DISSERTATION ZUR ERLANGUNG DES DOKTORGRADES

DER FAKULTÄT FÜR BIOLOGIE

DER LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Functional analysis of Tyrosine-1 of mammalian RNA polymerase II CTD

NILAY SHAH

April 2017

Completed at the Helmholtz Center Munich

German Research Center for Environment and Health (GmbH)

Institute of Functional Epigenetics

Department of Molecular Epigenetics

Date of submission:	**26th of April 2017**
First Examiner:	**Prof. Dr. Dirk Eick**
Second Examiner:	**Prof. Dr. Angelika Böttger**
Date of the oral examination:	**23rd of October 2017**

Dedicated to my parents

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass die vorliegende Dissertation von mir selbstständig und ohne unerlaubte Hilfe angefertigt ist.

Erklärung

Hiermit erkläre ich, dass die Dissertation nicht ganz oder in wesentlichen Teilen einer anderen Prüfungskommission vorgelegt worden ist.

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

München, im April 2017

_____

Nilay Shah

Summary

The largest subunit of RNA polymerase II (Pol II), Rpb1, contains an unusual carboxy-terminal domain (CTD), composed of tandem repeats with the consensus sequence of YSPTSPS. The CTD is evolutionary conserved from yeast to humans and is considered as a master regulator of eukaryotic transcription. The CTD undergoes dynamic post-translational modifications and serves as a binding platform for the recruitment of proteins that are involved in various stages of transcription.

Previous work from our laboratory using a Rpb1 knockout-knockin system demonstrated that in mammalian cells the substitution of Tyr1 by phenylalanine (Y1F) in repeats 4-51 of the CTD prevents the formation of the actively transcribing Pol IIO form and leads to the degradation of the CTD (Descostes et al., 2014). In this work, I tried to elucidate the role of Tyr1 of mammalian CTD in the process of transcription, by designing and analyzing mutants, in which Tyr1 was exchanged by phenylalanine only in the defined sections of the CTD.

All tyrosine mutants in this work stably expressed the actively transcribing form of Pol II. The mutant YFFF, in which the last 3/4th of the CTD heptads had Y1F mutations, demonstrated several interesting transcription phenotypes. First, the mutant showed pervasive read-through (RT) transcription that extends up to several hundred kbs downstream of 3' end sites at a genome-wide scale. Second, the 3'-end processing and polyadenylation of mRNA in the mutant YFFF was not unaffected, suggesting global termination defects in the mutant. Third, the mutant YFFF demonstrated reduced promoter-proximal pausing and a delayed pausing at 3' end of genes in ChIP-seq analysis. Finally, mass spectrometry (MS) analysis revealed the loss of interaction of two large complexes, Mediator and Integrator, with Pol II in the mutant YFFF, providing a hint for the role of these complexes in the regulation of transcription termination and promoter-proximal pause/release. In conclusion, this work unravels the role of Tyr1 of the CTD in the regulation of transcription-coupled processes and paves the way for understanding of the highly complex and poorly understood mechanism of transcription termination.

Table of Contents

1. Introduction

Gene expression is a fundamental and highly complex process that is tightly regulated in every living organism. First, the genetic information stored in the deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA); which is then translated into proteins. DNA-dependent RNA Polymerases (RNAPs) are the enzymes that transcribe DNA into RNA, by a process known as transcription.

1.1. RNA Polymerases – discovery and functions

RNAP was first discovered independently by Jerard Hurwitz and Samuel Weiss, in the year 1961 (Hurwitz, 2005). Bacteria and archaea contain a single RNAP (Kusser et al., 2008), while in eukaryotes, three specialized forms of RNA Polymerases - Pol I, Pol II and Pol III transcribe distinct classes of genes (Werner et al., 2011).

In the year 1969, Robert Roeder and William Rutter reported the discovery of three chromatographically separable forms of eukaryotic RNA Polymerase (Roeder et al., 1969). The classification was based on the differential response of these enzymes to inhibition by α-amanitin, a toxic bicyclic octapeptide from the *Amanita phalloides* mushroom. Simultaneously, in the year 1970, it was found that α-amanitin was highly specific inhibitor of one of the two RNA Polymerase activities present in calf thymus (Kedinger et al., 1970). Pol I is completely resistant to α-amanitin, while Pol III shows 50% inhibition at the concentration of 20 µg/ml. Pol II is the most sensitive to α-amanitin and can be completely inhibited at the concentration of 0.5 µg/ml (Schwartz et al., 1974).

Pol I transcribes most of the ribosomal RNA (rRNA) and contributes to more than 50% of total transcripts in the cell (Russell et al., 2005), while Pol III transcribes transfer RNA (tRNA) and several non-coding RNAs (ncRNAs) (Dieci et al., 2007). Pol II, the most widely studied polymerase transcribes messenger RNA (mRNA) and several small nuclear RNA (Nikolov et al., 1997). RNA Pol I, II and III contain 14, 12 and 17 subunits, respectively (Vannini et al., 2012), of which, five subunits – Rpb5, Rpb6, Rpb8, Rpb10 and

Rpb12 are common in all three polymerases (Hahn, 2004; Kimura et al., 2001).

1.2. RNA Polymerase II carboxy-terminal domain (CTD)

The largest subunit of RNA Polymerase II (Pol II), Rpb1, carries a unique and flexible structure at its carboxy-terminal domain that is termed as CTD. The CTD is composed of multiple tandem heptad repeats with the consensus sequence, tyrosine - serine - proline - threonine - serine - proline - serine ($Y_1S_2P_3T_4S_5P_6S_7$) (Heidemann et al., 2013). This highly repetitive domain, which is the focus of my PhD thesis, was first described independently by Jeffrey Corden (Corden et al., 1985) and L.A. Allison (Allison et al., 1985) in the year 1985.

The sequence YSPTSPS of the CTD is highly conserved across various taxa's of organism, but the number of repeats varies remarkably. There are 26 repeats in *S. cerevisiae* (Allison et al., 1988), 29 repeats in *S. pombe* (Schwer et al., 2011) and 52 repeats in mammalian CTD (Corden et al., 1985) (**Figure 1**). The CTD undergoes various post-translational modifications and serves as a docking platform to recruit cellular factors at appropriate stages of the transcription cycle. In next chapters, I will describe in detail, the genetic analysis of the CTD in yeast and mammals, its post-translational modifications and its role in the regulation of transcription processes.

1.3. Genetic analysis of the CTD in yeast and mammals

Since the discovery of the CTD, various laboratories have performed extensive genetic studies on the CTD in yeast and mammals. In doing so, different aspects of the CTD like, the structure and composition of the CTD, minimal length requirements of the CTD, phenotypes linked with mutation of individual amino acids in heptad repeats and functional units of the CTD have been studied.

1.3.1. Structure and composition of the CTD in yeast and mammals

The structure of the CTD displays several intriguing features. Although the length of the CTD varies remarkably in yeast and mammals, most of the heptads in yeast and mammals follow the consensus sequence, YSPTSPS, order (**Figure 1**).

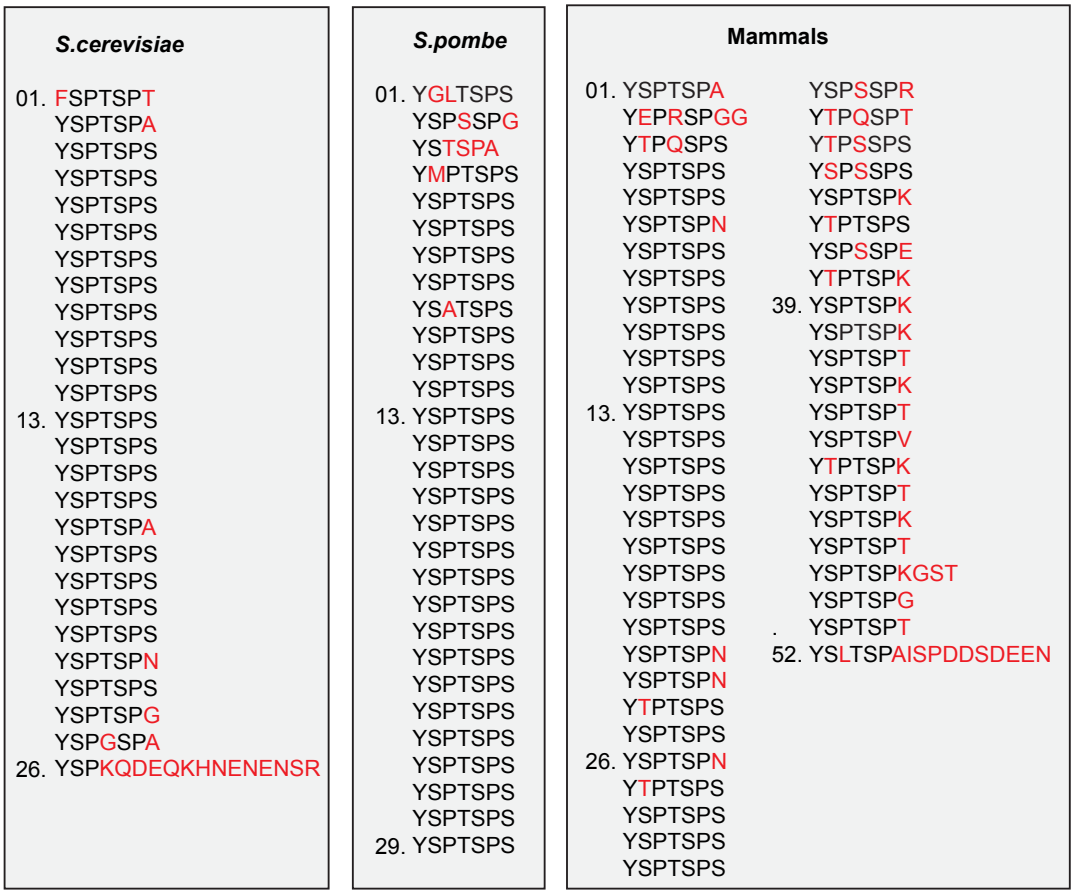


Figure 1: Composition of CTD sequences in *S.cerevisae* (left), *S.pombe* (middle) and mammals (right). Consensus sequences (YSPTSPS) are shown in black. Divergences from the consensus heptads are marked in red. The number on the left represents the repeat number of heptad. Modified from (Eick et al., 2013).

In mammals, there are 21 consensus heptads, most of which are present in the proximal half of the CTD, whereas most of the less conserved, divergent heptads are present in the distal half of the CTD. An evolutionary tree based on the Rpb1 domain derived by Stiller and Hall revealed that the CTDs of many organisms, mostly multicellular forms, deviate from the canonical structure. Thus, it appears that the evolution of higher eukaryotic taxa is associated with specific alterations of the CTD resulting in deviations from the

consensus repeat structure (Yang et al., 2014). Whether yeast and mammalian cells require all the heptads of CTD for viability, and what is the minimal length requirement of CTD in yeast and mammalian cells will be discussed in the following chapter.

1.3.2. Minimal length requirement of the CTD in yeast and mammals

Deletion of the entire CTD is lethal in yeast and mammalian cells, indicating that CTD is essential for cell growth and functions (Allison et al., 1988; Chapman et al., 2005; Meininghaus et al., 2000). However, deletion of a significant number of CTD repeats is well tolerated. The CTD of *S.cerevisiae* contains 26 repeats, but only eight consensus heptad-repeats are required for cell viability (West et al., 1995). In *S.pombe* (29 repeats), the CTD with 10-13 repeats resulted in slow growth and cold-sensitive phenotypes, while the CTD with 16 or more repeats was sufficient for wild-type like growth (Schneider et al., 2010). The minimum number of CTD repeats needed for growth in mammalian cells is not exactly determined, but the CTD with only 31 repeats shows intermediate phenotype with reduced cell proliferation and viability (Meininghaus et al., 2000). Furthermore, the mice homozygous containing 39 repeats are smaller than the wild-type littermates and have a high degree of neonatal lethality (Litingtung et al., 1999).

Interestingly, in mammalian cells, a CTD-less Polymerase (termed as Δ5 mutant) is transcriptionally inactive on chromatin templates. Pol II Δ5 mutant can bind to the c-myc promoter with same efficiency as the wild type Pol II CTD, but does not form a stable initiation complex and does not transcribe promoter proximal sequences (Lux et al., 2005). This suggests that the CTD is required for the transition from transcription initiation to promoter-proximal pausing. In mammalian cells, repeats 1-3 and repeat 52 are important for the stability of CTD. Deletion of repeats 1-3 or 52; or replacement of repeats 1-3 with consensus heptad-repeats, lead to cleavage and degradation of the CTD (Chapman et al., 2005). Intriguingly, the CTD mutant that had repeats 4-51 replaced with consensus repeats shows wild type like growth and can fulfill all essential functions for proliferation. This implies that the non-consensus repeats are dispensable and not essential for cell proliferation. However, the

arginine residue in repeat 31 of the CTD has been described in the regulation of the snRNA and snoRNA gene expression (Sims et al., 2011), suggesting that non-consensus repeats may have gene specific functions. Thus, the deletion studies of the CTD in yeast and mammalian cells show that not all heptads of the CTD are essential and cells are able to grow with less than the natural number of heptad repeats. The substitution of which amino acids within the heptad-repeat of the CTD are tolerated in yeast and mammalian cells is described next.

1.3.3. Phenotype linked to the mutation of individual amino acid within the heptad-repeat

The role of amino acids within the heptad-repeats for cell growth and viability can be studied by mutation of the respective amino acids in all repeats. In *S.cerevisiae*, the work by West et al., showed that the substitution of Tyr1 to phenylalaine (Y1F) leads to a lethal phenotype (West et al., 1995). Further, the substitution of Ser2 or Ser5 with alanine (S2A or S5A) or glutamate (S2E or S5E) did not support cell viability (West et al., 1995). This shows that the phosphorylation of both residues is essential for CTD function. The lethal phenotype arising upon the substitution of Ser2 and Ser5 by glutamate is an indication that glutamate cannot substitute for the phosphate group and/or that both, phosphorylation and dephosphorylation of Ser2 and Ser5 of the CTD are critical. The replacement of Thr4 to alanine (T4A) and Ser7 to alanine (S7A) is tolerated *in S.cerevisiae* (Stiller et al., 2000), indicating that modifications of these residues are not essential for cell viability.

More recently, similar study has been performed for *S.pombe* CTD by the laboratory of Schwer and Shuman. Surprisingly, the substitution of Tyr1 to phenylalanine (Y1F) is not lethal, indicating that tyrosine phosphorylation is not essential for the growth of fission yeast. However, the phenyl ring of Tyr1 is essential, as the replacement of tyrosine with partial isosteric leucine (Y1L mutant) is detrimental to *S.pombe* (Schwer et al., 2011). Similar to the phenotype in *S.cerevisiae*, the substitution of Ser5 to alanine (S5A) was lethal in *S.pombe*; whereas, Ser2 to alanine (S2A) mutants grew well at 30°C. This signifies that in *S.pombe*, not, Ser2-P, but Ser5-P is essential. Mutations of

Thr4 to alanine (T4A) and Ser7 to alanine (S7A) is tolerated in *S.pombe* (Schwer et al., 2011).

In mammalian cells, the repeats 1-3 and repeat 52 are essential for the CTD stability (Chapman et al., 2005; Chapman et al., 2004) therefore; amino acid substitutions were performed in repeats 4-51. CTD mutants comprising replacements of Ser2, Thr4 and Ser5, with alanine in repeats 4 to 51 showed a severe growth defect with a strongly reduced cell count after 4 days. Furthermore, Thr4 to serine and Ser7 to alanine mutants revealed an attenuated phenotype, but were also not viable (Hintermair et al., 2012). The replacement of Tyr1 by phenylalanine (Y1F) is lethal in mammalian cells. Furthermore, Y1F mutation prevents the formation of the hyper-phosphorylated Pol IIO form and induces the degradation of Pol II to the truncated Pol IIB form (Descostes et al., 2014). This indicates that in mammalian cells, phosphorylation of Tyr, Ser2, Thr4, Ser5 and probably also Ser7 is essential for CTD functions. Further manipulation of the CTD developed the concept of the functional unit of the CTD. An overview of the different mutants in yeast and mammalian cells is shown in the **Table 1**.

1.3.4. Functional unit of the CTD

The strong periodicity of the heptad repeats structure is one of the eye-catching features of the CTD. The laboratory of Stiller and Cook disrupted the strong periodicity of the CTD by introducing a single alanine residue between every single heptad (Stiller et al., 2004). This mutation was lethal in yeast cells. However, the introduction of an alanine residue between di-heptads had little effect on yeast viability, suggesting that the strong periodicity of the CTD can be disrupted and is not essential for cell growth. It also indicates that all essential CTD functions in yeast are accomplished through interaction of protein factors with a motif/s that lie within di-heptads (Stiller et al., 2004). These experiments introduced the concept of functional unit of the CTD in yeast, which lies within paired-heptads. Further substitution and insertion mutants within or between di-heptads, narrowed down the functional unit. Specifically, the three SP motifs, S_2P-S_5P-S_7P must be present and the two-tyrosine residues must be spaced seven amino acids apart in either Y_1-Y_8 (Y**SP**T**SP**SY**SP**) or Y_8-Y_{15} (**SP**T**SP**SY**SP**TSPSY) orientation (Liu et al., 2008).

The distance between the functional units of the CTD is also very critical. Placing two alanine residues between functional units was not lethal, but resulted in a slow growth rate at high and low temperatures (Stiller et al., 2004). Introducing further alanine residues between functional units resulted in a progressive decline in the cell viability (Liu et al., 2010).

Similar experiments in *S.pombe* revealed that the functional unit also lies within a di-heptad and is composed of a decapeptide unit (YSPTSPSYSP) (Schwer et al., 2012).

In mammals, unpublished work from our laboratory demonstrates that the minimal functional unit of the mammalian CTD lies within a penta-heptad repeat (Shah et al., unpublished), suggesting that the minimal functional unit in mammals is larger than in the yeast. An overview of the chapter 1.3, comparing the genetic analysis of the CTD in yeast and mammals is shown in the **Table 1**.

Table 1: An overview of genetic manipulations in yeast and mammals. The total length of CTD, minimum number of CTD repeats required for viability, phenotypes of CTD substitution mutants and minimal functional unit in yeast and mammals are compared.

	Mammals	***S.cerevisiae***	***S.pombe***
Total no. of CTD repeats	52	26	29
Minimum no. of CTD repeats required for viability	-	8	10-13
Y1F	Lethal	Lethal	Viable
S2A	Lethal	Lethal	Viable
S2E	-	Lethal	Lethal
T4A	Lethal	Viable	Viable
S5A	Lethal	Lethal	Lethal
S5E	-	Lethal	Lethal
S7A	Lethal	Viable	Viable
Functional Unit	(YSPTSPS)5	(YSPTSPS)2	(YSPTSPS)2

1.4. Post-translational modifications of the CTD

CTD undergoes dynamic post-translational modifications during the process of transcription. A number of enzymes are known to modify the CTD during the transcription cycle. An overview of these enzymes is presented in the **Table 2**. The role of the enzymes in the process of transcription will be discussed in the chapter 1.5.

Table 2: An overview of known CTD-modifying enzymes in mammals and yeast (Adapted from (Zaborowska et al., 2016).

		Organisms			
Modification	**Enzymes**	**Mammals**	***S.cerevisiae***	***S.pombe***	**References**
Tyr1-P	Kinases	cAbl			(Baskaran et al., 1993)
Ser2-P	Kinases	Cdk9, Cdk12, Cdk13, Brd4	Bur1, Ctk1	Cdk9, Lsk1	(Bartkowiak et al., 2010; Blazek et al., 2011; Bres et al., 2008; Cho et al., 2001; Devaiah et al., 2012; Greifenberg et al., 2016; Viladevall et al., 2009; Wood et al., 2006)
	Phosphatases	Fcp1	Fcp1	Fcp1	(Cho et al., 2001; Hausmann et al., 2002; Lin et al., 2002)
Thr4-P	Kinases	Plk3			(Hintermair et al., 2012)
Ser5-P	Kinases	Cdk7, Cdk8, Cdk13	Kin28	Mcs6	(Akhtar et al., 2009; Glover-Cutter et al., 2009; Greifenberg et al., 2016; Komarnitsky et al., 2000; Rodriguez et al., 2000)
	Phosphatases	Ssu72, Scp1, RPAP2	Rtr1, Ssu72		(Egloff et al., 2012; Hausmann et al., 2005; Krishnamurthy et al., 2004; Mosley et al., 2009; Zhang et al., 2012)
Ser7-P	Kinases	Cdk7	Kin28		(Akhtar et al., 2009; Glover-Cutter et al., 2009)
	Phosphatases	Ssu72	Ssu72		(Zhang et al., 2012)
Arginine Methylation		PRMT5, CARM1			(Sims et al., 2011; Zhao et al., 2016)
Lys7-acetylation		p300			(Schroder et al., 2013)

Tyrosine, threonine and the three serine of a heptad can be phosphorylated, while prolines can undergo isomerization between the -cis and -trans configuration. The dynamic interplay of CTD-modifying enzymes generates a characteristic pattern of CTD modifications throughout the transcription cycle of protein-coding genes in both, *S.cerevisiae* and mammals. This dynamic phosphorylation pattern of the CTD is described next.

1.4.1. Dynamic phosphorylation patterns of the CTD

Several labs have used the monoclonal antibodies and analyzed the CTD modification pattern at genome-wide scale in yeast (Bataille et al., 2012; Kim et al., 2010; Mayer et al., 2012; Mayer et al., 2010) and mammalian cells (Descostes et al., 2014; Hintermair et al., 2012; Koch et al., 2011) (**Figure 2**).

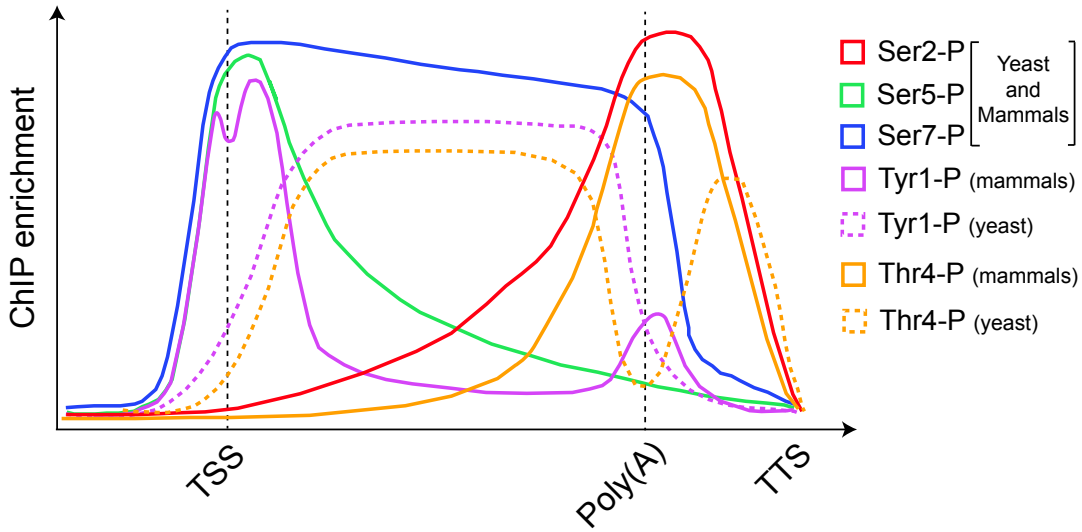


Figure 2: Dynamic phosphorylation pattern of the CTD along the transcription cycle in yeast and mammals. Schematic representation of ChIP enrichment for Ser2-P (red), Ser5-P (green), Ser7-P (blue), Tyr1-P (purple) and Thr4-P (yellow) is shown for yeast and mammals. TSS (transcription start site), poly(A) site and TTS (Transcription termination site) are shown. Modified from (Eick et al., 2013).

ChIP analysis indicated that the levels of Ser5-P are highest near the transcription start sites (TSS) and reduced towards the 3' end, whereas the Ser2-P levels are higher towards the 3' end of genes. Signals for Ser7-P are high near the TSS and continue to remain high towards the 3' end of genes. In general, the patterns of Ser2-P, Ser5-P and Ser7-P are similar in yeast and

mammals, while Tyr1-P and Thr4-P shows notable differences. In mammals, Tyr1-P is enriched near the TSS and is associated with antisense promoter transcription and active enhancers (Descostes et al., 2014), while in yeast, Tyr1-P is enriched over the gene bodies and is lower at the 5' and 3' ends of genes (Mayer et al., 2012). In mammals, Thr4-P is enriched in 3' end region of the gene (Hintermair et al., 2012), while in yeast; Thr4-P is associated with transcribed region of gene (Mayer et al., 2012). Recent study from Churchman lab, indicates that in yeast, Thr4-P mark also peaks after poly(A) sites (Harlen et al., 2016). For the non-consensus residues, genome-wide analysis revealed that the high levels of acetylated Lys7 and mono- or dimethylated Lys7 are found near the TSS of genes (Dias et al., 2015; Voss et al., 2015).

Monoclonal antibodies are reliable tools for the measurement of changes in the CTD modification and allow understanding of the functional role of Pol II CTD in gene expression. Nevertheless, monoclonal antibodies fail to provide information about the spatial patterns and signatures of CTD modifications. Recently published studies used genetic manipulation of the CTD combined with mass spectrometry analysis to provide several important information about the relative levels and spatial patterns of the CTD phosphorylation in yeast and mammals (Schuller et al., 2016b; Suh et al., 2016). First, both the studies show that the CTD can be uniformly phosphorylated across all repeats. This is irrespective of whether the repeats are similar to the one in the wild-type CTD or mutated to facilitate proteomic analysis. Second, mono-phosphorylated peptides were much more abundant than double- or triple-phosphorylated peptides. Third, both the studies reveal that Ser5-P and Ser2-P are much more abundant modifications than Tyr1-P, Thr4-P or Ser7-P mark. In human cells, Ser5-P and Ser2-P are found in similar quantities on mono-phosphorylated peptides and contributes for nearly 75% of the total phospho-counts (Schuller et al., 2016b). In yeast, Ser5-P is about four times more abundant than Ser2-P (Suh et al., 2016). The higher Ser5-P/Ser2-P ratio for the yeast CTD may result from dramatically different gene lengths in yeast and human (Schuller et al., 2016b). Thus, this approach provides an insight into the abundance and spatial patterns of CTD phosphorylation in vivo and can serve as a tool to study the CTD heptad-specific phosphorylation

patterns after knock out of specific CTD kinases or phosphatases. Additionally, such an approach can also be applied as a general method to study the modification of other proteins with low-complexity (LC) domains, such as members of the FET-protein family (Schuller et al., 2016a).

1.5. RNA Pol II CTD and regulation of transcription processes

The eukaryotic Pol II transcription cycle is a highly coordinated process that exhibits regulation at multiple steps and can be divided into three distinct stages – initiation, elongation and termination (**Figure 3**). Briefly, the transcription cycle begins with the hypo-phosphorylated Pol II gaining access to the core promoter and the formation of a functional pre-initiation complex (PIC) (Step 1). The CTD is phosphorylated, DNA is unwound and Pol II escapes/clears the core promoter and initiates the process of transcription (Step 2). Early elongating Pol II pauses at the promoter-proximal pause site (Step 3). The paused Pol II is then hyper-phosphorylated, escapes the promoter-proximal pausing and enters the process of productive elongation (Step 4). Elongating Pol II traverses through the entire gene body and pauses near the termination site for the 3' end processing of RNA (Step 5). Following the 3' end processing of the RNA, Pol II dissociates off the template DNA and the free Pol II can reinitiate the new cycle of transcription (Step 6). The next chapters will describe the current understanding about the distinct stages of the transcription cycle, factors and mechanisms controlling these processes and the role of the CTD in the process of transcription.

1.5.1. Transcription initiation

The process of transcription initiation begins with the formation of the pre-initiation complex (PIC). In eukaryotes, six general transcription factors (GTFs), TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, along with the transcription co-activator, Mediator complex, are recruited to the core promoter and together with Pol II forms the PIC (Buratowski, 2009; Hahn, 2004; Nechaev et al., 2011).

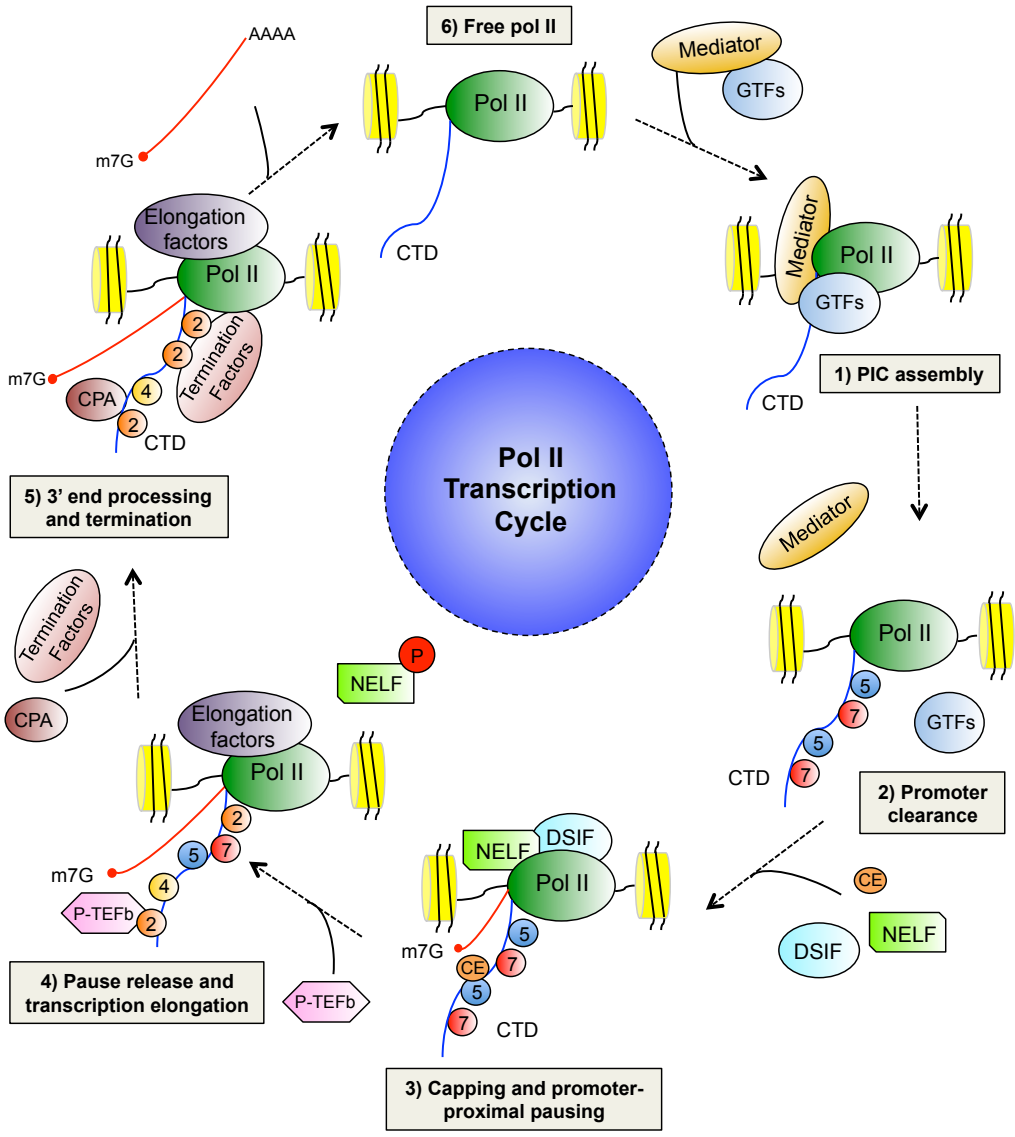


Figure 3: RNA Polymerase II transcription cycle. 1) General transcription factors (GTFs), Mediator complex and Pol II are recruited at the core promoter for the assembly of PIC. 2) Ser5-P and Ser7-P of the CTD leads to destabilization of the PIC. Mediator-Pol II connections are disrupted and Pol II escapes the promoter. 3) Recruitment of the capping enzyme and the elongating Pol II pauses near promoter-proximal sites. 4) Recruitment of P-TEFb facilitates the Ser2-P of the CTD and triggers the transition of Pol II from early elongation to productive elongation. Dynamic phosphorylation and dephosphorylation of the CTD during elongation phase. 5) Recruitment of elongation factors, 3' end processing and the termination factors. 6) Pre-mRNA is cleaved and Pol II dissociate off the template DNA. Hypo-phosphorylated Pol II can initiate a new round of transcription cycle.

