thus, optimization of the phosphopeptide purification step might lead to higher detectable numbers of mutliphosphoylated CTD-peptides in the near future.

In this approach, collision-induced dissociation (CID) technique was used for gas phase fragmentation of the peptides and the identification of phosphorylated peptides and subsequent phosphosite localization has been achieved by tandem MS (MS/MS) linked to MS^3 and multistage activation (MSA).

Trypsin generally produces double positive charged (2+) peptides (Steen et al., 2004). In CID, protonated peptides are accelerated by an electric potential in the vaccum of the MS and then forced to collide with an inert neutral gas (helium in this approach). Due to the collisions, the kinetic energy of the peptide ion is partially converted into internal energy that is spread over the molecule, breaking bonds and causing the peptide ion to fragments (Roepstorff et al., 1984; Biemann et al., 1988). Importantly, the phosphate group of a phosphopeptide is relatively labile providing a low-energy pathway like CID that competes with backbone fragmentation. Consequently, a CID spectrum of a phosphopeptide typically shows an intense neutral loss peak that locates at 98 Da or 80 Da lower than the precursor mass, reflecting the loss of H_3PO_4 and HPO_3, respectively (Boersema et al., 2009).

Generally, peptides are fragmented at the amid bonds along the backbone which results in the appearance of amino acid sequence-informative b- and y-ions with the charge retained at the N- or C-terminal end, respectively. Therefore, if a population of precursor peptide ions dissociate to produce a series of consecutive b- and y-ions,

the amino acid sequence of the peptide can be determined from the difference between these consecutive ions, including the mapping of the phosphorylation site. Since the residue masses of phosphorylated serine (167 Da), phosphorylated threonine (181 Da), and phosphorylated tyrosine (243 Da) are unique, the sequence ions of phosphorylated peptides are directly indicative of the existence of phosphopeptides and the location of the phosphorylation sites within the peptide (Palumbo et al., 2010). Figure 29 shows how phosphosites are mapped using b- and y-ions. For phosphorylation site localization, fragmentation spectra of the non-phosphorylated peptide and its phosphorylated analogue are compared (compare Figure 29A and 29B). Fragment ions that contain the same masses in both spectra indicate that these peptide fragments are not phosphorylated, thereby excluding Y_3, S_4, T_6 and S_7 as phosphosites (b_3 to b_9 in Figure 29A and 29B). Analysing the C-terminal fragment ions, y_3-y_6 have masses corresponding to the fragment ions that are unmodified implying that these peptide fragments are not phosphorylated as well.

Consequently, the determination of the backbone fragment ion series leaves T_{10} as the only option for containing the detected phosphorylation group within the phosphorylated peptide. Additionally, the site-determining ions b_{10} and y_7 show shifted mass peaks compared to the corresponding fragment ions in the unmodified peptide either due to the carriage of the phosphate group (+ 80 Da) or due to the loss of the phosphate group (- 18 Da) confirming the phosphorylation site to be T_{10}. Figure 29C gives an example where no clear phosphosite determination can be achieved. In more detail, fragment ions y_6-y_9 reflect unphosphorylated peptide fragments restricting the phosphorylation site to S_6, S_7 or S_8. With the additional dissection of the ion b-series b_2-b_6 residues Y_1, S_3, S_4 and S_6 can be excluded as potential phosphosites. However two residues, S_7 and S_8 remain as potential carriers of the phosphate group, since no complete b-and y-ion series are obtained and no site-determining ions with an intact phosphate group for direct validation are found for this phosphopeptide (Figure 29C).
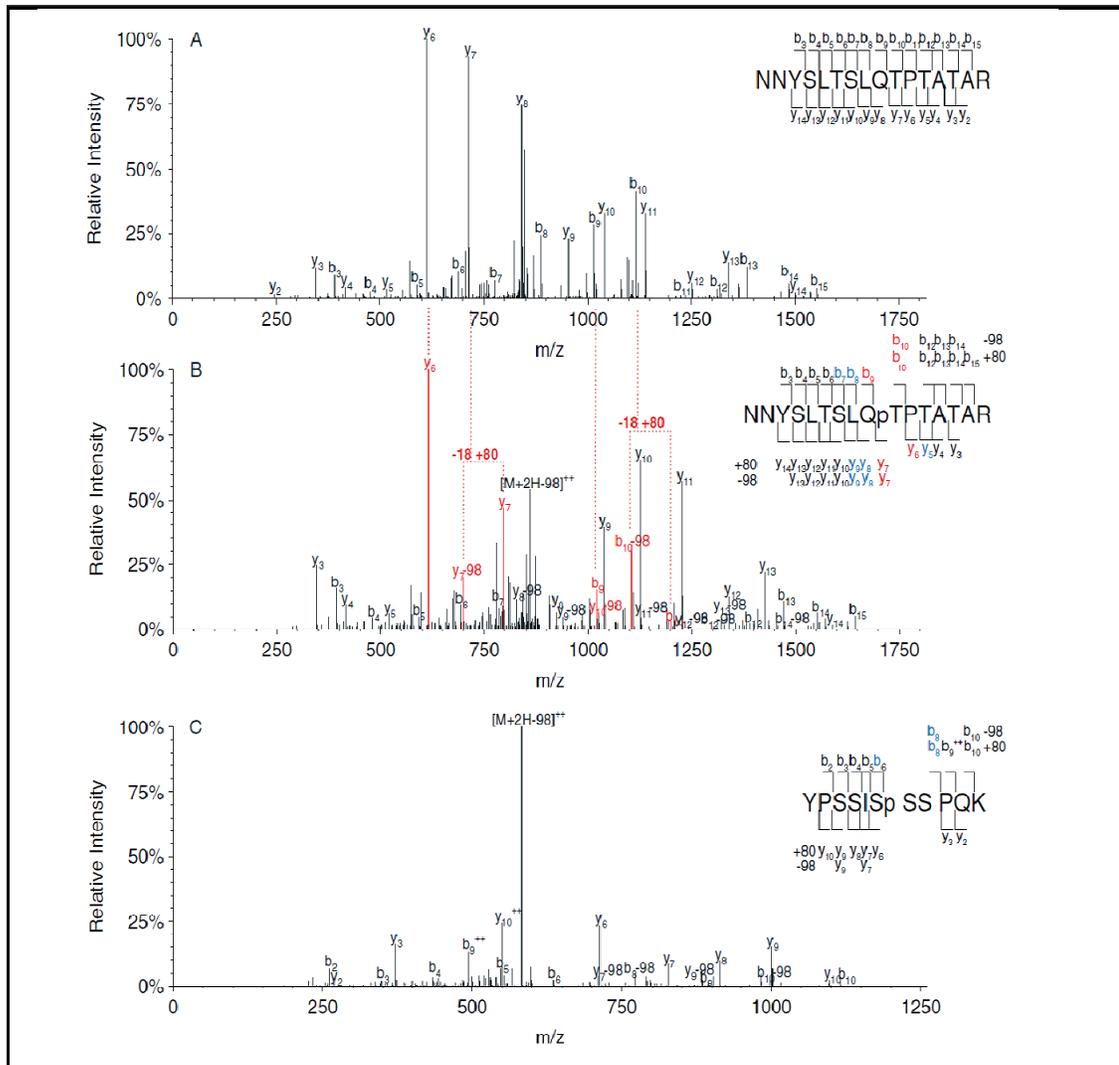
Figure 29 Phosphorylation site localization. A) CID spectrum of a non-phosphorylated peptide compared to the CID spectrum of its phosphorylated counterpart. B) Indicated on the peptide sequence are the fragment ions that were found, including ions that lost 98 Da or were 80 Da heavier than in the non-phosphorylated peptide. Highlighted in red are the site-determining ions and the corresponding peaks in the spectrum. In blue are indicated fragment ions that confirm the site localization. C) A CID spectrum of a phosphopeptide for which a precise phosphorylation site could not unambiguously be determined. Highlighted in blue are fragment ions that indicate that the phosphorylation is on either S_7 or S_8 (copied from *Boersema et al., (2009), Phosphopeptide fragmentation and analysis by mass spectrometry; J. Mass. Spectrom. 2009, 44, 861-878*).

In this work, to minimize incorrect phosphosite mapping within CTD-peptides, at least six replicates for each mutant were analysed leading to large MS data outcome. Moreover, to gain more complete b- and y-ion series of corresponding CTD-peptides, MS^3 and multistage activation were performed. The initial CID-MS/MS 98 Da neutral loss product ion was automatically subjected to either multistage CID-MS/MS (i.e. MS^3) or 'pseudo-MS^3' in ion trap mass spectrometers in

a data dependent mode of operation. Accordingly, CID-MS^3 involved isolation and further fragmentation of the neutral loss species, whereas pseudo MS^3 provided simultaneous activation of the precursor ion and the resultant 98 Da neutral loss product ion during a single CID-MS/MS event. In more detail, the stepwise MS procedure in this work included that the three most intense peptide ions with charge state higher than 1 were sequentially isolated to a target value of 10,000, fragmented in the linear ion trap by collision induced dissociation (CID). The pseudoMS^3 (pdMS^3) or multistage activation (MSA) was selected to automatically select and further fragment the fragment ion originating from the loss of one or two phosphate groups from the parent ion (see also Material and Methods).

3.2 Different Pol II forms

The human genome comprises approximately of 25000 genes which are transcribed by only a small fraction out of about 300000 Pol II molecules found in one single cell (Kimura et al., 1999). Theoretically, every single active or inactive Pol II molecule could feature a diverse phosphorylation profile along its CTD. However, Pol II accumulates in only two main forms as shown in SDS-PAGE electrophoresis, suggesting that the majority of Pol II within a cell is divided into only a few sub-populations carrying similar modifications patterns. In this work, however, many different phosphorylation patterns have been identified within every single CTD peptide, implying a great diversity of possible phosphorylation signatures along every single CTD of Pol II *in vivo*. The two main Pol II forms are Pol IIA form containing a hypophosphorylated CTD and Pol II0 form consisting of a hyperphosphorylated CTD, respectively. It has been suggested that the mammalian Pol II0 form carries on average one phosphate per repeat (Payne and Dahmus, 1993) although the number of phosphosites of the CTD at a certain stage during the transcription cycle remains elusive.

Pol IIA preferentially associates with the preinitiation complex at the promoter site and, therefore, any phosphorylation of the CTD before this site would potentially block the recruitment of Pol II and subsequent transcription initiation. It has been proposed that the hyperphosphorylation of CTD triggers the massive shift from IIA to

IIO (Palancade and Bensaude, 2003). However, the appearance of only a few Pol II subpopulations in the SDS-PAGE suggests that CTD-phosphorylation rather contributes to the shift of IIA to IIO in an indirect way by influencing the cis-/trans-isomerization of P_3 and P_6 within each single heptad repeat. Accordingly, different phosphorylation patterns might either stabilise a certain conformation of the proline residues or serve as binding motifs for corresponding proline isomerases.

In this work both IIO and IIA forms were analysed via MS and surprisingly many phosphorylated CTD peptides were found in the hypophosphorylated IIA form. Moreover frequently, highly phosphorylated CTD repeats carrying three or even four phosphosites at the same time were detected within the IIA form, as well. However, our MS data cannot decipher if the overall phosphorylation frequency along the whole CTD is different between IIO and IIA forms, as only small CTD fragments, mainly 14 or 21 amino acids long, originating from the same CTD could be analysed. Since many phosphorylated CTD peptides were found in both forms all data were collected and implemented in the subsequent bioinformatic analysis as one big data set. However, it will be interesting to investigate the phosphorylation signatures obtained from both forms in more detail in order to find out if possible variations in phosphorylation patterns between these two forms might explain the separation into these two main Pol II forms in the SDS-PAGE. Nevertheless, one reason for detecting similar highly phosphorylated CTD peptides in both forms could be that no complete separation of both forms had taken place in the SDS-PAGE. Instead, both forms might accumulate in both bands due to the high amount of material used in this approach. Therefore, next to the comparison of distinct phosphorylation signatures among both forms the total amount of detected peptides within the different phosphorylation levels (1P, 2P, 3P and 4P) of both forms should be analysed as well.

Interestingly, in the SDS-PAGE of the α-HA-IP a third Pol II form arose, IIO low, which was located between IIO and IIA (see also Figure 15 in results part). Since the α-HA recognizes Rpb1 outside of the CTD this antibody had no preference towards IIO or IIA during the purification step. This enrichment is most probably attributed to the fact that a greater amount of the IIA form was purified with this antibody, as the IIA form consists the inactive Pol II pool that was over-expressed in the CTD mutants. Both, IIO and IIO low gave weaker signals than the IIO form purified with the α-S_2/S_5 (see also Figure 15 in results part). Less amount of purified material of

the hyperphosporylated Pol II form might explain the further separation into two subpopulations both carrying a hyperphosphorylated CTD. Another reason of the occurrence of two different IIO forms could be that the CTD was more accessible to CTD-phosphatases as the α-HA binds outside the CTD. Consequently, the IIO low form might carry a slightly lower phosphorylated CTD form due to dephosphorylation by CTD phosphatases. Data obtained from the IIO low form revealed that the same phosphorylation frequencies within CTD peptides could be found in this form.

A more detailed analysis of phosphorylation signatures and overall phosphorylation counts of the IIO low form might uncover differences in phosphorylation patterns compared to the other two Pol II forms obtained in this work.

3.3 The most distal CTD repeat (repeat 52) exhibits unique features

In this work, MS analysis of recombinant CTDs that were mutated, in order to obtain information of phosphosites along the whole CTD sequence, revealed the existence of the full repertoire of possible phosphosites in the CTD *in vivo*. In more detail, MS data obtained from three different mutants that displayed 100% sequence coverage impressively showed that, indeed, all possible phosphosites were found to be phosphorylated *in vivo*. In other words, next to all potential phosphosites within the consensus repeats, S and T residues found in different positions within non-consensus repeats were all phosphorylated as well. Another interesting observation was that all repeats were also detected in their unphosphorylated state, with the exception of repeat 52, the most distal to its core subunit (Rpb1). This result suggests with the possible exception of repeat 52, that no constant phosphorylation state is maintained within distinct CTD repeats and highlights the importance of phosphorylation/de-phosphorylation cycles for the Pol II progression through transcription.

Comparative studies of different organisms revealed the development of divergent sequences following the most distal CTD repeat (Allison et al., 1988). The mammalian most distal CTD repeat consists of a total of 17 amino acids, containing two potential casein kinase II (CKII) sites, and is essential for mediating the binding

and phosphorylation of the CTD by the Abl1/c-Abl and Abl2/Arg tyrosine kinases (Pinna et al., 1990; Baskaran et al., 1997 and 1999). The consensus sequence for CKII phosphorylation is (S/T)XX(D/E) and found once in the non-consensus repeat 37 and twice in the most distal repeat of the Pol II CTD (Pinna et al., 1990; Chapman et al., 2004). Accordingly, it has been shown that CKII phosphorylates CTD (Dahmus et al., 1981), but not a stretch of consensus repeats (Bregman et al., 2000). Interestingly, stoichiometric analysis suggested that only one of these sites is being phosphorylated *in vivo*, most likely the outermost C-terminal serine (serine 13) of the CTD most distal repeat, as it was concluded due to the preference of CKII for sites surrounded by acidic residues (Kuenzel et al., 1987; Payne et al., 1989). Bensaude and co-workers established polyclonal antibodies that specifically recognized the CKII-phosphorylated or non-phosphorylated most distal CTD repeat 52, respectively. The CKII-phospho-specific antibody positively reacted with all forms of the largest subunit of Pol II under all conditions, whereas no reactivity was observed with the non-phospho specific antibody (Chapman et al., 2004). These results suggested that the most distal CTD repeat 52 is permanently phosphorylated by CKII *in vivo* and that CKII phosphorylation of this repeat takes place soon after translation of Rpb1, since a non-CKII-phosphorylated form could not be found *in vivo* (Chapman et al., 2004). Mass spec data of CTD mutants produced in this thesis, confirmed the detection of a permanent phosphorylated most distal repeat of the CTD of Pol II *in vivo*. However, whereas previous mutation studies suggested that the serine in position 13 and not serine in position 9 within the last repeat is constitutively phosphorylated (Chapman et al., 2004), the new obtained mass spec data showed both residues to be phosphorylated *in vivo*. Interestingly either serine 9 or serine 13 were found to be phosphorylated at a time and no double phosphorylated form of repeat 52 could be observed via MS analysis.

3.4 Phosphorylation frequencies within CTD-peptides

To date, the profound question on whether individual CTD repeats are independently phosphorylated, remains yet unanswered. Regarding the similarity of sequence among most of the CTD repeats, it has been doubted that kinases and phosphatases are capable of distinguishing these repeats sufficiently in order to modify adjacent heptad repeats individually. However, obtained mass spec data generated during this work demonstrated that neighbouring heptad repeats originating from the same CTD exhibit different phosphorylation patterns, while unphosphorylated and phosphorylated repeats have been detected as well.

When analysing phosphorylation frequencies of different peptides of CTD fragments it turned out that the coexistence of four phosphosites displayed the upper detection limit in the MS analysis. In contrast, in MS analysis of yeast CTD mutants only mono- and di-phosphorylated CTD peptides have been detected so far (data not shown). The yeast data suggested that the complexity of a so-called CTD code in yeast might be simpler, containing less phosphorylation combinations among different CTD residues and lower phosphorylation frequencies along the overall length of CTD compared to a CTD code existing in higher eukaryotic organisms like mammals.

Interestingly, the MS data presented here show that the same highest phosphorylation frequency could be found in mono-, di- and tri-heptads, where no more than four phosphosites were scored at the same time, suggesting that there could be a limit to the extent of phosphorylation within a certain sequence length. One reason for that could be that too many phosphosites located in close proximity to each other would lead to unfavoured structural changes due to strong charge repulsions between the negative charged phosphate groups. Moreover, it would be expected that longer CTD peptides, e.g. tetra- or penta-CTD heptads, would be found to carry more than four phosphosites explained by the longer distances between the individual phosphorylation residues. However, no highly phosphorylated tetra- or penta-CTD peptides could be detected, although there is an apparent weak data outcome of longer peptides in the MS analysis. Accordingly, CTD peptides containing two or three heptad repeats provided by far the highest amount of obtained MS information.