The mediator was first isolated as a complex of 20 subunit proteins in yeast (Kim et al., 1994) and later the mammalian counterparts of the mediator

complex were identified (Jiang et al., 1998). The mediator complex, which consists of four modules (head, middle, tail and CKM) is recruited by activator proteins and has a key role in mediating the recruitment of Pol II and other PIC factors to the promoter to stimulate transcription initiation (Fan et al., 2006; Kornberg, 2005; Malik et al., 2005). Mediator complex enables the Pol II recruitment via interactions with the unphosphorylated CTD (Kim et al., 1994; Myers et al., 1998; Naar et al., 2002). Precisely, the Pol II CTD interacts strongly with the Mediator middle module and weakly to the head module (Robinson et al., 2012; Tsai et al., 2013). In addition, Mediator complex can interact with the TFIIH component of the GTF, most likely via an interaction with MED11 subunit of the Mediator (Esnault et al., 2008).

After the formation of a PIC at the gene promoter, several important steps are critical for the Pol II enzyme to escape the promoter. The two-helicase subunits of TFIIH, XPB and XPD, are required for melting of the DNA and to make the template accessible for the transcription (Tirode et al., 1999). Pol II must break the contact with Mediator and the PIC to escape the promoter and transit to the early elongation. Here, the phosphorylation of Pol II CTD plays a key role in disrupting the Mediator-Pol II interaction and thereby facilitating the promote escape (Sogaard et al., 2007). Cdk7, the kinase subunit of the basal transcription factor TFIIH (Kin 28 in yeast), phosphorylates the CTD at serine-5 and serine-7 positions (Akhtar et al., 2009; Glover-Cutter et al., 2009; Kim et al., 2009a). The ability of TFIIH to phosphorylate the Pol II CTD is enhanced by the Mediator complex (Boeing et al., 2010; Nair et al., 2005). Once the Pol II escapes the promoter, it initiates the process of transcription and enters the early elongation phase.

1.5.2. Transcription elongation

During early elongation, Ser5-P of the CTD serves as a signal for the recruitment of various mRNA processing and histone modifying factors. When ~25 nucleotides of nascent RNA are synthesized, the cap structure is attached to its 5’ end. The capping enzyme specifically recognizes the Ser5-P CTD repeats and catalyzes the addition of a methylguanosine cap to the 5’ end of the nascent mRNA (Fabrega et al., 2003; Komarnitsky et al., 2000). In yeast, the Ser5-P binds and recruits the H3K4 methyltransferase, Set1, to the

CTD (Ng et al., 2003). Set1 establishes two distinct chromatin zones on the genes, with trimethylated H3K4 near the promoter regions and dimethylated H3K4 downstream of promoters. H3K4me2 recruits Set3 complex, a histone deacetylase that suppress nucleosome acetylation and remodeling (Kim et al., 2009b).

In many metazoan genes, the early elongating Pol II pauses around 20-60 nucleotides (nt) downstream of the transcription start site (TSS) and is defined as promoter-proximal pausing. The paused Pol II remains stably associated with the nascent RNA and is fully capable of resuming elongation, however further signals are needed for the pause release and transition of Pol II to a productive elongation (Adelman et al., 2012; Zhou et al., 2012). Two key regulators of promoter-proximal pausing are the DRB sensitivity factor (DSIF) and the Negative elongation factor (NELF) complexes. Both these factors associate with the early elongating Pol II and transiently halt the elongation (Gaertner et al., 2014; Wu et al., 2003; Yamaguchi et al., 1999). Pausing is overcome through P-TEFb (positive transcription factor b), which comprises of the cyclin-dependent kinase Cdk9 and Cyclin T subunits (Marshall et al., 1995; Wada et al., 1998). P-TEFb phosphorylates the repressive DSIF/NELF complex, causing the NELF to dissociate and promoting Pol II into transcription elongation (Fujinaga et al., 2004; Kim et al., 2001; Yamada et al., 2006). P-TEFb also phosphorylates the serine-2 residues of the CTD, creating a platform for the binding of RNA processing and chromatin modifying factors (Adelman et al., 2012; Peterlin et al., 2006).

As the Pol II elongates the transcript, the activity of the phosphatase Ssu72 is stimulated, which gradually dephosphorylates the Ser5-P residues of the CTD (Reyes-Reyes et al., 2007). The double phosphorylation marks of Ser2-P and Ser5-P helps in the recruitment of the Set2 methyltransferase, which trimethylates H3K36 (Kizer et al., 2005). Set2 mediated H3K36 methylation recruits histone deacetylases, Rpd3, which deacetylates the histones within the transcribed regions, thereby preventing cryptic transcription initiation within the intragenic sites (Carrozza et al., 2005). Recently, it was shown that in yeast, Set2-mediated H3K36 methylation regulates the selective suppression of antisense transcription (Venkatesh et al., 2016). In mammalian cells, the elongation factor Spt6 interacts with Ser2-P through its tandem SH2

domain and facilitates the recruitment of RNA export factor REF1/Aly (Yoh et al., 2007). The double phosphorylation marks of Ser2-P and Ser5-P aids in the recruitment of the splicing factor U2AF65, which recruits the Prp19 complex to the nascent RNA and enhances splicing in a CTD-dependent manner (David et al., 2011). In addition, the transcription termination factors, Pcf11 and Rtt103, preferentially bind to the Ser2-P CTD (Lunde et al., 2010; Meinhart et al., 2004). This is consistent with the increased levels of Ser2-P towards the 3' end of the transcription unit. Recently, it was shown that the Thr4-P of the CTD peaks after the poly(A) site and also aids in the recruitment of Rtt103 (Harlen et al., 2016).

Pol II productively traverses across the gene body and pause near 3' end of the gene before the process of transcription termination. Pol II termination is a very complex process and tightly coupled to the 3' end processing of the nascent RNA. Generally, 3' end processing of the RNA precedes the termination process and CTD plays an important role in both the processes. The mechanisms describing the 3' end processing of RNA will be discussed next, followed by the process of transcription termination.

1.5.3. 3' end processing of RNA

There are three well-described mechanisms for 3' end processing of Pol II transcripts (Eick et al., 2013). First is the poly(A)-dependent pathway, which is specific for most of the protein coding genes. Second, the Sen1 dependent pathway is used for many small noncoding RNAs like, snoRNAs or CUTs (cryptic unstable transcripts). And third is the Integrator-dependent pathway, and is used for the processing of the most of snRNA genes.

1.5.3.1. Poly(A)-dependent pathway

In eukaryotes, most of the protein coding mRNA precursors have a highly conserved poly(A) signal, AAUAAA, positioned 10-30 nucleotides upstream of the cleavage site, and a G/U-rich sequence around 10-30 nucleotides downstream of the cleavage site. 3' end processing of RNA is a two step reaction in which, first, transcription of a poly(A) site is followed by pausing of the Pol II and cleavage of the nascent transcript; and second, the polyadenylation of the upstream cleaved transcript (Kuehner et al., 2011; Proudfoot, 2016; Rosonina et al., 2006) (**Figure 4**).

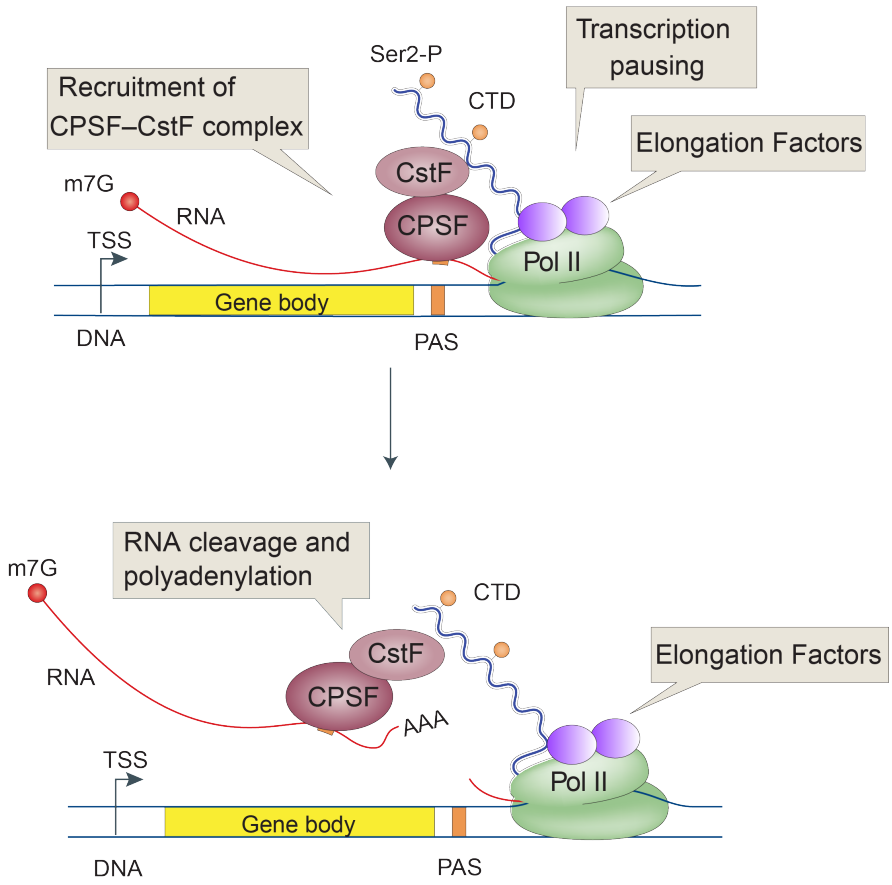


Figure 4: Model for 3'end processing of protein-coding RNAs. Cleavage and polyadenylation specificity factors (CPSF), Cleavage stimulation factor (CstF) and poly(A) polymerase (PAP) trigger endonucleolytic cleavage and polyadenylation of the mRNA. Modified from (Porrua et al., 2015)

Multisubunit factors such as, Cleavage stimulation factor (CstF), Cleavage and Polyadenylation Specificity Factor (CPSF) and poly(A) polymerase (PAP), catalyze the reactions. CstF consists of three main subunits, CstF50, CstF64 and CstF77, whereas, CPSF contains five main subunits, CPSF30, CPSF73, CPSF100, CPSF160 and Fip-1 (Richard et al., 2009). CstF50 binds to the Pol II CTD (Fong et al., 2001) and CstF64 binds to the downstream G/U-rich processing signal (MacDonald et al., 1994). CstF77 bridges the CstF64 and CstF50 subunits (Takagaki et al., 1994) and also makes direct contact with the CPSF160, contributing to the stabilization of CPSF-CstF complex (Murthy et al., 1995). CPSF30 subunit of the CPSF complex binds to the body of the polymerase and travels with the Pol II (Nag et al., 2007). As the Pol II traverses across the poly(A) signal, CPSF160 specifically recognizes and binds to the AAUAAA sequence (Keller et al., 1991; Murthy et al., 1995). It has been proposed that CPSF binding to the poly(A) signal induced Pol II

pausing. When CstF binds to the downstream G/U-rich signals, CPSF binds to CstF, releases its hold on the Pol II body and leads to the CPSF-mediated endonucleolytic cleavage of the nascent transcript (Kuehner et al., 2011). CPSF73 is responsible for the endonucleolytic cleavage of the nascent transcript (Mandel et al., 2006; Ryan et al., 2004). Following the cleavage, the poly(A) polymerase subsequently polyadenylates the 3' end of the mRNA.

1.5.3.2. Sen1-dependent pathway

In yeast, the 3' end processing and transcription termination of many noncoding RNAs, including small nuclear RNAs, small nucleolar RNAs and cryptic unstable transcripts relies on a mechanism that requires the activity of the NNS complex (Arigo et al., 2006; Steinmetz et al., 2001). The essential NNS complex contains two RNA-binding proteins, Nrd1 and Nab3, and the DNA helicase Sen1. The presence of short sequence motifs on the nascent RNA is recognized by Nrd1 and Nab3, and is crucial to engage the NNS complex onto the nascent RNAs (Carroll et al., 2007). The NNS complex interacts with Ser5-P CTD of Pol II through the CID of Nrd1 (Kubicek et al., 2012). Interestingly, transcription termination by the NNS complex does not appear to be associated with endonucleolytic cleavage of the nascent RNA but rather functions by a mechanism, in which the DNA helicase activity of Sen1 dislodges Pol II from the template DNA (Porrua et al., 2013). No clear homologues of Nrd1 and Nab3 have been described so far in metazoan. Senataxin (SETX), the putative homolog of Sen1, which possesses a conserved helicase domain (Moreira et al., 2004), has been described to promote efficient transcription termination by resolving RNA/DNA hybrid structures (Skourti-Stathaki et al., 2011).

1.5.3.3. Integrator-dependent pathway

A third, integrator dependent processing-termination pathway has been described for snRNA genes. Genes encoding snRNAs contain a conserved 13 - 16 nucleotide sequence element (termed the 3' box) that is required both for 3' end processing and for transcription termination. The 3' box, is recognized by the Integrator (INT) complex, and cleaved by its catalytic subunits INT9 and INT11, which are respectively, homologues of the CPSF73 and CPSF100 subunits of the CPSF complex (Baillat et al., 2005). Here, Ser7-P of the CTD plays an important role and facilitates the recruitment of

RPAP2, a putative Ser5 phosphatase, which in turn dephosphorylates Ser5 of the CTD and also recruits the Integrator complex (Egloff et al., 2012). This is consistent with the requirement of Ser7-P for the snRNA gene expression (Egloff et al., 2007).

1.5.4. Transcription termination

Following 3' end processing of RNA, Pol II must be released from the DNA template for the efficient transcription termination. Although, the molecular mechanisms that lead to the timely dissociation of the Pol II remains poorly understood, two models have emerged to explain the mechanistic basis of transcription termination: the allosteric and the torpedo models (**Figure 5**).

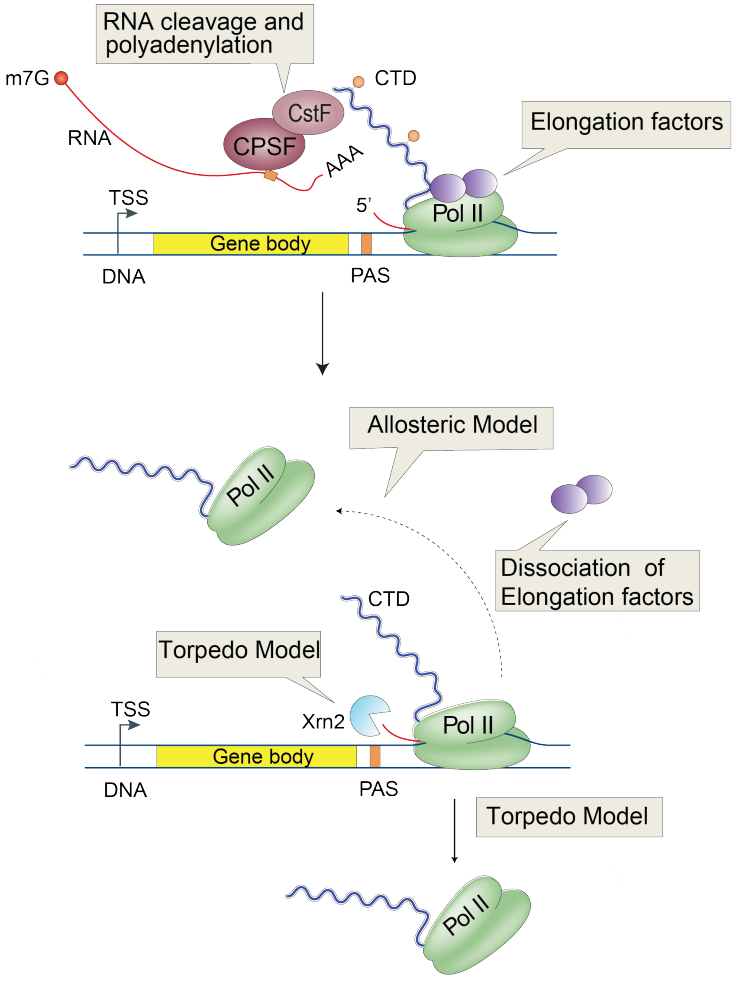


Figure 5: Transcription termination models. Following 3' end processing of mRNA, Pol II becomes termination prone. Allosteric model proposes the dissociation of elongation factors, destabilization of the complex and gradual release of Pol II from DNA template. Torpedo model proposes downstream degradation of cleaved product by exonuclease, Xrn2, and eventual release of Pol II from DNA template. Modified from (Porrua et al., 2015).

1.5.4.1. The Allosteric Model

The allosteric model or the anti-terminator model proposes that transcription through the poly(A) site leads to conformational changes of the elongation complex (EC), by dissociation of elongation factors and/or association of termination factors. Such changes could destabilize the complex and have marked pausing effects on transcription, resulting in the gradual release of the Pol II from the DNA template (Logan et al., 1987; Rosonina et al., 2006; Zhang et al., 2015). Several 3' end processing and termination factors, including Pcf11, become associated with Pol II elongation complex at the 3' end of the gene in a Ser2-P CTD-dependent manner (Ahn et al., 2004; Licatalosi et al., 2002). The CID of Pcf11 bridges the CTD to mRNA, and can dismantle the EC in vitro (Zhang et al., 2005). In drosophila, depletion of Pcf11 causes transcriptional read-through, and drosophila Pcf11 can also dismantle the EC in vitro (Zhang et al., 2006). In support of the allosteric model, several proteins have been described that has no direct role in cleavage/polyadenylation and yet, influences the process of termination. As an example, the transcriptional coactivator PC4 (human homolog of Sub1), interacts with CstF64 and inhibits premature termination and the dissociation of PC4 at the poly(A) site renders Pol II competent for termination (Calvo et al., 2001). Recent in vitro study shows that transcription termination of protein-coding genes requires the poly(A) signal but not the poly(A) site cleavage and that the conformational changes of the elongation complex triggers transcription termination (Zhang et al., 2015).

1.5.4.2. Torpedo model

The second model, the torpedo model, proposes that the endonucleolytic cleavage of the nascent pre-mRNA at the poly(A) site creates an entry site for a 5' -> 3' exonuclease, which degrades the downstream-cleaved product. The exonuclease continues degrading the transcript, until it catches up with the Pol II and acts as a molecular trigger to release Pol II from the DNA template (Connelly et al., 1988; Proudfoot, 2016). Direct evidence of the torpedo model came from the studies that demonstrated a role of the yeast 5' -> 3' exonuclease Rat1 and its human homologue, Xrn2, in Pol II termination. Depletion of Rat1 in *S.cerevisiae* or Xrn2 in mammalian cells led to a substantial loss in termination (Kim et al., 2004b; West et al., 2004).

However, exactly how Rat1 and Xrn2 displace the Pol II from the DNA template remains unknown. Interestingly, the depletion of Xrn2 by RNAi often causes a marginal termination defects, however, combining RNAi treatment with the expression of dominant negative Xrn2 (active-site mutant) causes the general inhibition of Pol II termination and shifts the termination zone further downstream of genes (Fong et al., 2015). In yeast, 3' end processing factors are required for the normal Rat1 recruitment. Rat1 in complex with its activating protein Rai1 copurifies with the third member of the complex Rtt103, which is a Ser2-P CTD binding protein and possibly accounts for the Rat1 recruitment to Pol II (Kim et al., 2004a). In yeast, Pcf11 is required for the Rat1 association with active genes (Luo et al., 2006).

However, so far the exact mechanism by which the poly(A) site dependent Pol II termination is regulated in vivo is not clear. The laboratory of David Bentley proposed a hybrid model of transcription termination (Luo et al., 2006). Accordingly, the complex that carries out poly(A) site cleavage comprises Pol II, 5' → 3' exonuclease and termination factors bound to the phosphorylated CTD. This complex catalyzes the cleavage at the poly(A) site, degrades the nascent RNA and then the allosteric changes are transmitted to the Pol II catalytic site, causing the Pol II to release from the template DNA.

1.6. Aim and scope of present work

The carboxy-terminal domain (CTD) of RNA Polymerase II (Pol II) consists of a unique and flexible structure that is highly conserved through evolution, apparently increasing in length and diversifying in structure with the complexity of the organism (Stiller et al., 2002). Of the 52 repeats in mammalian CTD, 21 repeats follow the evolutionary conserved, consensus heptapeptide Y_1-S_2-P_3-T_4-S_5-P_6-S_7 structure. The CTD undergoes various post-translational modifications to coordinate transcription-coupled processes such as transcription initiation, elongation, 3' processing as well as transcription termination. Tyrosine-1 phosphorylation (Y1P) of the CTD has diverse functions in yeast and mammals. In yeast, Y1P is enriched over the gene body and impairs the recruitment of transcription termination factors to Pol II (Mayer et al., 2012). A study from our laboratory showed that Y1P is associated with promoters and active enhancers in mammalian

cells (Descostes et al., 2014). Furthermore, in order to investigate the functional significance of Tyr1, a CTD mutant was generated that had tyrosine replaced by phenylalanine in repeats 4-51. However, the mutant inhibited the formation of the actively transcribing Pol IIO form and lead to the degradation of the CTD. Therefore, the functional analysis of such a mutant was challenging and the understanding about the role of Tyr1 in the regulation of transcription processes remained elusive.

The main aim of this project was to overcome this challenge and address the following questions, 1) Do the CTD need Tyr1 in all 52 repeats to stably express the actively transcribing form of Pol II? 2) What is the role of Tyr1 of CTD in the regulation of transcription-coupled processes? And 3) Are there any position specific functions of the heptad-repeats of CTD? In other words, what are the functional significance of Tyr1 in the proximal, the middle and/or the distal heptads of the CTD?

2. Results

2.1. RNA Pol II CTD tyrosine mutants

Tyrosine1 in mammalian CTD is a highly conserved residue, which is present in all 52 repeats. The significance of Tyr1 residues in the process of transcription can be studied by characterizing the mutant that lacks Tyr1. The replacement of Tyr1 by phenylalanine (Y1F) in repeats 4-51 prevented the formation of the actively transcribing Pol IIO form and leads to the degradation of the CTD in mammalian cells (Descostes et al., 2014). Consequently, the functional analysis of this mutant remained elusive. Here, we developed an alternative strategy and introduced Y1F mutations in only parts of the CTD.

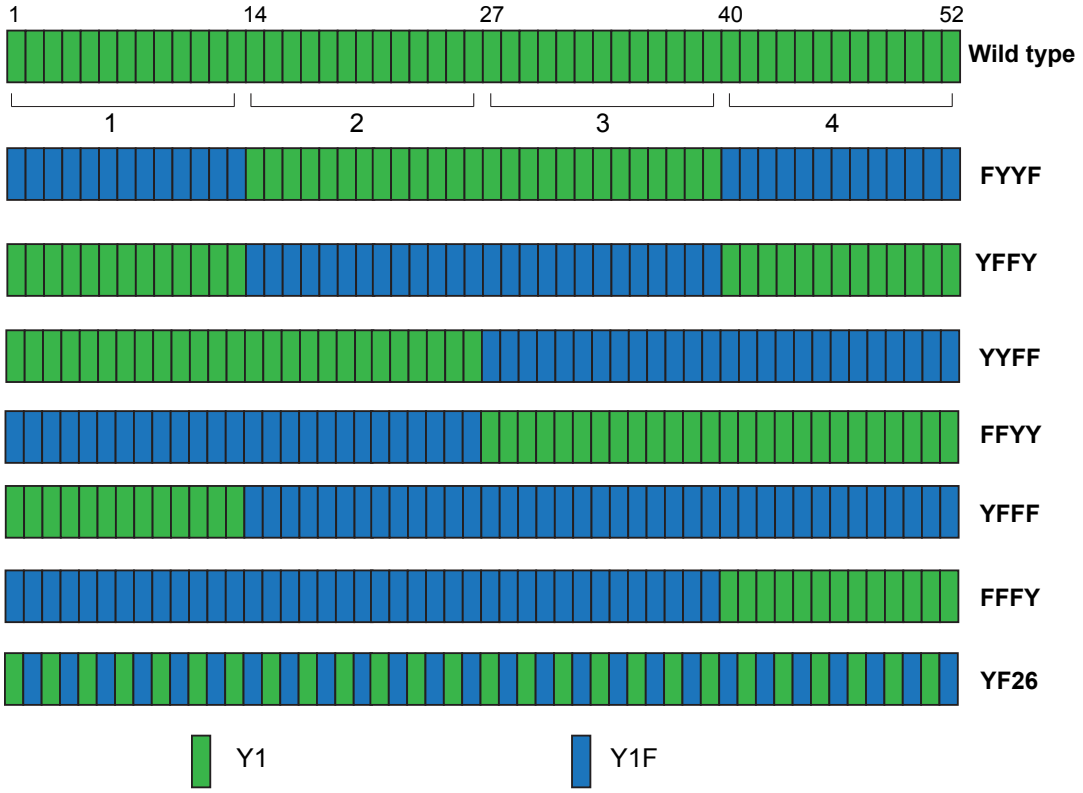


Figure 6: Schematic representation of wild type CTD and seven tyrosine mutants. Each box represents a single heptad repeat of the CTD. Green box represents the wild-type sequence; while blue box represents heptads in which tyrosine-1 is replaced by phenylalanine. Heptads 1, 14, 27, 40 and 52 of the wild type CTD are marked. The four clusters of the wild type CTD are numbered from 1-4.

A set of seven mutants was designed by dividing 52 heptads of the wild type (WT-CTD) into four clusters of 13 heptads each (**Figure 6**). Cluster 1 consists of heptads 1-13; cluster 2 - heptads 14-26; cluster 3 - heptads 27-39

and cluster 4 - heptads 40-52. Y1F mutations were introduced in one or more of these clusters and mutants were generated that had Y1F mutations in the proximal, the middle and/or the distal heptads of the CTD. The rationale was to generate a set of mutants, whose analysis could help us unravel the functional significance of Tyr1 residues in different heptads of the CTD. Mutants, FYYF, YFFY, YYFF, FFYY and YF26 have in total 26 heptads with Y1F mutations, whereas mutants YFFF and FFFY carries 39 mutated heptads each. GeneArt in Regensburg synthesized DNA fragments of CTD mutants.

2.1.1. Establishing cell lines expressing tyrosine mutants

All seven CTD constructs were cloned into the episomal expression vector pRX4-267 (Meininghaus et al., 2000) using a two-step cloning strategy (Materials and Methods 4.2.1.1).

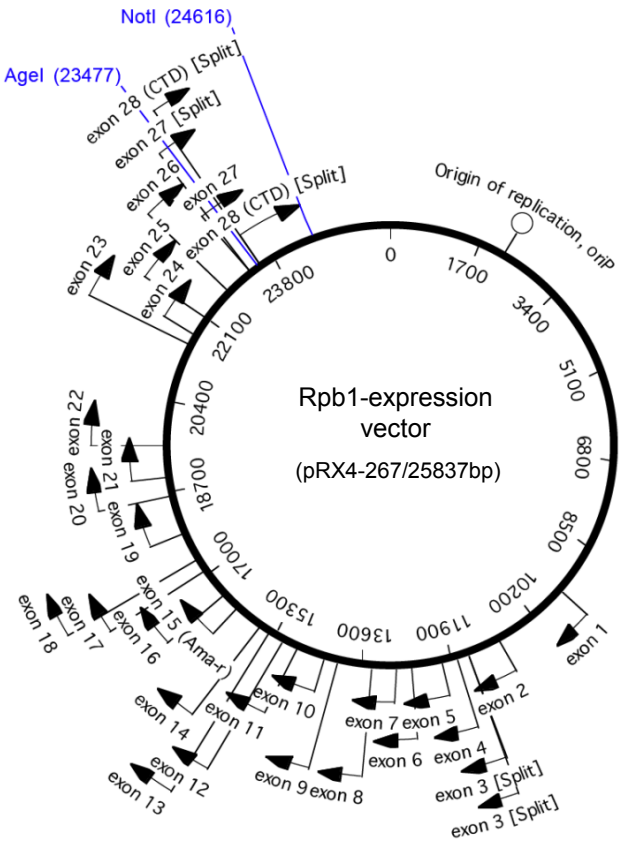


Figure 7: Scheme of end vector pRX4-267. The expression vector pRX4-267 encodes the entire mouse Rpb1 gene comprising 28 exons. The sequence encoding CTD can be easily replaced by the given artificial CTD sequences using restriction enzymes, AgeI and NotI.

The vector pRX4-267 contains important features for stable transfection and conditional expression of the mutants. The vector contains full length, haemagglutinin-(HA) tagged mouse Rpb1 gene comprising 28 exons (**Figure 7**). CTD is encoded by the exon 28 and can be replaced by any artificial CTD sequence using restriction sites. The Rpb1 gene has a point mutation (N793D) that confers resistance towards α-amanitin (Bartolomei et al., 1987). In the presence of α-amanitin, endogenous Pol II is inhibited, thereby allowing the possibility to study the properties of the mutant Pol II *in vivo*. The vector uses replication origin of the Epstein-barr virus (EBV) and can be episomally maintained in human cells expressing EBV-nuclear antigen 1 (EBNA1). An expression vector with WT-CTD sequence was and used as a positive control. It will be referred to as 'rWT' (recombinant wild-type) in the thesis.

2.1.2. Conditional expression of recombinant polymerases

The expression vector containing the mutant CTD sequence was transfected into Raji cells via electroporation. Cells were cultured for 2-3 weeks in RPMI growth media supplemented with 1mg/ml of G418 and 1 µg/ml tetracycline, for the selection of positively transfected cells. At a cell viability of 90-95%, expression of the recombinant Pol II was induced in the mutants by removal of tetracycline from the medium. Cells were cultured in the presence of 2µg/ml of α-amanitin, 24 hours after the induction. Whole cell lysates were prepared for western blot analysis, 48 hours after the addition of α-amanitin. All the seven-tyrosine mutants expressed the hypo-phosphorylated (IIA) and the hyper-phosphorylated (IIO) forms of Pol II using α-HA antibody for detection (3F10) (**Figure 8**).

Expression of the Pol IIO form indicates that the mutants are transcriptionally active. Interestingly, the Pol IIA band migrated faster in all mutants. This could be due to the increased loading of SDS of mutant Rpb1 protein. The mutant FFFY showed reduced expression of the Pol IIA and Pol IIO form. This may reflect weak induction or suggests that this mutant could not support its own expression. rWT cells were used as a positive control and expressed both forms of Pol II, while no signal was detected in the untransfected Raji cells, which served as a negative control.

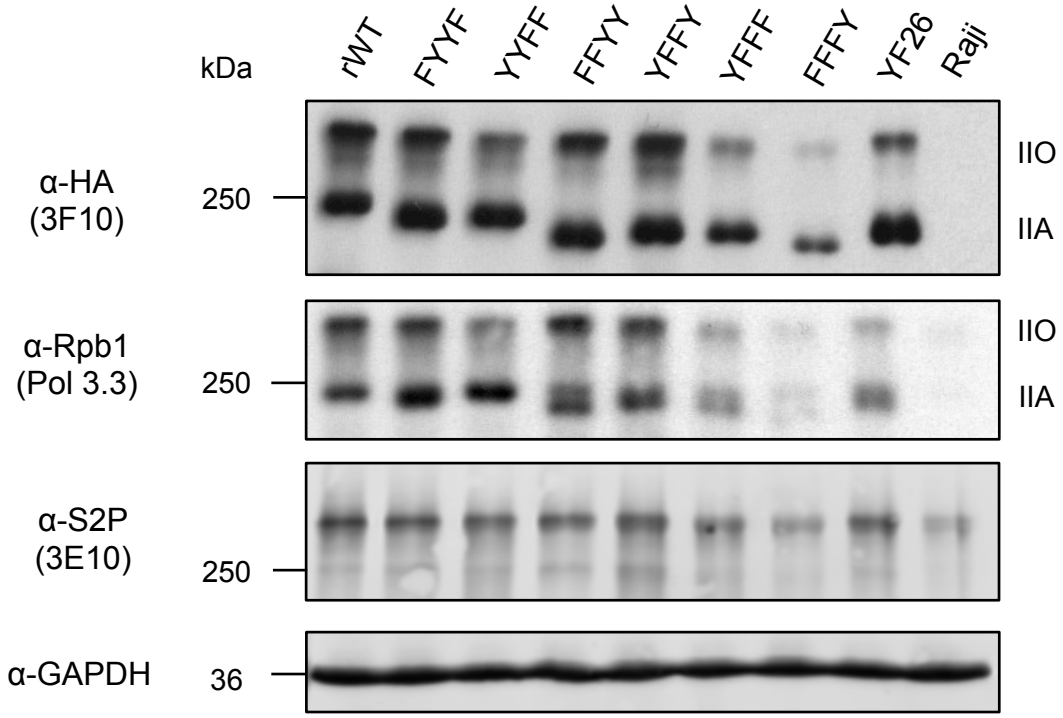


Figure 8: Expression of CTD tyrosine mutants. Cells were lysed 72 hours after induction and 48 hours after the addition of α-amanitin. Expression of the recombinant Pol II was analyzed using α-HA antibody (3F10). α-Rpb1 antibody (Pol 3.3) detects the expression level of the endogenous and the recombinant Pol II. rWT was used as a positive control and untransfected Raji cells as a negative control. The serine-2 phosphorylation level in the mutants was analyzed using the α-S2P antibody (3E10). α-GAPDH serves as a loading control.

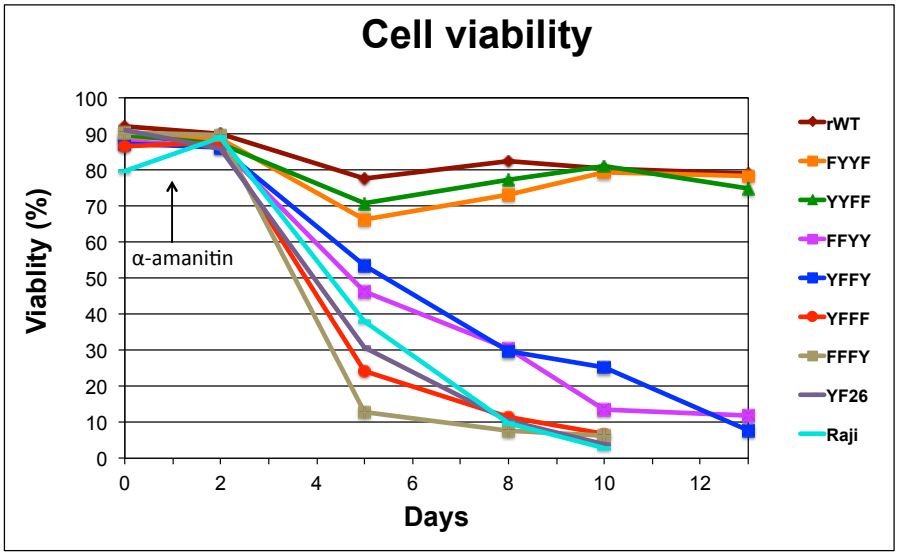
Expression of the endogenous and recombinant Pol II can be detected by α-Rpb1 antibody (Pol 3.3). In mutants, FFYY and YFFF, two bands of the Pol IIA form were detected. The upper band represents the endogenous Pol IIA, while the lower band, the Pol IIA form of the recombinant polymerase. Serine-2 phosphorylation of the CTD is associated with transcription elongation (Ahn et al., 2004). All tyrosine mutants were phosphorylated at serine-2 residues when probed with a α-S2P antibody (3E10), suggesting that mutants are elongation-competent. Thus, we successfully established a set of tyrosine mutant cell lines that are transcriptionally competent.

2.1.3. Growth kinetics and phenotypic analysis

The growth behavior of the tyrosine mutants was monitored by measuring the percentage of cell viability and cell proliferation for a period of two weeks (**Figure 9**). For each mutant, $2x10^7$ cells were induced and the number of

living cells (N_L) and dead cells (N_D) were calculated at intervals of 2 to 3 days intervals using trypan-blue staining. α-amanitin was added 24 hours after the induction. For cell proliferation, cumulative living cell number was calculated by multiplying the total number of living cells (N_L) with the factor by which the culture was split over the course of the experiment.

(A)



(B)

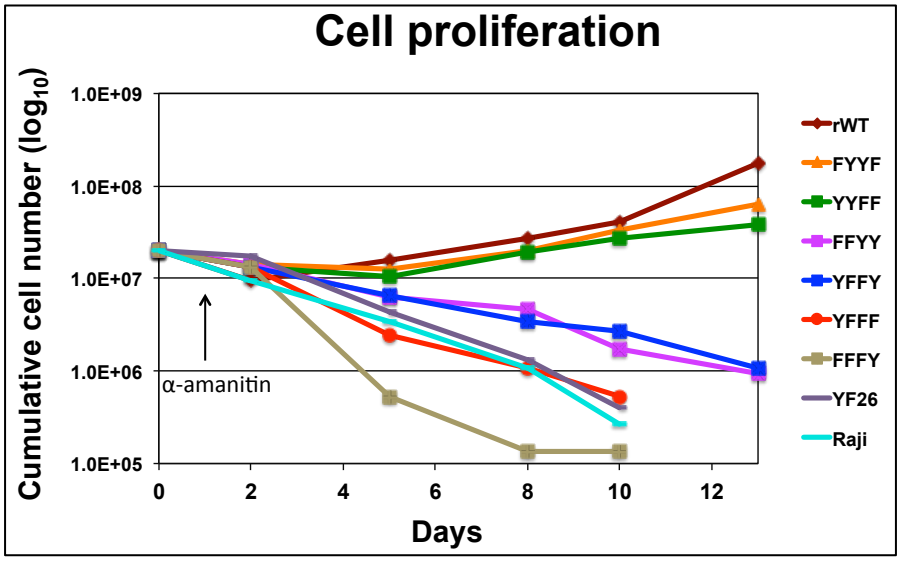


Figure 9: Growth kinetics of tyrosine mutants. Graphs representing (A) percentage of cell viability and (B) cell proliferation of mutants over a period of two weeks. Number of living cells (N_L) and of dead cells (N_D) was calculated at indicated days using trypan blue staining.

All the mutants initially displayed a small decline in cell growth and viability after induction and addition of α-amanitin. This is probably the negative effect of the switch between the endogenous and the recombinant Pol II, as cells

expressing the recombinant Pol II need an adaptation phase for growth. Percentage of living cells (**Figure 9**A) and cumulative living cell number (**Figure 9**B) of the mutants, rWT, FYYF and YYFF, gradually increased eight days post induction. Simultaneously percentage of living cell and cumulative living cell number of the remaining five mutants and the untransfected Raji, cells continued to decline. After two weeks, rWT and the two mutants, FYYF and YYFF, were viable and demonstrated an increase in cell number. However, the cell proliferation rate of both mutants was less than rWT. The remaining five mutants and Raji cells displayed a lethal phenotype with less than 10% of living cells after two weeks (**Figure 9**A). Interestingly, the two viable mutants, FYYF and YYFF, have tyrosine conserved in cluster 2 (heptads 14-26) of CTD, while all the lethal mutants carry Y1F mutations in cluster 2 (**Figure 6**).

Following the successful characterizing of the tyrosine mutants for the expression and growth phenotype, the transcriptomes of the mutants were analyzed in order to gain a deeper insight into their functional significance.

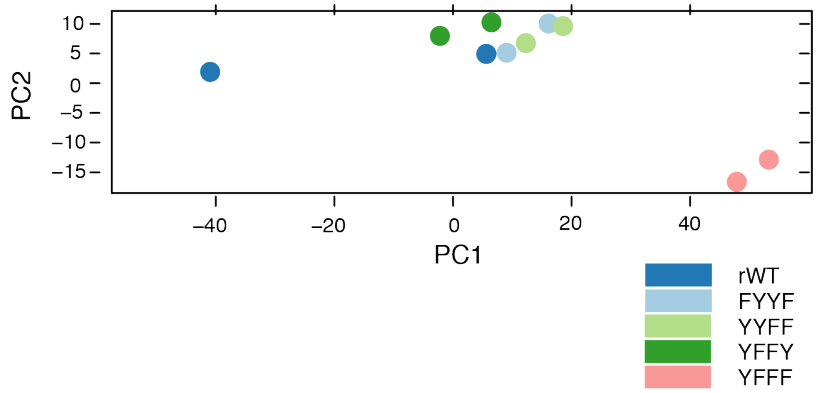
2.2. Total RNA-seq analysis

RNA-seq analysis of total RNA was performed for four of the seven mutants. Two of these mutants, FYYF and YYFF, were viable and two mutants, YFFY and YFFF, displayed a lethal phenotype (**Figure 9**). Cells were lysed in trizol reagent 72 hours after induction and 48 hours after the addition of α-amanitin. Total RNA was extracted according to the manufacturer's protocol (Trizol RNA isolation, ThermoFisher Scientific, USA). Ribosomal RNA was removed from total RNA using Ribo-Zero rRNA removal Kit (EpiCenter, USA) and the libraries were prepared using ScriptSeq total RNA Library Prep Kit (EpiCenter, USA) (Materials and methods 4.2.5.1). The libraries were sequenced on the Illumina HiSeq2000 platform and the data was processed for the analysis (Materials and methods 4.2.6.2). Samples were prepared using two biological replicates of each rWT and the four selected mutants (total 10 samples).

2.2.1. Quality control of RNA-seq libraries and differential gene expression (DGE) analysis

A first useful step in RNA-seq analysis is to assess the similarities and variations between the samples using the Principal Component Analysis (PCA). It is a statistical procedure that uses large sets of data (here, differential gene expression values) to define new sets of variables (Principal components) between the samples and the replicates (**Figure 10**A).

(A)



(B)

A \ B	rWT	YYFF	YFFY	FYYF	YFFF
rWT		41	26	33	810
YYFF	2		0	0	153
YFFY	2	3		3	458
FYYF	2	0	0		196
YFFF	48	19	27	26	

> 3 fold upregulated in B relative to A
> 3 fold downregulated in A relative to B

Figure 10: Principal component analysis and differential gene expression analysis (A) Principal component analysis (PCA) of rWT and tyrosine mutants representing variation and similarities between the samples and the replicates. Each point represents one sample. **(B)** Number of differentially expressed genes in rWT and mutants. Number of genes upregulated in B in comparison to A are shown in green boxes while the number of genes downregulated in A relative to B are shown in red boxes. Differential gene expression threshold > 3-fold change

The biological replicates of mutants, FYYF, YYFF and YFFY clustered in close proximity, suggesting a high reproducibility between the replicates and less variations in the transcriptome. Unexpectedly, one replicate of rWT was separated from the second replicate on the PC1 scale. The two biological replicates of the mutant YFFF clustered close to each other, but separated from the cluster of three other mutants and the rWT sample. This suggests that both replicates of the mutant YFFF are similar, but differ from rWT and other mutants.

Next the differential gene expression (DGE) analysis that was performed to identify the number of genes that was at least three-fold differentially expressed in rWT and mutants (**Figure 10**B). A total of 858 genes were differentially expressed in the mutant YFFF as compared to rWT (810 upregulated, 48 downregulated). In contrast, the number of differentially expressed genes in other mutants was significantly lower: 43 genes in the mutant YYFF (41 upregulated, 2 downregulated), 28 genes in the mutant YFFY (26 upregulated, 2 downregulated) and 35 genes in the mutant FYYF (33 upregulated, 2 downregulated).

Overall the data suggest that Y1F mutations in the CTD lead to the differential expression of genes, with most of them being upregulated. Furthermore, the mutant YFFF has the highest number of differentially expressed genes and its transcriptome seems to have the highest variations in comparison to rWT and the other tyrosine mutants.

2.2.2. Tyrosine mutants exhibit a read-through (RT) transcription phenotype at 5’ and 3’ end of genes

Next, the total RNA-seq signals at individual genes in rWT and mutants were compared. A very interesting transcription phenotype was observed in mutants as exemplified at the gene locus LY96 (**Figure 11**).

Strong RNA-seq signals were detected over the annotated transcription unit of the gene, LY96, in rWT and tyrosine mutants. In rWT, there were almost no RNA-seq signals in the region downstream of the annotated 3’ end site (red line). In contrast, in tyrosine mutants, RNA-seq signals were detected beyond the annotated 3’ gene boundary. These signals correspond to a 3’ end read-through (RT) transcription phenotype. Intriguingly, a very strong 3’

end RT transcription phenotype was observed in the mutant YFFF compared to other mutants. RT transcription in mutant YFFF extended more than 100 kb downstream of the annotated 3' end site and was observed not only at a single gene locus, but for a large number of genes as confirmed by the metagene analysis (**Figure 12**).

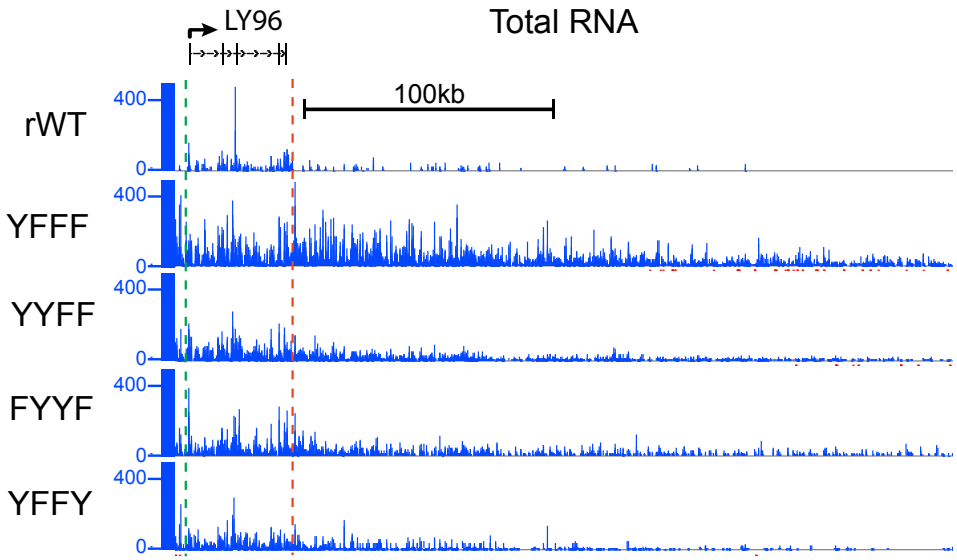


Figure 11: Signals of total RNA-seq at the Ly96 gene locus in rWT and mutants. A screenshot from the IGB genome browser, comparing RNA-seq signals at the gene LY96 between rWT and tyrosine mutants. The annotated transcription unit of the gene is marked between the green and the red lines.

Metagene profiles of rWT and mutants YYFF, FYYF, and YFFY showed comparable RNA-seq signals near the TSS, 3' end and in 20 kb regions around the annotated gene boundaries. In contrast, the mutant YFFF displayed a massive accumulation of RNA-seq signals in regions 20 kb upstream and downstream of the gene body. Accumulation of RNA-seq signals downstream of the annotated 3' end region confirmed the 3' end RT transcription phenotype and could be the consequence of either a defect in the 3' end processing of RNAs and/or the failure of Pol II to terminate transcription in mutant YFFF. Accumulation of RNA-seq signals upstream of the TSS in the mutant YFFF represents an increase in upstream antisense transcription and/or 3' end RT transcription of the gene located upstream of the corresponding gene.

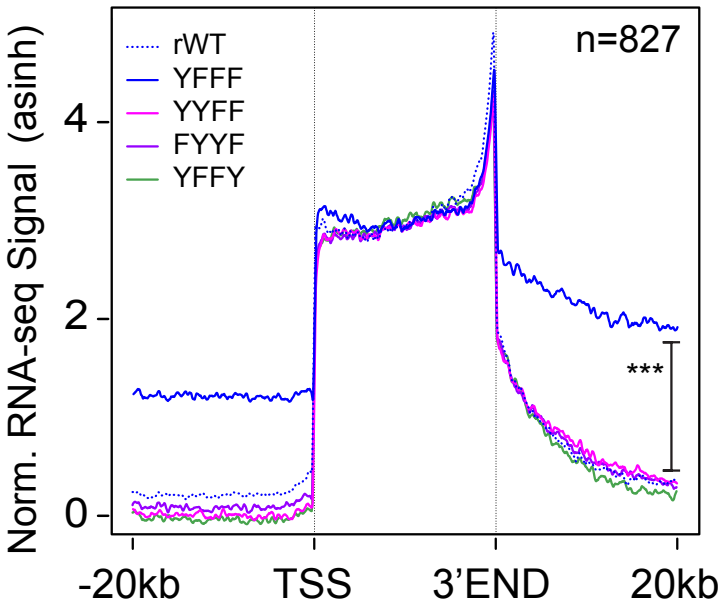


Figure 12: Metagene profiles of total RNA-seq in rWT and tyrosine mutants. Metagene profiles representing the average distribution of the read counts from total RNA-seq in rWT (dotted blue), and mutants YYFF (magenta), FYYF (purple), YFFY (green) and YFFF (blue). Profiles are distributed over the gene bodies between the transcription start site (TSS) and 3' end site and 20 kb around the annotated gene boundaries. A total of 827 genes from the hg 19 (human genome 19) that had no other annotation in regions 20 kb around the gene boundaries were selected for the analysis. Signals are normalized to the gene body.

Thus, the first round of RNA-seq analysis allowed us to isolate the mutant YFFF, in which a large number of genes are differentially expressed and a stringent RT transcription phenotype at 5' and 3' ends of genes occur. This observation was very exciting and prompted us to characterize mutant YFFF in more detail.

2.2.3. Library Preparation Kit: ScriptMiner vs Illumina

The data obtained from the libraries using the ScriptSeq total RNA library preparation kit showed high background noise in the intergenic regions. Therefore, it was important to first improve the quality of the libraries. Hence, the RNA-seq of total RNA in the wild-type Raji cells was compared using two different library preparation kits – Scriptseq and Illumina. The data produced using the Illumina RNA library preparation Kit lead to the reduction of noise in the intergenic regions (**Figure 13**).

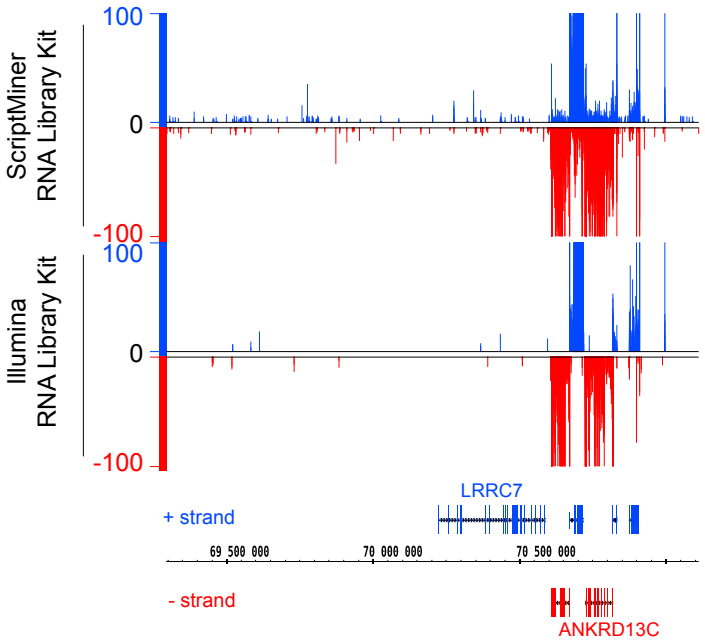


Figure 13: Total RNA-seq library preparation kit: ScriptMiner vs Illumina. Comparison of total RNA-seq in wild-type Raji cells using two different library preparation kits: ScriptMiner and Illumina. The libraries prepared using the Illumina library preparation kit lead to the reduction of artifacts caused by high background noise in the intergenic regions.

2.3. The mutant YFFF exhibits an increase in antisense and 3' end read-through transcription

After improving the quality of the libraries, RNA-seq of total RNA in rWT and the mutant YFFF was repeated. Now, strand-specific RNA-seq libraries were prepared using the library preparation kit from Illumina (Materials and methods 4.2.5.2). Principal component analysis revealed that biological replicates clustered together, indicating a high reproducibility (**Figure 14**A). Replicates of the mutant YFFF were clustered far away from replicates of rWT confirming a high difference in their transcriptomes.

Next, RNA-seq signals at several individual genes in rWT and the mutant YFFF were compared. A representative gene, PDCD6IP, is shown in the **Figure 14**B. The gene is transcribed on the plus strand (blue signals) and its annotated transcription unit is marked between the green and the red line. Both, rWT and the mutant YFFF showed high RNA-seq signals over the transcription unit of the gene. In addition, the mutant YFFF displayed significant 3' end RT transcription that extended more than 100 kb downstream of the annotated 3' end site (orange arrow).

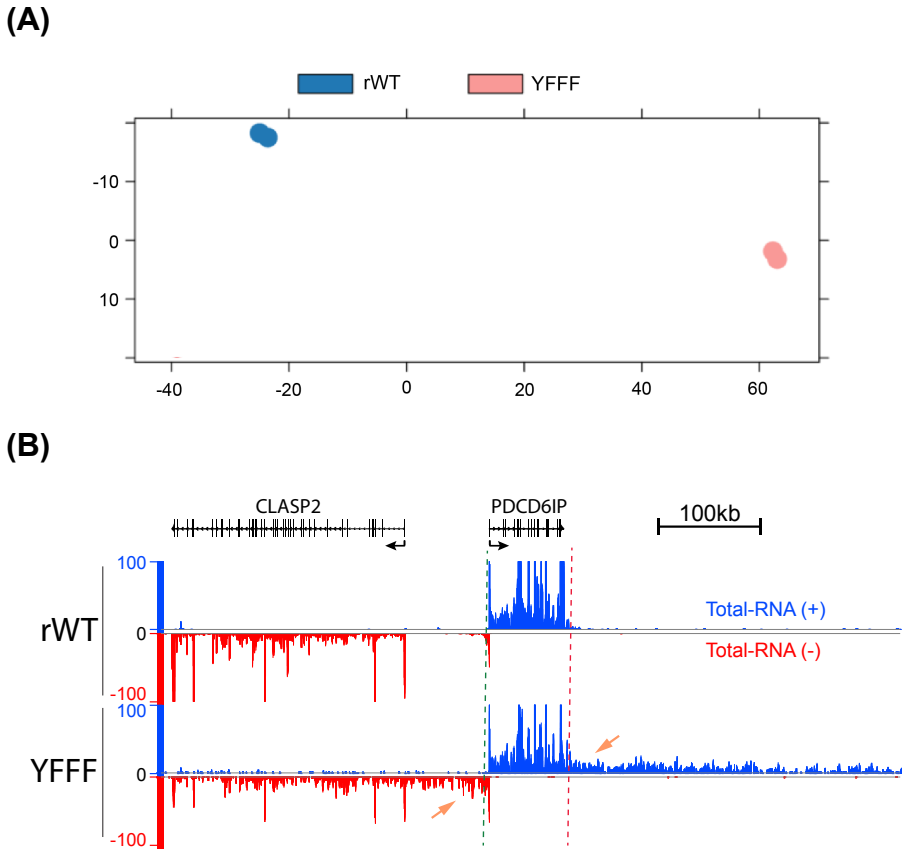


Figure 14: Total RNA-seq in rWT and the mutant YFFF. (A) Principal component analysis (PCA) of rWT and the mutant YFFF representing variation and similarities between the samples and the biological replicates. Each point represents one sample. (B) A screenshot from the Integrated Genome Browser (IGB) showing RNA-seq signals over two genes, CLASP2 and PDCD6IP, in rWT and the mutant YFFF. CLASP2 gene is transcribed on the minus strand (red signals) and the PDCD6IP on the plus strand (blue signals) of the genome. Downstream 3' end read-through transcription and an increased antisense transcription in the mutant YFFF are marked by orange arrows.

The promoter of the PDCD6IP gene gives rise to bidirectional transcription in rWT. In addition to the transcription in the sense direction (blue signals), RNA-seq signals were detected upstream of the PDCD6IP gene in antisense direction (red signals). This antisense transcription is a common phenomenon in mammalian cells (Lehner et al., 2002). In the mutant YFFF, antisense transcription signals upstream of the PDCD6IP gene were more pronounced than in rWT. Thus, the PDCD6IP gene in the mutant YFFF displayed RT transcription upstream of 5' end in antisense direction and downstream of 3' end in sense direction. A large number of genes in the mutant YFFF displayed such transcriptional phenotype (**Figure 15**).

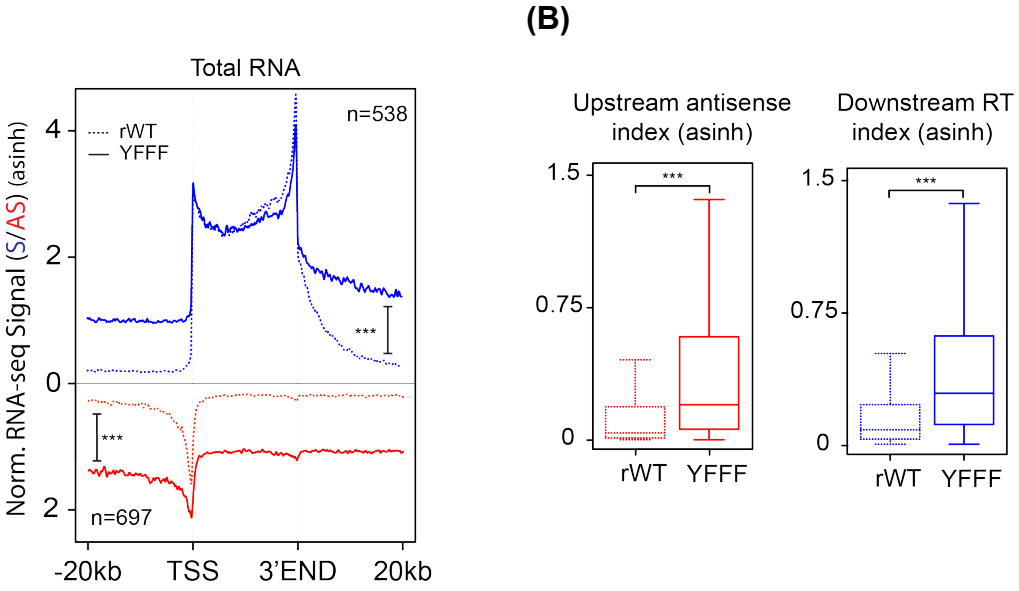


Figure 15: Total RNA-seq in rWT and the mutant YFFF. (A) Average metagene profiles of total RNA-seq representing sense (blue) and antisense (red) transcription in rWT and the mutant YFFF (n=538). Signals in the profiles are distributed over the gene bodies and in regions 20 kb around the annotated gene boundary. Signals are normalized to the gene body (B) Box plot quantification of upstream antisense transcription index (red) and downstream read-through transcription index (blue) in rWT and the mutant YFFF.

Thus repeating RNA-seq analysis of total RNA improved the quality of the samples and strengthened the initial observation of a massive 3' end RT transcription phenotype. It further detected RT transcription upstream of 5' end of genes in antisense direction in the mutant YFFF. It is important to note that not all the genes in the mutant YFFF had the described transcription phenotype. Several genes were normally transcribed and did not display any transcription defects.

2.4. Transcription phenotypes at 5' and 3' end of genes in the mutant YFFF are coupled

Next, we asked if the 3' end RT transcription phenotype in the mutant YFFF was linked to increased upstream antisense transcription. To address this question, the ratio of increased RNA-seq signals downstream of 3' end of genes in the mutant YFFF over rWT was determined. A heatmap was generated in which the genes were ranked in the order of decreasing 3' end RT ratio. For the corresponding genes, the ratio of 5' end upstream antisense

transcription in the mutant YFFF over rWT was plotted in the heatmap. Furthermore, genes in the heatmap were divided into four groups (A-D), from high to low 3' end RT index (**Figure 16**).

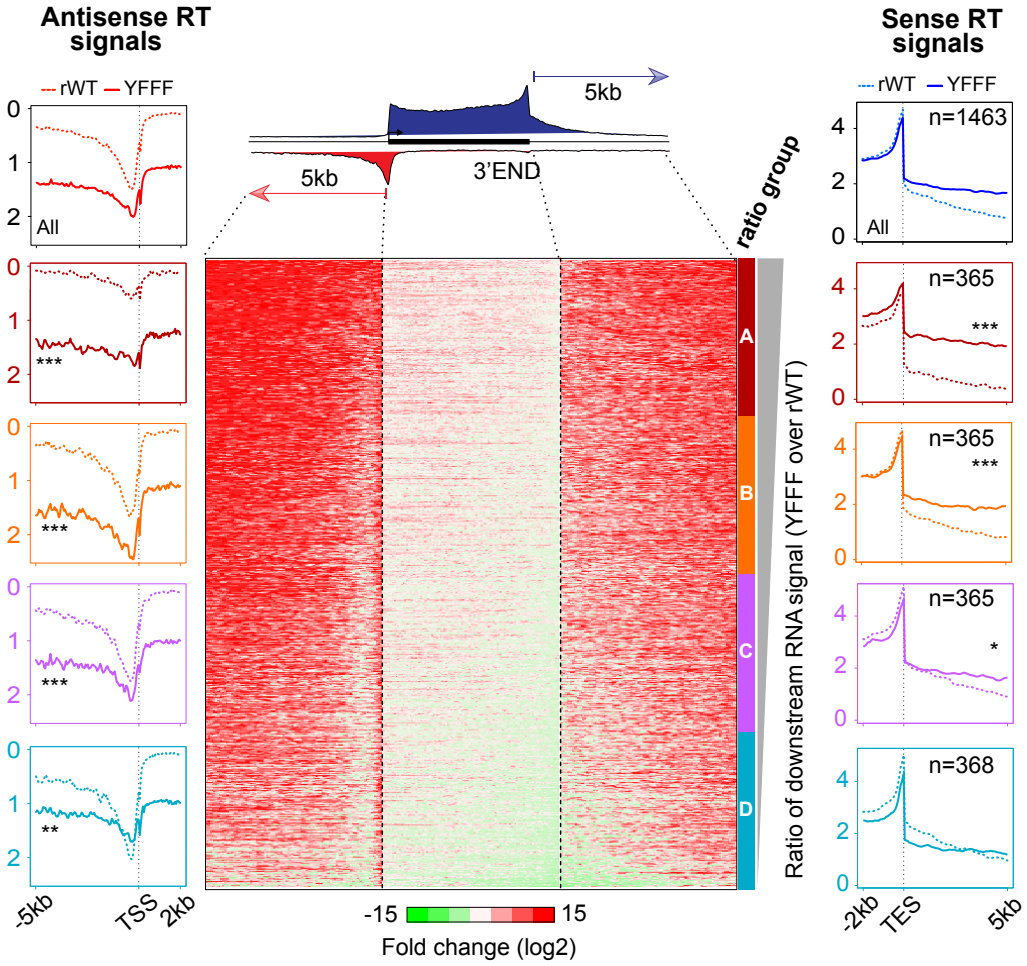


Figure 16: Comparison of upstream antisense and downstream 3' end read-through transcription phenotype in the mutant YFFF. Genes were ranked in the order of decreasing 3' end RT ratio in the mutant over rWT within 5 kb of the annotated 3' ends. Upstream antisense transcription ratio for the corresponding genes is plotted on the left side of the heatmap. Genes in the heatmap were divided into 4 equal sized groups (A-D) (colored profiles on the right). The profiles at the top represent global average profiles, whereas 4 profiles below, corresponds to genes in groups A-D. Scale and the color key depicting the log2 fold change ratio are shown at the bottom.