3.5 Establishment of CTD mutants

Two different sets of CTD mutants were established during this work. In the first round, four different CTD mutants were designed (M-6K; M-3K5R; M-8K; M-4K2R; see also Appendix) and the obtained fragmented CTD peptides after protein digestion ranged from mono- to hexa-peptides. In designing the CTD mutants it was important to find a balance on the one hand, not making too many mutations within the CTD sequence, in order to avoid risking cell viability, while, on the other hand, obtaining CTD peptides after trypsin digestion that did not exceed a certain length which could lead to weak MS output. Accordingly, CTD mutant M-8K, for instance, carried most mutations (13 mutations) leading to the highest number of fragmented peptides (18) ranging from mono- to penta-heptads. However, since it turned out that only di- and tri- heptads led to high MS data output, mapping of phosphosites did not cover the whole CTD length of the CTD mutants established by this first approach. Consequently, in a second round, five new mutants with even more mutations were designed, in order to retrieve shorter CTD peptides along the whole CTD sequence after protein digestion. Three out of five CTD mutants of round 2 were fragmented in only di-and tri-heptads after trypsinization and high numbers of phosphorylation sites could be mapped along the whole CTD. Importantly, although these mutants were frequently mutated within their CTD sequences, they all displayed full viability under normal cell culture conditions.

3.6 Minimal functional unit of the CTD

One reason for the high plasticity of the mammalian CTD of Pol II could originate from the genetic studies performed by Stiller and Greenleaf regarding the minimal functional unit of CTD in yeast. Results from yeast CTD mutants revealed that the minimal functional unit of CTD requires three consecutive serine residues in a 2-5-2 configuration, as well as paired tyrosines spaced 7 amino acids apart (Y_1-Y_8). Accordingly, the functional unit in yeast is embedded within di-heptads. Based on these findings in yeast, mutations that were introduced in the CTD mutants used in this work are not interfering with the minimal functional unit of the CTD. More specifically, the minimal distances between single mutations lay within di-heptads

along the whole CTD sequence of all newly established CTD mutants. In order to confirm that the same requirements of a minimal functional unit within the CTD is needed in both yeast and mammals, our lab is working towards this end. The project aims to establish two CTD mutants that display the minimal functional unit by either deleting residues 12-14 or by replacing residues 12-14 with alanine in every second CTD repeat in accordance to the genetic studies previously performed in yeast.

When comparing dominant phosphorylation patterns within the defined functional units of the CTD, one interesting observation was that newly identified phosphorylated residues, like Y_1-P and S_7-P, arose after the transition from highly abundant di-phosphorylated signatures to highly tri-phosphorylated signatures. Therefore, some phosphosites might be redundant in a low phospho-frequent background, but turn into dominant phosphosites when the overall phosphorylation frequency along a certain length of the CTD has increased. Accordingly, phosphorylation of distinct residues within the CTD might be triggered or enhanced by surrounding phosphosites of adjacent CTD residues. In the short functional unit (YSPTSPSYSP) the most dominant 2P-combination was also observed in 3 out of the 4 most abundant 3P-combinations suggesting an additive 'switching' mechanism between the different phosphorylation states within the CTD. However, in the longer version of the defined functional unit comprising of 14 residues, none of the five most abundant 2P-combinations could be found within the three most frequent 3P-signatures. These findings suggest that next to an additive pathway a two direction 'switching' mechanism including a dephosphorylation step followed by phosphorylation of new CTD residues might exist as well. Certainly, more data are required to decipher distinct mechanisms for switching between different phosphorylation states within defined CTD sequences, since the total count numbers of higher phosphorylated CTD peptides were much lower compared to the numbers obtained for mono- and di-phosphorylated CTD repeats.

3.7 Phosphorylation signatures within the consensus heptad repeat

Posttranslational modifications could have evolved in order to offer diversity to the otherwise evolutionary highly conserved CTD sequence/structure. The existence,

frequency and distinct combinations of phosphorylation sites across specific CTD lengths could offer valuable insights into unravelling possible different functional states of Pol II. In this MS analysis, all possible 1P-, 2P- and 3P-combinations within the consensus heptad sequence $Y_1S_2P_3T_4S_5P_6S_7$ could be found, suggesting the occurrence of highly dynamic phosphorylation and dephosphorylation processes of CTD residues *in vivo*. At the 1P-level all 5 possible phosphosites were mapped in comparably high amounts, providing the idea that all potential phosphosites play equally important roles in contributing to a 'so called' CTD code. Towards this end, it would be interesting to investigate if T_4-P and S_7-P residues could also be part of distinct binding motifs that serve in specific protein-CTD binding interactions. So far only CTD peptides containing S_2 and S_5 phosphosites have been mainly used in the assembly of crystal structures of specific CTD-protein interactions. At the 2P- and 3P-levels all possible phoshorylation combinations were found, implying that there are no phosphosites excluding each other and that the CTD uses its full repertoire of possible phosphorylation signatures within consensus repeats at different phosphorylation levels *in vivo*. Importantly, total count numbers of detected 1P-, 2P- and 3P-signatures showed that the 1P-CTD peptides were the dominant phosphorylation form within the CTD. The predominant existence of 1P-signatures could imply that mono-phosphorylated heptad repeats resemble the main scaffold of phosphorylated CTD and the addition of further phosphorylation groups, leading to 2P-or 3P-CTD repeats, are more likely transient stages in response to certain stimuli within the cell fulfilling transient tasks, like binding to specific factors at a certain stage during the transcription cycle.

3.8 Dominant phosphorylation signatures within CTD repeats

On big goal of this study was to identify dominant phosphorylation patterns within CTD repeats along the whole CTD domain. The proximal part consists mainly of consensus repeats and phosphorylation signatures were analysed within mono- and di-consensus repeats, as well as within defined functional units that originated from genetic studies of the CTD performed in yeast. Additionally, non-consensus WT-CTD

repeats located in the distal part of the CTD were also included in the mapping of highly abundant phosphorylation signatures. Accordingly highly abundant 2P-signatures and to a lesser extent dominant 3P-signatures could be found within the CTD. The obtained data of dominant CTD signatures within defined areas of the CTD could be used for subsequent CTD binding studies in order to identify highly specific novel protein-CTD interactions. In more detail, our lab is working on the establishment of a far western-approach, where biotin labeled synthetic CTD-peptides containing specific phosphorylation patterns will be used as bait proteins for 'fishing' interacting proteins originating from mammalian cell extracts. Subsequently, IP experiments will be performed followed by MS analysis to identify cellular factors that interact with the synthetic modified peptides. Consequently, non-consensus CTD peptides located in the distal part of the CTD will also be included in the binding assays. It has been speculated that proteins with core functions in the transcription cycle bind to consensus sequences within the proximal part of the CTD, whereas the distal part of the CTD, with its non-consensus repeats, most likely serves as a binding platform for protein factors activated upon cellular stress responses or might have tissue specific tasks. It would be of profound interest to determine whether such stress-related or tissue specific protein factors tend to bind to non-consensus CTD peptides containing dominant phosphorylation signatures using a pull-down binding approach.

3.9 Phosphorylation profiles are location dependent

When comparing identical CTD sequences, like mono-consensus heptads originating from different repeats along the CTD, a high variability of the abundance of distinct phosphorylation signatures has been documented. These data suggest that CTD repeats positioned at different locations within the CTD are preferentially differently phosphorylated. Accordingly, it could, therefore, be hypothesized that both CTD-kinases and phosphatases might recognize, or are specifically recruited by additional protein factors, to distinct areas along the CTD upon which they exert their catalytic activity. Interestingly, this hypothesis could be further supported by differences observed in the phosphorylation patterns between CTD repeats located

in great distances to each other, as well as between adjacent CTD repeats. These distinct phosphorylation profiles of individual CTD repeats might be linked to the recruitment of specific CTD-interacting protein factors. It appears that the more complex the binding motif of a CTD-binding factor is, the more restricted its binding area within the CTD is. In line with this, the binding motif of mammalian capping enzyme Mce1 contains four phosphosites and was found at only one position along the whole CTD, whereas CID-binding motifs of several CTD-interacting factors requiring only one or two CTD phosphosites could be mapped within several CTD repeats located at different areas along the CTD. A noteworthy study showed that mammalian cells containing only 31 CTD repeats were capable of transcription, splicing and polyadenylation, but exhibited defects in the transport of the matured mRNAs, suggesting that the missing CTD repeats are responsible for the binding of protein factors responsible for this specific function (Custodio et al., 2007). These data implied that the requirement of a certain length of CTD is dependent from the respective protein factor the CTD is interacting with. Accordingly, in mammalian cells, about 20 repeats were necessary for correct splicing and cleavage of the 3' end, whereas 16 repeats were sufficient for Pol II to support its own expression (de la Mata and Kornblihtt, 2006; Rosonina and Blencowe, 2004; Chapman et al., 2007). Consequently, a certain length of CTD might be the prerequisite for serving as an optimal binding platform for CTD-kinases, as well as for CTD- or phospho-CTD-associated factors. In line with this, the core protein complexes that have evolved to interact with canonical essential units define the minimal CTD length whereas additional CTD repeats more distal to the core CTD region might serve as binding sites for proteins with accessory functions that are not essential to cell viability. Consequently, genetic studies in yeast have shown that cells containing shortened CTDs revealed higher temperature sensitivity, implying that additional CTD repeats are necessary for optimal survival under stress conditions (West and Corden, 1995). The highly conserved CTD length, e.g. 52 repeats in human and mouse, might be determined by the highest possible number of CTD-interacting factors that can bind the CTD at the same time. The native length of CTD provides an optimal binding platform for core factors mainly bound to its proximal part, as well as for additional protein complexes that might play an important role in cell growth, differentiation or in cellular responses upon stress signals and which preferentially interact with the distal part of the protein. Accordingly, changes in the sequence or length of the CTD could

influence single CTD-protein interactions as well as the overall composition of CTD-associated proteins along the CTD.

3.10 New insights into the CTD code

One key finding in this work was the predominant occurrence of mono-phosphorylated CTD repeats. In more detail, the portion of 1P-populations among all detected phosphoforms (1P, 2P 3P and 4P) within mono- and di-consensus repeats was 94,2% and 83,1%, respectively. Consequently, 1P-CTD repeats most probably reflect the main scaffold upon which higher phosphorylated forms are built up in a more temporary fashion. Especially, the switch between 1P- and 2P-levels seems to appear quite often whereas 3P- and 4P-combinations were found in only very low amounts (Figure 30A). Accordingly, at the 1P- and 2P-levels of mono- and di-consensus heptads, all possible phosphorylation patterns could be found reflecting how dynamic and diverse CTD phosphorylations are *in vivo*. CTD residues Y_1, S_2, T_4, S_5 and S_7 were found to be phosphorylated in similar amounts in a 1P-background suggesting that all phosphosites within a CTD repeat fulfil important tasks *in vivo*. On the contrary, at the 2P-level, both low and high abundant phosphorylation patterns were detected and subsequently predominant 2P-signatures could be mapped. Interestingly, the majority of the dominant 2P-combinations contained two out of the three phosphosites S_2P, T_4P and S_5P. At the 3P-level, only a few more abundant phosphorylation signatures were found displaying much lower detection counts compared to the most abundant 2P-signatures. Interestingly, when analysing the more abundant 3P-combinations, new phosphosites like Y1P and S7P were implemented that played only a minor role in high abundant 2P-signatures. 4P-combinations within mono- and di-consensus heptads were only found in substoichiometric amounts (0,005% in mono-heptads and 0,1% in di-heptads) and no predominant 4P-signatures could be assigned.

Another important finding in this work was that different predominant 2P-signatures were detected within consensus repeats located at different regions within the CTD. Since these repeats have identical sequences these results suggest that CTD phosphorylation is location dependent. In other words, the CTD is divided into many

sub-compartments revealing different phosphorylation patterns defined by their position within the CTD (Figure 30B).
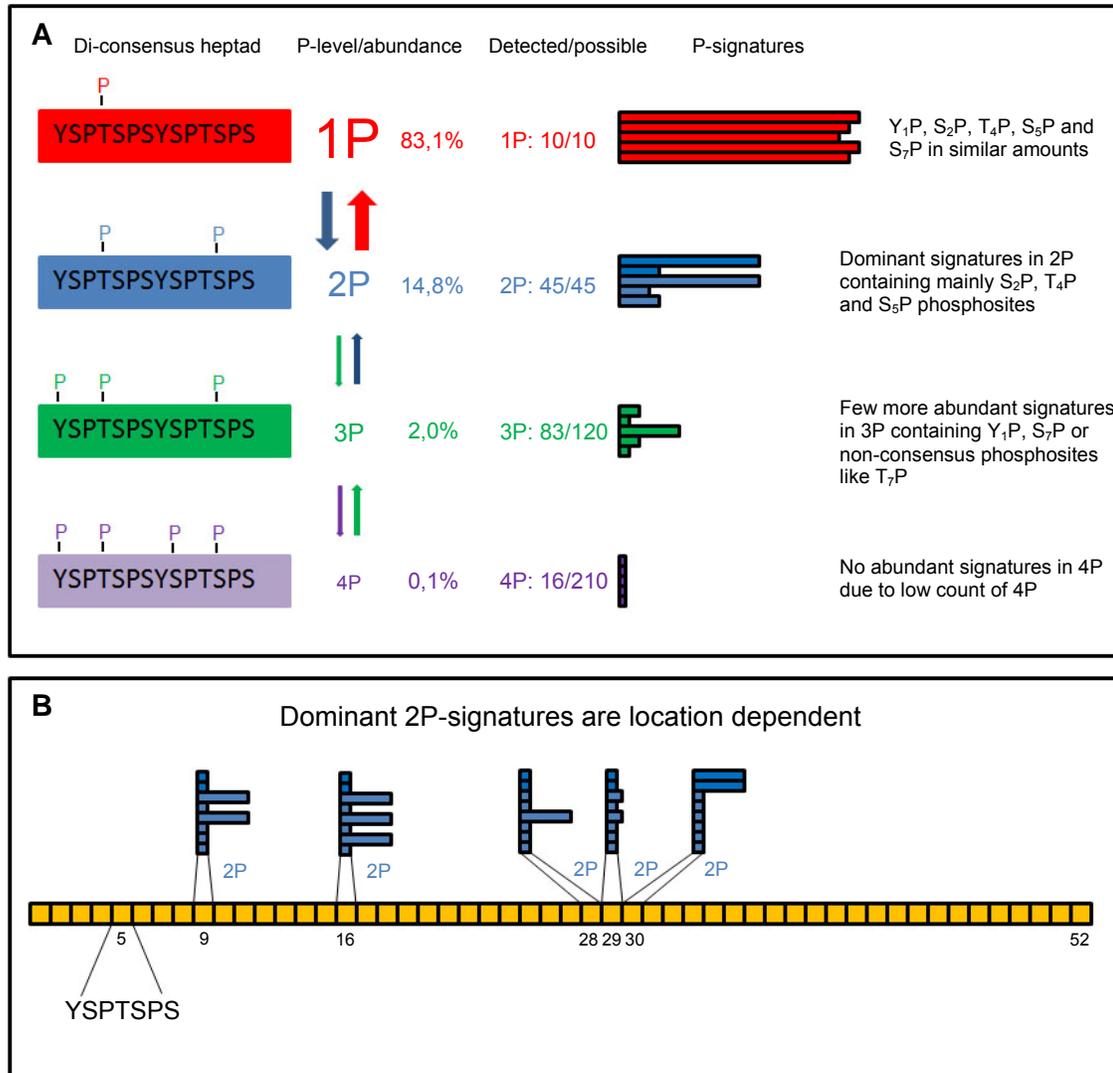
Figure 30 Key findings of this work. A) Different phosphorylation levels (1P/red, 2P/blue, 3P/green and 4P/purple) of di-consensus heptad YSPTSPSYSPTSPS are shown. Portion of different phosphorylation forms are indicated (1P: 83,1%, 2P: 14,8%, 3P: 2,0%, 4P: 0,1%). Detected and possible phosphorylation combinations within each phosphorylation state are shown (1P: 10/10, 2P: 45/45, 3P: 83/120, 4P: 16/210). Abundance of different phosphorylation (P-) signature profiles within the four possible phosphorylation levels are displayed and described. B) Different 2P-signature profiles obtained from consensus repeats located within different regions along the CTD (CTD repeat 9, 16, 28, 29, 30) are shown.

3.11 Outlook

This work provides new insights into phosphorylation signatures of Pol II CTD *in vivo.* For the first time CTD phosphosites could be mapped along the whole CTD molecule. In this study CTD peptides have been extracted from a total Pol II pool originating from cells cultured under normal cell conditions. In a future attempt, it will be interesting to compare CTD patterns that are linked to different cell states during cell cycle. Additionally, investigation of possible dynamic changes in phosphorylation patterns along the CTD upon cellular stress induction might contribute to a more detailed understanding of how a specific CTD code is defined upon a certain 'demand'. Furthermore, the MS approach established in this work could also be used for the mapping of CTD phosphosites in other mammalian cell culture cells other than Raji cells. Accordingly, insights into the CTD code of different tissue-specific cells, e.g. in mouse, might further expand the repertoire of distinct phosphorylation signatures within the CTD.

ChIP data revealed that different CTD phosphosites reach their peaks at different regions along actively transcribed genes suggesting that distinct phosphorylation patterns of CTD are related to specific transcription states. One big challenge for the future will be to extract Pol II located at specific sites denoting actively transcribed genes and subsequently map phosphosites of CTDs that are linked to transcription initiation, transcription elongation or transcription termination. However, one aspect wherein a great deal of improvement is still desired is the volume of sample needed since approximately 300 million cells are required for efficiently mapping hundreds of phosphosites by MS analysis.