The analysis revealed that a high 3' end RT transcription ratio for genes in the group A and B correlated with a high upstream RT transcription ratio at 5' ends. Whereas a lower 3' end RT ratio for genes in the group C and D was accompanied by a lower 5' end upstream antisense transcription ratio. Interestingly, several genes in the group D showed lower 3' end RT ratio. In the metagene profile on the right of group D, the RNA-seq signal in the mutant

YFFF dropped immediately downstream of the annotated 3' end site and continued to remain low. While in rWT, the signal intensity dropped gradually in the window of 5 kb. Thus this lower 3' end RT ratio may represent weak expression of several genes in the group D. Overall, the comparative analysis suggested that the RT phenotype at 5' and 3' end of genes in the mutant YFFF is probably coupled.

2.5. Read-through transcription in the mutant YFFF results in transcriptional interference with neighboring genes

One functional consequence of the RT transcription could be the transcriptional interference with neighboring genes. Several examples for genes were observed in the mutant YFFF, in which RT transcription resulted in transcriptional interference with the neighboring gene. An example is shown in the **Figure 17**.

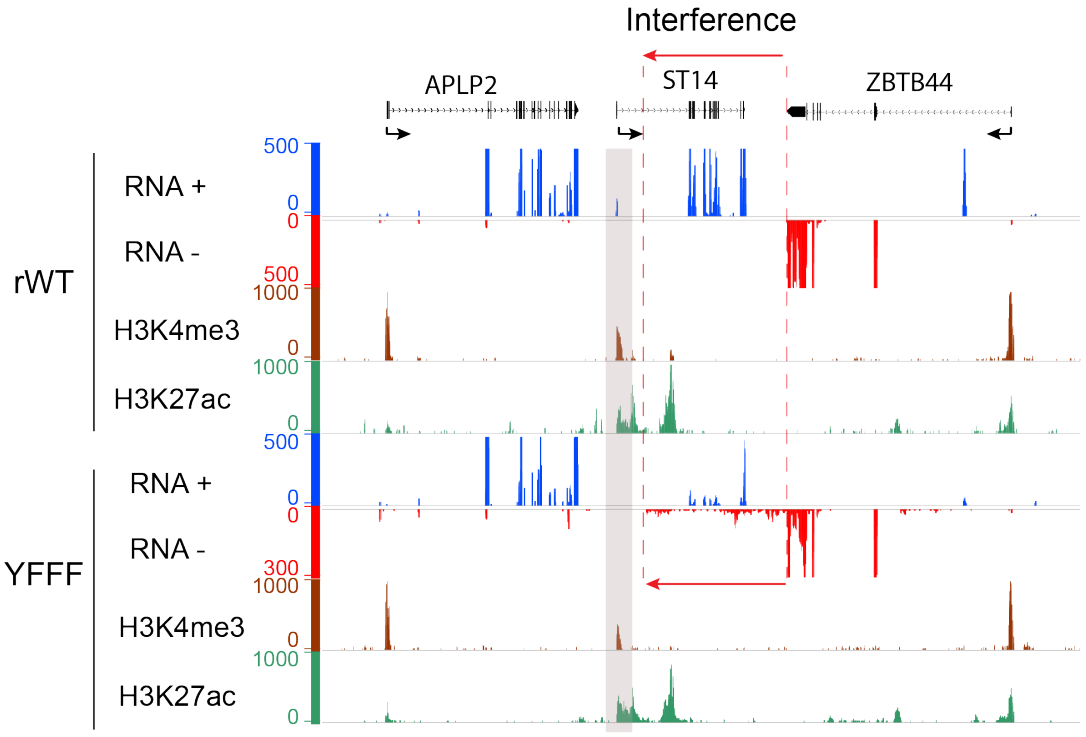


Figure 17: Transcriptional interference in the mutant YFFF. A screenshot from the Integrated Genome Browser (IGB) comparing RNA-seq signals and epigenetic marks, H3K4me3 and H3K27ac, over the genes APLP2, ST14, and ZBTB44 in rWT and the mutant YFFF. Genes APLP2 and ST14 are transcribed on the plus strand (blue signals), while ZBTB44 is transcribed on the minus strand (red signals) of the genome. The red arrow indicates ZBTB44 3' end RT transcription and suppression of ST14 gene transcription.

Two genes, ST14 and ZBTB44, are oriented head to head in the genome. ZBTB44 is transcribed on the minus strand of the genome (red signals) and displayed 3' end RT transcription in the mutant YFFF. Epigenetic modifications H3K4me3 and H3K27ac, which mark active promoters and enhancers, were mapped (Materials and methods 4.2.4). The information about the quantity of the antibodies and the number of cells used for ChIP is in the **Supplementary table 6**. The absence of these marks in the region downstream of the ZBTB44 gene confirmed 3' end RT transcription in the mutant YFFF, rather than a new-initiation event. RT transcription downstream of the ZBTB44 gene interferes with the transcription of the ST14 gene. The presence of H3K4me3 marks at the ST14 gene promoter suggested that the gene is transcribed in the mutant YFFF, but transcription over the gene body is largely repressed, indicating transcriptional interference.

2.6. Read-through transcription phenotype is linked to tyrosine mutations in CTD

We next asked, if the observed RT transcription phenotype in the mutant YFFF is specifically associated with tyrosine mutations or due to the general structural alteration of the last 39 heptads of the CTD. To address this question, three additional mutants were designed, in which the last 39 heptads of the CTD have replaced either serine-2, threonine-4 or serine-5 by alanine. The mutants were termed S2AAA, TAAA, and S5AAA, respectively and served as controls for the mutant YFFF (**Figure 18**A).

The three new mutants were cloned into the expression vector pRX4-267, transfected in Raji cells via electroporation and selected for a period of 2-3 weeks in RPMI growth media. Whole cell extracts were prepared 72 hours after the induction and 48 hours after addition of α-amanitin and proteins were analyzed by western blot analysis (**Figure 18**B). All new mutants expressed the hypophosphorylated (IIA) and the hyperphosphorylated (IIO) forms of Pol II when probed with the α-HA antibody (3F10), suggesting that the mutants are transcriptionally active. Weak expression of the Pol IIO forms in the mutant YFFF probably reflects weak induction of the recombinant Pol II. Mutants S2AAA and S5AAA showed a continuous decline in the cumulative living cell number over the period of two weeks after induction in the presence

of α-amanitin (**Figure 18**C). In contrast, the mutant TAAA displayed attenuated growth phenotype.

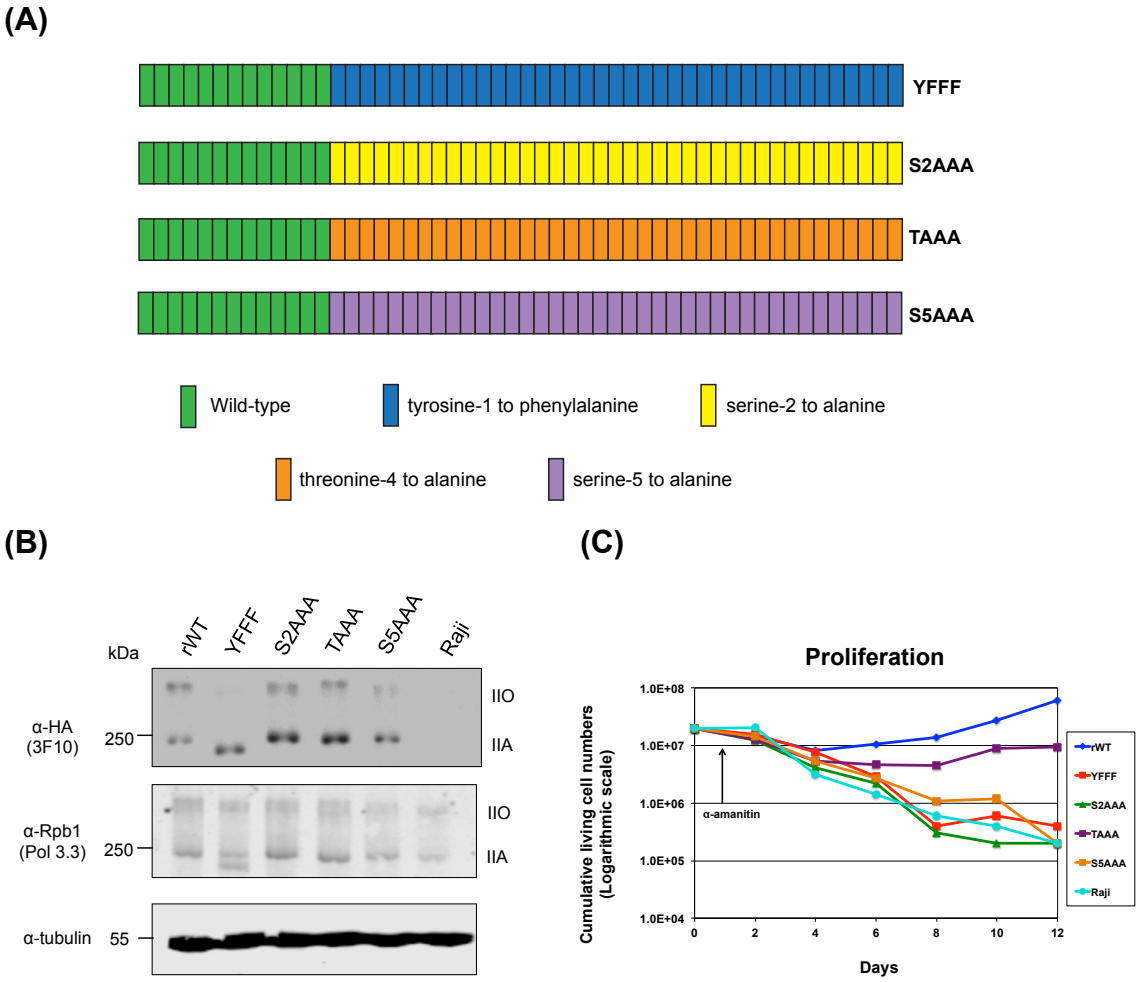


Figure 18: Control mutants for the mutant YFFF (A) A schematic representation of CTD mutants. Each box represents a single heptad of the CTD. Color code of heptad repeats are as follows: green - wild type; blue - tyrosine-1 replaced by phenylalanine; yellow - serine 2 replaced by alanine; orange - threonine 4 replaced by alanine; purple - serine 5 replaced by alanine. **(B)** Western blot analysis showing the expression of the recombinant Pol II with α-HA antibody and the endogenous plus total Pol II with the α-Rpb1 antibody (Pol 3.3). α-tubulin is used as a loading control. **(C)** Graph representing cell proliferation of mutants over a period of two weeks.

Next, in the RNA-seq analysis, the mutant S2AAA did not display any apparent RT transcription at 5' or 3' end of genes, as exemplified at the PDCD6IP gene locus (**Figure 19**). The comparison of mutants YFFF and S2AAA suggests that the RT transcription phenotype observed in the mutant YFFF is specifically associated with mutations of tyrosine residues and not due to general mutations in the last 39 heptads of the CTD.

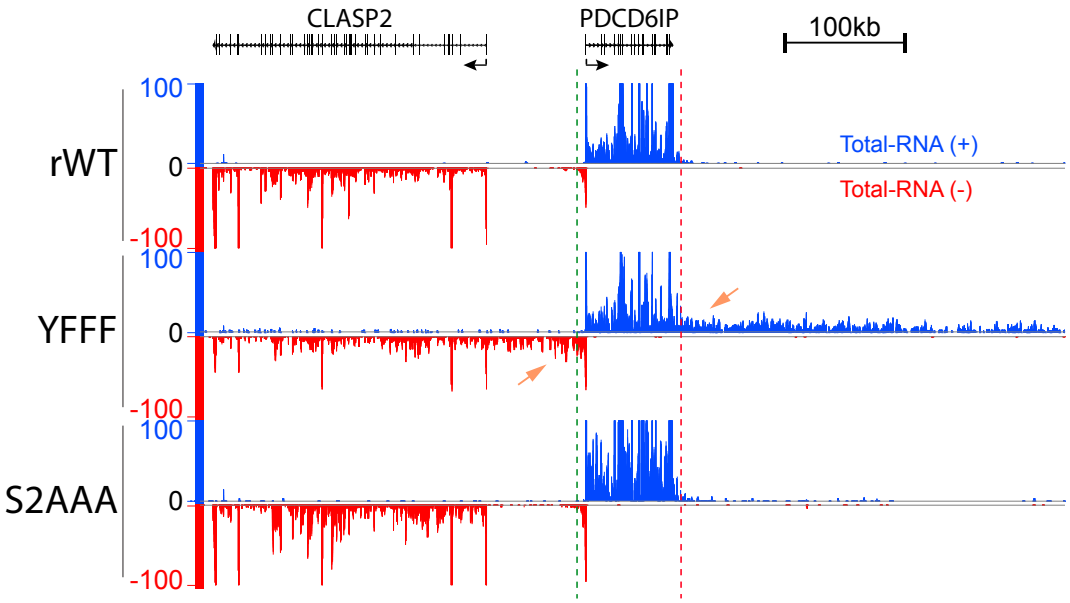


Figure 19: Comparison of RNA-seq signals at the PDCD6IP gene locus in rWT and mutants YFFF and S2AAA. A screenshot from the Integrated Genome Browser (IGB), comparing RNA-seq signals over genes CLASP2 and PDCD6IP in rWT and mutants YFFF and S2AAA. The CLASP2 gene is transcribed on the minus strand (red signals) and the PDCD6IP on the plus strand (blue signals). The annotated transcription unit of the PDCD6IP gene is marked between green and red lines. Downstream 3' end read-through transcription and an increase in antisense transcription in the mutant YFFF are marked by orange arrows.

We next asked if the 3' end RT phenotype in the mutant YFFF is associated with defects in the 3' end processing of RNAs.

2.7. PolyA$^+$-RNA-seq analysis

3' end processing of RNA is an important step in the transcription cycle. Cleavage and polyadenylation specificity factors (CPSF) recognize the AAUAAA sequence of RNA and together with the cleavage stimulation factor (CstF) act to promote RNA cleavage. Furthermore, poly(A) polymerase is recruited and facilitates polyadenylation of the 3' end of RNA (Proudfoot, 2011). To address if the 3' end RT phenotype in the mutant YFFF is due to defects in 3' end processing of RNAs, polyA$^+$-RNA-seq was performed (Materials and Methods 4.2.5.2).

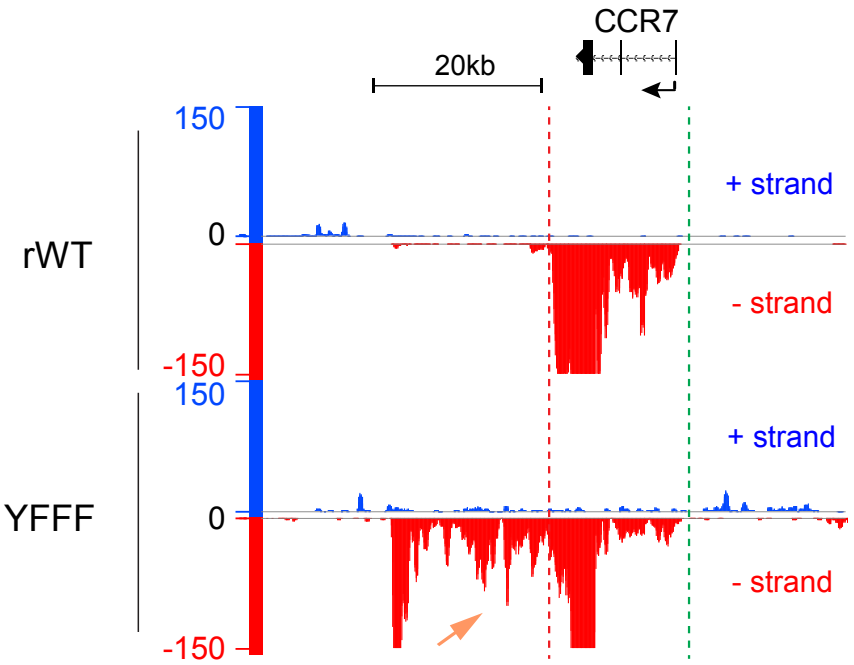


Figure 20: Comparison of polyA$^+$-RNA-seq signals for the CCR7 gene in rWT and the mutant YFFF. A screenshot from the IGB browser comparing RNA-seq signals over the gene locus CCR7 in rWT and the mutant YFFF. The gene is transcribed on the minus strand (red signals) and its annotated transcription unit is marked between green and red lines.

A representative gene, CCR7, is shown in the **Figure 20**. RNA-seq signals of polyA$^+$-RNAs were detected over the annotated transcription unit of the gene CCR7. Although there was a difference in the signal intensity at the 3' end of the gene, the enrichment of signals indicated that probably the processing and polyadenylation of the CCR7 mRNA occurred efficiently in the mutant YFFF. The mutant YFFF in addition, displayed a massive 3' end RT transcription phenotype.

In metagene analysis, signals of polyA$^+$-RNAs were comparable over the annotated transcription unit in rWT and the mutant YFFF. This suggests that the processing and polyadenylation of a large number of mRNAs in the mutant YFFF is efficient (**Figure 21**A). However, high RNA-seq signals were detected in the regions 20 kb upstream and downstream of the annotated gene boundaries in the mutant YFFF, a phenotype similar to the one observed in the analysis of total RNA.

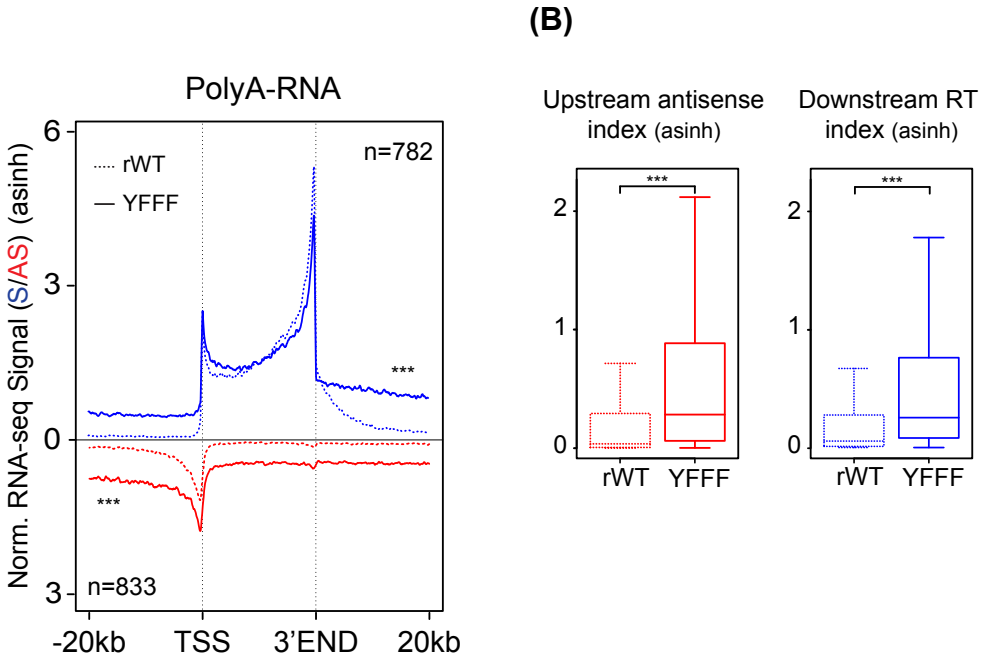


Figure 21: PolyA-RNA-seq analysis in rWT and the mutant YFFF. (A) Average profiles of read-counts of polyA$^+$-RNA-seq representing transcription in sense (blue) and antisense (red) direction of the gene body in rWT and the mutant YFFF. Signals in the profiles are distributed over gene bodies and in regions 20 kb around annotated gene boundaries. Signals are normalized to the gene body. (B) Box plot quantification of upstream antisense transcription index (red) and downstream read-through transcription index (blue) in rWT and the mutant.

Thus, the polyA$^+$-RNA-seq experiment suggested that 3' ends of RNAs in the mutant YFFF are properly processed and polyadenylated. The 3' end RT transcription phenotype in the mutant YFFF is probably then not due to defects in the 3' end processing of RNAs, but due to the failure of Pol II to terminate transcription. Whether the long RT transcripts in the mutant YFFF can be properly processed and if Pol II can terminate at all remains unclear.

2.8. ChIP-seq analysis

Next, ChIP-seq (chromatin immunoprecipitation) experiments were performed to map the occupancy of Pol II in rWT and the mutant YFFF (Materials and methods 4.2.4).

2.8.1. The mutant YFFF displays increased Pol II occupancy downstream of 3' ends of genes

Pol II occupancy in rWT and the mutant YFFF was compared at individual genes using the IGB genome browser. A representative gene, UMPS, is shown in the **Figure 22**.

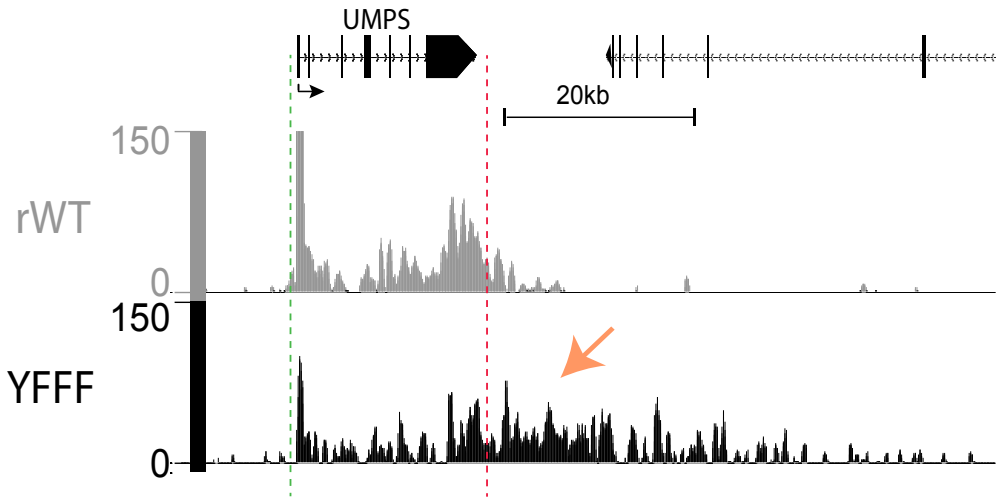


Figure 22: Pol II occupancy at the UMPS gene locus in rWT and the mutant YFFF. A screenshot from the Integrated Genome Browser (IGB) comparing the occupancy of Pol II over the gene locus UMPS. Green and red lines mark the annotated transcription unit of the gene. Orange arrow indicates Pol II occupancy downstream of the 3' end site in the mutant YFFF.

In rWT, the Pol II occupancy signals were detected over the transcription unit of the gene UMPS, but not in the region downstream of the annotated 3' end site. In contrast, in the mutant, in addition to the gene body, substantial Pol II density was detected in regions downstream of the annotated 3' end site. High Pol II density downstream of the annotated 3' end site in the mutant YFFF was observed at a genome-wide scale as revealed in the metagene analysis (**Figure 23**A) and confirmed by boxplot quantifications (**Figure 23**B). In addition, a peak near the 3' end of genes was visible in rWT, corresponding to the paused Pol II; a prerequisite for transcription termination. This peak was strongly increased and shifted by around 2.6 kb downstream of the annotated 3' end site in the mutant YFFF, indicative of a delay in Pol II pausing (**Figure 23**C).

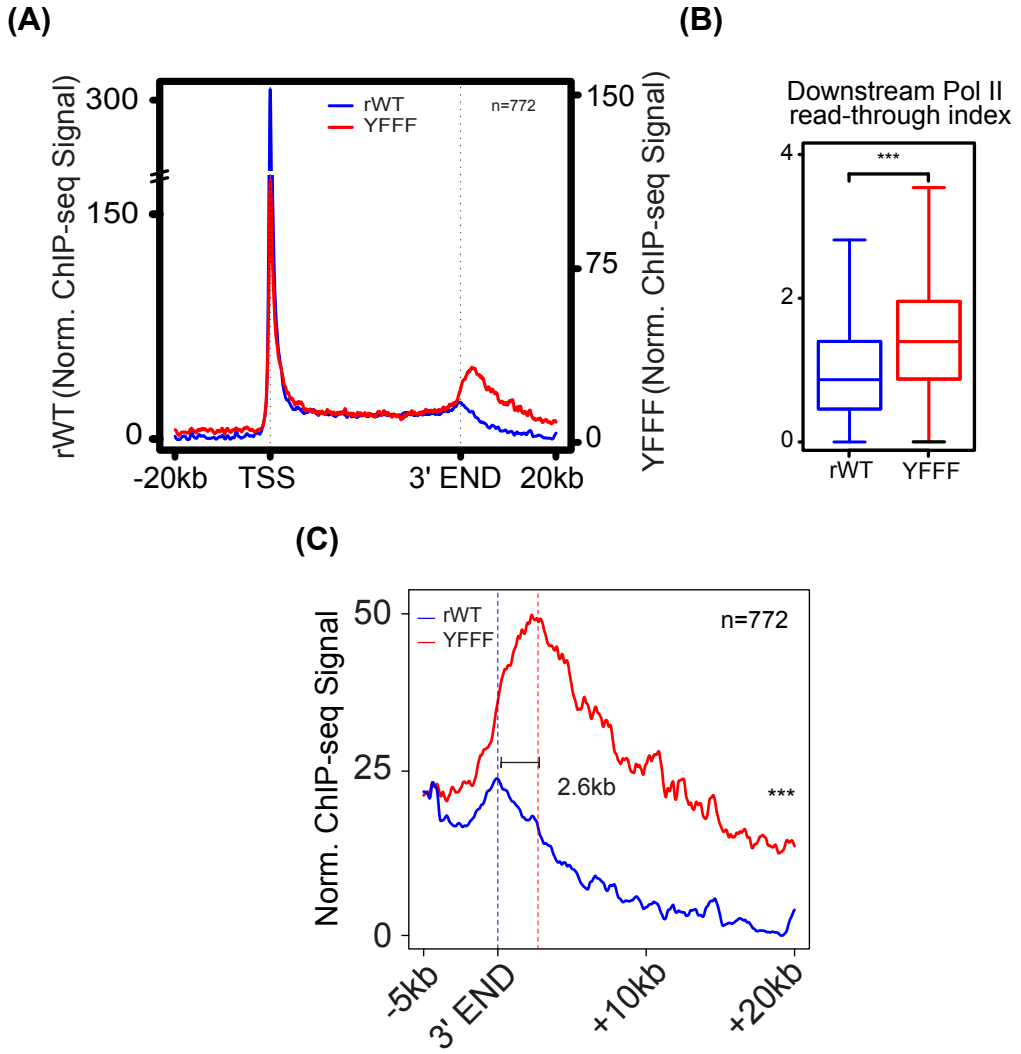


Figure 23: ChIP-seq analysis in rWT and the mutant YFFF. (A) Average metagene profiles of Pol II density in rWT (blue) and the mutant (red). A total of 772 genes from the hg19 were selected for analysis. Signals in the profiles are distributed over the gene bodies and in regions 20 kb around the annotated gene boundaries. Signals are normalized to the gene body. **(B)** Box plot quantification of downstream Pol II read-through transcription index in the rWT and the mutant YFFF. **(C)** Average Pol II density profiles of rWT and the mutant YFFF near the 3' end site.

Thus, ChIP-seq analysis demonstrated that the mutant Pol II fails to dissociate off the template DNA and traverses up to several hundred kbs beyond of the annotated 3' end site corroborating the observation of global termination defects.

2.8.2. The mutant YFFF shows reduced Pol II occupancy near TSS

Additionally, ChIP-seq analysis displayed a substantial loss in the Pol II density near TSS in the mutant YFFF, as exemplified in the **Figure 24**A.

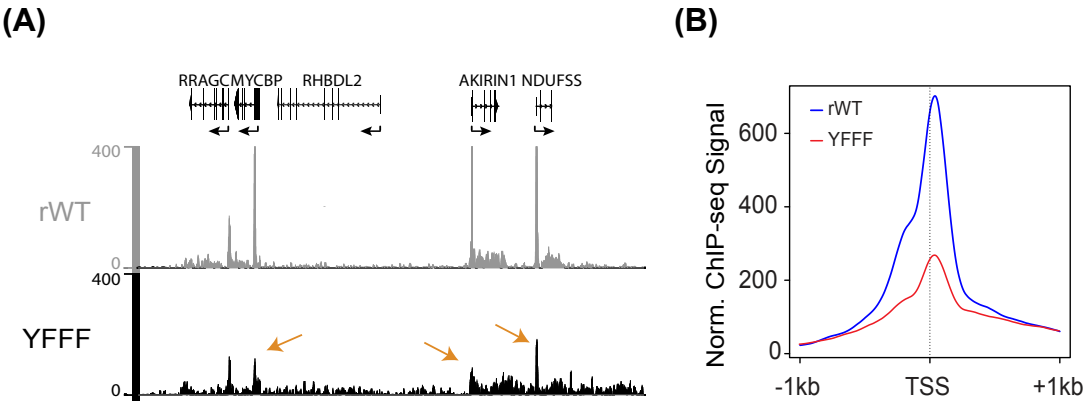


Figure 24: Pol II occupancy near TSS in rWT and the mutant YFFF (A) A screenshot from the Integrated Genome Browser (IGB) comparing the Pol II density over genes. Reduced Pol II occupancy near the TSS of genes in the mutant YFFF is marked by an orange arrow. (B) Metagene profiles of average Pol II density in rWT (blue) and the mutant YFFF (red) at TSS and in the regions 1 kb around the TSS.

Pol II density was detected over the annotated transcription unit of the genes, RRAGC, MYCBP, AKIRIN1 and NDUFSS in both the cell lines. The mutant YFFF displayed reduced Pol II occupancy near the TSS of genes MYCBP, AKIRIN1 and NDUFSS (orange arrows) compared to rWT. After normalizing Pol II density to the gene bodies, the mutant YFFF showed a massive reduction in the Pol II density near the TSS at a genome-wide scale (**Figure 24**B). This reduced Pol II occupancy at the TSS in the mutant prompted us to investigate promoter-proximal pausing in the mutant YFFF.

2.9. Promoter-proximal pausing

In many metazoan genes, Pol II pauses around 20-60 nucleotides downstream of the TSS before transiting into productive elongation. This is an important regulatory step in the transcription cycle and described as promoter-proximal pausing (Adelman et al., 2012). The Pol II pausing scores in rWT and the mutant YFFF were calculated by dividing Pol II occupancy signals at the promoters (-300 bp to +100 bp) by the signals in the gene body. The heatmap of Pol II density was generated by the order of increasing pausing scores in rWT (**Figure 25**A). A similar heatmap representing the pausing score of corresponding genes in the mutant YFFF was generated

(**Figure 25**A). Genes were classified into three groups, 1) low paused genes, 2) moderately paused genes and 3) highly paused genes.

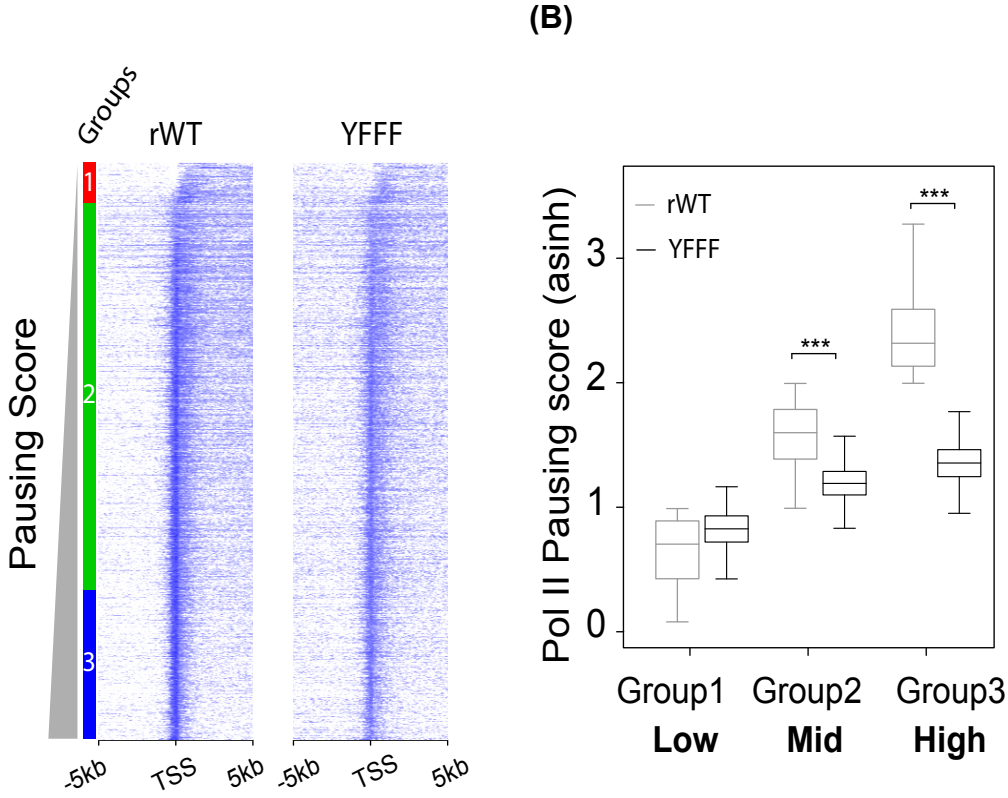


Figure 25: Promoter-proximal pausing in rWT and the mutant YFFF (A) Heatmaps representing Pol II pausing score in rWT and the mutant. Genes are ranked in the order of increasing pausing score in rWT and classified into three groups. (B) Quantification of Pol II pausing score for corresponding groups of genes in rWT (gray) and the mutant (black).

Comparison of Pol II pausing score between rWT and the mutant showed that genes that are moderately and highly paused in rWT, displayed significantly low pausing score in the mutant YFFF (**Figure 25**B). This indicates an early release of promoter-proximal paused Pol II in the mutant YFFF. Genes that have low pausing in rWT did not display any significant change.

Taken together, the mutant YFFF revealed a global defect in transcription termination at 5' and 3' ends of genes and an early release of promoter-proximal paused Pol II. However, the reason for these aberrant transcription phenotypes was not known. To address this, pursued with a Mass Spectrometry (MS) approach. The rationale was to see if the mutant Pol II reveals gain and/or loss of binding to certain transcription factor/s, potentially explaining the cause of pervasive transcription and reduced pausing.

2.10. Mass spectrometric analysis

Pol II from rWT and the mutant YFFF was immunoprecipitated using α-HA antibody (12CA5), 72 hours after the induction (Materials and methods 4.2.3.2). The immunoprecipitated samples were subjected to either in-gel trypsin digest or on-beads trypsin digest (Materials and methods 4.2.3.4 and 4.2.3.5) before processing them for the LC-MS/MS analysis (Materials and Methods 4.2.3.6).

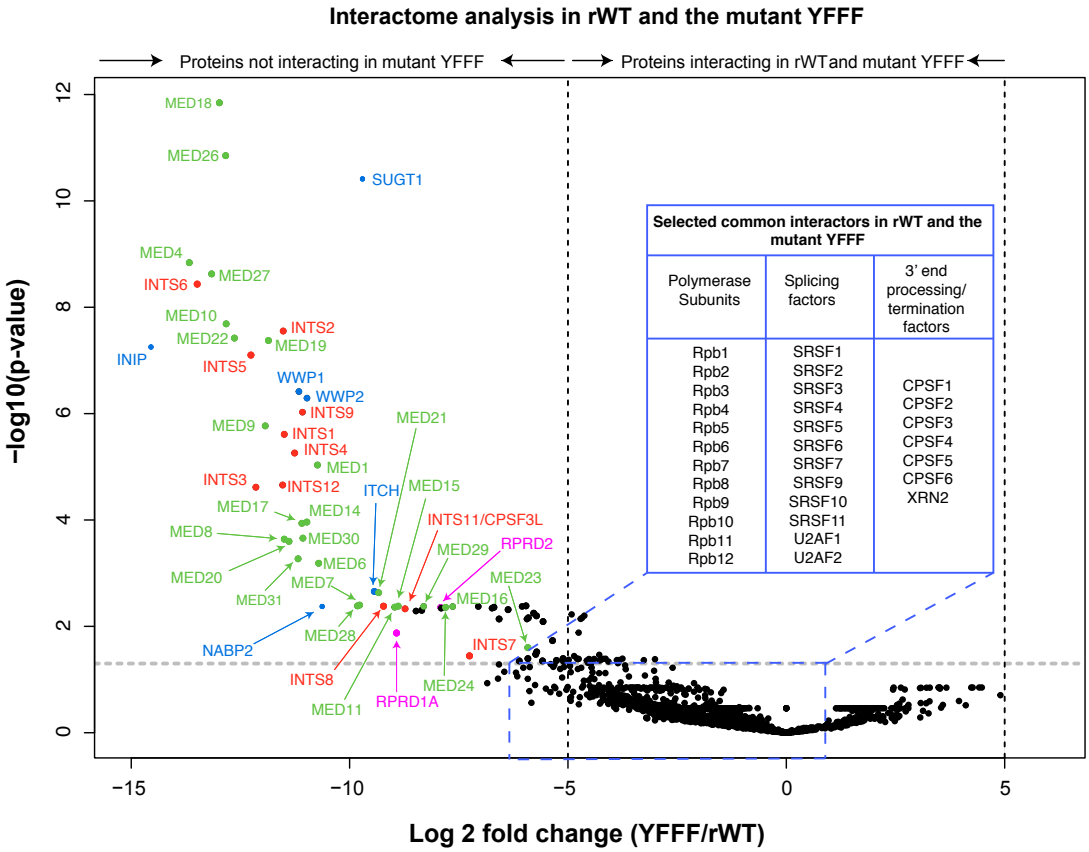


Figure 26: Volcano plot comparing Pol II interactome in rWT and the mutant YFFF. The table on the right lists selected proteins and complexes that interact with the Pol II of both, rWT and the mutant. Represented on the left are the proteins that do not interact with the mutant Pol II. Highlighted are 25 subunits of Mediator complex (green); 11 subunits of Integrator complex (red); CTD phosphatase (Magenta); E3-ubiquitin ligase, components of SOSS complex and others (blue). Threshold: Log2 fold change ≥ 5; p-value < 0.05. Data is based on five independent biological replicates.

A volcano plot was generated that provided information about the common interactors in rWT and the mutant YFFF. In addition, the information about proteins, which showed loss of interaction with the Pol II in mutant YFFF, was also revealed (**Figure 26**). Peptides of all 12 subunits of the Pol II (Rpb1-12)

and several splicing factors were detected as common interactors. In addition, the peptides corresponding to the 3' end processing factors, such as the cleavage and polyadenylation specificity factor (CPSF) and a 5'->3' exonuclease, Xrn2, were also present in comparable amounts in both samples. In total six subunits of the CPSF were detected. A table showing the peptide counts of the selected common interactors for each biological replicate is in the **Supplementary table 1**. The ratio of the log2 fold change (mutant YFFF/rWT) for these proteins is in the **Supplementary table 2**.

Next, a total of 69 proteins were found that showed loss of interaction with the Pol II in mutant YFFF. These proteins are shown on the left side of the volcano plot. A table showing the peptide counts of the 69 proteins and its log2fold change ratio is shown in the **Supplementary table 3** and 4, respectively.

Of the 69 proteins more than 50% belonged to two large cellular complexes, the Mediator (green) and the Integrator (red). Both, complexes are known to interact with Pol II CTD (Baillat et al., 2005). The Mediator is a protein complex that is composed of four core structural modules, the Head, the Middle, the Tail and a dissociable CKM (Allen et al., 2015). In the MS analysis, 25/30 subunits of the mediator complex were detected in rWT, but not in the mutant YFFF (Supplementary table 3). Interestingly, in rWT, the peptides for the entire CKM module were lacking. Integrator is a metazoan-specific, 14 subunits protein complex that interacts with CTD in a Ser2-P and Ser7-P dependent manner (Egloff et al., 2010). In our experiments, 11 subunits of the Integrator complex were enriched in rWT but not in the mutant YFFF.

Pol II in the mutant YFFF also showed the loss of interaction with RPRD1A and RPRD2. These proteins serve as scaffolds that recruit serine-5 phosphatase, RPAP2 (Ni et al., 2014). In addition, E3 ubiquitin ligases like - ITCH, WWP1 and WWP2 as well as components of the SOSS complex, INIP and NABP2, showed impaired interaction with the Pol II in mutant YFFF.

Next the Pol II interactome in the mutant S2AAA was studied (**Figure 27**) to investigate, if the loss of interaction of the Mediator and Integrator with Pol II in the mutant YFFF is specific for tyrosine mutations, or can be observed also for the mutant S2AAA.

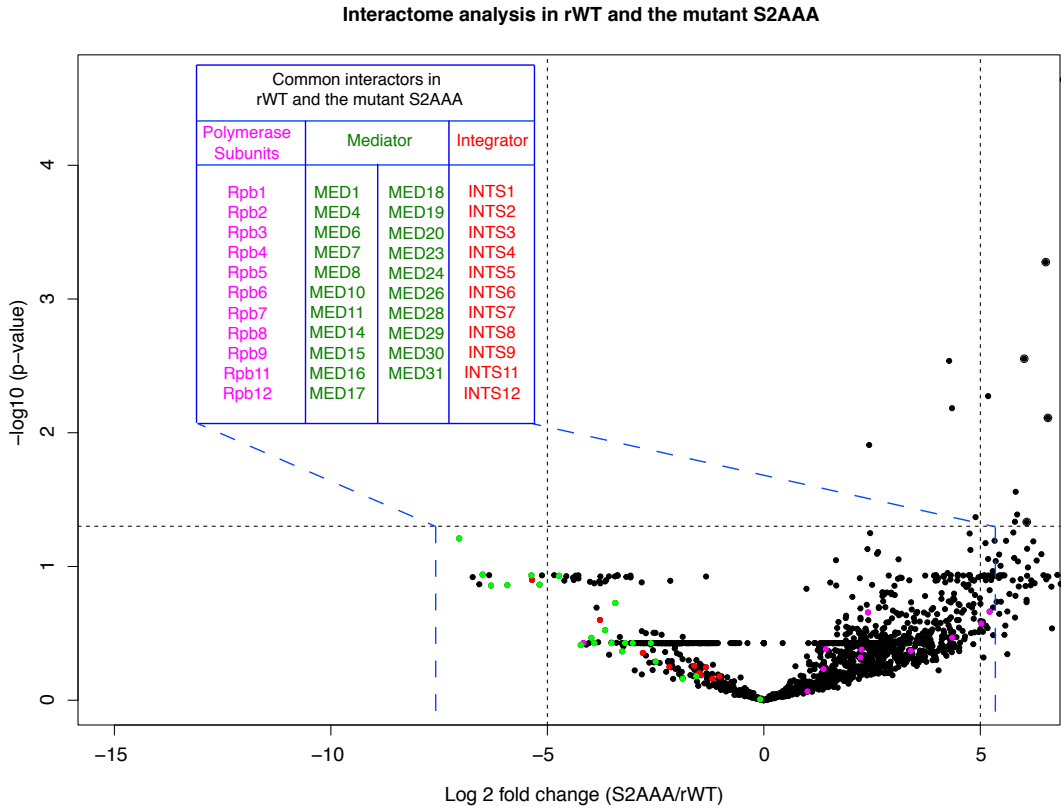


Figure 27: Volcano plot comparing the Pol II interactome in rWT and mutant S2AAA. A table on the left lists selected proteins and complexes that interact with Pol II in rWT and mutant S2AAA cells. Highlighed are subunits of Pol II (magenta), Mediator (green) and Integrator (red). Threshold: Log2 fold change ≥ 5; p-value < 0.05.

MS data revealed that the peptides corresponding to all 12 subunits of Pol II (magenta), 21 subunits of the Mediator (green) and 11 subunits of Integrator (11 subunits) were detected in the Pol II interactome of rWT and mutant S2AAA. A table showing the ratio of log2 folds change of these proteins is shown in the **Supplementary table 5**. MS data suggested that the Mediator and Integrator can interact with Pol II in the mutant S2AAA and that the loss of binding to Pol II in the mutant YFFF is specific for the tyrosine mutations.

Finally, we asked if the failure to detect peptides of Mediator and Integrator subunits in the mutant YFFF was due to the loss of the expression of both complexes. To address this, western blot analysis was performed to detect the expression of the Mediator subunit 15 (MED15) and the Integrator subunit 11 (INTS11) (**Figure 28**).

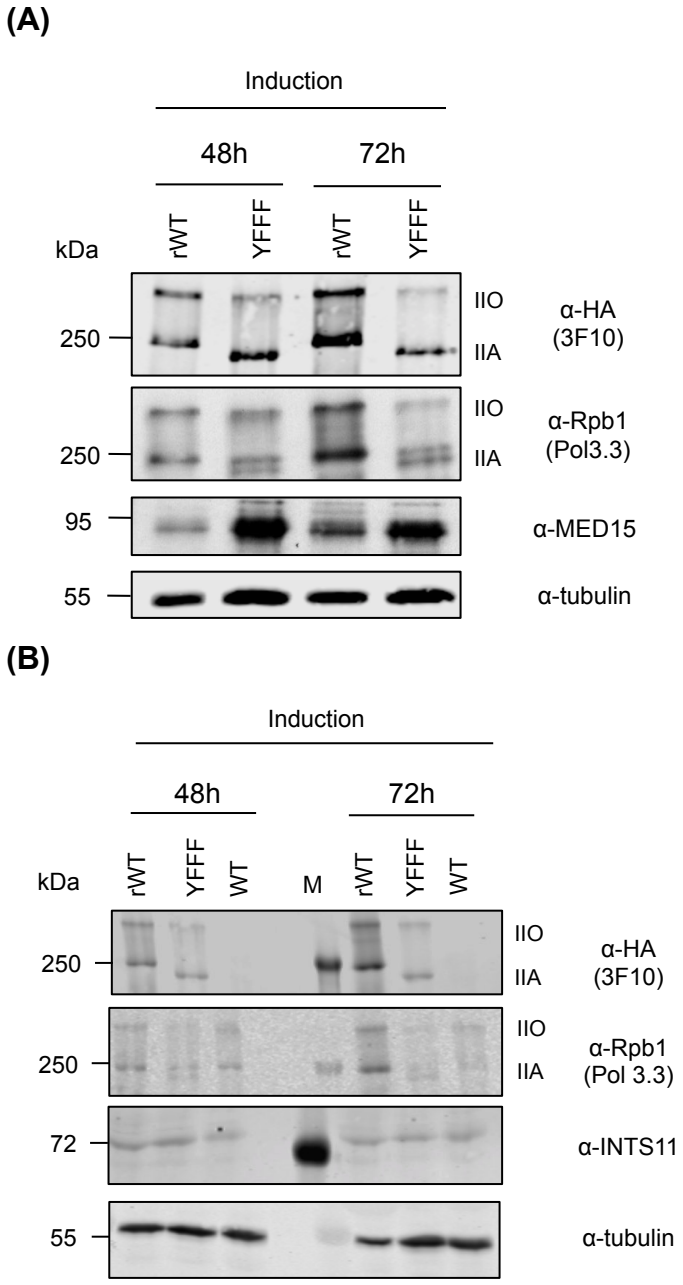


Figure 28: Expression of the Mediator subunit 15 (MED15) and the Integrator subunit 11 (INTS11) in rWT and the mutant YFFF. Cells were lysed 48 and 72 hours after the induction and the expression of the recombinant Pol II was analyzed using α-HA antibody (3F10). Expression of the recombinant and the endogenous Pol II was detected with α-Rpb1 (Pol 3.3) antibody. α-MED15 antibody and α-INTS11 antibody were used to detect the expression of the Mediator subunit 15 and Integrator subunit 11, respectively. α-tubulin was used as a loading control.

Both proteins, MED15 (**Figure 28**A) and INTS11 (**Figure 28**B), were expressed in rWT and mutant YFFF cells, 48 and 72 hours after induction. Interestingly, the expression level of the MED15 subunit was stronger in the mutant. This could be due to the stabilization of the MED15 subunit. Together,

western blot and the MS analysis demonstrated that the Mediator and the Integrator are expressed, but did not interact with Pol II in the mutant YFFF.
Thus, the MS analysis revealed that the Mediator and the Integrator interact with the CTD in a tyrosine-dependent manner. Tyrosine residues in the first 13 heptads of the CTD are not sufficient for the recruitment of these complexes.

3. Discussion

This thesis unravels the functional significance of Tyr1 of mammalian RNA polymerase II CTD in the regulation of global termination of gene transcription. We show that Tyr1 strongly limits the extent of pervasive transcription at 5' and 3' end of genes and also contributes to the regulation of promoter-proximal pause/release. The study further provides the clue for the involvement of the Mediator and/or the Integrator complexes in the regulation of transcription termination and/or Pol II pausing in a tyrosine-dependent manner.

3.1. Generation and characterization of tyrosine mutants

The functional role of Tyr1 residues of CTD can be studied by characterizing transcriptional active Pol II mutants that lack Tyr1. The replacement of Tyr1 by phenylalanine (Y1F) in all 26 heptads leads to the degradation of CTD in chicken DT40 cells. Strikingly, the mutant in which only a single phenylalanine was reverted back to tyrosine in the last repeat, prevented the degradation of CTD in chicken cells (Hsin et al., 2014). A previous study from our laboratory reported that Y1F replacements in heptads 4-51 lead to the degradation of CTD in mammalian cells, making it difficult to characterize the function of Tyr1 in this mutant (Descostes et al., 2014).

In my thesis, I developed a strategy to overcome this challenge. My first objective was to generate a set of Y1F mutants that are transcriptionally active. I also wanted to unravel the functional significance of Tyr1 residues in different parts of CTD. Considering these, I designed seven CTD constructs, in which Y1F mutations were introduced in either 26 or 39 heptads of CTD (**Figure 6**). Tyrosine was replaced by phenylalanine as both amino acids have an aromatic core structure. However, tyrosine has an -OH group and can be phosphorylated. I cloned these constructs into the episomal expression vector pRX4-267, transfected them into Raji cells and made use of the tetracycline inducible, Rpb1 knockout-knockin system to study the properties of the recombinant Pol II.

Following induction of the recombinant Pol II and shutdown of the endogenous Pol II by α-amanitin, all mutants stably expressed two forms of Pol II (**Figure 8**), which differ in the apparent molecular weight and in their phosphorylation state (Dahmus, 1981). Work by Dahmus and colleagues showed that the lower band, Pol IIA, is hypo-phosphorylated, whereas, the upper band, Pol IIO arises from the hyper-phosphorylation of CTD (Cadena et al., 1987). Pol IIA associates with the preinitiation complex at promoters. The transition from initiation to elongation is accompanied by the hyper-phosphorylation of Pol IIA and conversion to the transcriptionally active, Pol IIO form (Kang et al., 1993). The formation of the Pol IIO form by the tyrosine mutants suggested that all seven mutants were transcriptionally active. Pol II in all seven mutants is phosphorylated at Ser2 of CTD, suggesting that all polymerases were elongation-competent (Ahn et al., 2004). This was a very important result as it provided me with a set of Y1F mutants that could be characterized further.

Next, I tested the cell lines expressing Pol II CTD tyrosine mutants for their ability to proliferate in the presence of α-amanitin (**Figure 9**). The mutants, FYYF and YYFF carry 26 heptads with Y1F mutations and were viable after two weeks of α-amanitin selection. This suggests that Y1F mutations in the first and the last 13 heptads or the last 26 heptads of CTD are tolerated in mammalian cells. In contrast, the remaining five mutants displayed a lethal phenotype. The lethal mutants as well as the untransfected Raji cells displayed a cell viability of around 90% after one day of α-amanitin selection. Their cell viability declined to less than 50% after four days of selection. The mutants FFYY and YFFY also carry 26 heptads with Y1F mutations, but displayed a lethal phenotype. This suggests that not the number, rather the position of Y1F mutations affect cell viability. Intriguingly the viable mutants had conserved Tyr1 in heptads 14-26, whereas the mutants with a lethal phenotype had mutation in these heptads. This observation suggests that Tyr1 in heptads 14-26 might be particularly critical for cell growth, and perhaps recruit essential transcription factor/s for proper gene expression. However, this assumption needs further investigations. The mutant YF26 had Y1F substitutions in every alternate repeat and displayed a lethal phenotype. This is consistent with the functional unit studies of CTD in yeast. John Stiller

and colleagues showed that a single functional unit in yeast CTD is embedded within paired-heptads (Stiller et al., 2004). Essentially, the yeast CTD functional unit requires paired tyrosines spaced 7 amino acids apart (Y1-Y8) and the three SP motifs in a 2-5-9 direction ($Y_1S_2P_3T_4S_5P_6S_7Y_8S_9P_{10}$). Disruption of this motif is lethal in yeast. The lethality of the mutant YF26 in mammalian cells indicates that similar to yeast, the mammalian CTD probably needs tyrosine residues spaced seven amino acids apart. Finally, the mutants with 39 mutated heptads in either the proximal (FFFY) or the distal part (YFFF) displayed a lethal phenotype. This suggests that mammalian cells may require more than 13 Tyr1 residues to support cell growth.

3.2. Transcriptome analysis of tyrosine mutants

Next, I wanted to investigate the transcriptome of tyrosine mutants. For this, I collaborated with the laboratory of Jean-Christophe Andrau in Montpellier. We first performed the RNA-seq analysis of total RNA for four of the seven mutants, FYYF, YYFF, YFFY and YFFF. Two mutants displayed a viable phenotype and two other a lethal phenotype. In Principal Component Analysis (PCA), the biological replicates of the mutant YFFF clustered far away from the replicates of rWT and other tyrosine mutants (**Figure 10A**). Surprisingly, one replicate of rWT was separated from the second replicate on a PC1 scale. This could probably be due to the differences in the expression value of few genes between two replicates. In differential gene expression analysis a number of genes in tyrosine mutants were differentially expressed, with most of them being upregulated (**Figure 10B**). Interestingly, the mutant YFFF displayed the highest number of differentially expressed genes. The mutant YFFF also displayed a strong transcription read-through (RT) phenotype at 5’ and 3’ end of genes in our RNA-seq analysis. Such a phenotype was seen, but to a lesser extent, in other tyrosine mutants (**Figure 12**). Here, we performed the non-strand specific RNA-seq analysis and hence could not discriminate between sense and antisense transcription. As the mutant YFFF displayed the highest number of differentially expressed genes and has the strongest RT phenotype, we pursued to characterize this mutant in detail.

3.3. The mutant YFFF displays an increase in antisense and 3' end RT transcription phenotype

We next performed the strand-specific RNA-seq analysis of total RNA in rWT and the mutant YFFF. Here, the mutant YFFF not only confirmed the 3' end RT phenotype, but also displayed an increase in antisense transcription upstream of 5' ends of genes (**Figure 15**). Such a phenotype was not observed in the mutant S2AAA, suggesting that the RT phenotype is specifically associated with the loss of tyrosine residues and do not arise due to the structural alterations in the last 39 heptads of the CTD. We further compared the RT transcription at 5' and 3' end of genes in the mutant YFFF (**Figure 16**). Genes with higher 3' end RT transcription displayed higher upstream RT transcription at 5' end of genes, while genes with lower 3' end RT transcription displayed lower upstream RT transcription at 5' end of genes in the mutant YFFF. This suggests that the RT transcription at 5' and 3' ends of genes in the mutant YFFF is probably associated with each other. It is important to note that not all genes in the mutant YFFF displayed aberrant RT transcription phenotype. Several genes were normally transcribed and did not display any transcription defects. One such representative gene is shown in the **Supplementary Figure 1**. Thus, our RNA-seq data suggests that the RT phenotype in the mutant YFFF was a global, but not a general phenotype. High levels of Pol II downstream of the annotated 3' end site were detected in the mutant YFFF in our ChIP-seq analysis (**Figure 23**) further supporting the observation of RT transcription phenotype.

Antisense transcription is defined as the transcription from the strand opposite to the protein coding or sense strand. In mammalian cells, many polymerases bound to promoters are associated with upstream antisense transcription. The divergent initiation of transcription by Pol II at TSS is a general phenomenon for mammalian promoters (Core et al., 2008; Seila et al., 2008). Divergent transcription generates upstream antisense RNAs that are typically short and relatively unstable (Wu et al., 2013). Recent studies show that upstream antisense RNAs are cleaved and polyadenylated shortly after initiation (Almada et al., 2013). Longer non-coding antisense transcripts can be detected upstream of genes in conditions of inhibition of the RNA

degradation machinery (Lepoivre et al., 2013). In the mutant YFFF, upstream antisense transcripts seem to be polyadenylated (**Figure 21**) and extend beyond the natural antisense transcription.

RT transcription downstream of annotated 3' end of genes has been reported in various situations of cellular stress or inactivation of certain transcription factors. In clear cell renal cell carcinoma (ccRCC) high levels of RT transcription is prevalent (Grosso et al., 2015). The authors identified Setd2 inactivation as a major driving force of impaired transcription termination and high levels of RT transcription. In a different study, the induction of osmotic stress in cultured cells resulted in a large number of genes failing to terminate at the 3' end of genes (Vilborg et al., 2015). The RT transcripts derived from protein-coding genes were called downstream of genes (DoGs) transcript and often transcribed over long non-coding regions. Infection by viruses may be considered as an extreme form of cellular stress. Herpes simplex virus infection causes a massive disruption of host gene transcription termination (Rutkowski et al., 2015). Although, the prevalence of RT transcription described in these studies is not as massive as observed in the mutant YFFF, it would be interesting to see if and how tyrosine phosphorylation of CTD is affected in infected cells.

RT transcription downstream of the annotated 3' end gene boundaries often interferes with the transcription of the downstream genes. In our RNA-seq analysis, high RNA-seq signals downstream of the ZBTB44 gene seem to interfere with the transcription of the ST14 gene in the mutant YFFF (**Figure 17**). To investigate whether the high RNA-seq signals downstream of the ZBTB44 gene are due to new-initiation, we mapped the epigenetic modification marks H3K4me3 and H3K27ac. The H3K4me3 maps active promoters and H3K27ac enhancers. Both marks were absent in the region downstream of the ZBTB44 gene, suggesting that the high RNA-seq signals do not correspond to new initiation, but are caused by the 3' end RT transcription. Interestingly, the H3K4me3 mark was present at the ST14 gene promoter, suggesting that the gene is transcribed in the mutant YFFF. However, weak RNA-seq signals over the gene body suggested repressed transcription. How the expression of the ST14 gene in the mutant YFFF is affected remains unanswered.

3.4. 3’ end processing of mRNAs is not affected in the mutant YFFF

One important question was whether the RT phenotype in the mutant YFFF is due to an improper 3’ end processing of RNAs. There are 3 well-described mechanisms for the 3’ end processing of Pol II transcripts (Introduction 1.5.3). Here, we focused our analysis on the 3’ end processing of protein-coding genes, during which cleavage of the nascent transcript by CPSF73 is followed by 3’ polyadenylation of the cleaved transcript. According to the torpedo model of termination, cleavage of the nascent transcript creates an entry site for a 5’->3’ exonuclease, Xrn2, which degrades the downstream-cleaved product and releases Pol II from the template DNA. The laboratory of Nicholas Proudfoot analyzed the knockdown of Cleavage and Polyadenylation factors (CPA) as well as Xrn2 in the process of termination (Nojima et al., 2015). The knockdown of CPSF73 lead to termination defects, but the knockdown of CstF64 or Xrn2 did not show any significant termination defects at a genome-wide scale. In contrast, a study from the laboratory of David Bentley showed that the inactivation of Xrn2 leads to substantial defects in transcription termination for most protein-coding genes (Fong et al., 2015). They reported that the expression of an inducible dominant-negative Xrn2 mutant (D235A), in combination with shRNA-mediated knockdown of the endogenous Xrn2, caused a general inhibition of Pol II termination and shifted the termination zone further downstream of genes.

We observe that the Pol II in mutant YFFF can efficiently recruit Xrn2 as well as the CPSF complex, including the CPSF73 in our MS data. However, the massive RT transcription phenotype in the mutant YFFF suggests that these proteins probably contribute to the regulation of termination, but do not contribute to the removal of Pol II from the DNA template. Further, polyA$^+$-RNAs were efficiently enriched in the mutant YFFF as in rWT cells (**Figure 21**), indicating that the 3’ end processing of RNAs is not affected. Together, our data demonstrated that the mutant YFFF has defects in transcription termination and displayed RT transcription downstream of the functional 3’ end site. Whether the long RT transcripts in the mutant YFFF can efficiently terminate remains elusive, but appears very unlikely.

3.5. Impaired recruitment of the Mediator and the Integrator to Pol II in the mutant YFFF

Although we could not elucidate the exact underlying mechanism of the global termination defect in the mutant YFFF, our MS data provided a possible explanation to it. Pol II in the mutant YFFF revealed a strong impairment in the recruitment of Mediator (MED) and Integrator (INT) complexes (**Figure 26**). Peptides corresponding to 25 subunits of the Mediator and 11 subunits of the Integrator were detected in the Pol II interactome of rWT, but not of mutant YFFF. The recruitment of Mediator and Integrator to Pol II in the mutant S2AAA suggested that the loss of interaction of these complexes is specific for tyrosine mutations. Here, I will describe how the Mediator and the Integrator interact with the Pol II CTD. The roles of these complexes in transcription termination will be discussed in the chapter 3.6.

The Mediator was first isolated as a complex of 20 proteins in yeast (Kim et al., 1994) and its mammalian counterpart was identified later (Jiang et al., 1998). The Mediator is composed of four distinct modules termed as the Head, Middle, Tail and a dissociable CKM kinase module. Mediator complex interacts with the unphosphorylated Pol II CTD (Kim et al., 1994; Myers et al., 1998; Naar et al., 2002; Tsai et al., 2013) and can also interact with the general transcription factor, TFIIH (Esnault et al., 2008). CTD contacts the Head as well as the Middle module of the Mediator (Robinson et al., 2012; Tsai et al., 2013). The Kornberg lab published the X-ray crystal structure of the Head module of the Mediator bound to a four-heptad long CTD peptide (Robinson et al., 2012). In the structure the CTD adopted an extended conformation and interacted with the MED6, MED8 and MED17 subunits of the Head module. MED8 and MED17 were positioned close to each other and created a binding pocket, which was rich in hydrophobic residues. Tyr8 of the used CTD peptide was buried within this binding pocket. Furthermore, Tyr15 of the CTD peptide formed the hydrogen bond with the Arg173 residue of the MED6. In a different study, Francisco Asturias and colleagues examined the Mediator-CTD interaction using Electron Microscopy. They used GST-tagged unphosphorylated full-length CTD and observed that the CTD in addition to the Head module also makes contacts with the Middle module of the Mediator

(Tsai et al., 2013). As the X-ray structure revealed that the Tyr1 is directly involved in the Mediator-CTD interaction, the lack of Tyr1 residues in the mutant YFFF could explain the impaired recruitment of the Mediator to Pol II in our MS data. This further suggests that the Mediator does not bind to the first 13 heptads of the CTD and provides a hint for the heptad specific functions of the CTD, wherein specific heptads recruit specific transcription factors. However, the heptad-specific binding of Mediator requires further analysis.