Another noteworthy aspect of future research is the lack of information on the stoichiometry of phosphorylation of distinct CTD peptides. There are several established MS-based approaches available for analysing the stoichiometry of phosphorylation. In an isotope-free approach the big challenge is to calculate the 'flyability' ratio for each of the phosphopeptide versions with respect to that of their unphosphorylated counterpart. The flyability ratio can be analysed by using the corresponding set of synthetic peptides (Steen et al., 2005). In a different approach, the use of a known amount of stable isotope-labeled standard peptides to quantify the abundance of different versions of the phosphopeptide might be the most

accurate existing method for measuring stoichiometry of phosphorylation (Mayya et al., 2006a; Mayya et al., 2006b). However, the cost of synthesizing thousands of stable isotope-labeled standard peptides that have been purified and quantified accurately is a substantial limitation in performing such experiments.

One important point which is currently under investigation is how different neighbouring amino acid residues influence nearby phosphosites within the CTD. In this approach, identical sequences located within the same CTD repeats of different CTD mutants that differ in adjacent CTD residues will be analysed. Moreover, it is important to show that consensus repeats located at the same position in different mutants display similar 2P-signatures profiles and thereby strengthening the idea of a location dependency of CTD phosphorylation obtained by comparing consensus heptads along the CTD in this work. Additionally, the key finding in this work, that mono-phosphorylated CTD peptides are the predominant phosphorylation form *in vivo*, has to be confirmed by the analysis of synthetic peptides containing one to four phosphosites thereby excluding a technical bias towards low phosphorylated CTD peptides.

Furthermore, separated data sets linked to the different Pol II forms which were detected by SDS-PAGE will be created. In this way useful information on possible dominant Pol II forms could be acquired. Last but not least, high abundant phosphorylation signatures will be used for pull down experiments 'fishing' for novel specific CTD-protein interactions by implementing synthetic peptides carrying the corresponding phosphosites.

4. Material and Methods

4.1 Materials

4.1.1 Chemicals

1,4-Dithiothreitol (DTT)	*Carl Roth GmbH&CoKG, Karlsruhe*
1 kb DNA ladder	*Invitrogen, Karlsruhe*
3-(N-Morpholino)-propanesulfonic Acid (MOPS)	*Sigma-Aldrich Chemie, GmbH, Deisenhofen*
Agarose	*Invitrogen, Karlsruhe*
Albumin Fraktion V (BSA)	Carl Roth GmbH&CoKG, Karlsruhe
$\alpha-$Amanitin	*Roche Molecular Biochemicals, Mannheim*
Bromophenol Blue (BPB)	*Sigma-Aldrich Chemie GmbH, Deisenhofen*
Dimethyl Sulfoxide (DMSO)	*Sigma-Aldrich Chemie GmbH, Deisenhofen*
Ethanol (EtOH), absolute	*Merck, Darmstadt*
Ethidium Bromide (EtBr)	*Fluka Chemie GmbH, Buchs*
Ethylendiaminetetraacetic Acid (EDTA)	*Carl Roth GmbH&CoKG, Karlsruhe*
Fetal Bovine Serum (FBS)	*PAA Laboratories, Pasching, Österreich*
Formaldehyde (37 %)	*Carl Roth GmbH&CoKG, Karlsruhe*
Glycerol 86%	*Carl Roth GmbH&CoKG, Karlsruhe*
Glycine	*Carl Roth GmbH&CoKG, Karlsruhe*
Isopropanol, absolute	*Carl Roth, Karlsruhe*
L-Glutamine 200mM (100x)	*Gibco BRL Life Technologies, Eggenstein*
Methanol (MeOH), absolute	*Merck KGaA, Darmstadt*
Neomycin (G148)	*Promega Corp., Wisconsin, USA*
Orange G	*Sigma-Aldrich Chemie GmbH, Deisenhofen*
Peniciline/Streptomycin 10.000 U/ml	*Gibco BRL Life Technologies, Eggenstein*
Phenylmethanesulfonyl Fluoride (PMSF)	*ICN Biomedicals Inc., Fountain Pkwy, USA*

Polyacrylamide 30% (PAA)	*Carl Roth GmbH&CoKG, Karlsruhe*
Powdered Milk, blotting grade	*Carl Roth GmbH&CoKG, Karlsruhe*
Prestained Protein Ladder Plus	*Fermentas, St. Leon-Rot*
RPMI Medium 1640	*Gibco BRL Life Technologies, Eggenstein*
Sodium Dodecyl Sulfate (SDS)	*Carl Roth, Karlsruhe*
Tetramethylethylenediamine (TEMED)	*Carl Roth GmbH&CoKG, Karlsruhe*
Tetracycline	*Promega Corp., Wisconsin, USA*
Tris(hydroxymethyl)aminomethane (TRIS)	*Merck KGaA, Darmstadt*
Triton X-100	*Sigma-Aldrich Chemie GmbH, Steinheim*
Tryphan Blue	*Invitrogen, Karlsruhe*
Tween-20	*Sigma-Aldrich Chemie GmbH, Deisenhofen*

4.1.2 Consumables and kits

Agar plates	*Greiner GmbH, Frickenhausen*
Cover Slides	*Menzel, Braunschweig*
Cyrovials 1.5 ml	*Nunc GmbH, Wiesbaden*
DNA Mini/Maxi kits	*Qiagen GmbH, Hilden*
Electroporation cuvettes	*Peqlab, Erlangen*
Gel Blotting Paper GB003	*Schleicher & Schuell, Deutschland*
Hybond N+ Nylon Membrane	*GE Healthcare, München*
Laboratory Glassware	*Duran Productions GmbH & Co. KG, Mainz*
Nitrile Gloves	*Kimberly-Clark, Koblenz*
Parafilm	*Carl Roth GmbH&CoKG, Karlsruhe*
Pasteur Pipettes	*Hirschmann Laborgeräte, Eberstadt*
Phosphataseinhibitor cocktail	*Roche Diagnostics, Penzberg*
Pipette Tips ART 10, 20, 200, 1000)	*MolecularBio-Products, San Diego*
Plastic ware for cell culture	*Greiner Bio-One,Frickenhausen*
Proteaseinhibitor cocktail	*Roche Diagnostics, Penzberg*
Protein A-Sepharose beads	*GE Healthcare, München*
Protein G-Sepharose beads	*GE Healthcare, München*

Reaction Tubes 1.5 ml, 2 ml	*Eppendorf, Hamburg*
Reaction Tubes 15 ml, 50 ml	*Becton Dickinson Biosiences, Heidelberg*
Scalpel	*Braun, Tuttlingen*
Sterile filters	*Milipore GmbH, Eschborn*

4.1.3 Technical Instruments

-80°C freezer	*Colora Messtechnik GmbH, Lorch*
-20°C freezer	*Siemens, München*
Bacteria incubator	*Heraeus Sepatech GmbH, Osterode*
Bacteria shaker (Series 25)	*New Brunswick ScientificCo., NJ, USA*
Biofuge 13	*Heraeus Sepatech GmbH, Osterode*
Bio-Rad PowerPac 300	*Bio-Rad Laboratories GmbH, München*
Blotting chamber	*Bio-Rad Laboratories GmbH, München*
Branson Sonifier 250	*Heinemann Ultraschall- und Labortechnik*
Electrophoresis equipment	*Bio-Rad Laboratories GmbH, München*
Electroporator (eukaryotic cells)	*Bio-Rad Laboratories GmbH, München*
Eppendorf Centrifuge 5417R	*Eppendorf, Hamburg*
Eppendorf Thermomixer 5436	*Eppendorf-Netheler-Hinz GmbH, Hamburg*
Fridge KU 171	*Liebherr, Biberach*
Fuchs-Rosentahl chamber	*GLW Gesellschaft für Laborbedarf GmbH*
Gilson Pipettes 2,10, 20, 200,1000	*Gilson, Bad Camberg*
Hypercassette	*Amersham Pharmacia Biotech, Freiburg*
Inkubator Heraeus 6000	*Heraeus Sepatech GmbbH, Osterode*
Laminar Flow Hood	*BDK Luft-und Reinraumtechnik GmbH*
Magnet stirer M23	*GLW, Würzburg*
Megafuge 2.0	*Heraeus Sepatech GmbH, Osterode*
Microwave	*Panasonic, Hamburg*
Multi-calimatic pH-meter	*Knick GmbH & Co. KG, Berlin*
Nanodrop 1000	*Thermo Scientific, Braunschweig*
Odyssey Infrared Imaging System	*Odyssey LI-COR*
PipetMan P	*Gilson, Bad Camberg*

Rollermixer SRT 2   *Dunn GmbH, Augsburg*

SDS-PAGE gel tank   *Amersham Pharmacia Biotech, Freiburg*

ST305 Electrophoresis Power Supply   *Gibco BRL Life Technologies, Eggenstein*

Telaval 31 Lichtmikroskop   *Carl Zeiss Jena GmbH, Göttingen*

UV lamp VL-4. LC   *PeqLab Biotechnologie GmbH, Erlangen*

Vortexer Vortex Genie 2   *Bender & Hobein GmbH, Ismaning*

Waterbath   *Thermo Fisher Scientific, Karlsruhe*

4.1.4 Buffer and Solutions

0,7% Agarose-TAE-Gel for DNA   2,1 g Agarose

300 ml 1x TAE

boil in microwave

cool to 65°C

EtBr (375 ng/µl)

10 x DNA Loading Dye   20 g Sucrose

100 mg Orange G

ad 50 ml H_2O

PBS   137 mM NaCl

2,7 mM KCl

4,3 mM Na_2HPO_4*$6H_2O$

1,4 mM KH_2PO_4

Lämmli-Buffer (2x)   2% SDS

100 mM DTT

10 mM EDTA

10% Glycerol

60 mM Tris/HCl pH 6,8

0,01% BPB

1 mM PMSF

Milk powder solution	5% powdered milk in 1 x TBST
NP40-Lysisbuffer	50 mM Tris-Cl, pH 8,0
	1% NP40
	150 mM NaCl
	Phosphataseinhibitor (1:200)
	Proteaseinhibitor (1:200)
IP-Washingbuffer	50 mM Tris-Cl, pH 8,0
	150 mM NaCl
TiO_2-Loading Buffer	80% Acetonitrile
	5% TFA
	1 M Glycolic acid
TiO_2-Washing Buffer 1	80% Acetonitrile
	1% TFA
TiO_2-Washing Buffer 2	10% Acetonitrile
	0,2% TFA
TiO_2-Elution Buffer	40 µl Ammonium solution (28%)
	in 960 µl H_2O, pH 11,3
2xTris/SDS pH 8,8	22,68 g Tris/Base
	2,5 ml SDS 20%
	add 250 ml H_2O
	pH 8,8 (with HCL)
2xTris/SDS pH 6,8	7,56 g Tris/Base
	2,5 ml SDS 20%
	add 250 ml H_2O
	pH 6,8 (with HCL)

Running-gel (6,5%)	8,6 ml PAA 30%
	20 ml 2xTris/SDS pH 8,8
	11 ml H_2O
	334 µl APS
	20 µl TEMED
Stacking-gel (3%)	1,5 ml PAA 30%
	7,5 ml 2xTris/SDS pH 6,8
	5,9 ml H_2O
	90 µl APS
	20 µl TEMED
SDS-PAGE-running buffer (10x)	60,4 g Tris/Base
	288 g Glycin
	5 ml SDS 20%
	add 2 l H_2O
Western-transfer buffer (10x)	60,4 g Tris/Base
	288 g Glycin
	5 ml SDS 20%
	200 ml Methanol
	add 2 l H_2O
Western-blocking-reagent	10% (v/v) TBS
	0,1% (v/v) Tween 20
	5% (w/v) Magermilchpulver
PBS/Tween	99,9% (v/v) PBS
	0,1% (v/v) Tween 20

4.1.5 Antibodies

Primary antibodies:

Pol3.3:

Monoclonal antibody (mouse) that recognizes a conserved epitope of the large subunit of Pol II (Rpb1) outside of the CTD.

(originally produced from *E.K. Bautz, Universität Heidelberg. Received as a supernatant solution from E. Kremmer, Helmholtz Zentrum, München*)

8WG16:

Monoclonal antibody (mouse) that recognizes the unphosphorylated CTD of the large subunit of Pol II (Rpb1).

(Received as a supernatant solution from *E. Kremmer, Helmholtz Zentrum, München*)

1C7:

Monoclonal antibody (rat) that recognizes the unphosphorylated CTD of the large subunit of Pol II (Rpb1).

(Received as a supernatant solution from *E. Kremmer, Helmholtz Zentrum, München*)

3F10 (HA):

Monoclonal antibody (rat) that recognizes an epitope contained in the haemagglutinin polypeptide of the human influenza virus (*Roche Diagnostics, GmbH, Mannheim*)

α-Tyr-1-P, α-Ser-2-P, α-Ser-5-P, α-Ser-7-P, and α-Thr-4-P:

Monoclonal antibodies (rat) that recognize the phosphorylated form of their respective amino acids within the CTD of the large subunit of Pol II (Rpb1).

(Received as a supernatant solution from *E. Kremmer, Helmholtz Zentrum, München*)

Secondary antibodies:

Alexa Fluor 680 Goat α-Rat IgG (H+L)	*Molecular Probes*
IR Dye 800 CW α-Rat IgG (H+L)	*Rockland Inc, Rockland*
Alexa Fluor 680 Goat α-Mouse IgG (H+L)	*Molecular Probes*
IR Dye 800 CW α-Mouse IgG (H+L)	*Rockland Inc, Rockland*

4.2 Materials for cloning

4.2.1 Oligonucleotides

Newly synthesized CTD sequences:

All mutated CTD sequences as well as the wildtype-CTD sequence were synthesized by *GeneArt Regensburg*.

Primers for CTD sequencing in the final expression vector LS*mock:

wt fwd: 5'CTCCTGCTGACGCACCTGTTCT3'

CTD fwd: 5'CCTTTGTCTTTTCCTATAGGTGGTGC3'

CTD rev: 5'GTCAGACAACCTCGGTGGCCTGTGTG3'

4.2.2 Plasmids used during this work

RX2-287 vector (subcloning vector):
Vector containing last exon (CTD) of the α-amanitin resistant Pol II Rpb1 gene.

RX4-267 (LS*Mock - expression vector):
A tetracycline-regulated expression vector containing the α-amanitin-resistant and HA-tagged mouse Rpb1 gene.

4.2.3 Cloning strategy

All synthesized CTD must be flanked by an AvrII restriction site upstream to the CTD sequence and by a NotI restriction site downstream to the CTD sequence. Synthesized CTDs must not contain AvrII, NotI, AgeI, BspEI, NgoMIV, NheI, SpeI and ClaI restriction sites within the sequence.

First the newly synthesized CTD cassette will be cloned into the subcloning vector RX2-287 vector using the restriction sites AvrII and NotI. RX2-287 will be cut with BspEI and NotI and cloned into the AgeI/NotI site of RX4-267 (Figure 31).
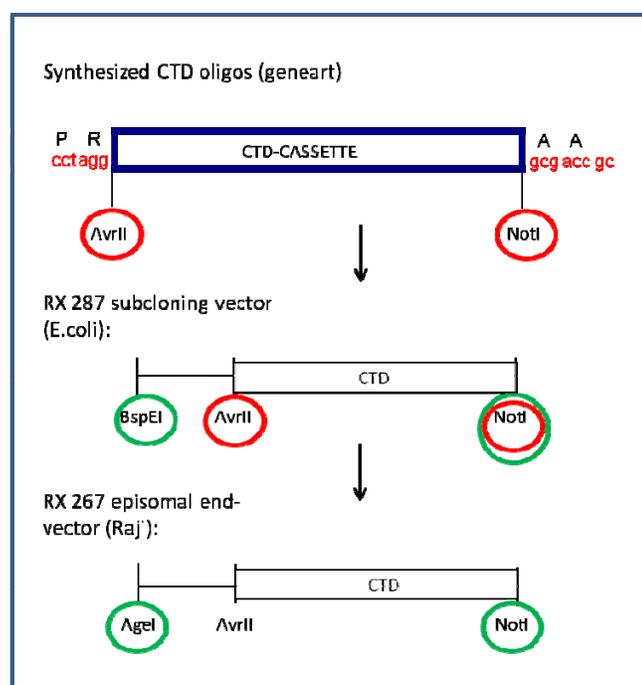
Figure 31 Cloning strategy. Schematic view of the two step cloning procedure using two different vectors and different restriction sites. In red letters: recognition sites of AvrII (left) and NotI (right). AvrII recognition site encodes proline and arginine (P and R). NotI recognition site encodes two alanines (A). These two sites flank the CTD open reading frame. Restriction sites in red circles are used for the first cloning step into the subcloning vector (RX 287) and restriction sites in green circles are used for the second cloning step into the final expression vector (RX4 267). All cloning steps and vectors used in this work have been established by *Chapman et al.*

4.2.4 Bacteria

<u>DH10B:</u> *E.coli* strain purchased from *Invitrogen GmbH, Karlsruhe.* Used for the cloning of all plasmid DNA.

4.3 Human cell lines

4.3.1 Basic cell lines

<u>Raji:</u>

Cell type: human Burkitt lymphoma

DSMZ no.: ACC 319

Origin: established from the left maxilla of a 12-year-old African boy with Burkitt lymphoma in 1963; first continuous human hematopoietic cell line; classified as risk category 1 according to the German Central Commission for Biological Safety (ZKBS)

References: Pulvertaft, Lancet: 238-240 (1964), PubMed ID 14086209

 (1965), PubMed ID 4284655

4.3.2 Stably transfected cell lines

Name	Plasmid	Cell line	Resistance
WT rec	LS*mock	Raji	neomycine
M-6K	--	--	--
M-3K5R	--	--	--
M-8K	--	--	--
M-4K2R	--	--	--
M-13K2R	--	--	--
M-9K4R	--	--	--
M-12K2R	--	--	--
M-8K4R	--	--	--
M-9K2R	--	--	--

4.4 Methods

4.4.1 Bacterial cell culture

4.4.1.1 The maintenance and preparation of bacterial plasmids

Bacteria were cultured either on LB-agar plates in a bacterial incubator, or in liquid LB medium in a thermoshuttler at 37°C overnight. Liquid cultures were produced through infection of 200-400 ml LB medium with a single bacterial colony picked from an agar plate. Transformed bacteria were selected for the antibiotic resistance of the transformed plasmid through the addition of antibiotic (Ampicillin, Kanamycin or Spectionomycin) to growth media: liquid culture medium, 100 µg/ml; agar plates, 50µg/ml.

LB-medium: 20 mM $MgSO_4$; 10 mM KCL; 1% Bacto-Tryptone; 0,5% Bactoyeast extract; 0,5% NaCl

LB-agar: 20 mM $MgSO_4$; 10 mM KCL; 1% Bacto-Tryptone; 0,5% Bactoyeast extract; 0,5% NaCl; 1,2% Bacto-agar

4.4.1.2 Preparation of competent bacteria

To increase the efficiency of plasmid DNA uptake (transformation), bacteria were treated with solutions of di-valent cations. An LB plate was first inoculated with a probe from a bacterial stock and grown overnight at 37°C. A single colony was then used to inoculate 2.5 ml of LB medium, which was then incubated overnight in a loose-capped vessel, with shaking. The following day, the entire overnight culture was used to inoculate 250 ml of LB medium containing 20 mM $MgSO_4$. Bacteria were grown in a 1 dm^3 flask, with shaking, at 37°C until the absorbance at 600 nm (A_{600}) reached 0,4-0,6 (approx. 5-6 h).

Bacteria were pelleted at 4,500 x g, 5' at 4°C. Medium was discarded and the bacteria re-suspended in 0,4 volume (of the original culture volume) of ice-cold TFB1. Bacteria were incubated a further 5' on ice before centrifugation (as above),

and re-suspension in 1/25 of the original culture volume of ice-cold TFB2 and incubated a further 15-60' on ice. Aliquots of 50-100 µl bacteria were snap-frozen in liquid nitrogen and removed to storage at -80°C.

NOTE: All vessels and pipettes should be pre-chilled.

TFB1:   30 mM Potassium acetate; 10 mM $CaCl_2$; 50 mM $MnCl_2$; 100 mM RbCl; 15% glycerol. Adjust to pH 5,8 with 1 M acetic acid; sterile filter (0,2 µm)

TFB2:   10 mM MOPS pH 6,5; 75 mM $CaCl_2$; 10mM RbCl; 15% glycerol. Adjust to pH 6,5 with 1 M KOH; sterile filter (0,2 µm)

4.4.1.3 Transformation of bacteria

For standard sub-cloning and production of large amounts of cloned DNA, the recombination-deficient *Escherichia Coli* strain, DH10B was used. For transformation, 1 ng to 0,1 µg of plasmid DNA or 10 µl of ligation mixture was added to 50-100 µl of competent bacteria suspension and incubated on ice for 30'. Cells were then subjected to heat shock at 42°C for 30'' before returning to ice for 2'. 200-400 µl (4 volumes) of LB recovery medium was added to the cells and incubated at 37°C for 1h for cells to express resistance genes conferred by the plasmid.

Finally, 50-200 µl of suspension was plated onto LB-agar plates containing ampicillin, kanamycin or spectinomycin and grown at 37°C for 16-18 h.

4.4.1.4 Miniprep of plasmid DNA

Single colonies obtained from transformation of bacteria following ligation reactions were used to inoculate 2 ml LB, and grown overnight. Using the alkaline lysis method (Birnboim, 1983), plasmid DNA was isolated from a bacteria culture. This method relies on the fact that high molecular weight linear chromosomal DNA will be denatured when cells are lysed at pH 12-12,6, whereas low molecular weight supercoiled plasmid DNA remains unaffected. Neutralisation of pH in the presence of high salt concentrations subsequently precipitates chromosomal DNA, which can be

separated from the mix. The following protocol was routinely used: 2 ml of LB media containing appropriate antibiotic was inoculated with a single colony of transformed bacteria in a 20 ml loose-capped tube. The culture was incubated overnight at 37°C with vigorous shaking. 1,5 ml of culture was then decanted into a microfuge tube and centrifugated at maximum speed for 1' in a microfuge. Medium was removed from the pellet and re-suspended in 150 µl of alkaline lysis solution 1 by vigorous pipetting up and down, ensuring complete dispersal of bacterial cells. 150 µl of alkaline lysis solution 2 was added, and mixed gently by inverting 5 times before incubation at room temperature for 5'. To precipitate, 150 µl of pre-chilled alkaline lysis solution 3 was added, and then mixed by inverting, before incubation on ice for 20'. After centrifugation at maximum speed for 10', the resulting supernatant was transferred to a fresh microfuge tube and mixed with 450 µl isopropanol (1:1) to precipitate plasmid DNA, and centrifuged at maximum speed for 10'. The pellet was then washed once with 5 ml of 70% ethanol and centrifuged at maximum speed for 5'. Supernatant was thoroughly removed and pellets re-suspended either in 200 µl H_2O or in 100 µl TE buffer.

For screening of these crude preparations with restriction enzymes, master mixes of enzymes and appropriate restriction enzyme buffers (New England Biolab/Promega) containing RNase A were prepared, such that 10 µl of mix could be added to 10 µl of miniprep to give 0,5-3 units of enzyme (depending on efficiency), 1x restriction enzyme buffer and 20 µg/ml RNase A. Digests were incubated at appropriate temperatures for at least 2 h before separation on agarose gels.

Solution 1: 50 mM Tris-Cl; pH 8; 10 mM EDTA
Solution 2: 0,2 M NaOH; 1% SDS
Solution 3: 3M potassium acetate; 2M glacial acetic acid

4.4.1.5 Maxiprep of plasmid DNA

Large quantities of plasmids were purified using Qiagen Maxiprep protocols based on a modified alkaline lysis procedure. Plasmid DNA is recovered by running the bacterial lysate through an anion exchange column under appropriate low-salt and pH conditions. Following washing, the DNA can be eluted by a high-salt buffer.

A single colony was picked from a freshly streaked selection plate and incubated in 200-400 ml of selective LB medium overnight at 37°C on an orbital shaker. Cells were harvested by centrifugation at 6000 x g for 15' at 4°C. Supernatant was discarded and pellets were re-suspended in 10 ml of chilled alkaline lysis solution 1 containing 100 µg/ml RNAase A. 10 ml of alkaline lysis solution 2 was added to the suspension and mixed gently by inverting. Cells were lysed for 5 min at room temperature before stopping the reaction by addition of 10 ml alkaline lysis solution 3 and mixed again by inverting. Mixtures were incubated on ice for 20' to aid precipitation of cell debris, genomic DNA and SDS. The sample was then centrifuged at 20.000 x g for 30' at 4°C. The supernatant was then passed through a filter paper to remove any residual precipitate and applied to a Qiagen column equilibrated with 10 ml of equilibration buffer and washed twice with 30 ml wash buffer. The bound plasmid was eluted in 15 ml elution buffer and precipitated by adding 10.5 ml (0,7 volumes) of room temperature isopropanol, followed by centrifugation at 20.000 x g for 30' at 4°C. Pellets were re-suspended in 300 µl TE or H_2O and transferred to a microfuge tube.

Equilibration buffer: 50mM MOPS, pH 7; 750 mM NaCl; 15% ethanol
Wash buffer: 50mM MOPS, pH 7; 1 M NaCl; 15% ethanol
Elution buffer: 50mM Tris-Cl, pH 8,5; 1,25 M NaCl; 15% ethanol

4.4.2 Eukaryotic cell culture

4.4.2.1 Human cell culture methods

Raji is a human, EBV positive Burkitt lymphoma cell line that is cultivated at 37°C in a 5% CO_2-containing atmosphere in cell culture bottles under sterile conditions. The suspension cells were splitted in a 1:3 ratio with fresh media every three days, receiving a cell density of about 250000 cells/ml.

Media for cell culture: RPMI-Medium with following additives: 10% FCS,

100 U Penicillin/ml, 100 µg Streptomycin/ml und 2 mM L-Glutamine.

Medium for the transfected cells additionally contained:

1 mg G418/ml.

α-amanitin (stock solution: 1mg/ml in H_2O) was added in a final concentration of 2 µg/ml to inhibit the endogenous Pol II. Protein extracts for western analysis were prepared 24 - 48 h after treatment with α-amanitin.

4.4.2.2 Determination of the living cell number

The living cell number was calculated using a Fuchs-Rosenthal-Zählkammer.

40 µl of cell suspension was mixed with 0,5% Trypanblau-solution (w/v in PBS) containing the same volume. Dead cells absorb the stain over their membranes and can be distinguished from the colourless, living cells. The calculation of the living cell number results from the counting of four big squares using the formula: c (cells/ml) = average value of the cells per big square x 10^4.

4.4.2.3 Unfreezing of cells

To revive stocks from storage, cryotubes were thawed at room temperature for 3' before transfer to 15 ml fresh RPMI medium. Cells and medium was mixed through inversion several times, before re-centrifugation (300 x g, 4'), and re-suspension in 10 ml fresh culture medium containing 20% FCS.

4.4.2.4 Refreezing of cells

Cells were split 1:1 with fresh medium one day before storage. Cells were pelleted (300 x g, 4'), and medium discarded, before resuspension in storage medium and transfer of 1 ml aliquots to 1,5 ml cryotubes. The tubes were wrapped with paper avoiding shock freezing and first stored overnight at - 80°C before transfer to storage facilities in liquid nitrogen.

Storage medium: 10% RPMI 1640; 10% DMSO; 80% FCS

4.4.2.5 Stable transfection of B-cells

For the production of stable cell lines, plasmid vectors were introduced into cell lines, and a selection performed utilizing an antibiotic resistance gene expressed by the introduced vector. For best results, cells should be split 1:1 one day prior to transfection. 2×10^7 cells are required per transfection. Cells were collected by centrifugation (300 x g, 4'), and re-suspended at a density of 2×10^7 cells / 500 μl in PBS. Cells were transferred to a 4 mm electroporation cuvette, and gently mixed with 10 μg of plasmid DNA through knocking against the wall of the cuvette. The cell/plasmid DNA mixture was incubated at room temperature for 20' before starting the electroporation. Electroporation was performed using a voltage of 250 V and capacitance charge of 950 μF. Both pulse buttons were pushed simultaneously till the noise signal appeared and then immediately 1 ml FCS was added to the cuvette and mixed with the transfected cells through pipetting up and down. Transfected probes were incubated for 5' at room temperature before transfer to a flask containing 5 ml fresh medium. It is important to leave the white smear (toxic) consisting of dead cell material behind.

Two days following transfection, the appropriate selection reagent for the vectors used was applied to the cell culture medium. Selection typically requires 2 – 4 weeks to recover a viability of 95%. This technique produces a polyclonal cell line.

Neomycin (G418): Stock: 100 mg/ml in PBS; Working concentration 1mg/ml

4.4.3 Molecular techniques for cloning

4.4.3.1 Digestion of DNA using restriction endonucleases

Restriction enzymes were used as described by the manufacturer. For the analysis of plasmid 'mini-prep' DNA, multiple digests were performed in a compatible buffer. For subsequent gel extraction 5 μg of DNA was digested.

4.4.3.2 Ligation of DNA fragments

Equimolar amounts of DNA fragments (100 – 200 ng) were mixed together and heated to 45°C for 5', before transfer to ice. Ligase buffer and T4 DNA ligase were added, as recommended by the manufacturer, and incubated overnight at 16°C in the cold room. The entire ligation mixture was used for transformation.

4.4.3.3 DNA agarose-gel electrophoresis

For a 0.7 % DNA agarose-gel, 2,1 g agarose was boiled in 300 ml 1 x TAE buffer. 3 µl/100 ml ethidium bromide (EtBr, stock: 10 mg/µl) was added after cooling of the gel to 65°C. DNA was mixed with 10 x DNA loading dye and loaded on the gel. The gel was run with a 1 kb DNA ladder in 1 x TAE buffer (80 V, 3 hours). A gel photo was taken under UV light.

4.4.3.4 DNA-gel extraction for cloning

Purification of a mixture of different-sized DNA was achieved via separation over a 1 x TAE 0,7% agarose gel, containing ethidium bromide (0,5 µg/ml). A band of interest was removed from the gel using a scalpel, melted at 65°C and further processed by using the QIAquick gel extraction Kit. This kit provides spin column, buffers and collection tubes for silica-membrane-based purification of DNA fragments from gels. DNA ranging from 70 bp to 10 kb is purified using a simple and fast bind-wash-elute procedure and an elution volume of 30-50 µl. An integrated pH indicator allows easy determination of the optimal pH for DNA binding to the spin column. The procedure can be fully automated on the QIAcube. The final DNA concentration was measured by using the nanodrop system.

For gel extraction of DNA bigger than 10 kb the QIAEX II gel extraction kit was used. This method provides a suspension of silica particles to which DNA fragments bind in the presence of chaotropic salts. QIAEX II suspension is added to solubilized agarose gel slices and binds DNA. The particles are collected by a brief centrifugation, washed, and DNA up to 50 kb is eluted in Tris buffer or water.

4.4.4 Methods for the analysis of protein

4.4.4.1 Cell lysis

For subsequent western analysis Raji cells were transferred to 15 ml falcons and centrifuged at 300 x g for 4'. The supernatant was removed and the cell pellet was washed with PBS. After a second centrifugation step (300 x g, 4') 100 - 300 μl 2x Lämmli-buffer was added and the viscous lysate was repeatedly drawn through a narrow pipette tip to shear genomic DNA, before denaturing by boiling at 95°C for 5'. After briefly cooling on ice, samples were centrifuged (15000 x g, 4') to clear insoluble contaminants. Finally the lysates were either directly loaded on to the gel or stored at –20°C.

For subsequent immunoprecipitation, Raji cells were harvested by centrifugation (300 x g, 4'), washed twice with cold PBS, re-centrifuged and re-suspended in ice-cold NP40 lysis buffer (100 μl lysis buffer/1,5 x 10^6 cells). Cells were mixed by sporadic vortexing while incubation on ice for at least 30' before sonication (duration time 1'; output 7; duty cycle 70%) and subsequent incubation for another 20' on ice. Cells were mixed before pelleting of the nuclei (472 x g, 15', 4°C). The supernatant was carefully removed for following immunoprecipitation, or storage at -20°C.

2 x Lämmli buffer: 2% SDS; 100 mM DTT; 10 mM EDTA; 10% Glycerol; 60 mM Tris/HCl pH 6,8; 0,01% Bromphenol Blue; 1 mM PMSF

NP40 lysis buffer: 50 mM Tris-Cl, pH 8,0; 1% NP40; 150 mM NaCl

4.4.4.2 Immunoprecipitation

First, 4 ml of antibody solution (20-40 μg/ml) was incubated with 60 μl of Protein A sepharose beads (stored in 20% ethanol) for 3 – 4 hours in the cold room on a rotary shaker. After two washing steps with PBS 10 ml of lysates (100 μl lysis buffer/1,5 x

10^6 cells) was added to the antibody-beads mixture and incubated in the cold room on a rotary shaker overnight. After incubation time, the probes were centrifuged (300 x g, 4'), the supernatant was removed and the specific protein-antibody-beads mixture was transferred to 1,5 ml eppendorf tubes. The samples were washed three times with washing buffer and finally resuspended in 30 µl Lysis buffer and 30 µl 2 x Lämmli buffer (final volume: 60µl). The probes were boiled at 95°C for 8' under high shaking to separate the proteins from the protein-A beads. After centrifugation (15000 x g, 4') the whole supernatant was loaded on to the SDS-PAGE.

4.4.4.3 SDS-PAGE and transfer

Lysate samples were loaded and separated on a denaturing SDS-polyacrylamide gel with a 6.5 % running gel and a 4 % stacking gel in 1 x SDS running buffer (SDS-PAGE, 30 mA for 3 hours). 3.5 µl of a pre-stained protein ladder was used as a running marker. The gel was transferred to a nitrocellulose membrane by semi-dry blotting in 1 x transfer buffer (450 mA, 1.5 hours). The membrane was blocked in milk powder solution (1 hour at RT) and incubated with specific primary antibodies, diluted in antibody stock solution, on a roller mixer at 4°C over night. Primary antibodies were removed and membranes were washed three times with 1 x TBST for 5'. Membranes were then either incubated with horseradish peroxidase (HRP)-conjugated secondary antibodies when the enhanced chemiluminescence (ECL) kit was used for signal detection or incubated with fluorescent-tagged secondary antibodies when the the Odyssey-Licor-Scanner system was chosen for detection. Secondary antibodies were diluted in milk powder solution and hybridized to membranes for 1.5 hours at RT. Membranes were washed three times with 1 x TBST for 5' and briefly poured in H_2O. When using the ECL-detection method, reagents were mixed (1:1) and incubated with membranes for 2'. Membranes were transferred to a detection cassette, a film was exposed and signals were detected with a developing machine.

Due to the use of fluorescent-coupled antibodies instead of peroxidase-linked antibodies, it is possible to check one and the same lysate with two different antibodies as long as they originate from different organisms (e.g. rat and mouse)

and are recognized by the appropriate secondary antibodies emitting different wavelenghts (680nm and 800nm). Since the membranes can be examined through the different channels either individually or simultaneously, both quantitative and qualitative conclusions can be made. In addition, varieties in the intensity of different, defined areas on the membrane can be visualized and therefore it is possible to make statements about e.g. the phosphorylation scale of Pol II (within the same channel) or about the amount of binding motifs of different antibodies comparing both channels.

Finally, it is very important to test the different antibodies used in this system for possible cross-reactions that can occur between them and would falsify the results obtained with the Odyssey system.

4.4.4.4 Coomassie staining

For Coomassie staining of SDS-PAGE-gels the EZBlue™ Gel Staining Reagent from Sigma was used. EZBlue™ Coomassie Brilliant Blue G-250 colloidal protein stain improves protein electrophoresis results while significantly reducing staining time. As a colloidal stain, it reacts only with proteins, not the gel itself. Background staining is reduced, so protein bands can be visualized very fast. EZBlue is extremely sensitive, detecting as little as 5 ng of protein.