The Integrator is a metazoan specific multisubunit complex. The initial purification identified 12 subunits of the Integrator complex (INTS1 to INTS12) (Baillat et al., 2005) and the subsequent proteomic analysis further identified two additional subunits, INTS13 and INTS14 (Baillat et al., 2015). Although the co-crystal structure of the Integrator bound to a CTD peptide is not available, Shona Murphy and colleagues showed that serine-7 phosphorylation of the CTD is crucial for the recruitment of the Integrator complex (Egloff et al., 2012). Ser7-P of the CTD recruits Ser5-P phosphatase RPAP2, which dephosphorylates Ser5-P and aids in the recruitment of Integrator. Tyr1 residues may not directly recruit the Integrator, but might provide the structural conformation to the CTD for the Integrator recruitment. The replacement of Tyr1 by phenylalanine in the CTD increases the hydrophobicity and this might alter the CTD conformation important for the recruitment of the Integrator.

3.6. Roles of the Mediator and the Integrator in transcription termination

The Mediator complex has been described to regulate transcription initiation as well as promoter-proximal pause/release, but little is known about its role in the process of termination. However, the MED18 subunit of the complex has been shown to regulate transcription termination in yeast (Mukundan et al., 2011). This study reported that the MED18 was highly enriched at the 3’ end of genes. In the absence of MED18 the recruitment of termination factors was compromised leading to a read-through transcription phenotype. In our proteomic studies, MED18 is one of the most strongly affected subunits that lost binding to Pol II in the mutant YFFF. However, whether the termination

defects in the mutant YFFF is a consequence of impaired recruitment of the MED18 remains elusive.

In contrast to the Mediator, the Integrator has been well characterized in the control of transcription termination of diverse classes of genes (Skaar et al., 2015). The laboratory of Ramin Shiekhattar showed that the Integrator is recruited to snRNA genes and mediates the snRNA 3' end processing (Baillat et al., 2005). 3' end processing of snRNA is mainly catalyzed by the subunits INTS9 and INTS11, which share sequence homology with CPSF73 and CPSF100 subunits of the CPSF complex, respectively. In addition, the Integrator regulates the transcription termination of histone genes (Skaar et al., 2015). Integrator bind to the 3' end of histone genes and the depletion of Integrator subunits by siRNA-mediated knockdowns resulted in defects in histone mRNA processing and termination failure.

Thus we speculate that the impaired recruitment of Mediator and/or Integrator complexes to Pol II in the mutant YFFF leads to a phenotype of global termination defects. Whether one or both of these complexes contribute to the termination defects in the mutant YFFF is not yet clear and needs further experimental evidences.

3.7. An early release of promoter-proximal paused Pol II in the mutant YFFF

Pol II in the mutant YFFF displayed reduced occupancy near the TSS in our ChIP-seq analysis (**Figure 24**), indicative of either reduced Pol II initiation and/or early release of promoter-proximal paused Pol II. To investigate this, we analyzed promoter-proximal pausing in the mutant YFFF. Genes that were highly paused in rWT displayed significantly lower Pol II pausing score in the mutant YFFF, suggesting an early release of promoter-proximal paused Pol II (**Figure 25**).

Pol II pauses around 20-60 nucleotides downstream of the transcription start site before transiting into productive elongation. This rate-limiting step of transcription is described as promoter-proximal pausing and is controlled by multiple transcription factors. The DRB-sensitivity factor (DSIF) and negative elongation factor (NELF) functions together to induce Pol II pausing, while the cyclin-dependent kinase P-TEFb plays an important role in the release of the

paused Pol II (Adelman et al., 2012). Mediator has been reported to regulate Pol II pause/release. Although the exact the molecular mechanism remains unclear, one proposed mechanism is that the Mediator recruits Super Elongation Complexes (SECs) containing P-TEFb and contributes to the release of paused Pol II. The Conaway lab showed that the N-terminal domain (NTM) of MED26 subunit of the Mediator complex served as a docking site for the recruitment of ELL and P-TEFb containing SEC (Takahashi et al., 2011). They reported that the MED26 knockdown interfered with the recruitment of SEC to the c-MYC and HSP70 genes and in concomitant reduced transcription. This suggests that the MED26 might contribute to the transition of promoter-proximal paused Pol II into productive elongation.

Two independent studies have implicated the role of Integrator in the regulation of promoter-proximal pause/release (Gardini et al., 2014; Stadelmayer et al., 2014). However, both studies described opposite roles. The laboratory of Ramin Shiekhattar described the role of Integrator in pause release at immediate early genes (IEGs) (Gardini et al., 2014). In this study, ChIP-seq analysis revealed high levels of Integrator at the transcription start site and within the gene body of IEGs upon epidermal growth factor stimulation. The knockdown of Integrator subunits (INTS1 and INTS11) led to a reduction in Pol II levels over the gene body, while Pol II levels at the 5' end were increased. This suggested that Pol II failed to escape pausing and transit to productive elongation. Mechanistically, the authors showed that Integrator is required for the recruitment of SEC-containig P-TEFb. The knockdown of Integrator abolished the recruitment of SEC, thereby preventing Pol II from entering into productive elongation. The work from Monsef Benkirane described the opposite phenotype of Integrator knockdown (Stadelmayer et al., 2014). In this study, the Integrator was shown to interact with NELF in mass spectrometry analysis. Upon the knockdown of INTS11 there was an increase in Pol II levels over the gene body indicative of faster release of paused Pol II. Thus this study implicates the role of Integrator in inducing Pol II pausing.

The studies described here analyzed promoter-proximal pause/release after the knockdown of individual subunits of Mediator and Integrator. In contrast,

we observe a complete loss of both these complexes in our MS data. We assume that the loss of binding of either Mediator or Integrator or both to Pol II may cause such pause defects in the mutant YFFF. However, this assumption needs further experimental evidences.

3.8. Conclusions

Transcription termination occurs when the transcribing Pol II is released from the template DNA. Until today, the mechanism that leads to the timely and efficient termination of transcription remains poorly understood and new concepts are emerging to explain this process.

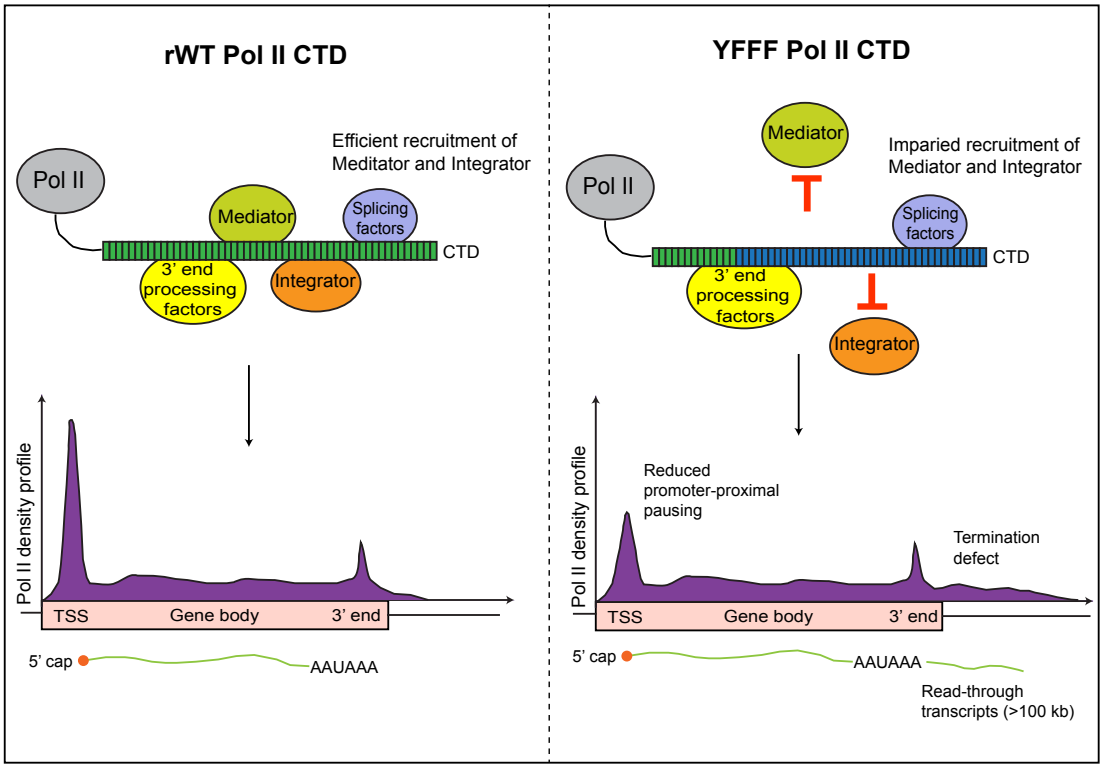


Figure 29: Proposed roles of Tyr1 in the regulation of transcription-coupled processes. A model recapitulating the transcriptional phenotypes in rWT (left) and the mutant YFFF (right). Mass spectrometry data is represented at the top. Pol II density profiles are highlighted in purple and mRNA transcripts are shown in green.

Our work provides a novel insight in the process of transcription termination and proposes the involvement of Tyr1 of Pol II CTD in this process. The work further implicates Tyr1 in the regulation of promoter-proximal pause/release. In rWT, Pol II recruits Mediator, Integrator, 3' end processing and splicing factors for the regulation of transcription-coupled processes (**Figure 29**). The

binding of 3' end processing and splicing factors to Pol II in the mutant YFFF was not impaired and the 3' end processing of RNAs was also not affected. However, the mutant YFFF showed impaired recruitment of Mediator and Integrator complexes. The lack of recruitment of these two large complexes could explain the phenotypes of an early release of the paused Pol II and a global termination defect. Our experiments also provided clues for the heptad-specific functions of the CTD. First, the strong termination defect was observed only in the mutant YFFF and not other mutants. Not all genes in the mutant YFFF displayed a termination defect, suggesting that different genes might require different heptads of the CTD to regulate various transcription events. Our MS data indicated that the Mediator and the Integrator might not interact with the first 13 heptads of the CTD.

3.9. Outlook

This thesis established Tyr1 of Pol II CTD as a key player in the regulation of transcription termination and Pol II pausing and also provided a clue for the involvement of the Mediator and the Integrator in these processes. However, the study raised few important questions that are yet to be answered. 1) What is the mechanism by which the Mediator and/or the Integrator regulate transcription termination and Pol II pause/release? 2) Does the Mediator and the Integrator, both, contribute to these processes? 3) Does the Mediator and the Integrator bind the same or different heptads in the CTD?

A strategy to answer these questions would be to generate sophisticated CTD mutants that might show loss of interaction with either the Mediator or the Integrator complex. The functional study of such mutants would help us unravel the transcriptional phenotypes associated with the loss of one of the two complexes. If we could not get our hands on such CTD mutants, then we might assume that both these complexes interact with same heptads in the CTD. The question to pursue then would be to ask if Mediator and Integrator do bind to the same heptads in a mutually exclusive manner?

The strategy in this thesis of combining genetics with next generation sequencing can serve as a general experimental workflow to investigate the heptad specific functions of CTD in the future. Such an analysis has been challenging so far due to the highly repetitive structure of the CTD. Our work

provides hints that Tyr1 residues in the distal heptads of CTD play a crucial role in limiting global pervasive transcription. In this context, it will be interesting to analyze the tyrosine mutants, especially the mutant FFFY, to ask if Tyr1 residues in the proximal and distal parts of the CTD regulate specific transcription-coupled processes. The growth phenotype of CTD mutants suggested that Tyr1 residues in heptads 14-26 might be critical for cell growth. The study of additional CTD mutants that can help us understand the functional significance of Tyr1 in these heptads will further contribute to our understanding of heptad specific functions of CTD.

4. Materials and Methods

4.1. Materials

4.1.1. List of Chemicals

1 kb DNA ladder	-	Invitrogen, Karlsruhe
1,4-Dithiothreitol (DTT)	-	Carl Roth GmbH & Co.KG, Karlsruhe
6X DNA loading dye	-	Thermo Fisher Scientific, Karlsruhe
Acetone	-	Merck KGaA, Darmstadt
Acetonitrile	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Agarose	-	Lonza Cologne GmbH
Albumin Fraction V (BSA)	-	Carl Roth GmbH & Co.KG, Karlsruhe
Ammonium Bicarbonate	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Ammonium peroxydisulphate	-	Carl Roth GmbH & Co.KG, Karlsruhe
Bromophenol Blue (BPB)	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Dimethyl pimelimidate	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Dimethyl Sulfoxide (DMSO)	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Dulbeco's PBS (DPBS)	-	Gibco Life Technologies, Eggenstein
Ethanol, absolute (EtOH)	-	Merck KGaA, Darmstadt
Ethanolamine	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Ethidium Bromide (EtBr)	-	Applichem, Darmstadt
Ethylendiaminetetraacetic Acid (EDTA)	-	Carl Roth GmbH & Co.KG, Karlsruhe
Fetal Bovine Serum (FBS)	-	PAA Laboratories, Pasching, Austria

Formaldehyde (37 %)	-	Carl Roth GmbH & Co.KG, Karlsruhe
Glycerol 86%	-	Carl Roth GmbH & Co.KG, Karlsruhe
HEPES	-	Carl Roth GmbH & Co.KG, Karlsruhe
Hydrogen Chloride (HCl)	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Isopropanol, absolute	-	Carl Roth, Karlsruhe
L-Glutamine 200mM (100x)	-	Gibco Life Technologies, Eggenstein
Methanol (MeOH), absolute	-	Merck KGaA, Darmstadt
Milk powder, blotting grade	-	Carl Roth GmbH & Co.KG, Karlsruhe
Neomycin (G148)	-	Promega Corp., Wisconsin, USA
Nonidet P-40 (NP40)	-	Carl Roth GmbH & Co.KG, Karlsruhe
Penicillin/Streptomycin	-	Gibco Life Technologies, Eggenstein
Polyacrylamide 30% (PAA)	-	Carl Roth GmbH & Co.KG, Karlsruhe
Prestained protein ladder plus	-	Fermentas, St. Leon-Rot
RPMI Medium 1640	-	Gibco Life Technologies, Eggenstein
Sodium Azide (NaN_3)	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Sodium Borate	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Sodium chloride (NaCl)	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Sodium Dodecyl Sulfate (SDS)	-	Carl Roth, Karlsruhe
Tetracycline	-	Promega Corp., Wisconsin, USA
Tetramethylethylenediamine (TEMED)	-	Carl Roth GmbH & Co.KG, Karlsruhe
Trifluoroacetic acid	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Tris	-	Carl Roth GmbH & Co.KG, Karlsruhe
Triton X-100	-	Sigma-Aldrich Chemie GmbH, Steinheim
Trizol	-	Ambion, Life technologies, Eggenstein

Tryphan Blue	-	Invitrogen, Karlsruhe
Trypsin	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
Tween-20	-	Sigma-Aldrich Chemie GmbH, Deisenhofen
α - amanitin	-	Roche Molecular Biochemicals, Mannheim

4.1.2. Lab consumables

0.2 ml PCR tubes	-	Thermo Fisher Scientific, Karlsruhe
0.4 cm Electroporation cuvettes	-	Bio-rad laboratories GmbH, Munich
Agar plates	-	Greiner GmbH, Frickenhausen
Amersham Hyperfilm ECL	-	GE Healthcare, Munich
Amershan Protran premium 0.45 µM nitrocellulose Membrane	-	GE Healthcare, Munich
Cell culture flasks	-	Greiner Bio-One,Frickenhausen
Cyrovials 1.5 ml	-	Nunc GmbH, Wiesbaden
Falcon Tubes 15 ml, 50 ml	-	Corning GmbH, Kaiserslautern
Filter papers	-	Machery-nagel GmbH & Co.K.G, Düren
Laboratory Glassware	-	Duran Productions GmbH & Co. KG, Mainz
Microcentrifuge tubes 1.5 ml, 2 ml	-	Eppendorf, Hamburg
Nitrile Gloves	-	Shield scientific, Netherlands
Parafilm	-	Carl Roth GmbH&CoKG, Karlsruhe
Pasteur Pipettes	-	Hirschmann Laborgeräte, Eberstadt
Phosphatase inhibitor cocktail	-	Roche Diagnostics, Penzberg
Pipette Tips 10, 20, 200, 1000 µl	-	Molecular Bio-Products, San Diego
Plastic Pipettes	-	Greiner Bio-one GmbH, Frickenhausen
Protease inhibitor cocktail	-	Roche Diagnostics, Penzberg
Protein A-Sepharose beads	-	GE Healthcare, Munich
Protein G-Sepharose beads	-	GE Healthcare, Munich

Scalpel	-	Braun, Tuttlingen
Sterile flip filters	-	Milipore GmbH, Eschborn
Whatmann gel blotting paper GB003	-	Schleicher & Schuell, Germany

4.1.3. Consumable Kits

Agilent RNA 6000 pico kit	-	Agilent Technologies, USA
DNA Mini/Maxi kits	-	Qiagen GmbH, Hilden
TrueSeq ChIP Library Preparation Kit	-	Illumina, USA
QIAquick Gel extraction	-	Qiagen GmbH, Hilden
QIAEX II Gel extraction	-	Qiagen GmbH, Hilden
Ribo-Zero-rRNA removal kit	-	Epicenter, USA
ScriptSeq RNA library Preparation kit	-	Epicenter, USA
TrueSeq Small RNA library Preparation kit	-	Illumina, USA

4.1.4. Instruments

-20°C freezer	-	Siemens, Munich
-80°C freezer	-	Colora Messtechnik GmbH, Lorch
Agilent 2100 Bioanalyzer	-	Agilent Technologies, USA
Bacteria incubator	-	Heraeus Sepatech GmbH, Osterode
Bacteria shaker (Series 25)	-	New Brunswick ScientificCo., NJ, USA
Bio-Rad PowerPac 300	-	Bio-Rad Laboratories GmbH, Munich
Bio-Rad PowerPac basic	-	Bio-Rad Laboratories GmbH, Munich
Bioruptor Pico sonicator	-	Diagenode Inc., USA
Blotting chamber	-	Bio-Rad Laboratories GmbH, Munich

Branson Sonifier 250	-	Heinemann Ultraschall- und Labortechnik
DNA gel analyser	-	Peqlab biotechnologies GmbH, Erlangen
Electroporator (eukaryotic cells)	-	Bio-Rad Laboratories GmbH, Munich
Eppendorf Centrifuge 5417R	-	Eppendor-Netheler-Hinz GmbH, Hamburg
Eppendorf centrifuge 5424	-	Eppendor-Netheler-Hinz GmbH, Hamburg
Eppendorf Thermomixer 5436	-	Eppendorf-Netheler-Hinz GmbH, Hamburg
Fridge KU 171	-	Liebherr, Biberach
Fuchs-Rosenthal chamber	-	GLW Gesellschaft für Laborbedarf GmbH
Hyper-cassette	-	GE healthcare, Munich
Illumina HiSeq 2000 support	-	Illumina, USA
Laminar Flow Hood	-	BDK Luft-und Reinraumtechnik GmbH
Magnet stirrer M23	-	GLW, Würzburg
Microwave	-	Panasonic, Hamburg
Multi-calimatic pH-meter	-	Knick GmbH & Co. KG, Berlin
Nanodrop 1000	-	Thermo Scientific, Braunschweig
Odyssey Infrared Imaging System	-	Odyssey LI-COR
Roller mixer SRT 2	-	Dunn GmbH, Augsburg
Rotina 380 centrifuge	-	Andreas Hettich GmbH & Co.KG, Tuttlingen
SDS-PAGE gel tank	-	Amersham Pharmacia Biotech, Freiburg
Ultimate 3000 RSLCnano System	-	Thermo Fisher Scientific, Karlsruhe
Water bath	-	Thermo Fisher Scientific, Karlsruhe

Zeiss primovert microscope	-	Carl Zeiss microscopy GmbH, Munich

4.1.5. Buffer and Solutions

0.7% Agarose solution	-	2.1 gms of agarose powder in 300 ml of 1X TAE buffer. Boil in a microwave and cool to room temperature. Add 3 µl/100 ml of EtBr.
2X Laemmlie buffer	-	2% SDS 100 mM DTT 10 mM EDTA 20% Glycerol 60 mM Tris/HCl; pH 6.8 0.01% Bromophenolbue
6X Laemmlie buffer	-	9% SDS 375 mM Tris/HCl; pH 6.8 9% β-mercaptoethanol 50% Glycerol 0.06% Bromophenolblue
Blocking buffer	-	5 gms of milk in 100 ml of 1X TBS-T
IP Lysis buffer	-	50 mM Tris/HCl, pH 8.0 1% NP40 150 mM NaCl
IP Wash buffer	-	50 mM Tris/HCl, pH 8.0 150 mM NaCl
2xTris/SDS pH 8.8	-	90.72 gms Tris base 10 ml of 20% SDS Adjust volume to 1000 ml with H_2O Adjust pH to 8.8
2xTris/SDS pH 6.8	-	30.24 gms Tris base 10 ml of 20% SDS Adjust volume to 1000 ml with H_2O Adjust pH to 6.8
Separating gel (6.5%)	-	4.3 ml 30% PAA

		10 ml 2xTris/SDS pH 8.8
		5.5 ml H_2O
		167 µl APS
		17 µl TEMED
Stacking gel (4%)	-	1.5 ml 30% PAA
		7.5 ml 2xTris/SDS pH 6.8
		5.9 ml H_2O
		90 µl APS
		20 µl TEMED
Western running buffer (10x)	-	30.2 g Tris/Base
		144 g Glycin
		5 ml 20% SDS
		Make volume to 1 l with H_2O
Western transfer buffer (10x)	-	30.2 g Tris/Base
		144 g Glycine
		200 ml Methanol
		Make volume to 1 l with H_2O
Cross-linking solution	-	11% formaldehyde
		100 mM NaCl
		1 mM EDTA
		0.5 mM EGTA
		50 mM Hepes, pH 7.8
Protein stripping buffer	-	7 µl/ml β-mercaptoethanol
		2% SDS

4.1.6. Antibodies

4.1.7. Primary antibodies

3F10
A rat monoclonal antibody raised against an epitope contained in the haemagglutinin Polypeptide of the human influenza virus (Roche Diagnostics, GmbH, Mannheim. Received as a supernatant solution from E. Kremmer, Helmholtz Zentrum, Munich)

12CA5
A rat monoclonal antibody raised against an epitope contained in the haemagglutinin Polypeptide of the human influenza virus. Received as a supernatant solution from E. Kremmer, Helmholtz Zentrum, Munich)

ab9110
A rabbit Polyclonal antibody raised against an epitope contained in haemagglutinin Polypeptide of the human influenza virus. (Abcam, Cambridge, United Kingdom)

Pol 3.3
A mouse monoclonal antibody that recognizes a conserved epitope of the largest 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, Munich)

3E8
A rat monoclonal antibody that recognizes the Ser2P within the CTD of the large subunit (Rpb1) of RNAPII (Received as a supernatant solution from E. Kremmer, Helmholtz Zentrum, Munich)

MED15
A rabbit Polyclonal antibody that recognizes MED15 subunit of Mediator complex (Proteintech, United Kingdom)

INTS11
A rabbit Polyclonal antibody that recognizes Integrator subunit 11 of human Integrator complex (Bethyl laboratories, Biomol, Hamburg)

ab8580
A rabbit polyclonal antibody that recognizes histone H3 trimethyl K4. Abcam, Cambridge, United Kingdom)

ab4729
A rabbit polyclonal antibody that recognizes histone H3 acetyl K27. Abcam, Cambridge, United Kingdom)

5C4
A mouse monoclonal antibody that recognizes an epitope in GAPDH protein (Received from E. Kremmer, Helmholtz Zentrum, Munich)

DM1A
A mouse monoclonal antibody that recognizes an epitope located in C-terminal end of the α-tubulin isoform (Sigma-Aldrich GmbH, Deisenhofen).

4.1.8. Secondary antibodies

Alexa Fluor 680 Goat anti-Rat IgG (H+L)	-	Invitrogen
IR Dye 800 CW anti-Mouse IgG (H+L)	-	Rockland Inc, Rockland
Anti-mouse IgG HRP conjugate	-	Promega
Anti-rabbit IgG HRP conjugate	-	Promega
Goat IgG Anti-rat (H+L)-HRPO	-	Dianova

4.1.9. Materials for cloning

Restriction enzymes

AvrII, NotI and BspE1	-	New England Biolabds GmbH, Frankfurt am Main
Age1 and Not1	-	Thermo Fisher Scientific GmbH, Munich

Synthesis of CTD sequences

GeneArt, Regensburg, sequenced all CTD sequences used for cloning.

Plasmids used during cloning

RX2-287 (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

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

WT fwd - 5'CTCCTGCTGACGCACCTGTTCT3'

CTD fwd - 5'CCTTTGTCTTTTCCTATAGGTGGTGC3'

CTD rev - 5'GTCAGACAACCTCGGTGGCCTGTGTG3'

Bacteria

DH10B: E.coli strain purchased from Invitrogen GmbH, Karlsruhe was used for the cloning of all plasmid DNA.

4.1.10. Human cell lines

Basic cell line: Raji

Cell type: human Burkitt lymphoma

DSMZ no.: ACC 319

Origin: established in 1963 from a 11 year old African boy with Burkitt lymphoma. Cell line carries latent Epstein-Barr virus (EBV) genome and is positive for EBNA. Cells are cultured as suspension cells. Classified as risk category 1 according to the German Central Commission for Biological Safety (ZKBS). A list of stably transfected cell lines generated for the project is listed in Table 3.

Table 3: List of stably transfected cell lines

Name	Plasmid	Cell line	Resistance
rWT	RX4-267	Raji	G418
YYFF	-	-	-
FFYY	-	-	-
YFFY	-	-	-
FYYF	-	-	-
YFFF	-	-	-
FFFY	-	-	-
YF26	-	-	-
S2AAA	-	-	-
TAAA	-	-	-
S5AAA	-	-	-

4.2. Methods

4.2.1. Molecular techniques for cloning

4.2.1.1. Cloning strategy

All CTD sequences were synthesized by GeneArt and flanked by an AvrII and NotI restriction sites upstream and downstream of the sequences. Synthesized CTD does not contain AvrII, NotI, AgeI, BspEI, NgoMIV, NheI,

SpeI and ClaI restriction sites within the sequence. CTD sequences were cloned into final expression vector, RX4-267, by two-step cloning strategy.

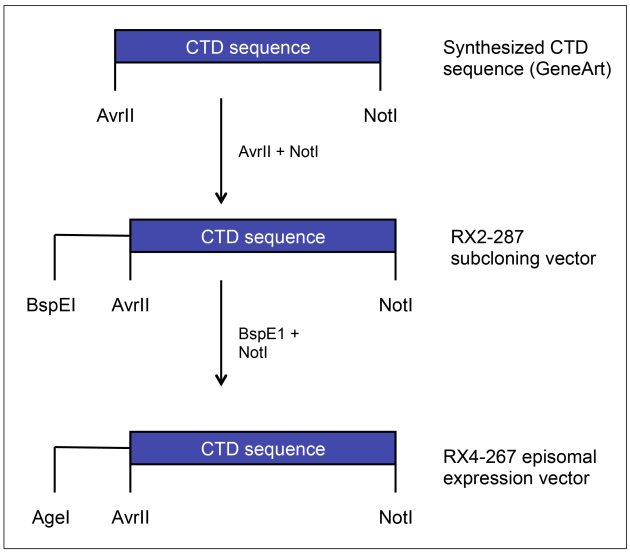


Figure 30: A schematic representation of two-step cloning procedure. CTD sequence (blue) synthesized by GeneArt is flanked by AvrII and NotI restriction sites. The flanked sequence is ligated into sub-cloning vector RX2-287. CTD sequence from RX2-287 is flanked by BspEI and NotI restriction sites and eventually ligated to AgeI/NotI sites of the final expression vector RX4-267.

First the newly synthesized CTD sequence was excised from the start vector using the restriction sites AvrII and NotI and eventually cloned into the sub-cloning vector RX2-287. The CTD sequence from the RX2-287 vector was excised using BspEI and NotI restriction sites and finally cloned into the AgeI/NotI site of the end vector RX4-267.

4.2.1.2. Transformation of Bacteria

For the transformation, recombinant-deficient *Escherichia Coli* strain DH10B, was used. 0.5 µg of plasmid DNA or 20 µl of ligation mixture was added to 100 µl of competent bacteria and incubated on ice for 20 minutes. Cells were then heat shocked at 42°C for 30 sec and kept on ice for 2 minutes. 400 µl (4 volumes) of LB medium was then added to the cells and incubated at 37°C for 90 minutes. Finally, 50 - 400 µl of suspension was plated on the agar plates containing ampicillin, kanamycin or spectinomycin resistance and the incubated at 37°C for 16-18h.

4.2.1.3. Miniprep of plasmid DNA

Plasmid DNA was isolated from bacterial culture using Qiagen miniprep kit. Single colony obtained following transformation of bacteria was inoculated in

2 ml of LB media containing appropriate antibiotics in a 14 ml loose-capped tube. The culture was incubated for 16-18h at 37°C with vigorous shaking in an orbital shaker. Later, 1.5 ml of the culture was transferred into 1.5 ml eppendorf tube and centrifuged at 10,000g for 1 min. The supernatant was discarded and pellet was re-suspended in 150 µl of buffer P1 (resuspension buffer). Next, 150 µl of buffer P2 (lysis buffer) was added and mixed gently before incubating for 5 min at RT. Later, 150 µl of pre-chilled buffer P3 (neutralization buffer) was added, mixed by inverting, and incubated on ice for 20 min. The mixture was then centrifuged at 10,000g for 5 min and the resulting supernatant was transferred into a fresh corresponding 1.5 ml eppendorf. To precipitate the plasmid DNA, 450 µl of isopropanol (1:1) was added to the supernatant and centrifuged at 10,000g for 5 min. Supernatant was discarded and the pellet was washed with 1 ml of 70% ethanol and centrifuged at 10,000g for 5 min. Supernatant was removed and the pellet was re-suspended in 200 µl of ddH_20.

4.2.1.4. Maxiprep of plasmid DNA

Large quantities of plasmid DNA were purified using Qiagen plasmid purification Maxiprep protocol. A single colony was picked from the selection plate and incubated in 400ml of LB media containing appropriate antibiotics at 37°C for 16-18h in an orbital shaker. Cells were harvested by centrifugation at 4000g for 15 min at 4°C. Supernatant was discarded and the pellet was re-suspended in 10 ml of chilled buffer P1 (resuspension buffer) containing 100 µg/ml of RNAse A. 10 ml of buffer P2 (lysis buffer) was added and cells were lysed by incubating for 5 min at room temperature. 10 ml of pre-chilled buffer P3 (neutralization buffer) was added to the mixture and incubated on ice for 20 min to enhance the precipitation of genomic DNA, protein and cell debris. The mixture was centrifuged at 4000g for 30 min at 4°C and the supernatant containing the plasmid DNA was passed through Qiagen column that was pre-equilibrated using 10 ml of equilibration buffer QBT. After washing twice with 30 ml of wash buffer, the plasmid DNA was eluted with 15 ml of elution buffer QF. DNA was precipitated by adding 10.5 ml of isopropanol to the eluted DNA, followed by centrifugation at 4000g for 30 min at 4°C. Supernatant was discarded and DNA pellet was washed with 5 ml of 70% ethanol, followed by centrifugation at 4000g for 10 min at 4°C.

Supernatant was discarded and the DNA was dissolved in 300 µl of TE buffer and transferred to a microcentrifuge tube. DNA concentration was measured using nanodrop.