Following electrophoresis the gel was placed in a tray and washed three times with water for 5' each with agitation. 40 ml EZBlue was added and the gel was incubated for 1 hour under shaking. Coomassie stain was removed and the gel was rinsed with water several times up to 1 hour to intensify protein bands. Gel was stored in water in the fridge covered by parafilm.

4.4.4.5 Protein in-gel digestion with trypsin

In-gel digestion comprises the four main steps destaining, reduction and alkylation (R&A) of the cysteines in the protein, proteolytic cleavage of the protein and extraction of the generated peptides.

After excision of the protein band of interest from the Coomassie stained gel, it was cut into three equal pieces and placed in a 96-well plate covered with water. For

destaining, the gel pieces were incubated with ammonium bicarbonate (ABC) and acetonitrile (ACN) in a 1:1 ratio three times for 20' at 37°C under shaking. After washing with ABC and subsequent dehydration of the gel pieces with ACN, dithiothreitol (DTT) was added and the probes were incubated for 1 hour at RT. The reduction reaction was stopped by adding iodoacetamide (IAA) and the gel pieces were incubated 30' at RT in darkness that led to the irreversible alkylation of the SH groups and the cysteines were transformed to the stable S-carboxyamidomethylcysteine (CAM). The gel pieces were dehydrated by repeated treatment with ACN and after removing of all the supernatant the probes were dried by applying vacuum for 5'. For digestion of the proteins, the enzyme trypsin was used and diluted to approximately 10 ng/µl using 20mM ABC. Trypsin solution was added covering the gel pieces. The gel pieces were incubated at 4°C for 30' taking up the trypsin solution. Supernatant was removed, 100 µl 20mM ABC were added and the probes were incubated at 37°C under shaking overnight. Next day, peptides were extracted by incubation with 50% ACN/ 0.25% TFA twice for 10' followed by two subsequent incubation steps with 100% ACN for 10' each. The supernatants containing the extracted peptides were transferred to a cool 1.5 ml tube on ice.

4.4.4.6 Purification of phosphorylated peptides using titanium dioxid (TiO$_2$)

Solution containing the extracted peptides was evaporated to approximately 5 µl dryness and then incubated in 100 µl loading buffer for 1 hour. 10 µl of TiO$_2$ beads slurry solution (slurry concentration: 0,03 mg/µl beads in 100% ACN) were added and the tubes were placed on the shaker (high shaking) at RT for 10'. TiO$_2$ beads bound to the phosphopeptides were pelleted by centrifugation (table centrifuge < 1') and supernatant (unbound fraction) was transferred to a new 1,5 ml tube and stored at -20°C. Beads were washed with 50 µl loading buffer, mixed for 15'' by pipetting up and down and centrifuged (table centrifuge < 1'). After removing the supernatant 2 additional washing steps were performed (washing buffer 1 + washing buffer 2). After the final washing step, the supernatant was completely removed without removing the beads and the lid of the tube was left open in the fume hood for at least 10' to allow drying. For eluting the phosphopeptides, 50 µl elution buffer was added, mixed

well and the solution was left shaking for 10' to allow an efficient elution. Probes were centrifuged for 1' (table centrifuge) and the supernatant was transferred to a new 1,5 ml tube without any beads. Eluate containing the purified phosphopeptides was evaporated to 5-10 µl dryness. 30 µl 0,1% formic acid (FA) was added and the probes were stored at -20°C.

4.4.4.7 Liquid chromatography– tandem mass spectrometry (LC-MS/MS)

For LC-MS/MS purposes, ~ 28- 30 µl of purified phosphopeptides were injected in an Ultimate 3000 HPLC system (LC Packings) and desalted on-line in a C18 micro column (300 µm i.d. x 5mm cm, PepMap100 C18 5 µm, 100 Å from LC Packings). Desalted sample was then separated in a C18 analytical column (75 µm i.d. x 10 cm, packed in-house with Reprosil Pur C18 AQ 2.4 µm from Doctor Maisch) with a 60 min gradient from 5 to 30% and 10 minutes from 30 to 95% acetonitrile in 0.1% formic acid. The effluent from the HPLC was directly electrosprayed into a LTQ Orbitrap XL mass spectrometer. The MS instrument was operated in data dependent mode to automatically switch between full scan MS and MS/MS acquisition. Survey full scan MS spectra (from m/z 300 – 2000) were acquired in the Orbitrap with resolution R=60,000 at m/z 400 (after accumulation to a 'target value' of 500,000 in the linear ion trap). The three most intense peptide ions with charge state higher than 1 were sequentially isolated to a target value of 10,000, fragmented in the linear ion trap by collision induced dissociation (CID). The pseudoMS^3 (pdMS^3) or multistage activation (MSA) was selected to automatically select and further fragment the fragment ion originating from the loss of one or two phosphate groups from the parent ion. For all measurements with the Orbitrap detector, 3 lock-mass ions from ambient air (m/z=371.10123, 445.12002, 519.13882) were used for internal calibration as described (Olsen 2005). Typical mass spectrometric conditions were: spray voltage, 1.5 kV; no sheath and auxiliary gas flow; heated capillary temperature, 200°C; normalized collision energy 35% for CID in LTQ. The ion selection threshold was 10,000 counts for MS2. An activation q = 0.25 and activation time of 30 ms were used. Mass spectrometry data was analyzed using Proteome Discoverer 1.3 (MS tol, 10ppm; MS/MS tol, 0.8Da, Variable modifications, Oxidation

(M) and Phosphorylation (S,T,Y); Fixed modifications, Carbamidomethyl (C); FDR Peptide, 0,01) using a DB including the sequences of the wild type and the mutated RNA polymerase II.

4.4.4.8 Data analysis software program

SEQUEST is a tandem mass spectrometry database search program originally developed in 1993 in the Yates lab at the University of Washington. It correlates tandem mass spectra of peptides against peptide sequences from a sequence database. Classical SEQUEST applies a two-stage scoring method for each search. The first stage applies the preliminary score to filter through all candidate peptides in the sequence database. The best scoring candidate peptides are then re-scored using the cross correlation algorithm. The sensitivity of the cross correlation algorithm is enhanced by the correction factor that is applied in its calculation (copied from http://proteomicsresource.washington.edu/sequest.php).

5. Bibliography

A

Akhtar, M. S., M. Heidemann, et al. (2009). "TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II." Mol Cell 34(3): 387-93.

Allison, L. A., M. Moyle, et al. (1985). "Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases." Cell 42(2): 599-610.

Allison, L. A., J. K. Wong, et al. (1988). "The C-terminal domain of the largest subunit of RNA polymerase II of Saccharomyces cerevisiae, Drosophila melanogaster, and mammals: a conserved structure with an essential function." Mol Cell Biol 8(1): 321-9.

Ansari, A. and M. Hampsey (2005). "A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping." Genes Dev 19(24): 2969-78.

Archambault, J., R. S. Chambers, et al. (1997). "An essential component of a C-terminal domain phosphatase that interacts with transcription factor IIF in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A 94(26): 14300-5.

Arigo, J. T., K. L. Carroll, et al. (2006). "Regulation of yeast NRD1 expression by premature transcription termination." Mol Cell 21(5): 641-51.

B

Baillat, D., M. A. Hakimi, et al. (2005). "Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II." Cell 123(2): 265-76.

Barilla, D., B. A. Lee, et al. (2001). "Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in Saccharomyces cerevisiae." Proc Natl Acad Sci U S A 98(2): 445-50.

Bartkowiak, B. and A. L. Greenleaf "Phosphorylation of RNAPII: To P-TEFb or not to P-TEFb?" Transcription 2(3): 115-119.

Bartkowiak, B., P. Liu, et al. "CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1." Genes Dev 24(20): 2303-16.

Bartolomei, M. S. and J. L. Corden (1987). "Localization of an alpha-amanitin resistance mutation in the gene encoding the largest subunit of mouse RNA polymerase II." Mol Cell Biol 7(2): 586-94.

Bartolomei, M. S., N. F. Halden, et al. (1988). "Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II." Mol Cell Biol 8(1): 330-9.

Baskaran, R., G. G. Chiang, et al. (1997). "Tyrosine phosphorylation of RNA polymerase II carboxyl-terminal domain by the Abl-related gene product." J Biol Chem 272(30): 18905-9.

Baskaran, R., M. E. Dahmus, et al. (1993). "Tyrosine phosphorylation of mammalian RNA polymerase II carboxyl-terminal domain." Proc Natl Acad Sci U S A 90(23): 11167-71.

Baskaran, R., S. R. Escobar, et al. (1999). "Nuclear c-Abl is a COOH-terminal repeated domain (CTD)-tyrosine (CTD)-tyrosine kinase-specific for the mammalian RNA polymerase II: possible role in transcription elongation." Cell Growth Differ 10(6): 387-96.

Bataille, A. R., C. Jeronimo, et al. "A universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes." Mol Cell 45(2): 158-70.

Beaudenon, S. L., M. R. Huacani, et al. (1999). "Rsp5 ubiquitin-protein ligase mediates DNA damage-induced degradation of the large subunit of RNA polymerase II in Saccharomyces cerevisiae." Mol Cell Biol 19(10): 6972-9.

Becker, R., B. Loll, et al. (2008). "Snapshots of the RNA processing factor SCAF8 bound to different phosphorylated forms of the carboxyl-terminal domain of RNA polymerase II." J Biol Chem 283(33): 22659-69.

Bensaude, O., F. Bonnet, et al. (1999). "Regulated phosphorylation of the RNA polymerase II C-terminal domain (CTD)." Biochem Cell Biol 77(4): 249-55.

Biemann, K. (1988). "Contributions of mass spectrometry to peptide and protein structure." Biomed Environ Mass Spectrom 16(1-12): 99-111.

Birse, C. E., L. Minvielle-Sebastia, et al. (1998). "Coupling termination of transcription to messenger RNA maturation in yeast." Science 280(5361): 298-301.

Blazek, D., J. Kohoutek, et al. "The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes." Genes Dev 25(20): 2158-72.

Bodenmiller, B., L. N. Mueller, et al. (2007). "Reproducible isolation of distinct, overlapping segments of the phosphoproteome." Nat Methods 4(3): 231-7.

Boeing, S., C. Rigault, et al. "RNA polymerase II C-terminal heptarepeat domain Ser-7 phosphorylation is established in a mediator-dependent fashion." J Biol Chem 285(1): 188-96.

Boersema, P. J., S. Mohammed, et al. (2009). "Phosphopeptide fragmentation and analysis by mass spectrometry." J Mass Spectrom 44(6): 861-78.

Bregman, D. B., R. G. Pestell, et al. (2000). "Cell cycle regulation and RNA polymerase II." Front Biosci 5: D244-57.

Bres, V., S. M. Yoh, et al. (2008). "The multi-tasking P-TEFb complex." Curr Opin Cell Biol 20(3): 334-40.

Brookes, E., I. de Santiago, et al. "Polycomb associates genome-wide with a specific RNA polymerase II variant, and regulates metabolic genes in ESCs." Cell Stem Cell 10(2): 157-70.

Bucheli, M. E. and S. Buratowski (2005). "Npl3 is an antagonist of mRNA 3' end formation by RNA polymerase II." Embo J 24(12): 2150-60.

Bucheli, M. E., X. He, et al. (2007). "Polyadenylation site choice in yeast is affected by competition between Npl3 and polyadenylation factor CFI." Rna 13(10): 1756-64.

Buratowski, S. (2003). "The CTD code." Nat Struct Biol 10(9): 679-80.

Buratowski, S. (2005). "Connections between mRNA 3' end processing and transcription termination." Curr Opin Cell Biol 17(3): 257-61.

Buratowski, S. (2009). "Progression through the RNA polymerase II CTD cycle." Mol Cell 36(4): 541-6.

C

Cagas, P. M. and J. L. Corden (1995). "Structural studies of a synthetic peptide derived from the carboxyl-terminal domain of RNA polymerase II." Proteins 21(2): 149-60.

Carrozza, M. J., B. Li, et al. (2005). "Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription." Cell 123(4): 581-92.

Chapman, R. D., M. Conrad, et al. (2005). "Role of the mammalian RNA polymerase II C-terminal domain (CTD) nonconsensus repeats in CTD stability and cell proliferation." Mol Cell Biol 25(17): 7665-74.

Chapman, R. D., M. Heidemann, et al. (2007). "Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7." Science 318(5857): 1780-2.

Chapman, R. D., M. Heidemann, et al. (2008). "Molecular evolution of the RNA polymerase II CTD." Trends Genet 24(6): 289-96.

Chapman, R. D., B. Palancade, et al. (2004). "The last CTD repeat of the mammalian RNA polymerase II large subunit is important for its stability." Nucleic Acids Res 32(1): 35-44.

Cho, E. J., M. S. Kobor, et al. (2001). "Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain." Genes Dev 15(24): 3319-29.

Cho, E. J., T. Takagi, et al. (1997). "mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain." Genes Dev 11(24): 3319-26.

Cho, H., T. K. Kim, et al. (1999). "A protein phosphatase functions to recycle RNA polymerase II." Genes Dev 13(12): 1540-52.

Clemente-Blanco, A., N. Sen, et al. "Cdc14 phosphatase promotes segregation of telomeres through repression of RNA polymerase II transcription." Nat Cell Biol 13(12): 1450-6.

Comer, F. I. and G. W. Hart (2001). "Reciprocity between O-GlcNAc and O-phosphate on the carboxyl terminal domain of RNA polymerase II." Biochemistry 40(26): 7845-52.

Connelly, S. and J. L. Manley (1988). "A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II." Genes Dev 2(4): 440-52.

Corden, J. L. (1990). "Tails of RNA polymerase II." Trends Biochem Sci 15(10): 383-7.

Corden, J. L., D. L. Cadena, et al. (1985). "A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II." Proc Natl Acad Sci U S A 82(23): 7934-8.

Core, L. J. and J. T. Lis (2008). "Transcription regulation through promoter-proximal pausing of RNA polymerase II." Science 319(5871): 1791-2.

Coudreuse, D., H. van Bakel, et al. "A gene-specific requirement of RNA polymerase II CTD phosphorylation for sexual differentiation in S. pombe." Curr Biol 20(12): 1053-64.

Cramer, P., K. J. Armache, et al. (2008). "Structure of eukaryotic RNA polymerases." Annu Rev Biophys 37: 337-52.

Cramer, P., D. A. Bushnell, et al. (2001). "Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution." Science 292(5523): 1863-76.

Cramer, P., A. Srebrow, et al. (2001). "Coordination between transcription and pre-mRNA processing." FEBS Lett 498(2-3): 179-82.

Custodio, N., M. Vivo, et al. (2007). "Splicing- and cleavage-independent requirement of RNA polymerase II CTD for mRNA release from the transcription site." J Cell Biol 179(2): 199-207.

D

Dahmus, M. E. (1981). "Phosphorylation of eukaryotic DNA-dependent RNA polymerase. Identification of calf thymus RNA polymerase subunits phosphorylated by two purified protein kinases, correlation with in vivo sites of phosphorylation in HeLa cell RNA polymerase II." J Biol Chem 256(7): 3332-9.

Dahmus, M. E. (1996). "Phosphorylation of mammalian RNA polymerase II." Methods Enzymol 273: 185-93.

Dahmus, M. E. (1996). "Reversible phosphorylation of the C-terminal domain of RNA polymerase II." J Biol Chem 271(32): 19009-12.

Daulny, A., F. Geng, et al. (2008). "Modulation of RNA polymerase II subunit composition by ubiquitylation." Proc Natl Acad Sci U S A 105(50): 19649-54.

David, C. J. and J. L. Manley "The RNA polymerase C-terminal domain: a new role in spliceosome assembly." Transcription 2(5): 221-5.

de la Mata, M. and A. R. Kornblihtt (2006). "RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20." Nat Struct Mol Biol 13(11): 973-80.

Devaiah, B. N., B. A. Lewis, et al. "BRD4 is an atypical kinase that phosphorylates serine2 of the RNA polymerase II carboxy-terminal domain." Proc Natl Acad Sci U S A 109(18): 6927-32.

Dichtl, B., D. Blank, et al. (2002). "A role for SSU72 in balancing RNA polymerase II transcription elongation and termination." Mol Cell 10(5): 1139-50.

Dichtl, B., D. Blank, et al. (2002). "Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination." Embo J 21(15): 4125-35.

Dieci, G., G. Fiorino, et al. (2007). "The expanding RNA polymerase III transcriptome." Trends Genet 23(12): 614-22.

Drouin, S., L. Laramee, et al. "DSIF and RNA polymerase II CTD phosphorylation coordinate the recruitment of Rpd3S to actively transcribed genes." PLoS Genet 6(10): e1001173.

E

Egloff, S. and S. Murphy (2008). "Cracking the RNA polymerase II CTD code." Trends Genet 24(6): 280-8.

Egloff, S., D. O'Reilly, et al. (2007). "Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression." Science 318(5857): 1777-9.

Egloff, S., S. A. Szczepaniak, et al. "The integrator complex recognizes a new double mark on the RNA polymerase II carboxyl-terminal domain." J Biol Chem 285(27): 20564-9.

Egloff, S., J. Zaborowska, et al. "Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes." Mol Cell 45(1): 111-22.

F

Fabrega, C., V. Shen, et al. (2003). "Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II." Mol Cell 11(6): 1549-61.

Feaver, W. J., J. Q. Svejstrup, et al. (1994). "Relationship of CDK-activating kinase and RNA polymerase II CTD kinase TFIIH/TFIIK." Cell 79(6): 1103-9.

Fong, N. and D. L. Bentley (2001). "Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD." Genes Dev 15(14): 1783-95.

Fuchs, S. M., R. N. Laribee, et al. (2009). "Protein modifications in transcription elongation." Biochim Biophys Acta 1789(1): 26-36.

G

Ganem, C., F. Devaux, et al. (2003). "Ssu72 is a phosphatase essential for transcription termination of snoRNAs and specific mRNAs in yeast." Embo J 22(7): 1588-98.

Gebara, M. M., M. H. Sayre, et al. (1997). "Phosphorylation of the carboxy-terminal repeat domain in RNA polymerase II by cyclin-dependent kinases is sufficient to inhibit transcription." J Cell Biochem 64(3): 390-402.

Ghazy, M. A., X. He, et al. (2009). "The essential N terminus of the Pta1 scaffold protein is required for snoRNA transcription termination and Ssu72 function but is dispensable for pre-mRNA 3'-end processing." Mol Cell Biol 29(8): 2296-307.

Ghosh, A., S. Shuman, et al. "Structural insights to how mammalian capping enzyme reads the CTD code." Mol Cell 43(2): 299-310.

Ghosh, A., S. Shuman, et al. (2008). "The structure of Fcp1, an essential RNA polymerase II CTD phosphatase." Mol Cell 32(4): 478-90.

Gilbert, W., C. W. Siebel, et al. (2001). "Phosphorylation by Sky1p promotes Npl3p shuttling and mRNA dissociation." Rna 7(2): 302-13.

Glover-Cutter, K., S. Larochelle, et al. (2009). "TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II." Mol Cell Biol 29(20): 5455-64.

Govind, C. K., H. Qiu, et al. "Phosphorylated Pol II CTD recruits multiple HDACs, including Rpd3C(S), for methylation-dependent deacetylation of ORF nucleosomes." Mol Cell 39(2): 234-46.

Grummt, I. and C. S. Pikaard (2003). "Epigenetic silencing of RNA polymerase I transcription." Nat Rev Mol Cell Biol 4(8): 641-9.

Gudipati, R. K., T. Villa, et al. (2008). "Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice." Nat Struct Mol Biol 15(8): 786-94.

Guo, Z. and J. W. Stiller (2004). "Comparative genomics of cyclin-dependent kinases suggest co-evolution of the RNAP II C-terminal domain and CTD-directed CDKs." BMC Genomics 5: 69.

H

Hausmann, S., H. Erdjument-Bromage, et al. (2004). "Schizosaccharomyces pombe carboxyl-terminal domain (CTD) phosphatase Fcp1: distributive mechanism, minimal CTD substrate, and active site mapping." J Biol Chem 279(12): 10892-900.

Hausmann, S. and S. Shuman (2002). "Characterization of the CTD phosphatase Fcp1 from fission yeast. Preferential dephosphorylation of serine 2 versus serine 5." J Biol Chem 277(24): 21213-20.

Heidemann, M., C. Hintermair, et al. "Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription." Biochim Biophys Acta 1829(1): 55-62.

Heine, G. F., A. A. Horwitz, et al. (2008). "Multiple mechanisms contribute to inhibit transcription in response to DNA damage." J Biol Chem 283(15): 9555-61.

Hengartner, C. J., V. E. Myer, et al. (1998). "Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases." Mol Cell 2(1): 43-53.

Hintermair, C., M. Heidemann, et al. "Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation." Embo J 31(12): 2784-97.

Hsin, J. P. and J. L. Manley "The RNA polymerase II CTD coordinates transcription and RNA processing." Genes Dev 26(19): 2119-37.

Hsin, J. P., A. Sheth, et al. "RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3' end processing." Science 334(6056): 683-6.

Huibregtse, J. M., J. C. Yang, et al. (1997). "The large subunit of RNA polymerase II is a substrate of the Rsp5 ubiquitin-protein ligase." Proc Natl Acad Sci U S A 94(8): 3656-61.

J

Jani, D., S. Lutz, et al. (2009). "Sus1, Cdc31, and the Sac3 CID region form a conserved interaction platform that promotes nuclear pore association and mRNA export." Mol Cell 33(6): 727-37.

Jasnovidova, O. and R. Stefl "The CTD code of RNA polymerase II: a structural view." Wiley Interdiscip Rev RNA 4(1): 1-16.

Jenuwein, T. and C. D. Allis (2001). "Translating the histone code." Science 293(5532): 1074-80.

K

Kanagaraj, R., D. Huehn, et al. "RECQ5 helicase associates with the C-terminal repeat domain of RNA polymerase II during productive elongation phase of transcription." Nucleic Acids Res 38(22): 8131-40.

Kanin, E. I., R. T. Kipp, et al. (2007). "Chemical inhibition of the TFIIH-associated kinase Cdk7/Kin28 does not impair global mRNA synthesis." Proc Natl Acad Sci U S A 104(14): 5812-7.

Kelly, W. G., M. E. Dahmus, et al. (1993). "RNA polymerase II is a glycoprotein. Modification of the COOH-terminal domain by O-GlcNAc." J Biol Chem 268(14): 10416-24.

Keogh, M. C., S. K. Kurdistani, et al. (2005). "Cotranscriptional set2 methylation of histone H3 lysine 36 recruits a repressive Rpd3 complex." Cell 123(4): 593-605.

Keogh, M. C., V. Podolny, et al. (2003). "Bur1 kinase is required for efficient transcription elongation by RNA polymerase II." Mol Cell Biol 23(19): 7005-18.

Kim, H., B. Erickson, et al. "Gene-specific RNA polymerase II phosphorylation and the CTD code." Nat Struct Mol Biol 17(10): 1279-86.

Kim, M., S. H. Ahn, et al. (2004). "Transitions in RNA polymerase II elongation complexes at the 3' ends of genes." Embo J 23(2): 354-64.

Kim, M., N. J. Krogan, et al. (2004). "The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II." Nature 432(7016): 517-22.

Kim, M., H. Suh, et al. (2009). "Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2, 5, and 7." J Biol Chem 284(39): 26421-6.

Kim, M., L. Vasiljeva, et al. (2006). "Distinct pathways for snoRNA and mRNA termination." Mol Cell 24(5): 723-34.

Kim, S., J. Yamamoto, et al. "Evidence that cleavage factor Im is a heterotetrameric protein complex controlling alternative polyadenylation." Genes Cells 15(9): 1003-13.

Kim, T. and S. Buratowski (2009). "Dimethylation of H3K4 by Set1 recruits the Set3 histone deacetylase complex to 5' transcribed regions." Cell 137(2): 259-72.

Kizer, K. O., H. P. Phatnani, et al. (2005). "A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation." Mol Cell Biol 25(8): 3305-16.

Kobor, M. S., J. Archambault, et al. (1999). "An unusual eukaryotic protein phosphatase required for transcription by RNA polymerase II and CTD dephosphorylation in S. cerevisiae." Mol Cell 4(1): 55-62.

Koch, F., R. Fenouil, et al. "Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters." Nat Struct Mol Biol 18(8): 956-63.

Kohoutek, J. and D. Blazek "Cyclin K goes with Cdk12 and Cdk13." Cell Div 7: 12.

Kong, S. E., M. S. Kobor, et al. (2005). "Interaction of Fcp1 phosphatase with elongating RNA polymerase II holoenzyme, enzymatic mechanism of action, and genetic interaction with elongator." J Biol Chem 280(6): 4299-306.

Kornberg, R. D. (2005). "Mediator and the mechanism of transcriptional activation." Trends Biochem Sci 30(5): 235-9.

Krishnamurthy, S., M. A. Ghazy, et al. (2009). "Functional interaction of the Ess1 prolyl isomerase with components of the RNA polymerase II initiation and termination machineries." Mol Cell Biol 29(11): 2925-34.

Krishnamurthy, S., X. He, et al. (2004). "Ssu72 Is an RNA polymerase II CTD phosphatase." Mol Cell 14(3): 387-94.

Krogan, N. J., M. Kim, et al. (2003). "Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II." Mol Cell Biol 23(12): 4207-18.

Kubicek, K., H. Cerna, et al. "Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1." Genes Dev 26(17): 1891-6.

Kuenzel, E. A., J. A. Mulligan, et al. (1987). "Substrate specificity determinants for casein kinase II as deduced from studies with synthetic peptides." J Biol Chem 262(19): 9136-40.

L

Larsen, M. R., T. E. Thingholm, et al. (2005). "Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns." Mol Cell Proteomics 4(7): 873-86.

Li, H., Z. Zhang, et al. (2007). "Wwp2-mediated ubiquitination of the RNA polymerase II large subunit in mouse embryonic pluripotent stem cells." Mol Cell Biol 27(15): 5296-305.

Li, J., D. Moazed, et al. (2002). "Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation." J Biol Chem 277(51): 49383-8.

Licatalosi, D. D., G. Geiger, et al. (2002). "Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II." Mol Cell 9(5): 1101-11.

Litingtung, Y., A. M. Lawler, et al. (1999). "Growth retardation and neonatal lethality in mice with a homozygous deletion in the C-terminal domain of RNA polymerase II." Mol Gen Genet 261(1): 100-5.

Liu, P., A. L. Greenleaf, et al. (2008). "The essential sequence elements required for RNAP II carboxyl-terminal domain function in yeast and their evolutionary conservation." Mol Biol Evol 25(4): 719-27.

Liu, P., J. M. Kenney, et al. "Genetic organization, length conservation, and evolution of RNA polymerase II carboxyl-terminal domain." Mol Biol Evol 27(11): 2628-41.

Liu, Y., C. Kung, et al. (2004). "Two cyclin-dependent kinases promote RNA polymerase II transcription and formation of the scaffold complex." Mol Cell Biol 24(4): 1721-35.

Lu, K. P., G. Finn, et al. (2007). "Prolyl cis-trans isomerization as a molecular timer." Nat Chem Biol 3(10): 619-29.

Lunde, B. M., S. L. Reichow, et al. "Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain." Nat Struct Mol Biol 17(10): 1195-201.

Lykke-Andersen, S. and T. H. Jensen (2007). "Overlapping pathways dictate termination of RNA polymerase II transcription." Biochimie 89(10): 1177-82.

M

MacKellar, A. L. and A. L. Greenleaf "Cotranscriptional association of mRNA export factor Yra1 with C-terminal domain of RNA polymerase II." J Biol Chem 286(42): 36385-95.

Malik, S. and R. G. Roeder (2005). "Dynamic regulation of pol II transcription by the mammalian Mediator complex." Trends Biochem Sci 30(5): 256-63.

Mamone, G., G. Picariello, et al. "Hydroxyapatite affinity chromatography for the highly selective enrichment of mono- and multi-phosphorylated peptides in phosphoproteome analysis." Proteomics 10(3): 380-93.

Mayya, V. and K. H. D (2006a). "Proteomic applications of protein quantification by isotope-dilution mass spectrometry." Expert Rev Proteomics 3(6): 597-610.

Mayya, V., K. Rezual, et al. (2006b). "Absolute quantification of multisite phosphorylation by selective reaction monitoring mass spectrometry: determination of inhibitory phosphorylation status of cyclin-dependent kinases." Mol Cell Proteomics 5(6): 1146-57.

Mayer, A., M. Heidemann, et al. "CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II." Science 336(6089): 1723-5.

Mayer, A., M. Lidschreiber, et al. "Uniform transitions of the general RNA polymerase II transcription complex." Nat Struct Mol Biol 17(10): 1272-8.

McCracken, S., N. Fong, et al. (1997). "The C-terminal domain of RNA polymerase II couples mRNA processing to transcription." Nature 385(6614): 357-61.

Meinhart, A. and P. Cramer (2004). "Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors." Nature 430(6996): 223-6.

Meininghaus, M., R. D. Chapman, et al. (2000). "Conditional expression of RNA polymerase II in mammalian cells. Deletion of the carboxyl-terminal domain of the large subunit affects early steps in transcription." J Biol Chem 275(32): 24375-82.

Minvielle-Sebastia, L., P. J. Preker, et al. (1994). "RNA14 and RNA15 proteins as components of a yeast pre-mRNA 3'-end processing factor." Science 266(5191): 1702-5.

Mortillaro, M. J., B. J. Blencowe, et al. (1996). "A hyperphosphorylated form of the large subunit of RNA polymerase II is associated with splicing complexes and the nuclear matrix." Proc Natl Acad Sci U S A 93(16): 8253-7.

Mosley, A. L., S. G. Pattenden, et al. (2009). "Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation." Mol Cell 34(2): 168-78.

Munoz, M. J., M. de la Mata, et al. "The carboxy terminal domain of RNA polymerase II and alternative splicing." Trends Biochem Sci 35(9): 497-504.

Muse, G. W., D. A. Gilchrist, et al. (2007). "RNA polymerase is poised for activation across the genome." Nat Genet 39(12): 1507-11.

Myers, L. C. and R. D. Kornberg (2000). "Mediator of transcriptional regulation." Annu Rev Biochem 69: 729-49.

N

Nakanishi, S., B. W. Sanderson, et al. (2008). "A comprehensive library of histone mutants identifies nucleosomal residues required for H3K4 methylation." Nat Struct Mol Biol 15(8): 881-8.

Nechaev, S. and K. Adelman "Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation." Biochim Biophys Acta 1809(1): 34-45.

Nedea, E., X. He, et al. (2003). "Organization and function of APT, a subcomplex of the yeast cleavage and polyadenylation factor involved in the formation of mRNA and small nucleolar RNA 3'-ends." J Biol Chem 278(35): 33000-10.

Neil, H., C. Malabat, et al. (2009). "Widespread bidirectional promoters are the major source of cryptic transcripts in yeast." Nature 457(7232): 1038-42.

Ng, H. H., F. Robert, et al. (2003). "Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity." Mol Cell 11(3): 709-19.

Ni, Z., J. B. Olsen, et al. "Control of the RNA polymerase II phosphorylation state in promoter regions by CTD interaction domain-containing proteins RPRD1A and RPRD1B." Transcription 2(5): 237-42.

Noble, C. G., D. Hollingworth, et al. (2005). "Key features of the interaction between Pcf11 CID and RNA polymerase II CTD." Nat Struct Mol Biol 12(2): 144-51.

O

O'Sullivan, J. M., S. M. Tan-Wong, et al. (2004). "Gene loops juxtapose promoters and terminators in yeast." Nat Genet 36(9): 1014-8.

P

Palancade, B. and O. Bensaude (2003). "Investigating RNA polymerase II carboxyl-terminal domain (CTD) phosphorylation." Eur J Biochem 270(19): 3859-70.

Palumbo, A. M., S. A. Smith, et al. "Tandem mass spectrometry strategies for phosphoproteome analysis." Mass Spectrom Rev 30(4): 600-25.

Pascual-Garcia, P., C. K. Govind, et al. (2008). "Sus1 is recruited to coding regions and functions during transcription elongation in association with SAGA and TREX2." Genes Dev 22(20): 2811-22.

Patturajan, M., X. Wei, et al. (1998). "A nuclear matrix protein interacts with the phosphorylated C-terminal domain of RNA polymerase II." Mol Cell Biol 18(4): 2406-15.

Payne, J. M. and M. E. Dahmus (1993). "Partial purification and characterization of two distinct protein kinases that differentially phosphorylate the carboxyl-terminal domain of RNA polymerase subunit IIa." J Biol Chem 268(1): 80-7.

Payne, J. M., P. J. Laybourn, et al. (1989). "The transition of RNA polymerase II from initiation to elongation is associated with phosphorylation of the carboxyl-terminal domain of subunit IIa." J Biol Chem 264(33): 19621-9.

Pei, Y., S. Hausmann, et al. (2001). "The length, phosphorylation state, and primary structure of the RNA polymerase II carboxyl-terminal domain dictate interactions with mRNA capping enzymes." J Biol Chem 276(30): 28075-82.

Peterlin, B. M. and D. H. Price (2006). "Controlling the elongation phase of transcription with P-TEFb." Mol Cell 23(3): 297-305.

Phatnani, H. P. and A. L. Greenleaf (2006). "Phosphorylation and functions of the RNA polymerase II CTD." Genes Dev 20(21): 2922-36.

Phatnani, H. P., J. C. Jones, et al. (2004). "Expanding the functional repertoire of CTD kinase I and RNA polymerase II: novel phosphoCTD-associating proteins in the yeast proteome." Biochemistry 43(50): 15702-19.

Pinna, L. A. (1990). "Casein kinase 2: an 'eminence grise' in cellular regulation?" Biochim Biophys Acta 1054(3): 267-84.

Proudfoot, N. J. (1989). "How RNA polymerase II terminates transcription in higher eukaryotes." Trends Biochem Sci 14(3): 105-10.

Q
Qiu, H., C. Hu, et al. (2009). "Phosphorylation of the Pol II CTD by KIN28 enhances BUR1/BUR2 recruitment and Ser2 CTD phosphorylation near promoters." Mol Cell 33(6): 752-62.

R
Rahman, S., M. E. Sowa, et al. "The Brd4 extraterminal domain confers transcription activation independent of pTEFb by recruiting multiple proteins, including NSD3." Mol Cell Biol 31(13): 2641-52.

Ranuncolo, S. M., S. Ghosh, et al. "Evidence of the involvement of O-GlcNAc-modified human RNA polymerase II CTD in transcription in vitro and in vivo." J Biol Chem 287(28): 23549-61.

Richard, P. and J. L. Manley (2009). "Transcription termination by nuclear RNA polymerases." Genes Dev 23(11): 1247-69.

Rickert, P., J. L. Corden, et al. (1999). "Cyclin C/CDK8 and cyclin H/CDK7/p36 are biochemically distinct CTD kinases." Oncogene 18(4): 1093-102.

Roepstorff, P. and J. Fohlman (1984). "Proposal for a common nomenclature for sequence ions in mass spectra of peptides." Biomed Mass Spectrom 11(11): 601.

Rosonina, E. and B. J. Blencowe (2004). "Analysis of the requirement for RNA polymerase II CTD heptapeptide repeats in pre-mRNA splicing and 3'-end cleavage." Rna 10(4): 581-9.

Russell, J. and J. C. Zomerdijk (2005). "RNA-polymerase-I-directed rDNA transcription, life and works." Trends Biochem Sci 30(2): 87-96.

S
Sadowski, M., B. Dichtl, et al. (2003). "Independent functions of yeast Pcf11p in pre-mRNA 3' end processing and in transcription termination." Embo J 22(9): 2167-77.