4.2.1.5. DNA digestion using restriction endonucleases

Several clones of Plasmid 'miniprep' DNA were digested with appropriate restriction enzymes, as test digest at various stages of cloning. Restriction enzymes from the New England biolabs (NEB) and Thermo Fisher Scientifics were used and the digest was performed in compatible buffers as described by the manufacturers. For the gel extraction, 5 µg of DNA was used.

4.2.1.6. Ligation of DNA fragments

200 ng of extracted insert was mixed with 200 ng of the vector and heated at 50°C for 5 min. The mixture was then incubated on ice and T4 DNA ligase and ligase buffers were added as recommended by the manufacturers. The ligation mix was incubated overnight at 16°C before using it for transformation.

4.2.1.7. DNA agarose gel electrophoresis

0.7% DNA agarose gel was prepared by boiling 2.1 g of agarose in 300 ml of 1X TAE buffer. The solution was allowed to cool before adding 3 µl/100 ml of ethidium bromide (EtBr) to it and casting it in Biorad chamber. DNA was mixed with 6X DNA loading dye before loading on the gel. The gel was run for 2 hours at 80V constant voltage. A gel photo was taken under the UV light using DNA gel analyzer.

4.2.1.8. DNA gel extraction

After DNA agarose gel electrophoreses, the band of interest was excised from the gel using a scalpel. From there, QIAquick gel extraction kit from qiagen was used to purify the DNA ranging from 70 bp to 10 kb. To purify the DNA, bigger than 10 kb, QIAEX II kit was used. The purified DNA was eluted in Tris buffer and its concentration was measured using nanodrop system.

4.2.2. Cell culture

4.2.2.1. Cell thawing

Cell stock aliquots frozen in cryotubes at -80°C were thawed at room temperature for 2-3 min, re-suspended in 10 ml pre-warmed RPMI 1640 complete media and centrifuged at 1200 rpm for 4 min. The cell pellet was

washed once with 1X PBS and re-suspended in 10 ml appropriate culture media. The cells were cultured at 37°C, 5% CO_2 incubator. A running culture was split 1:3 and fresh culture medium was added every two-three days.

Complete media: 10% FBS, 2 mM L-glutamine, 100 U Penicillin and 100 µg/ml Streptomycin

4.2.2.2. Cell freezing

Cells were split 1:1 with fresh culture medium, one day before freezing them. Cells were centrifuged at 1200 rpm for 4 min and the supernatant was discarded. Pellet was re-suspended in 1 ml of freezing medium and transferred to 1.5 ml of cryotubes. The cryotubes were first wrapped with paper to avoid shock freezing and stored overnight at -80°C, before transferring them to storage facilities in liquid nitrogen.

Freezing medium: 80% FBS, 10% DMSO and 10% RPMI 1640

4.2.2.3. Cell Counting

Cells were counted using a Fuchs-rosenthal counting chamber. 40 µl of cell suspension was mixed in a ratio of 1:1 with 0.04% trypan blue solution. Dead cells absorb the stain over their membranes and can be distinguished from the colorless living cells. Number of living cells in three big squares was calculated and the total number of living cells in 1 ml medium was calculated by multiplying the average number of cells per big square with 10^4.

4.2.2.4. Cell Transfection

RX4-267 expression vector carrying the mutant CTD is transfected into Raji cells via electroporation. Raji is an Epstein-Barr virus-positive Burkitt's lymphoma cell line. Cells were split 1:1 one day prior to transfection. $20*10^6$ cells were used for each transfection. Cells were washed twice with 1X PBS and resuspended in 500 µl of PBS. Cells were transferred in 4 mm of electroporation cuvette, and 10 µg of plasmid DNA was gently mixed with the cells before incubating the mixture for 20 min at room temperature. Next, electroporation was performed at voltage of 250 V and charge capacitance of 950 µF. Time was set to constant and both the pulse buttons were pushed till the noise signal appears. Immediately after that, 500 µl of FBS was added to the cuvette and mixed by gently pipetting the transfected cells up and down. Transfected cells were incubated for 5 min at room temperature and then transferred to a cell culture flask containing 5 ml of complete media

supplemented with 1 µg/ml of tetracycline. Two days after electroporation, 1 mg/ml of G418 was added to select for the positively transfected cells. Selection was carried out for 2-4 weeks until cell reaches 95% cell viability.

4.2.3. Protein analysis

4.2.3.1. Cell lysis

For western analysis, raji cells were first counted using trypan blue and transferred to 15 ml of falcons. The cells were centrifuged at 1200rpm for 4 min and washed once with 1X PBS. After washing, supernatant was removed and cells were lysed in appropriate volume of 2X Laemmli buffer (100 µl of 2X Laemmlie per $1*10^6$ cells). The viscous lysate was repeatedly drawn through a narrow pipette tip and transferred into microcentrifuge tube, before boiling samples at 95°C for 5 minutes. Cells were sonicated five times (10 pulses; duty cycle 50%; output 5) and then boiled again at 95°C for 5 minutes. Samples were then centrifuged at 10,000g for 4 minutes to clear the insoluble contaminants. The lysates were then either loaded directly on the gel or stored at -20°C.

For immunoprecipitation experiments, Raji cells were counted, harvested by centrifugation (1200 rpm, 4 minutes), washed twice with ice cold PBS and re-suspended in appropriate volume of NP-40 lysis buffer (100 µl lysis buffer per $1*10^6$ cells). Cells were incubated on rotary shaker for 30 minutes at 4°C. Later, the cells were sonicated for 3 times (duty cycle 50%; output 5) and incubated on rotary shaker for 30 minutes at 4°C. Cells were centrifuged at 10,000g for 15 minutes at 4°C. After centrifugation, supernatant was transferred to fresh microcentrifuge tube for subsequent immunoprecipitation or stored at -20°C.

2X Laemmli buffer: 60mM Tris/HCl pH 6.8, 2% SDS, 100mM DTT, 10mM EDTA, 10% glycerol and 0.01% bromophenol blue

NP40 lysis buffer: 50mM Tris/HCl pH 8.0, 150mM NaCl and 1% NP40.

4.2.3.2. Immunoprecipitation

For immunoprecipitation, appropriate concentration of antibody was mixed with 30 µl of Protein A sepharose beads and/or 30 µl of Protein G sepharose beads and incubated for 3-4 hours on rotary shaker at 4°C. Later, after

washing twice with ice-cold PBS, appropriate volume of cell lysate was added to the antibody-beads reaction mixture and incubated overnight at 4°C on rotary shaker. Next day, the samples were centrifuged at 1200 rpm for 1 minutes and the supernatant was collected. The samples were washed four times with wash buffer and finally re-suspended in 50 µl of 2X Laemmli buffer. The Laemmli lysate was boiled at 95°C for 7 minutes and the samples were centrifuged at 10,000 g for 4 minutes before loading on the gel.

Wash buffer: 50mM Tris/HCl pH 8.0, 150mM NaCl

4.2.3.3. IP for mass-spectrometric analysis

75-80*10^6 cells were used for the IP reaction and subsequent mass spectrometric (MS) analysis. Immunoprecipitation was performed as described in 4.2.3.2. After the IP reaction, the samples were loaded on to the SDS PAGE. Once the samples enter the separating gel, they were allowed to run for a short time of around 10 minutes. The gel pieces containing the IP samples were then cut into 2 equal halves from each lane and transferred into a clean PCR tubes. Gel pieces were hydrated in 100 µl of ddH_2O and processed for either in-gel trypsin digest or on-beads digest.

4.2.3.4. In-gel trypsin digest

The excised gel pieces were first rinsed twice with 100 µl of H_2O and then twice with 50 mM NH_4HCO_3 (ammonium bicarbonate) to remove un-polymerized acrylamide from the gel pieces. The excised gel pieces were then incubated with 50 mM NH_4HCO_3 and ACN (acetonitrile) in a ratio of 1:1 for 60 minutes at 37°C. The gel pieces were subsequently washed thrice with 50 mM NH_4HCO_3 and dehydrated by incubating with ultrapure ACN (three times, 10 minutes each). Dehydration of the gel pieces by ACN and subsequent swelling facilitate the permeabilization of the enzymes to the gel for the digestion of the proteins. After dehydration, gel pieces were incubated with 10mM DTT for 1 hour at room temperature. Reduction reaction was stopped by removal of DTT and then the gel pices were incubated with 55 mM of IAA (Iodoacetamiede) for 30 minutes at room temperature in darkness. IAA is an alkylating agent that led to the irreversible alkylation of the -SH groups and the cysteines were transformed to the stable S-carboxyamidomethylcysteine. The gel pieces were washed once with NH_4HCO_3, before dehydrating them by repeated treatment with ultrapure

ACN. All the supernatant from the probes were removed and the probes were dried applying vacuum for 5 minutes. In the meanwhile, trypsin stock solution was prepared at 25 ng/µl in 50 mM NH_4HCO_3. For the digestion of the proteins, 10-15 µl of trypsin was added to the gel pieces and incubated at 4°C for 45 minutes to allow the gel pieces to absorb trypsin. After 45 minutes, the unabsorbed trypsin was removed and the gel pieces were hydrated by adding 100 µl of NH_4HCO_3 and incubated for overnight at 37°C on a shaker.

Next day, supernatant was transferred to the cool corresponding 1.5 ml microcentrifuge tube. Peptides were extracted by incubating the gel pieces with 50% ACN and 0.25% TFA (trifluoroacetic acid) solution twice for 10 minutes each. The supernatant containing extracted peptides was transferred to cool corresponding 1.5 ml microcentrifuge tube. Later, the gel pieces were incubated with ultra pure ACN for three times, 10 minutes each and the supernatant was transferred to cool corresponding 1.5 ml microcentrifuge tube. Samples containing extracted peptides was evaporated to dryness in a speecvac and re-suspended in 30 µl of 0.1% TFA for subsequent C18 stage tips treatment for desalting and step elution of peptide mixtures. First, C18 stage tip columns were prepared by injecting C18 chromatography paper in 200 µl of pipette tip. C18 stagetip columns were activated by pre-treatment with 100% acetonitrile followed by 0.1% TFA. First, 20 µl of 100% ACN was allowed to pass through the column thrice. Next, 20 µl of 0.1% TFA was allowed to pass through the column thrice, before allowing the samples to pass through C18 stagetip columns. The columns were then washed with 0.1% TFA (3 times, 20 µl each). After washing, the peptides were eluted into fresh 1.5 ml eppendorf tube using 20 µl of 80% ACN and 0.25% TFA solution. Elution step was repeated two more times and the eluted peptides were again evaporated to dryness using speedvac.

4.2.3.5. On-beads digest

Following the standard immunoprecipitation procedure, beads were first washed with lysis buffer (three times) and then with 50mM NH_4HCO_3 (ammonium bicarbonate). For trypsin digest, beads were incubated with 100 µl of 10 ng/µl of trypsin solution in 1M Urea and 50mM NH_4HCO_3 for 30 minutes at 25°C. The supernatant was collected, beads washed twice with 50mM NH_4HCO_3 and all three supernatants collected together and incubated

overnight at 25°C after addition of 1mM DTT. 27mM of iodoacetamide (IAA) was then added to the samples and incubated at 25°C for 30 minutes in dark. Next, 1 µl of 1M DTT was added to the samples and incubated for 10 minutes to quench the IAA. Finally, 2.5 µl of trifluoroacetic acid (TFA) was added to the samples and desalted using C18 stage tips (Ishihama et al., 2006). Samples were evaporated to dryness, re-suspended in 30 µl of 0.1% formic acid solution and stored at -20°C until LC-MS analysis.

4.2.3.6. Processing of samples

For LC-MS/MS purposes, samples were desalted using C18 Stagetip and injected in an Ultimate 3000 RSLCnano system (Thermo), separated in a 15-cm analytical column (75µm ID with ReproSil-Pur C18-AQ 2.4 µm from Dr. Maisch) with a 50 min gradient from 5 to 60% acetonitrile in 0.1% formic acid. The effluent from the HPLC was directly electrosprayed into a QexactiveHF (Thermo) operated in data dependent mode to automatically switch between full scan MS and MS/MS acquisition. Survey full scan MS spectra (from m/z 375–1600) were acquired with resolution R=60,000 at m/z 400 (AGC target of $3x10^6$). The 10 most intense peptide ions with charge states between 2 and 5 were sequentially isolated to a target value of $1x10^5$, and fragmented at 27% normalized collision energy. Typical mass spectrometric conditions were: spray voltage, 1.5 kV; no sheath and auxiliary gas flow; heated capillary temperature, 250ºC; ion selection threshold, 33.000 counts. MaxQuant 1.5.2.8 was used to identify proteins and quantify by iBAQ with the following parameters: Database, Uniprot_Hsapiens_3AUP000005640_151111; MS tol, 10ppm; MS/MS tol, 0.5 Da; Peptide FDR, 0.1; Protein FDR, 0.01 Min. peptide Length, 5; Variable modifications, Oxidation (M); Fixed modifications, Carbamidomethyl (C); Peptides for protein quantitation, razor and unique; Min. peptides, 1; Min. ratio count, 2. Identified proteins were considered as interaction partners if their MaxQuant iBAQ values displayed a greater than log2 5-fold enrichment and p-value 0.05 (ANOVA) when compared to the rWT control.

4.2.4. ChIP-Seq

4.2.4.1. Cross-linking of cells

To cross-link the cells for ChIP, 1/10th volume of 10X crosslinking solution was added to the raji cells in culture medium. After 10 minutes of incubation at room temperature, glycine was added to a final concentration of 250mM to quench the remaining formaldehyde and stop the cross-linking. After five minutes of quenching, cells were washed twice with cold PBS. Cells were then sonicated as described in next paragraph or snap frozen in liquid nitrogen and stored at -80°C for sonication at a later stage.

Crosslinking solution: 100mM NaCl 1mM EDTA pH 8, 0.5mM EGTA pH 8, 50mM HEPES pH 7.8 and 11% formaldehyde

4.2.4.2. Chromatin-Immunoprecipitation

50 million cross-linked cells were resuspended in cold 2.5mL buffer LB1 at 4°C for 20 minutes on a rotating wheel. Nuclei were pelleted down by spinning at 1350 X g in a refrigerated centrifuge and washed in 2.5mL buffer LB2 for 10 minutes at 4°C on a rotating wheel followed by centrifugation to collect nuclei. Nuclei were then resuspended in 1mL of buffer LB3 and sonicated using Bioruptor Pico in 15mL tubes for 25 cycles of 30 sec ON and 30 sec OFF pulses in 4°C water bath. All buffers (LB1, LB2 and LB3) were complemented with EDTA free Protease inhibitor cocktail, 0.2mM PMSF and 1µg/mL Pepstatin just before use. After sonication, Triton X-100 was added to a final concentration of 1% followed by centrifugation at 20000 g and 4°C for 10 minutes to remove particulate matter. After taking a 50µl aliquot to serve as input, chromatin was aliquoted and snap-frozen in liquid nitrogen and stored at -80°C until use in ChIP assays. Input aliquots were mixed with equal volume of 2X elution buffer and incubated at 65°C for 12 hours. An equal volume of TE buffer, pH 8.0 was added, followed by treatment with RNase A (0.2µg/mL) at 37°C for one hour and Proteinase K (0.2µg/mL) for two hours at 55°C. DNA was isolated by phenol:chloroform:isoamylalcohol (25:24:1 pH 8) extraction, followed by Qiaquick PCR Purification kit. Purified DNA was then analyzed on a 2% agarose gel or on Bioanalyzer (Agilent, USA) using a High Sensitivity DNA Assay.

For Chromatin-Immunoprecipitation, Protein-G coated Dynabeads were incubated at 4°C in blocking solution (0.5% BSA in PBS) carrying specific antibodies. Sonicated chromatin was added to pre-coated beads and the mixture was incubated overnight at 4°C on a rotating wheel. Information about specific antibodies and the quantity of chromatin used for each ChIP is described in the **Supplementary table 6**. After incubation with chromatin, beads were washed 7 times with Wash buffer (50mM Hepes pH 7.6, 500mM LiCl, 1mM EDTA pH 8, 1% NP-40, 0.7% Na-Deoxycholate, 1X protease inhibitor cocktail) followed by one wash with TE-NaCl buffer and a final wash with TE buffer, pH 8.0. Immunoprecipitated chromatin was eluted by two sequential incubations with 50µl Elution buffer (50mM Tris pH 8, 10mM EDTA pH 8, 1% SDS) at 65°C for 15 minutes. The two eluates were pooled and incubated at 65°C for 12 hours to reverse-crosslink the chromatin followed by treatment with RNase A and Proteinase K. DNA was purified as described above for Input samples. Purified DNA was quantified with Qubit DS DNA HS Assay.

LB1:	50mM Hepes pH 7.5, 140mM NaCl, 1mM EDTA pH 8, 10% glycerol, 0.75% NP-40, 0.25% Triton X-100
LB2:	200mM NaCl, 1mM EDTA pH 8, 0.5mM EGTA pH 8, 10mM Tris pH 8
LB3:	1mM EDTA pH 8, 0.5mM EGTA pH 8, 10mM Tris pH 8, 100mM NaCl, 0.1% Na-Deoxycholate, 0.5% N-lauroylsarcosine
Elution buffer:	100mM Tris pH 8.0, 20mM EDTA, 2%SDS

4.2.4.3. Libraries for sequencing

Atleast 1ng of ChIP DNA was used to prepare sequencing library with Illumina ChIP Sample Library Prep Kit (Illumina, USA). After end-repair and adapter ligation, library fragments were size-selected using E-Gel SizeSelect 2% Agarose Gel, followed by 12 cycles of PCR amplification. Barcoded libraries from different samples were pooled together and sequenced on Illumina HiSeq2000 platform in paired-end sequencing runs.

4.2.5. RNA-seq:

4.2.5.1. Total RNA-seq

72 hours after induction, cells were lysed in trizol reagent and RNA was extracted from cells according to manufacturer's instructions. Any DNA contaminant was digested with rigorous Turbo DNase (ThermoFisher Scientific, USA) treatment according to manufacturers protocol. Purified RNA was quantified with Nanodrop 1000 and quality of RNA was assessed using RNA Nano or Pico Assay kit with Bioanalyzer (Agilent Technologies, USA). Only the RNA samples with RIN value above 8 were used for sequencing.

For strand-specific sequencing, ribosomal RNA was removed from total RNA with Ribo-Zero rRNA Removal Kit (EpiCenter, USA) according to manufacturer's instructions. Depletion of rRNA was confirmed by analyzing the samples with RNA Pico Assay on Bioanalyzer. Libraries were prepared either with ScriptSeq Total RNA Library prep kit (EpiCenter, USA) according to manufacturer's instructions or with Small RNA Library Prep Kit (Illumina, USA) using a modified protocol as follows: 50ng rRNA depleted total RNA was fragmented to ~150bp by digesting with 1U of RNaseIII (ThermoFisher Scientific, USA) for 10 minutes at 37°C in a 10µl reaction. Fragmentation reaction was stopped, by adding 90µl nuclease-free water and quickly adding 350µl RLT buffer from RNeasy Mini Kit (Qiagen, Germany). Fragmented RNA was purified using RNA Cleanup protocol according to manufacturer's instructions. 20ng of fragmented RNA was used as input for ligation of 3' and 5' adapters according to Small RNA Library Prep Protocol. cDNA was synthezed from adapter ligated RNA with 10 cycles of PCR amplification. However instead of performing a size-selection of agarose gel (as recommended by manufacturer) we used 1 volume of Ampure XP Beads (Beckman Coulter, USA) to clean up the amplified library and remove adapter dimers according to manufacturer's instructions. Purified libraries were then analyzed with HS DNA Asaay Kit on Bioanalyzer (Agilent Technologoes, USA) and sequenced on Illumina HiSeq2000 platform.

4.2.5.2. PolyA RNA-seq

Polyadenylated RNA was isolated from 5µg total RNA sample by two sequential purifications using Dynabeads mRNA Purification Kit (ThermoFisher Scientific, USA) according to manufacturer's instruction. Purified poly(A) RNA was analyzed on RNA Pico Assay on Bioanalyzer. Sequencing libraries were then prepared using Small RNA Library Prep Kit (Illumina, USA) using the modified protocol as described above for total RNA-seq.

4.2.6. Bioinformatics Analysis of Sequencing data

4.2.6.1. ChIP-Seq data processing

Raw sequencing reads were aligned to human genome (hg19) using Bowtie2. Sequence reads that aligned multiple times in genome with equal alignment score were discarded as well as the duplicate reads with identical coordinates (sequencing depth taken into account) were discarded to remove potential sequencing and alignment artifacts. Aligned reads were elongated *in silico* using the DNA fragment size inferred from paired-reads or an estimated optimal fragment size for orphan reads using an in-house developed R pipeline named PASHA (Fenouil et al., 2016). These elongated reads were then used to calculate the number of fragments that overlapped at a given nucleotide thus representing an enrichment score for each nucleotide in the genome. Wiggle files representing average enrichment score of every 50bp were generated. Sequencing data from Input samples were treated in the same way to generate Input wiggle files. All wiggle files were then rescaled to normalize the enrichment scores to reads per million. Enrichment scores from Input sample wiggle files were then subtracted from ChIP sample wiggle files. This allowed us to remove/reduce the over-representation of certain genomic regions due to biased sonication and DNA sequencing. Besides this, input subtraction also improves the signal/noise ratio especially for ChIPs with low enrichment. Rescaled and Input subtracted wiggle files from biological replicate experiments were then used to generate a wiggle file that represents the average signal from biological replicates.

4.2.6.2. RNA-seq data processing

Raw sequencing reads were aligned to human genome (hg19) using TopHat2. Sequence reads that aligned multiple times in genome with equal alignment score were discarded. Reads that align to Watson or Crick strands were seperated and processed them separately using PASHA (Fenouil et al., 2016) pipeline to generate strand-specific wiggle files. All wiggle files were then rescaled to normalize the enrichment scores to reads per million. Rescaled wiggle files from biological replicate experiments were then used to generate a wiggle file that represents the average strand-specific RNA signal from several biological replicates.

4.2.6.3. Gene Expression Analysis

Differential Gene Expression (DGE) analysis was performed using the DESeq package from Bioconductor. First, HTseq-count program from the HTSeq framework was used to count the sequence reads mapped to gene annotations. These counts were processed using the DESeq package to identify genes that are at least 3 fold differentially expressed relative to the reference sample.

4.2.6.4. Average Metagene Profiles

To generate average signal profiles, we selected the hg19 genes that do not have any other annotation within 20Kb around boundaries. Removal of the annotations too close to each other is necessary to avoid mixing signals from close-by annotations, which can cause misinterpretation of the results. ChIP-seq and strand-specific RNA-seq values from wiggle files were retrieved with in-house R and Perl scripts for selected genes and enhancer regions. An algorithm described previously (KOCH ET AL 2011) was used to rescale the genes to same length by interpolating the values on 1000 points and build a matrix on which each column is averaged and resulting values are used to plot average metagene profiles.

5. Bibliography

A

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nature reviews Genetics *13*, 720-731.

Ahn, S.H., Kim, M., and Buratowski, S. (2004). Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. Molecular cell *13*, 67-76.

Akhtar, M.S., Heidemann, M., Tietjen, J.R., Zhang, D.W., Chapman, R.D., Eick, D., and Ansari, A.Z. (2009). TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. Molecular cell *34*, 387-393.

Allen, B.L., and Taatjes, D.J. (2015). The Mediator complex: a central integrator of transcription. Nature reviews Molecular cell biology *16*, 155-166.

Allison, L.A., Moyle, M., Shales, M., and Ingles, C.J. (1985). Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. Cell *42*, 599-610.

Allison, L.A., Wong, J.K., Fitzpatrick, V.D., Moyle, M., and Ingles, C.J. (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. Molecular and cellular biology *8*, 321-329.

Almada, A.E., Wu, X., Kriz, A.J., Burge, C.B., and Sharp, P.A. (2013). Promoter directionality is controlled by U1 snRNP and polyadenylation signals. Nature *499*, 360-363.

Arigo, J.T., Eyler, D.E., Carroll, K.L., and Corden, J.L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. Molecular cell *23*, 841-851.

B

Baillat, D., Hakimi, M.A., Naar, A.M., Shilatifard, A., Cooch, N., and Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. Cell *123*, 265-276.

Baillat, D., and Wagner, E.J. (2015). Integrator: surprisingly diverse functions in gene expression. Trends in biochemical sciences *40*, 257-264.

Bartkowiak, B., Liu, P., Phatnani, H.P., Fuda, N.J., Cooper, J.J., Price, D.H., Adelman, K., Lis, J.T., and Greenleaf, A.L. (2010). CDK12 is a transcription

elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. Genes & development *24*, 2303-2316.

Bartolomei, M.S., and Corden, J.L. (1987). Localization of an alpha-amanitin resistance mutation in the gene encoding the largest subunit of mouse RNA polymerase II. Molecular and cellular biology *7*, 586-594.

Baskaran, R., Dahmus, M.E., and Wang, J.Y. (1993). Tyrosine phosphorylation of mammalian RNA polymerase II carboxyl-terminal domain. Proceedings of the National Academy of Sciences of the United States of America *90*, 11167-11171.

Bataille, A.R., Jeronimo, C., Jacques, P.E., Laramee, L., Fortin, M.E., Forest, A., Bergeron, M., Hanes, S.D., and Robert, F. (2012). A universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. Molecular cell *45*, 158-170.

Blazek, D., Kohoutek, J., Bartholomeeusen, K., Johansen, E., Hulinkova, P., Luo, Z., Cimermancic, P., Ule, J., and Peterlin, B.M. (2011). The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. Genes & development *25*, 2158-2172.

Boeing, S., Rigault, C., Heidemann, M., Eick, D., and Meisterernst, M. (2010). RNA polymerase II C-terminal heptarepeat domain Ser-7 phosphorylation is established in a mediator-dependent fashion. The Journal of biological chemistry *285*, 188-196.

Bres, V., Yoh, S.M., and Jones, K.A. (2008). The multi-tasking P-TEFb complex. Current opinion in cell biology *20*, 334-340.

Buratowski, S. (2009). Progression through the RNA polymerase II CTD cycle. Molecular cell *36*, 541-546.

C

Cadena, D.L., and Dahmus, M.E. (1987). Messenger RNA synthesis in mammalian cells is catalyzed by the phosphorylated form of RNA polymerase II. The Journal of biological chemistry *262*, 12468-12474.

Calvo, O., and Manley, J.L. (2001). Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination. Molecular cell *7*, 1013-1023.

Carroll, K.L., Ghirlando, R., Ames, J.M., and Corden, J.L. (2007). Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. Rna *13*, 361-373.

Carrozza, M.J., Li, B., Florens, L., Suganuma, T., Swanson, S.K., Lee, K.K., Shia, W.J., Anderson, S., Yates, J., Washburn, M.P.*, et al.* (2005). Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. Cell *123*, 581-592.

Chapman, R.D., Conrad, M., and Eick, D. (2005). Role of the mammalian RNA polymerase II C-terminal domain (CTD) nonconsensus repeats in CTD stability and cell proliferation. Molecular and cellular biology *25*, 7665-7674.

Chapman, R.D., Palancade, B., Lang, A., Bensaude, O., and Eick, D. (2004). The last CTD repeat of the mammalian RNA polymerase II large subunit is important for its stability. Nucleic acids research *32*, 35-44.

Cho, E.J., Kobor, M.S., Kim, M., Greenblatt, J., and Buratowski, S. (2001). Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. Genes & development *15*, 3319-3329.

Connelly, S., and Manley, J.L. (1988). A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. Genes & development *2*, 440-452.

Corden, J.L., Cadena, D.L., Ahearn, J.M., Jr., and Dahmus, M.E. (1985). A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. Proceedings of the National Academy of Sciences of the United States of America *82*, 7934-7938.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

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. The Journal of biological chemistry *256*, 3332-3339.

David, C.J., Boyne, A.R., Millhouse, S.R., and Manley, J.L. (2011). The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. Genes & development *25*, 972-983.

Descostes, N., Heidemann, M., Spinelli, L., Schuller, R., Maqbool, M.A., Fenouil, R., Koch, F., Innocenti, C., Gut, M., Gut, I.*, et al.* (2014). Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells. eLife *3*, e02105.

Devaiah, B.N., Lewis, B.A., Cherman, N., Hewitt, M.C., Albrecht, B.K., Robey, P.G., Ozato, K., Sims, R.J., 3rd, and Singer, D.S. (2012). BRD4 is an atypical kinase that phosphorylates serine2 of the RNA polymerase II carboxy-terminal domain. Proceedings of the National Academy of Sciences of the United States of America *109*, 6927-6932.

Dias, J.D., Rito, T., Torlai Triglia, E., Kukalev, A., Ferrai, C., Chotalia, M., Brookes, E., Kimura, H., and Pombo, A. (2015). Methylation of RNA

polymerase II non-consensus Lysine residues marks early transcription in mammalian cells. eLife *4*.

Dieci, G., Fiorino, G., Castelnuovo, M., Teichmann, M., and Pagano, A. (2007). The expanding RNA polymerase III transcriptome. Trends in genetics : TIG *23*, 614-622.

E

Egloff, S., O'Reilly, D., Chapman, R.D., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D., and Murphy, S. (2007). Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. Science *318*, 1777-1779.

Egloff, S., Szczepaniak, S.A., Dienstbier, M., Taylor, A., Knight, S., and Murphy, S. (2010). The integrator complex recognizes a new double mark on the RNA polymerase II carboxyl-terminal domain. The Journal of biological chemistry *285*, 20564-20569.

Egloff, S., Zaborowska, J., Laitem, C., Kiss, T., and Murphy, S. (2012). Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. Molecular cell *45*, 111-122.

Eick, D., and Geyer, M. (2013). The RNA polymerase II carboxy-terminal domain (CTD) code. Chemical reviews *113*, 8456-8490.

Esnault, C., Ghavi-Helm, Y., Brun, S., Soutourina, J., Van Berkum, N., Boschiero, C., Holstege, F., and Werner, M. (2008). Mediator-dependent recruitment of TFIIH modules in preinitiation complex. Molecular cell *31*, 337-346.

F

Fabrega, C., Shen, V., Shuman, S., and Lima, C.D. (2003). Structure of an mRNA capping enzyme bound to the phosphorylated carboxy-terminal domain of RNA polymerase II. Molecular cell *11*, 1549-1561.

Fan, X., Chou, D.M., and Struhl, K. (2006). Activator-specific recruitment of Mediator in vivo. Nature structural & molecular biology *13*, 117-120.

Fenouil, R., Descostes, N., Spinelli, L., Koch, F., Maqbool, M.A., Benoukraf, T., Cauchy, P., Innocenti, C., Ferrier, P., and Andrau, J.C. (2016). Pasha: a versatile R package for piling chromatin HTS data. Bioinformatics *32*, 2528-2530.

Fong, N., and Bentley, D.L. (2001). Capping, splicing, and 3' processing are independently stimulated by RNA polymerase II: different functions for different segments of the CTD. Genes & development *15*, 1783-1795.

Fong, N., Brannan, K., Erickson, B., Kim, H., Cortazar, M.A., Sheridan, R.M., Nguyen, T., Karp, S., and Bentley, D.L. (2015). Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II

Termination Suggest Widespread Kinetic Competition. Molecular cell *60*, 256-267.

Fujinaga, K., Irwin, D., Huang, Y., Taube, R., Kurosu, T., and Peterlin, B.M. (2004). Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. Molecular and cellular biology *24*, 787-795.

G

Gaertner, B., and Zeitlinger, J. (2014). RNA polymerase II pausing during development. Development *141*, 1179-1183.

Gardini, A., Baillat, D., Cesaroni, M., Hu, D., Marinis, J.M., Wagner, E.J., Lazar, M.A., Shilatifard, A., and Shiekhattar, R. (2014). Integrator regulates transcriptional initiation and pause release following activation. Molecular cell *56*, 128-139.