Schwer, B., A. M. Sanchez, et al. "Punctuation and syntax of the RNA polymerase II CTD code in fission yeast." Proc Natl Acad Sci U S A 109(44): 18024-9.

Schwer, B. and S. Shuman "Deciphering the RNA polymerase II CTD code in fission yeast." Mol Cell 43(2): 311-8.

Shaw, P. E. (2002). "Peptidyl-prolyl isomerases: a new twist to transcription." EMBO Rep 3(6): 521-6.

Shaw, P. E. (2007). "Peptidyl-prolyl cis/trans isomerases and transcription: is there a twist in the tail?" EMBO Rep 8(1): 40-5.

Shi, Y., S. Chan, et al. (2009). "An up-close look at the pre-mRNA 3'-end processing complex." RNA Biol 6(5): 522-5.

Shpakovski, G. V., J. Acker, et al. (1995). "Four subunits that are shared by the three classes of RNA polymerase are functionally interchangeable between Homo sapiens and Saccharomyces cerevisiae." Mol Cell Biol 15(9): 4702-10.

Sims, R. J., 3rd, R. Belotserkovskaya, et al. (2004). "Elongation by RNA polymerase II: the short and long of it." Genes Dev 18(20): 2437-68.

Sims, R. J., 3rd, L. A. Rojas, et al. "The C-terminal domain of RNA polymerase II is modified by site-specific methylation." Science 332(6025): 99-103.

Singh, B. N., A. Ansari, et al. (2009). "Detection of gene loops by 3C in yeast." Methods 48(4): 361-7.

Singh, B. N. and M. Hampsey (2007). "A transcription-independent role for TFIIB in gene looping." Mol Cell 27(5): 806-16.

Singh, N., Z. Ma, et al. (2009). "The Ess1 prolyl isomerase is required for transcription termination of small noncoding RNAs via the Nrd1 pathway." Mol Cell 36(2): 255-66.

Steen, H. and M. Mann (2004). "The ABC's (and XYZ's) of peptide sequencing." Nat Rev Mol Cell Biol 5(9): 699-711.

Steen, H., J. A. Jebanathirajah, et al. (2005). "Stable isotope-free relative and absolute quantitation of protein phosphorylation stoichiometry by MS." Proc Natl Acad Sci U S A 102(11): 3948-53.

Steinmetz, E. J. and D. A. Brow (1998). "Control of pre-mRNA accumulation by the essential yeast protein Nrd1 requires high-affinity transcript binding and a domain implicated in RNA polymerase II association." Proc Natl Acad Sci U S A 95(12): 6699-704.

Steinmetz, E. J. and D. A. Brow (2003). "Ssu72 protein mediates both poly(A)-coupled and poly(A)-independent termination of RNA polymerase II transcription." Mol Cell Biol 23(18): 6339-49.

Steinmetz, E. J., N. K. Conrad, et al. (2001). "RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts." Nature 413(6853): 327-31.

Stewart, M. "Nuclear export of mRNA." Trends Biochem Sci 35(11): 609-17.

Stiller, J. W. and M. S. Cook (2004). "Functional unit of the RNA polymerase II C-terminal domain lies within heptapeptide pairs." Eukaryot Cell 3(3): 735-40.

Suzuki, M. (1989). "SPXX, a frequent sequence motif in gene regulatory proteins." J Mol Biol 207(1): 61-84.

Svejstrup, J. Q., Y. Li, et al. (1997). "Evidence for a mediator cycle at the initiation of transcription." Proc Natl Acad Sci U S A 94(12): 6075-8.

T

Terzi, N., L. S. Churchman, et al. "H3K4 trimethylation by Set1 promotes efficient termination by the Nrd1-Nab3-Sen1 pathway." Mol Cell Biol 31(17): 3569-83.

Thiebaut, M., E. Kisseleva-Romanova, et al. (2006). "Transcription termination and nuclear degradation of cryptic unstable transcripts: a role for the nrd1-nab3 pathway in genome surveillance." Mol Cell 23(6): 853-64.

Thingholm, T. E., O. N. Jensen, et al. (2009). "Analytical strategies for phosphoproteomics." Proteomics 9(6): 1451-68.

Thingholm, T. E., M. R. Larsen, et al. (2008). "TiO(2)-based phosphoproteomic analysis of the plasma membrane and the effects of phosphatase inhibitor treatment." J Proteome Res 7(8): 3304-13.

Tietjen, J. R., D. W. Zhang, et al. "Chemical-genomic dissection of the CTD code." Nat Struct Mol Biol 17(9): 1154-61.

Tisseur, M., M. Kwapisz, et al. "Pervasive transcription - Lessons from yeast." Biochimie 93(11): 1889-96.

V

Vasiljeva, L., M. Kim, et al. (2008). "The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain." Nat Struct Mol Biol 15(8): 795-804.

Venters, B. J. and B. F. Pugh (2009). "How eukaryotic genes are transcribed." Crit Rev Biochem Mol Biol 44(2-3): 117-41.

Verdecia, M. A., M. E. Bowman, et al. (2000). "Structural basis for phosphoserine-proline recognition by group IV WW domains." Nat Struct Biol 7(8): 639-43.

Viladevall, L., C. V. St Amour, et al. (2009). "TFIIH and P-TEFb coordinate transcription with capping enzyme recruitment at specific genes in fission yeast." Mol Cell 33(6): 738-51.

Vojnic, E., B. Simon, et al. (2006). "Structure and carboxyl-terminal domain (CTD) binding of the Set2 SRI domain that couples histone H3 Lys36 methylation to transcription." J Biol Chem 281(1): 13-5.

W

Werner, M., P. Thuriaux, et al. (2009). "Structure-function analysis of RNA polymerases I and III." Curr Opin Struct Biol 19(6): 740-5.

Werner-Allen, J. W., C. J. Lee, et al. "cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72." J Biol Chem 286(7): 5717-26.

West, M. L. and J. L. Corden (1995). "Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations." Genetics 140(4): 1223-33.

West, S., N. Gromak, et al. (2004). "Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites." Nature 432(7016): 522-5.

Wood, A., J. Schneider, et al. (2003). "The Paf1 complex is essential for histone monoubiquitination by the Rad6-Bre1 complex, which signals for histone methylation by COMPASS and Dot1p." J Biol Chem 278(37): 34739-42.

Wood, A., J. Schneider, et al. (2005). "The Bur1/Bur2 complex is required for histone H2B monoubiquitination by Rad6/Bre1 and histone methylation by COMPASS." Mol Cell 20(4): 589-99.

Woychik, N. A. (1994). "Regulating the regulators." Trends Biochem Sci 19(3): 103-5.

X

Xiang, K., T. Nagaike, et al. "Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex." Nature 467(7316): 729-33.

Xu, Y. X., Y. Hirose, et al. (2003). "Pin1 modulates the structure and function of human RNA polymerase II." Genes Dev 17(22): 2765-76.

Y

Yeo, M., P. S. Lin, et al. (2003). "A novel RNA polymerase II C-terminal domain phosphatase that preferentially dephosphorylates serine 5." J Biol Chem 278(28): 26078-85.

Young, R. A. (1991). "RNA polymerase II." Annu Rev Biochem 60: 689-715.

Yuryev, A., M. Patturajan, et al. (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(14): 6975-80.

Z

Zeidan, Q., Z. Wang, et al. "O-GlcNAc cycling enzymes associate with the translational machinery and modify core ribosomal proteins." Mol Biol Cell 21(12): 1922-36.

Zhang, D. W., A. L. Mosley, et al. "Ssu72 phosphatase-dependent erasure of phospho-Ser7 marks on the RNA polymerase II C-terminal domain is essential for viability and transcription termination." J Biol Chem 287(11): 8541-51.

Zhang, D. W., J. B. Rodriguez-Molina, et al. "Emerging Views on the CTD Code." Genet Res Int 2012: 347214.

Zhang, J. and J. L. Corden (1991). "Phosphorylation causes a conformational change in the carboxyl-terminal domain of the mouse RNA polymerase II largest subunit." J Biol Chem 266(4): 2297-302.

Zhang, Y., Y. Kim, et al. (2006). "Determinants for dephosphorylation of the RNA polymerase II C-terminal domain by Scp1." Mol Cell 24(5): 759-70.

Zhang, Z., J. Fu, et al. (2005). "CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11." Genes Dev 19(13): 1572-80.

Zhang, Z. and D. S. Gilmour (2006). "Pcf11 is a termination factor in Drosophila that dismantles the elongation complex by bridging the CTD of RNA polymerase II to the nascent transcript." Mol Cell 21(1): 65-74.

Zhou, K., W. H. Kuo, et al. (2009). "Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5." Proc Natl Acad Sci U S A 106(17): 6956-61.

6. Appendix

A
Rpb1 CTD mutants with corresponding CTD peptide fragmentation pattern

```
>gi|6677795|ref|NP_033115.1|  DNA-directed  RNA  polymerase  II
subunit RPB1 [Mus musculus]M-6K
```

```
MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDPRQGVIERT
GRCQTCAGNMTECPGHFGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIKDILAKSKGQ
PKKRLTHVYDLCKGKNICEGGEEMDNKFGVEQPEGDEDLTKEKGHGGCGRYQPRIRRSGLELYAEWKH
VNEDSQEKKILLSPERVHEIFKRISDEECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGSARN
QDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFHVATMVDNELPGLPRAMQKSGRPLKSLK
QRLKGKEGRVRGNLMGKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFNIDRLQELVRR
GNSQYPGAKYIIRDNGDRIDLRFHPKPSDLHLQTGYKVERHMCDGDIVIFNRQPTLHKMSMMGHRVRI
LPWSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMVPRMIVTPQSNRPVMGIVQDTLTA
VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHSTHPDD
EDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYSNIQTVINNWL
LIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDA
RDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYG
PESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDATVRNSIN
QVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRFDYTNERALRRTLQEDLVKDVLSNAHIQNEL
EREFERMREDREVLRVIFPTGDSKVVLPCNLLRMIWNAQKIFHINPRLPSDLHPIKVVEGVKELSKKL
VIVNGDDPLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDWLLGEIESKFNQAIAHPGEMVG
ALAAQSLGEPATQMTLNTFHYAGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKD
ILCRLEHTTLRKVTANTAIYYDPNPQSTVVAEDQEWVNVYYEMPDFDVARISPWLLRVELDRKHMTDR
KLTMEQIAEKINAGFGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNM
LTDMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
DIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVNRQDTGPLM
KCSFEETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKYGMEIPTNIPGLGAA
GRSGMTPGAAGFSPSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPR
YSPTSPAYEPR
SPGGYTPQSPSYSPTSPSYSPTSPK
YSPTSPNYSPTSPSYSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPTSPSYSPASPSYSPTSPK
YSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPNYSPTSPNYTPTSPSYSPTSPK
YSPTSPNYTPTSPNYSPTSPSYSPTSPSYSPTSPSYSPSSPK
YTPQSPTYTPSSPSYSPSSPSYSPTSPK
YTPTSPSYSPSSPEYTPASPK
YSPTSPK
YTPTSPK
YSPTSPTYSPTTPK
YSPTSPTYSPTSPVYTPTSPK
YSPTSPTYSPTSPK
YSPTSPTYTPASPK
GSTYSPTSPGYSPTSPTYSLTSPAISPDDSDEEN
```

>gi|6677795|ref|NP_033115.1| DNA-directed RNA polymerase II
subnuit RPB1 [Mus musculus] **M-3K5R**

MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDPRQGVIERT
GRCQTCAGNMTECPGHFGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIKDILAKSKGQ
PKKRLTHVYDLCKGKNICEGGEEMDNKFGVEQPEGDEDLTKEKGHGGCGRYQPRIRRSGLELYAEWKH
VNEDSQEKKILLSPERVHEIFKRISDEECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGSARN
QDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFHVATMVDNELPGLPRAMQKSGRPLKSLK
QRLKGKEGRVRGNLMGKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFNIDRLQELVRR
GNSQYPGAKYIIRDNGDRIDLRFHPKPSDLHLQTGYKVERHMCDGDIVIFNRQPTLHKMSMMGHRVRI
LPWSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMVPRMIVTPQSNRPVMGIVQDTLTA
VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHSTHPDD
EDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYSNIQTVINNWL
LIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDA
RDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYG
PESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDATVRNSIN
QVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRFDYTNERALRRTLQEDLVKDVLSNAHIQNEL
EREFERMREDREVLRVIFPTGDSKVVLPCNLLRMIWNAQKIFHINPRLPSDLHPIKVVEGVKELSKKL
VIVNGDDPLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDWLLGEIESKFNQAIAHPGEMVG
ALAAQSLGEPATQMTLNTFHYAGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKD
ILCRLEHTTLRKVTANTAIYYDPNPQSTVVAEDQEWVNVYYEMPDFDVARISPWLLRVELDRKHMTDR
KLTMEQIAEKINAGFGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNM
LTDMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
DIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVNRQDTGPLM
KCSFEETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKYGMEIPTNIPGLGAA
GRSGMTPGAAGFSPSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPR
YSPTSPAYEPR
SPGGYTPQSPSYSPTSPR
YSPTSPSYSPTSPNYSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYTPTSPSYSPTSPSYSPTSPR
YSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPTSPNYSPTSPNYTPTSPR
YSPTSPSYSPTSPNYTPTSPNYSPTSPK
YSPTSPSYSPTSPSYSPSSPR
YTPQSPTYTPSSPSYSPSSPSYSPTSPK
YTPTSPSYSPSSPEYTPASPK
YSPTSPR
YSPTSPK
YSPTSPTYSPTTPK
YSPTSPTYSPTSPVYTPTSPK
YSPTSPTYSPTSPR
YSPTSPTYSPTSPK
GSTYSPTSPGYSPTSPTYSLTSPAISPDDSDEEN

>gi|6677795|ref|NP_033115.1| DNA-directed RNA polymerase II
subunit RPB1 [Mus musculus]**M-8K**

```
MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDPRQGVIERT
GRCQTCAGNMTECPGHFGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIKDILAKSKGQ
PKKRLTHVYDLCKGKNICEGGEEMDNKFGVEQPEGDEDLTKEKGHGGCGRYQPRIRRSGLELYAEWKH
VNEDSQEKKILLSPERVHEIFKRISDEECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGSARN
QDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFHVATMVDNELPGLPRAMQKSGRPLKSLK
QRLKGKEGRVRGNLMGKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFNIDRLQELVRR
GNSQYPGAKYIIRDNGDRIDLRFHPKPSDLHLQTGYKVERHMCDGDIVIFNRQPTLHKMSMMGHRVRI
LPWSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMVPRMIVTPQSNRPVMGIVQDTLTA
VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHSTHPDD
EDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYSNIQTVINNWL
LIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDA
RDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYG
PESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDATVRNSIN
QVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRFDYTNERALRRTLQEDLVKDVLSNAHIQNEL
EREFERMREDREVLRVIFPTGDSKVVLPCNLLRMIWNAQKIFHINPRLPSDLHPIKVVEGVKELSKKL
VIVNGDDPLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDWLLGEIESKFNQAIAHPGEMVG
ALAAQSLGEPATQMTLNTFHYAGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKD
ILCRLEHTTLRKVTANTAIYYDPNPQSTVVAEDQEWVNVYYEMPDFDVARISPWLLRVELDRKHMTDR
KLTMEQIAEKINAGFGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNM
LTDMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
DIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVNRQDTGPLM
KCSFEETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKYGMEIPTNIPGLGAA
GRSGMTPGAAGFSPSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPR
YSPTSPAYEPR
SPGGYTPQSPSYSPTSPK
YSPTSPSYSPTSPNYSPTSPSYSPTSPK
AYSPTSPSYSPTSPSYTPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPNYSPTSPNYTPTSPK
AYSPTSPSYSPTSPNYTPTSPNYSPTSPK
YSPTSPSYSPTSPSYSPSSPK
YTPQSPTYTPSSPSYSPSSPSYSPTSPK
YTPTSPSYSPSSPEYTPASPK
AYSPTSPK
YSPTSPK
YSPTSPTYSPTTPK
YSPTSPTYSPTSPVYTPTSPK
AYSPTSPTYSPTSPK
YSPTSPTYSPTSPK
GSTYSPTSPGYSPTSPTYSLTSPAISPDDSDEEN
```

>gi|6677795|ref|NP_033115.1| DNA-directed RNA polymerase II
subunit RPB1 [Mus musculus]**M-4K2R**

```
MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDPRQGVIERT
GRCQTCAGNMTECPGHFGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIKDILAKSKGQ
PKKRLTHVYDLCKGKNICEGGEEMDNKFGVEQPEGDEDLTKEKGHGGCGRYQPRIRRSGLELYAEWKH
VNEDSQEKKILLSPERVHEIFKRISDEECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGSARN
QDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFHVATMVDNELPGLPRAMQKSGRPLKSLK
QRLKGKEGRVRGNLMGKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFNIDRLQELVRR
GNSQYPGAKYIIRDNGDRIDLRFHPKPSDLHLQTGYKVERHMCDGDIVIFNRQPTLHKMSMMGHRVRI
LPWSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMVPRMIVTPQSNRPVMGIVQDTLTA
VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHSTHPDD
EDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYSNIQTVINNWL
LIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDA
RDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYG
PESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDATVRNSIN
QVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRFDYTNERALRRTLQEDLVKDVLSNAHIQNEL
EREFERMREDREVLRVIFPTGDSKVVLPCNLLRMIWNAQKIFHINPRLPSDLHPIKVVEGVKELSKKL
VIVNGDDPLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDWLLGEIESKFNQAIAHPGEMVG
ALAAQSLGEPATQMTLNTFHYAGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKD
ILCRLEHTTLRKVTANTAIYYDPNPQSTVVAEDQEWVNVYYEMPDFDVARISPWLLRVELDRKHMTDR
KLTMEQIAEKINAGFGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNM
LTDMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
DIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVNRQDTGPLM
KCSFEETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKYGMEIPTNIPGLGAA
GRSGMTPGAAGFSPSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPR
YSPTSPAYEPR
SPGGYTPQSPSYSPTSPSYSPTSPSYSPTSPNYSPTSPK
YSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPSSPSYSPTSPSYSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPTSPNYSPTSPNYTPTSPSYSPTSPK
YSPTSPNYTPTSPNYSPTSPSYSPTSPSYSPTSPSYSPSSPR
YTPQSPTYTPSSPSYSPSSPSYSPTSPK
YTPTSPSYSPSSPEYTPASPK
YSPTSPR
YSPTSPK
YSPTSPTYSPTTPK
YSPTSPTYSPTSPVYTPTSPK
YSPTSPTYSPTSPR
YSPTSPTYSPTSPK
GSTYSPTSPGYSPTSPTYSLTSPAISPDDSDEEN
```

>gi|6677795|ref|NP_033115.1| DNA-directed RNA polymerase II
subunit RPB1 [Mus musculus]**M-13K2R**

MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDPRQGVIERT
GRCQTCAGNMTECPGHFGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIKDILAKSKGQ
PKKRLTHVYDLCKGKNICEGGEEMDNKFGVEQPEGDEDLTKEKGHGGCGRYQPRIRRSGLELYAEWKH
VNEDSQEKKILLSPERVHEIFKRISDEECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGSARN
QDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFHVATMVDNELPGLPRAMQKSGRPLKSLK
QRLKGKEGRVRGNLMGKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFNIDRLQELVRR
GNSQYPGAKYIIRDNGDRIDLRFHPKPSDLHLQTGYKVERHMCDGDIVIFNRQPTLHKMSMMGHRVRI
LPWSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMVPRMIVTPQSNRPVMGIVQDTLTA
VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHSTHPDD
EDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYSNIQTVINNWL
LIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDA
RDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYG
PESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDATVRNSIN
QVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRFDYTNERALRRTLQEDLVKDVLSNAHIQNEL
EREFERMREDREVLRVIFPTGDSKVVLPCNLLRMIWNAQKIFHINPRLPSDLHPIKVVEGVKELSKKL
VIVNGDDPLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDWLLGEIESKFNQAIAHPGEMVG
ALAAQSLGEPATQMTLNTFHYAGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKD
ILCRLEHTTLRKVTANTAIYYDPNPQSTVVAEDQEWVNVYYEMPDFDVARISPWLLRVELDRKHMTDR
KLTMEQIAEKINAGFGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNM
LTDMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
DIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVNRQDTGPLM
KCSFEETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKYGMEIPTNIPGLGAA
GRSGMTPGAAGFSPSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPR
YSPTSPAYEPR
SPGGYTPQSPSYSPTSPR
AYSPTSPSYSPTSPK
YSPTSPSYSPTSPK
VYSPTSPSYSPTSPK
LYSPTSPSYSPTSPK
AAYSPTSPSYSPTSPK
AVYSPTSPSYSPTSPK
ALYSPTSPSYSPTSPK
VVYSPTSPSYSPTSPK
LLYSPTSPSYSPTSPK
LYSPTSPNYTPTSPK
LVYSPTSPSYSPTSPK
YTPTSPNYSPTSPK
YSPTSPSYSPTSPSYSPSSPR
YTPQSPTYTPSSPK
YSPSSPSYSPTSPK
YTPTSPSYSPSSPEYTPASPK
AALYSPTSPSYSPTSPK
YSPTSPTYSPTTPK
YSPTSPTYSPTSPVYTPTSPK
AYSPTSPTYSPTSPK
YSPTSPTYSPTSPK
GSTYSPTSPGYSPTSPR
YSLTSPAISPDDSDEEN

>gi|6677795|ref|NP_033115.1| DNA-directed RNA polymerase II
subunit RPB1 [Mus musculus]**M-9K4R**

```
MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDPRQGVIERT
GRCQTCAGNMTECPGHFGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIKDILAKSKGQ
PKKRLTHVYDLCKGKNICEGGEEMDNKFGVEQPEGDEDLTKEKGHGGCGRYQPRIRRSGLELYAEWKH
VNEDSQEKKILLSPERVHEIFKRISDEECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGSARN
QDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFHVATMVDNELPGLPRAMQKSGRPLKSLK
QRLKGKEGRVRGNLMGKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFNIDRLQELVRR
GNSQYPGAKYIIRDNGDRIDLRFHPKPSDLHLQTGYKVERHMCDGDIVIFNRQPTLHKMSMMGHRVRI
LPWSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMVPRMIVTPQSNRPVMGIVQDTLTA
VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHSTHPDD
EDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYSNIQTVINNWL
LIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDA
RDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYG
PESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDATVRNSIN
QVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRFDYTNERALRRTLQEDLVKDVLSNAHIQNEL
EREFERMREDREVLRVIFPTGDSKVVLPCNLLRMIWNAQKIFHINPRLPSDLHPIKVVEGVKELSKKL
VIVNGDDPLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDWLLGEIESKFNQAIAHPGEMVG
ALAAQSLGEPATQMTLNTFHYAGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKD
ILCRLEHTTLRKVTANTAIYYDPNPQSTVVAEDQEWVNVYYEMPDFDVARISPWLLRVELDRKHMTDR
KLTMEQIAEKINAGFGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNM
LTDMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
DIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVNRQDTGPLM
KCSFEETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKYGMEIPTNIPGLGAA
GRSGMTPGAAGFSPSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPR
YSPTSPAYEPR
SPGGYTPQSPSYSPTSPR
YSPTSPSYSPTSPNYSPTSPK
YSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPASPK
YSPTSPSYSPTSPSYSPSSPK
YSPTSPSYSPTSPSYSPLSPK
AYSPTSPSYSPTSPSYSPLSPR
YSPTSPNYTPTSPSYSPTSPK
YSPTSPNYTPTSPNYSPTSPK
YAPTSPSYSPASPSYAPSSPR
YTPQSPTYTPSSPK
YSPSSPSYSPTSPK
YTPTSPSYSPSSPEYTPTSPK
YSPTSPSYSPTSPK
YSPTSPTYSPTTPK
YSPTSPTYTPTSPVYTPTSPR
YSPTSPTYSPTSPK
YSPTSPTYSPTSPR
GSTYSPTSPGYSPTSPK
YSLTSPAISPDDSDEEN
```

>gi|6677795|ref|NP_033115.1| DNA-directed RNA polymerase II
subunit RPB1 [Mus musculus]**M-12K2R**

```
MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDPRQGVIERT
GRCQTCAGNMTECPGHFGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIKDILAKSKGQ
PKKRLTHVYDLCKGKNICEGGEEMDNKFGVEQPEGDEDLTKEKGHGGCGRYQPRIRRSGLELYAEWKH
VNEDSQEKKILLSPERVHEIFKRISDEECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGSARN
QDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFHVATMVDNELPGLPRAMQKSGRPLKSLK
QRLKGKEGRVRGNLMGKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFNIDRLQELVRR
GNSQYPGAKYIIRDNGDRIDLRFHPKPSDLHLQTGYKVERHMCDGDIVIFNRQPTLHKMSMMGHRVRI
LPWSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMVPRMIVTPQSNRPVMGIVQDTLTA
VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHSTHPDD
EDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYSNIQTVINNWL
LIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDA
RDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYG
PESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDATVRNSIN
QVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRFDYTNERALRRTLQEDLVKDVLSNAHIQNEL
EREFERMREDREVLRVIFPTGDSKVVLPCNLLRMIWNAQKIFHINPRLPSDLHPIKVVEGVKELSKKL
VIVNGDDPLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDWLLGEIESKFNQAIAHPGEMVG
ALAAQSLGEPATQMTLNTFHYAGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKD
ILCRLEHTTLRKVTANTAIYYDPNPQSTVVAEDQEWVNVYYEMPDFDVARISPWLLRVELDRKHMTDR
KLTMEQIAEKINAGFGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNM
LTDMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
DIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVNRQDTGPLM
KCSFEETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKYGMEIPTNIPGLGAA
GRSGMTPGAAGFSPSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPR
YSPTSPAYEPR
SPGGYTPQSPSYSPTSPR
YSPTSPSYSPTSPNYSPTSPK
YSPTSPSYSPTSPSYSPTSPK
AYSPTSPSYSPTSPK
AYSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPK
AAYSPTSPSYSPTSPSYSPTSPK
AYSPTSPSYSPTSPNYSPTSPK
YTPTSPSYSPTSPSYSPTSPK
AYTPTSPNYSPTSPSYSPTSPK
YSPTSPSYSPASPK
YTPQSPTYTPSSPK
YSPSSPSYSPTSPK
YTPTSPSYSPSSPEYTPASPR
AAYSPTSPSYSPTSPK
YSPTSPTYSPTTPK
YSPTSPTYSPTSPVYTPTSPK
AYSPTSPTYSPTSPK
YSPTSPTYSPTSPK
GSTYSPTSPGYSPTSPK
YSLTSPAISPDDSDEEN
```

>gi|6677795|ref|NP_033115.1| DNA-directed RNA polymerase II
subunit RPB1 [Mus musculus]**M-8K4R**

MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDPRQGVIERT
GRCQTCAGNMTECPGHFGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIKDILAKSKGQ
PKKRLTHVYDLCKGKNICEGGEEMDNKFGVEQPEGDEDLTKEKGHGGCGRYQPRIRRSGLELYAEWKH
VNEDSQEKKILLSPERVHEIFKRISDEECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGSARN
QDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFHVATMVDNELPGLPRAMQKSGRPLKSLK
QRLKGKEGRVRGNLMGKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFNIDRLQELVRR
GNSQYPGAKYIIRDNGDRIDLRFHPKPSDLHLQTGYKVERHMCDGDIVIFNRQPTLHKMSMMGHRVRI
LPWSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMVPRMIVTPQSNRPVMGIVQDTLTA
VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHSTHPDD
EDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYSNIQTVINNWL
LIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDA
RDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYG
PESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDATVRNSIN
QVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRFDYTNERALRRTLQEDLVKDVLSNAHIQNEL
EREFERMREDREVLRVIFPTGDSKVVLPCNLLRMIWNAQKIFHINPRLPSDLHPIKVVEGVKELSKKL
VIVNGDDPLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDWLLGEIESKFNQAIAHPGEMVG
ALAAQSLGEPATQMTLNTFHYAGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKD
ILCRLEHTTLRKVTANTAIYYDPNPQSTVVAEDQEWVNVYYEMPDFDVARISPWLLRVELDRKHMTDR
KLTMEQIAEKINAGFGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNM
LTDMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
DIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVNRQDTGPLM
KCSFEETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKYGMEIPTNIPGLGAA
GRSGMTPGAAGFSPSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPR
YSPTSPAYEPR
SPGGYTPQSPSYSPTSPK
YSPTSPSYSPTSPNYSPTSPK
YSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPASPK
YSPTSPSYSPTSPSYSPSSPK
AYSPTSPSYSPTSPSYSPLSPR
YSPTSPNYSPTSPNYTPTSPK
YSPTSPSYSPTSPNYTPTSPNYSPTSPK
YSPTSPSYSPTSPSYSPSSPR
YTPQSPTYTPSSPSYSPSSPSYSPTSPK
YTPTSPSYSPSSPEYTPASPR
YSPTSPK
YSPTSPR
YSPTSPTYSPTTPK
YSPTSPTYSPTSPVYTPTSPK
YSPTSPTYSPTSPR
YSPTSPTYSPTSPK
GSTYSPTSPGYSPTSPTYSLTSPAISPDDSDEEN

>gi|6677795|ref|NP_033115.1| DNA-directed RNA polymerase II
subunit RPB1 [Mus musculus]**M-9K2R**

MHGGGPPSGDSACPLRTIKRVQFGVLSPDELKRMSVTEGGIKYPETTEGGRPKLGGLMDPRQGVIERT
GRCQTCAGNMTECPGHFGHIELAKPVFHVGFLVKTMKVLRCVCFFCSKLLVDSNNPKIKDILAKSKGQ
PKKRLTHVYDLCKGKNICEGGEEMDNKFGVEQPEGDEDLTKEKGHGGCGRYQPRIRRSGLELYAEWKH
VNEDSQEKKILLSPERVHEIFKRISDEECFVLGMEPRYARPEWMIVTVLPVPPLSVRPAVVMQGSARN
QDDLTHKLADIVKINNQLRRNEQNGAAAHVIAEDVKLLQFHVATMVDNELPGLPRAMQKSGRPLKSLK
QRLKGKEGRVRGNLMGKRVDFSARTVITPDPNLSIDQVGVPRSIAANMTFAEIVTPFNIDRLQELVRR
GNSQYPGAKYIIRDNGDRIDLRFHPKPSDLHLQTGYKVERHMCDGDIVIFNRQPTLHKMSMMGHRVRI
LPWSTFRLNLSVTTPYNADFDGDEMNLHLPQSLETRAEIQELAMVPRMIVTPQSNRPVMGIVQDTLTA
VRKFTKRDVFLERGEVMNLLMFLSTWDGKVPQPAILKPRPLWTGKQIFSLIIPGHINCIRTHSTHPDD
EDSGPYKHISPGDTKVVVENGELIMGILCKKSLGTSAGSLVHISYLEMGHDITRLFYSNIQTVINNWL
LIEGHTIGIGDSIADSKTYQDIQNTIKKAKQDVIEVIEKAHNNELEPTPGNTLRQTFENQVNRILNDA
RDKTGSSAQKSLSEYNNFKSMVVSGAKGSKINISQVIAVVGQQNVEGKRIPFGFKHRTLPHFIKDDYG
PESRGFVENSYLAGLTPTEFFFHAMGGREGLIDTAVKTAETGYIQRRLIKSMESVMVKYDATVRNSIN
QVVQLRYGEDGLAGESVEFQNLATLKPSNKAFEKKFRFDYTNERALRRTLQEDLVKDVLSNAHIQNEL
EREFERMREDREVLRVIFPTGDSKVVLPCNLLRMIWNAQKIFHINPRLPSDLHPIKVVEGVKELSKKL
VIVNGDDPLSRQAQENATLLFNIHLRSTLCSRRMAEEFRLSGEAFDWLLGEIESKFNQAIAHPGEMVG
ALAAQSLGEPATQMTLNTFHYAGVSAKNVTLGVPRLKELINISKKPKTPSLTVFLLGQSARDAERAKD
ILCRLEHTTLRKVTANTAIYYDPNPQSTVVAEDQEWVNVYYEMPDFDVARISPWLLRVELDRKHMTDR
KLTMEQIAEKINAGFGDDLNCIFNDDNAEKLVLRIRIMNSDENKMQEEEEVVDKMDDDVFLRCIESNM
LTDMTLQGIEQISKVYMHLPQTDNKKKIIITEDGEFKALQEWILETDGVSLMRVLSEKDVDPVRTTSN
DIVEIFTVLGIEAVRKALERELYHVISFDGSYVNYRHLALLCDTMTCRGHLMAITRHGVNRQDTGPLM
KCSFEETVDVLMEAAAHGESDPMKGVSENIMLGQLAPAGTGCFDLLLDAEKCKYGMEIPTNIPGLGAA
GRSGMTPGAAGFSPSAASDASGFSPGYSPAWSPTPGSPGSPGPSSPYIPSPGGAMSPR
YSPTSPAYEPR
SPGGYTPQSPSYSPTSPK
YSPTSPSYSPTSPNYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPASPK
YSPTSPSYSPTSPSYSPTSPSYSPTSPK
YSPTSPSYSPTSPSYSPASPSYSPTSPK
AYSPTSPNYTPTSPSYSPTSPK
YSPTSPNYTPTSPNYSPTSPK
YSPTSPSYSPTSPSYSPSSPR
YTPQSPTYTPSSPK
YSPSSPSYSPTSPK
YTPTSPSYSPSSPEYTPASPK
YSPTSPR
YSPTSPK
YSPTSPTYSPTTPK
YSPTSPTYSPTSPVYTPTSPK
YSPTSPTYSPTSPR
YSPTSPTYSPTSPK
GSTYSPTSPGYSPTSPTYSLTSPAISPDDSDEEN

B *Curriculum vitae*

Roland Schüller
Born 30th of November, 1982 in Salzburg, Austria

School Education

1989 -1993	Volksschule Bad Vigaun, Salzburg

1993 - 2001   A-levels 2.0
                   Bundesrealgymnasium Hallein, Salzburg

2002 - 2008   Diploma 1.3
                   Faculty for Biology
                   Ludwig Maximilian University (LMU) of Munich,
                   Germany

Research Experience
2009 - 2013   Doctoral Thesis
                   Institute for Clinical Molecular Biology and Tumor
                   Genetics
                   Department of Molecular Epigenetics
                   Helmholtz Zentrum (HMGU), Germany
                   „Insights into the RNA Polymerase II CTD code"
                   Prof. Dr. Dirk Eick

C   Publications

Sims RJ 3rd, Rojas LA, Beck D, Bonasio R, **Schüller R**, Drury WJ 3rd, Eick D, Reinberg D. (2011) The C-terminal domain of RNA polymerase II is modified by site-specific methylation
Science. 2011, 332, 99-103.

D Acknowledgements

In the final section I would like to thank all the people who supported me and believed in me during my time as a PhD student.

Prof. Dr. Dirk Eick for his supervision, for always keeping me on the right way during this project, and for his inspiration and dedication towards science.

My colleagues, **Martin Heidemann, Corinna Hintermair, Kirsten Voss, Bastian Muehl, Kaspar Burger, Markus Kellner, Michaela Rohrmoser, Nilay Shah, Tim Decker and Anita Gruber-Eber** for good team work, for great lab atmosphere and for becoming friends.

Dr. Ignasi Forne, Dr. Tobias Straub, and Andrew Flatley for successful collaboration and fruitful discussion during this project.

Dr. Rob Chapman for introducing me into this project, and for all the tools he bequeathed me.

Dr. Marcus Conrad and his group, Ugur Hirlak, Harald Polzer, Diana Schmid, Verena Weth, Nadine Sandhoefer, Hanna Janke and many more, for all the great time spending together in the Helmholtz Zentrum.

My family who is the best family one can have.

My love, **Yolanda**, who makes me the most happiest man alive.