Glover-Cutter, K., Larochelle, S., Erickson, B., Zhang, C., Shokat, K., Fisher, R.P., and Bentley, D.L. (2009). TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II. Molecular and cellular biology *29*, 5455-5464.

Greifenberg, A.K., Honig, D., Pilarova, K., Duster, R., Bartholomeeusen, K., Bosken, C.A., Anand, K., Blazek, D., and Geyer, M. (2016). Structural and Functional Analysis of the Cdk13/Cyclin K Complex. Cell reports *14*, 320-331.

Grosso, A.R., Leite, A.P., Carvalho, S., Matos, M.R., Martins, F.B., Vitor, A.C., Desterro, J.M., Carmo-Fonseca, M., and de Almeida, S.F. (2015). Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. eLife *4*.

H

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. Nature structural & molecular biology *11*, 394-403.

Harlen, K.M., Trotta, K.L., Smith, E.E., Mosaheb, M.M., Fuchs, S.M., and Churchman, L.S. (2016). Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. Cell reports *15*, 2147-2158.

Hausmann, S., Koiwa, H., Krishnamurthy, S., Hampsey, M., and Shuman, S. (2005). Different strategies for carboxyl-terminal domain (CTD) recognition by serine 5-specific CTD phosphatases. The Journal of biological chemistry *280*, 37681-37688.

Hausmann, S., and Shuman, S. (2002). Characterization of the CTD phosphatase Fcp1 from fission yeast. Preferential dephosphorylation of serine 2 versus serine 5. The Journal of biological chemistry *277*, 21213-21220.

Heidemann, M., Hintermair, C., Voss, K., and Eick, D. (2013). Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. Biochimica et biophysica acta *1829*, 55-62.

Hintermair, C., Heidemann, M., Koch, F., Descostes, N., Gut, M., Gut, I., Fenouil, R., Ferrier, P., Flatley, A., Kremmer, E.*, et al.* (2012). Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. The EMBO journal *31*, 2784-2797.

Hsin, J.P., Li, W., Hoque, M., Tian, B., and Manley, J.L. (2014). RNAP II CTD tyrosine 1 performs diverse functions in vertebrate cells. eLife *3*, e02112.

Hurwitz, J. (2005). The discovery of RNA polymerase. The Journal of biological chemistry *280*, 42477-42485.

I

Ishihama, Y., Rappsilber, J., and Mann, M. (2006). Modular stop and go extraction tips with stacked disks for parallel and multidimensional Peptide fractionation in proteomics. J Proteome Res *5*, 988-994.

J

Jiang, Y.W., Veschambre, P., Erdjument-Bromage, H., Tempst, P., Conaway, J.W., Conaway, R.C., and Kornberg, R.D. (1998). Mammalian mediator of transcriptional regulation and its possible role as an end-point of signal transduction pathways. Proceedings of the National Academy of Sciences of the United States of America *95*, 8538-8543.

K

Kang, M.E., and Dahmus, M.E. (1993). RNA polymerases IIA and IIO have distinct roles during transcription from the TATA-less murine dihydrofolate reductase promoter. The Journal of biological chemistry *268*, 25033-25040.

Kedinger, C., Gniazdowski, M., Mandel, J.L., Jr., Gissinger, F., and Chambon, P. (1970). Alpha-amanitin: a specific inhibitor of one of two DNA-pendent RNA polymerase activities from calf thymus. Biochemical and biophysical research communications *38*, 165-171.

Keller, W., Bienroth, S., Lang, K.M., and Christofori, G. (1991). Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA. The EMBO journal *10*, 4241-4249.

Kim, H., Erickson, B., Luo, W., Seward, D., Graber, J.H., Pollock, D.D., Megee, P.C., and Bentley, D.L. (2010). Gene-specific RNA polymerase II phosphorylation and the CTD code. Nature structural & molecular biology *17*, 1279-1286.

Kim, J.B., and Sharp, P.A. (2001). Positive transcription elongation factor B phosphorylates hSPT5 and RNA polymerase II carboxyl-terminal domain

independently of cyclin-dependent kinase-activating kinase. The Journal of biological chemistry *276*, 12317-12323.

Kim, M., Ahn, S.H., Krogan, N.J., Greenblatt, J.F., and Buratowski, S. (2004a). Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. The EMBO journal *23*, 354-364.

Kim, M., Krogan, N.J., Vasiljeva, L., Rando, O.J., Nedea, E., Greenblatt, J.F., and Buratowski, S. (2004b). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature *432*, 517-522.

Kim, M., Suh, H., Cho, E.J., and Buratowski, S. (2009a). Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2, 5, and 7. The Journal of biological chemistry *284*, 26421-26426.

Kim, T., and Buratowski, S. (2009b). Dimethylation of H3K4 by Set1 recruits the Set3 histone deacetylase complex to 5' transcribed regions. Cell *137*, 259-272.

Kim, Y.J., Bjorklund, S., Li, Y., Sayre, M.H., and Kornberg, R.D. (1994). A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. Cell *77*, 599-608.

Kimura, M., Sakurai, H., and Ishihama, A. (2001). Intracellular contents and assembly states of all 12 subunits of the RNA polymerase II in the fission yeast Schizosaccharomyces pombe. European journal of biochemistry / FEBS *268*, 612-619.

Kizer, K.O., Phatnani, H.P., Shibata, Y., Hall, H., Greenleaf, A.L., and Strahl, B.D. (2005). A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. Molecular and cellular biology *25*, 3305-3316.

Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T.K., Zacarias-Cabeza, J., Spicuglia, S., de la Chapelle, A.L., Heidemann, M., Hintermair, C.*, et al.* (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nature structural & molecular biology *18*, 956-963.

Komarnitsky, P., Cho, E.J., and Buratowski, S. (2000). Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes & development *14*, 2452-2460.

Kornberg, R.D. (2005). Mediator and the mechanism of transcriptional activation. Trends in biochemical sciences *30*, 235-239.

Krishnamurthy, S., He, X., Reyes-Reyes, M., Moore, C., and Hampsey, M. (2004). Ssu72 Is an RNA polymerase II CTD phosphatase. Molecular cell *14*, 387-394.

Kubicek, K., Cerna, H., Holub, P., Pasulka, J., Hrossova, D., Loehr, F., Hofr, C., Vanacova, S., and Stefl, R. (2012). Serine phosphorylation and proline

isomerization in RNAP II CTD control recruitment of Nrd1. Genes & development *26*, 1891-1896.

Kuehner, J.N., Pearson, E.L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. Nature reviews Molecular cell biology *12*, 283-294.

Kusser, A.G., Bertero, M.G., Naji, S., Becker, T., Thomm, M., Beckmann, R., and Cramer, P. (2008). Structure of an archaeal RNA polymerase. Journal of molecular biology *376*, 303-307.

L

Lehner, B., Williams, G., Campbell, R.D., and Sanderson, C.M. (2002). Antisense transcripts in the human genome. Trends in genetics : TIG *18*, 63-65.

Lepoivre, C., Belhocine, M., Bergon, A., Griffon, A., Yammine, M., Vanhille, L., Zacarias-Cabeza, J., Garibal, M.A., Koch, F., Maqbool, M.A.*, et al.* (2013). Divergent transcription is associated with promoters of transcriptional regulators. BMC Genomics *14*, 914.

Licatalosi, D.D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J.B., and Bentley, D.L. (2002). Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. Molecular cell *9*, 1101-1111.

Lin, P.S., Dubois, M.F., and Dahmus, M.E. (2002). TFIIF-associating carboxyl-terminal domain phosphatase dephosphorylates phosphoserines 2 and 5 of RNA polymerase II. The Journal of biological chemistry *277*, 45949-45956.

Litingtung, Y., Lawler, A.M., Sebald, S.M., Lee, E., Gearhart, J.D., Westphal, H., and Corden, J.L. (1999). Growth retardation and neonatal lethality in mice with a homozygous deletion in the C-terminal domain of RNA polymerase II. Molecular & general genetics : MGG *261*, 100-105.

Liu, P., Greenleaf, A.L., and Stiller, J.W. (2008). The essential sequence elements required for RNAP II carboxyl-terminal domain function in yeast and their evolutionary conservation. Molecular biology and evolution *25*, 719-727.

Liu, P., Kenney, J.M., Stiller, J.W., and Greenleaf, A.L. (2010). Genetic organization, length conservation, and evolution of RNA polymerase II carboxyl-terminal domain. Molecular biology and evolution *27*, 2628-2641.

Logan, J., Falck-Pedersen, E., Darnell, J.E., Jr., and Shenk, T. (1987). A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proceedings of the National Academy of Sciences of the United States of America *84*, 8306-8310.

Lunde, B.M., Reichow, S.L., Kim, M., Suh, H., Leeper, T.C., Yang, F., Mutschler, H., Buratowski, S., Meinhart, A., and Varani, G. (2010). Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. Nature structural & molecular biology *17*, 1195-1201.

Luo, W., Johnson, A.W., and Bentley, D.L. (2006). The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model. Genes & development *20*, 954-965.

Lux, C., Albiez, H., Chapman, R.D., Heidinger, M., Meininghaus, M., Brack-Werner, R., Lang, A., Ziegler, M., Cremer, T., and Eick, D. (2005). Transition from initiation to promoter proximal pausing requires the CTD of RNA polymerase II. Nucleic acids research *33*, 5139-5144.

M

MacDonald, C.C., Wilusz, J., and Shenk, T. (1994). The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Molecular and cellular biology *14*, 6647-6654.

Malik, S., and Roeder, R.G. (2005). Dynamic regulation of pol II transcription by the mammalian Mediator complex. Trends in biochemical sciences *30*, 256-263.

Mandel, C.R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J.L., and Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature *444*, 953-956.

Marshall, N.F., and Price, D.H. (1995). Purification of P-TEFb, a transcription factor required for the transition into productive elongation. The Journal of biological chemistry *270*, 12335-12338.

Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., and Cramer, P. (2012). CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. Science *336*, 1723-1725.

Mayer, A., Lidschreiber, M., Siebert, M., Leike, K., Soding, J., and Cramer, P. (2010). Uniform transitions of the general RNA polymerase II transcription complex. Nature structural & molecular biology *17*, 1272-1278.

Meinhart, A., and Cramer, P. (2004). Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. Nature *430*, 223-226.

Meininghaus, M., Chapman, R.D., Horndasch, M., and Eick, D. (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. The Journal of biological chemistry *275*, 24375-24382.

Moreira, M.C., Klur, S., Watanabe, M., Nemeth, A.H., Le Ber, I., Moniz, J.C., Tranchant, C., Aubourg, P., Tazir, M., Schols, L.*, et al.* (2004). Senataxin, the ortholog of a yeast RNA helicase, is mutant in ataxia-ocular apraxia 2. Nat Genet *36*, 225-227.

Mosley, A.L., Pattenden, S.G., Carey, M., Venkatesh, S., Gilmore, J.M., Florens, L., Workman, J.L., and Washburn, M.P. (2009). Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. Molecular cell *34*, 168-178.

Mukundan, B., and Ansari, A. (2011). Novel role for mediator complex subunit Srb5/Med18 in termination of transcription. The Journal of biological chemistry *286*, 37053-37057.

Murthy, K.G., and Manley, J.L. (1995). The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. Genes & development *9*, 2672-2683.

Myers, L.C., Gustafsson, C.M., Bushnell, D.A., Lui, M., Erdjument-Bromage, H., Tempst, P., and Kornberg, R.D. (1998). The Med proteins of yeast and their function through the RNA polymerase II carboxy-terminal domain. Genes & development *12*, 45-54.

N

Naar, A.M., Taatjes, D.J., Zhai, W., Nogales, E., and Tjian, R. (2002). Human CRSP interacts with RNA polymerase II CTD and adopts a specific CTD-bound conformation. Genes & development *16*, 1339-1344.

Nag, A., Narsinh, K., and Martinson, H.G. (2007). The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. Nature structural & molecular biology *14*, 662-669.

Nair, D., Kim, Y., and Myers, L.C. (2005). Mediator and TFIIH govern carboxyl-terminal domain-dependent transcription in yeast extracts. The Journal of biological chemistry *280*, 33739-33748.

Nechaev, S., and Adelman, K. (2011). Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. Biochimica et biophysica acta *1809*, 34-45.

Ng, H.H., Robert, F., Young, R.A., and Struhl, K. (2003). Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. Molecular cell *11*, 709-719.

Ni, Z., Xu, C., Guo, X., Hunter, G.O., Kuznetsova, O.V., Tempel, W., Marcon, E., Zhong, G., Guo, H., Kuo, W.H.*, et al.* (2014). RPRD1A and RPRD1B are human RNA polymerase II C-terminal domain scaffolds for Ser5 dephosphorylation. Nature structural & molecular biology *21*, 686-695.

Nikolov, D.B., and Burley, S.K. (1997). RNA polymerase II transcription initiation: a structural view. Proceedings of the National Academy of Sciences of the United States of America *94*, 15-22.

Nojima, T., Gomes, T., Grosso, A.R., Kimura, H., Dye, M.J., Dhir, S., Carmo-Fonseca, M., and Proudfoot, N.J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. Cell *161*, 526-540.

P

Peterlin, B.M., and Price, D.H. (2006). Controlling the elongation phase of transcription with P-TEFb. Molecular cell *23*, 297-305.

Porrua, O., and Libri, D. (2013). A bacterial-like mechanism for transcription termination by the Sen1p helicase in budding yeast. Nature structural & molecular biology *20*, 884-891.

Porrua, O., and Libri, D. (2015). Transcription termination and the control of the transcriptome: why, where and how to stop. Nature reviews Molecular cell biology *16*, 190-202

Proudfoot, N.J. (2011). Ending the message: poly(A) signals then and now. Genes & development *25*, 1770-1782.

Proudfoot, N.J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. Science *352*, aad9926.

R

Reyes-Reyes, M., and Hampsey, M. (2007). Role for the Ssu72 C-terminal domain phosphatase in RNA polymerase II transcription elongation. Molecular and cellular biology *27*, 926-936.

Richard, P., and Manley, J.L. (2009). Transcription termination by nuclear RNA polymerases. Genes & development *23*, 1247-1269.

Robinson, P.J., Bushnell, D.A., Trnka, M.J., Burlingame, A.L., and Kornberg, R.D. (2012). Structure of the mediator head module bound to the carboxy-terminal domain of RNA polymerase II. Proceedings of the National Academy of Sciences of the United States of America *109*, 17931-17935.

Rodriguez, C.R., Cho, E.J., Keogh, M.C., Moore, C.L., Greenleaf, A.L., and Buratowski, S. (2000). Kin28, the TFIIH-associated carboxy-terminal domain kinase, facilitates the recruitment of mRNA processing machinery to RNA polymerase II. Molecular and cellular biology *20*, 104-112.

Roeder, R.G., and Rutter, W.J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature *224*, 234-237.

Rosonina, E., Kaneko, S., and Manley, J.L. (2006). Terminating the transcript: breaking up is hard to do. Genes & development *20*, 1050-1056.

Russell, J., and Zomerdijk, J.C. (2005). RNA-polymerase-I-directed rDNA transcription, life and works. Trends in biochemical sciences *30*, 87-96.

Rutkowski, A.J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C.C*., et al.* (2015). Widespread disruption of host transcription termination in HSV-1 infection. Nature communications *6*, 7126.

Ryan, K., Calvo, O., and Manley, J.L. (2004). Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. Rna *10*, 565-573.

S

Schneider, S., Pei, Y., Shuman, S., and Schwer, B. (2010). Separable functions of the fission yeast Spt5 carboxyl-terminal domain (CTD) in capping enzyme binding and transcription elongation overlap with those of the RNA polymerase II CTD. Molecular and cellular biology *30*, 2353-2364.

Schroder, S., Herker, E., Itzen, F., He, D., Thomas, S., Gilchrist, D.A., Kaehlcke, K., Cho, S., Pollard, K.S., Capra, J.A*., et al.* (2013). Acetylation of RNA polymerase II regulates growth-factor-induced gene transcription in mammalian cells. Molecular cell *52*, 314-324.

Schuller, R., and Eick, D. (2016a). Getting Access to Low-Complexity Domain Modifications. Trends in biochemical sciences *41*, 894-897.

Schuller, R., Forne, I., Straub, T., Schreieck, A., Texier, Y., Shah, N., Decker, T.M., Cramer, P., Imhof, A., and Eick, D. (2016b). Heptad-Specific Phosphorylation of RNA Polymerase II CTD. Molecular cell *61*, 305-314.

Schwartz, L.B., Sklar, V.E., Jaehning, J.A., Weinmann, R., and Roeder, R.G. (1974). Isolation and partial characterization of the multiple forms of deoxyribonucleic acid-dependent ribonucleic acid polymerase in the mouse myeloma, MOPC 315. The Journal of biological chemistry *249*, 5889-5897.

Schwer, B., Sanchez, A.M., and Shuman, S. (2012). Punctuation and syntax of the RNA polymerase II CTD code in fission yeast. Proceedings of the National Academy of Sciences of the United States of America *109*, 18024-18029.

Schwer, B., and Shuman, S. (2011). Deciphering the RNA polymerase II CTD code in fission yeast. Molecular cell *43*, 311-318.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.

Sims, R.J., 3rd, Rojas, L.A., Beck, D., Bonasio, R., Schuller, R., Drury, W.J., 3rd, Eick, D., and Reinberg, D. (2011). The C-terminal domain of RNA polymerase II is modified by site-specific methylation. Science *332*, 99-103.

Skaar, J.R., Ferris, A.L., Wu, X., Saraf, A., Khanna, K.K., Florens, L., Washburn, M.P., Hughes, S.H., and Pagano, M. (2015). The Integrator complex controls the termination of transcription at diverse classes of gene targets. Cell research *25*, 288-305.

Skourti-Stathaki, K., Proudfoot, N.J., and Gromak, N. (2011). Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Molecular cell *42*, 794-805.

Sogaard, T.M., and Svejstrup, J.Q. (2007). Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator. The Journal of biological chemistry *282*, 14113-14120.

Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S.*, et al.* (2014). Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. Nature communications *5*, 5531.

Steinmetz, E.J., Conrad, N.K., Brow, D.A., and Corden, J.L. (2001). RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. Nature *413*, 327-331.

Stiller, J.W., and Cook, M.S. (2004). Functional unit of the RNA polymerase II C-terminal domain lies within heptapeptide pairs. Eukaryotic cell *3*, 735-740.

Stiller, J.W., and Hall, B.D. (2002). Evolution of the RNA polymerase II C-terminal domain. Proceedings of the National Academy of Sciences of the United States of America *99*, 6091-6096.

Stiller, J.W., McConaughy, B.L., and Hall, B.D. (2000). Evolutionary complementation for polymerase II CTD function. Yeast *16*, 57-64.

Suh, H., Ficarro, S.B., Kang, U.B., Chun, Y., Marto, J.A., and Buratowski, S. (2016). Direct Analysis of Phosphorylation Sites on the Rpb1 C-Terminal Domain of RNA Polymerase II. Molecular cell *61*, 297-304.

T

Takagaki, Y., and Manley, J.L. (1994). A polyadenylation factor subunit is the human homologue of the Drosophila suppressor of forked protein. Nature *372*, 471-474.

Takahashi, H., Parmely, T.J., Sato, S., Tomomori-Sato, C., Banks, C.A., Kong, S.E., Szutorisz, H., Swanson, S.K., Martin-Brown, S., Washburn, M.P.*, et al.* (2011). Human mediator subunit MED26 functions as a docking site for transcription elongation factors. Cell *146*, 92-104.

Tirode, F., Busso, D., Coin, F., and Egly, J.M. (1999). Reconstitution of the transcription factor TFIIH: assignment of functions for the three enzymatic subunits, XPB, XPD, and cdk7. Molecular cell *3*, 87-95.

Tsai, K.L., Sato, S., Tomomori-Sato, C., Conaway, R.C., Conaway, J.W., and Asturias, F.J. (2013). A conserved Mediator-CDK8 kinase module association regulates Mediator-RNA polymerase II interaction. Nature structural & molecular biology *20*, 611-619.

V

Vannini, A., and Cramer, P. (2012). Conservation between the RNA polymerase I, II, and III transcription initiation machineries. Molecular cell *45*, 439-446.

Venkatesh, S., Li, H., Gogol, M.M., and Workman, J.L. (2016). Selective suppression of antisense transcription by Set2-mediated H3K36 methylation. Nature communications *7*, 13610.

Viladevall, L., St Amour, C.V., Rosebrock, A., Schneider, S., Zhang, C., Allen, J.J., Shokat, K.M., Schwer, B., Leatherwood, J.K., and Fisher, R.P. (2009). TFIIH and P-TEFb coordinate transcription with capping enzyme recruitment at specific genes in fission yeast. Molecular cell *33*, 738-751.

Vilborg, A., Passarelli, M.C., Yario, T.A., Tycowski, K.T., and Steitz, J.A. (2015). Widespread Inducible Transcription Downstream of Human Genes. Molecular cell *59*, 449-461.

Voss, K., Forne, I., Descostes, N., Hintermair, C., Schuller, R., Maqbool, M.A., Heidemann, M., Flatley, A., Imhof, A., Gut, M.*, et al.* (2015). Site-specific methylation and acetylation of lysine residues in the C-terminal domain (CTD) of RNA polymerase II. Transcription *6*, 91-101.

W

Wada, T., Takagi, T., Yamaguchi, Y., Watanabe, D., and Handa, H. (1998). Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. The EMBO journal *17*, 7395-7403.

Werner, F., and Grohmann, D. (2011). Evolution of multisubunit RNA polymerases in the three domains of life. Nature reviews Microbiology *9*, 85-98.

West, M.L., and Corden, J.L. (1995). Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. Genetics *140*, 1223-1233.

West, S., Gromak, N., and Proudfoot, N.J. (2004). Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature *432*, 522-525.

Wood, A., and Shilatifard, A. (2006). Bur1/Bur2 and the Ctk complex in yeast: the split personality of mammalian P-TEFb. Cell cycle *5*, 1066-1068.

Wu, C.H., Yamaguchi, Y., Benjamin, L.R., Horvat-Gordon, M., Washinsky, J., Enerly, E., Larsson, J., Lambertsson, A., Handa, H., and Gilmour, D. (2003). NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. Genes & development *17*, 1402-1414.

Wu, X., and Sharp, P.A. (2013). Divergent transcription: a driving force for new gene origination? Cell *155*, 990-996.

Y

Yamada, T., Yamaguchi, Y., Inukai, N., Okamoto, S., Mura, T., and Handa, H. (2006). P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. Molecular cell *21*, 227-237.

Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., and Handa, H. (1999). NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. Cell *97*, 41-51.

Yang, C., and Stiller, J.W. (2014). Evolutionary diversity and taxon-specific modifications of the RNA polymerase II C-terminal domain. Proceedings of the National Academy of Sciences of the United States of America *111*, 5920-5925.

Yoh, S.M., Cho, H., Pickle, L., Evans, R.M., and Jones, K.A. (2007). The Spt6 SH2 domain binds Ser2-P RNAPII to direct Iws1-dependent mRNA splicing and export. Genes & development *21*, 160-174.

Z

Zaborowska, J., Egloff, S., and Murphy, S. (2016). The pol II CTD: new twists in the tail. Nature structural & molecular biology *23*, 771-777.

Zhang, D.W., Mosley, A.L., Ramisetty, S.R., Rodriguez-Molina, J.B., Washburn, M.P., and Ansari, A.Z. (2012). Ssu72 phosphatase-dependent erasure of phospho-Ser7 marks on the RNA polymerase II C-terminal domain is essential for viability and transcription termination. The Journal of biological chemistry *287*, 8541-8551.

Zhang, H., Rigo, F., and Martinson, H.G. (2015). Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. Molecular cell *59*, 437-448.

Zhang, Z., Fu, J., and Gilmour, D.S. (2005). CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11. Genes & development *19*, 1572-1580.

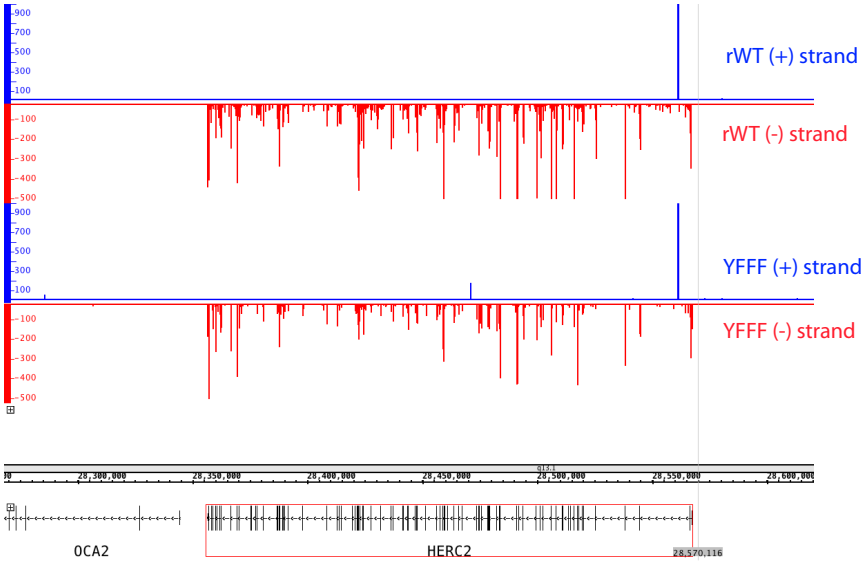
Zhang, Z., and Gilmour, D.S. (2006). Pcf11 is a termination factor in Drosophila that dismantles the elongation complex by bridging the CTD of RNA polymerase II to the nascent transcript. Molecular cell *21*, 65-74.

Zhao, D.Y., Gish, G., Braunschweig, U., Li, Y., Ni, Z., Schmitges, F.W., Zhong, G., Liu, K., Li, W., Moffat, J.*, et al.* (2016). SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination. Nature *529*, 48-53.

Zhou, Q., Li, T., and Price, D.H. (2012). RNA polymerase II elongation control. Annual review of biochemistry *81*, 119-143.

6. Supplementary

A. Figure:



Supplementary Figure 1: Comparison of total RNA-seq signals for the HERC2 gene in rWT and the mutant YFFF. The gene is transcribed on the minus strand of the genome (red signals).

B. Tables:

Supplementary table 1: Peptide counts of proteins interacting with Pol II of both, the rWT and the mutant YFFF for all 5 biological replicates. Samples in the experiment number 1 and 2 were subjected to on-beads trypsin digest, while 3,4 and 5 were subjected to in-gel trypsin digest. (*) represents the sample that had less peptide.

Peptide counts of selected proteins interacting with Pol II of both, the rWT and the mutant YFFF													
			rWT						**YFFF**				
	Experiment number		**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5***
	Uniprot ID	Gene Name	Peptide counts						Peptide counts				
	Polymerase Subunits												
1	P24928	RPB1	119	132	156	153	153		98	126	139	102	11
2	P30876	RPB2	33	53	73	68	70		26	49	61	35	1
3	P19387	RPB3	7	15	17	19	16		5	14	13	9	0
4	O15514	RPB4	0	1	3	4	11		0	0	1	2	0
5	P19388	RPB5	4	11	14	11	10		6	11	9	5	0

6	U3KPY1	RPB6	0	0	1	1	2		0	0	1	0	0
7	P62487	RPB7	1	0	3	4	4		0	1	1	0	0
8	P52434	RPB8	8	12	12	11	12		7	11	11	9	2
9	P36954	RPB9	2	5	7	9	8		2	5	5	3	0
10	P62875	RPB10	3	3	1	1	2		3	3	1	1	0
11	P52435	RPB11	5	7	6	4	9		5	6	6	4	0
12	P53803	RPB12	2	1	2	1	3		0	2	2	1	0
	Splicing factors												
13	Q07955	SRSF1	5	14	20	11	15		5	16	20	13	2
14	J3KP15	SRSF2	0	0	7	0	2		0	2	6	1	0
15	P84103	SRSF3	4	11	10	7	6		3	11	10	10	1
16	Q08170	SRSF4	3	8	7	5	4		2	8	8	4	1
17	Q13243	SRSF5	0	4	5	4	3		1	5	7	3	1
18	Q13247	SRSF6	3	9	9	8	8		3	8	9	9	1
19	Q16629	SRSF7	3	10	11	8	11		7	11	10	10	2
20	Q13242	SRSF9	1	9	16	11	12		4	11	20	17	2
21	O75494	SRSF10	1	9	11	8	10		1	9	12	9	0
22	Q5T760	SRSF11	0	1	2	0	0		0	2	4	0	0
23	Q01081	U2AF1	1	7	7	4	5		1	6	7	3	1
24	P26368	U2AF2	2	2	14	2	4		0	4	8	0	0
	3' end processing and termination factors												
25	Q10570	CPSF1	0	3	13	4	3		0	2	8	1	0
26	Q9P2I0	CPSF2	0	1	2	1	1		0	1	3	0	0
27	G5E9W3	CPSF3	0	0	2	1	2		0	0	1	0	0
28	B7Z7B0	CPSF4	0	1	0	0	1		0	0	1	1	0
29	O43809	CPSF5	0	3	3	0	0		0	3	5	0	0
30	F8WJN3	CPSF6	0	1	2	0	1		0	2	1	0	0
31	Q9H0D6	XRN2	0	8	23	12	14		1	8	17	2	0

Supplementary table 2: Log2Fold change (YFFF/rWT) of proteins interacting with Pol II of both, the rWT and the mutant YFFF.

	Log2Fold change (YFFF/rWT) of proteins interacting with Pol II of rWT and the mutant YFFF				
	Uniprot ID	**Gene Name**	**Description**	**Log2Fold Change (YFFF/rWT)**	**p-value**
	Polymerase Subunits				
1	P24928	RPB1	RNA Polymerase II subunit A	-2.985	1.14E-01
2	P30876	RPB2	RNA Polymerase II subunit B	-3.399	1.40E-01
3	P19387	RPB3	RNA Polymerase II subunit C	-4.302	1.74E-01
4	O15514	RPB4	RNA Polymerase II subunit D	-5.753	1.27E-01
5	P19388	RPB5	RNA Polymerase II subunit E	-3.989	2.23E-01
6	U3KPY1	RPB6	RNA Polymerase II subunit F	-4.930	2.12E-01
7	P62487	RPB7	RNA Polymerase II subunit G	-5.289	1.45E-01
8	P52434	RPB8	RNA Polymerase II subunit H	-2.692	1.14E-01
9	P36954	RPB9	RNA Polymerase II subunit I	-4.499	1.33E-01
10	P62875	RPB10	RNA Polymerase II subunit L	-3.389	3.17E-01
11	P52435	RPB11	RNA Polymerase II subunit J	-3.829	2.43E-01
12	P53803	RPB12	RNA Polymerase II subunit K	-6.601	9.63E-02
	Splicing Factors				
13	Q07955	SRSF1	Serine/Arginine-Rich Splicing Factor 1	-0.720	6.94E-01
14	J3KP15	SRSF2	Serine/Arginine-Rich Splicing Factor 2	2.484	6.05E-01
15	P84103	SRSF3	Serine/Arginine-Rich Splicing Factor 3	-1.116	6.05E-01
16	Q08170	SRSF4	Serine/Arginine-Rich Splicing Factor 4	-3.133	2.99E-01
17	Q13243	SRSF5	Serine/Arginine-Rich Splicing Factor 5	-2.073	6.31E-01
18	Q13247	SRSF6	Serine/Arginine-Rich Splicing Factor 6	-0.275	8.76E-01
19	Q16629	SRSF7	Serine/Arginine-Rich Splicing Factor 7	-0.720	6.11E-01
20	Q13242	SRSF9	Serine/Arginine-Rich Splicing	0.558	8.08E-01

			Factor 9		
21	O75494	SRSF10	Serine/Arginine-Rich Splicing Factor 10	-2.286	5.02E-01
22	Q5T760	SRSF11	Serine/Arginine-Rich Splicing Factor 11	0.217	9.57E-01
23	Q01081	U2AF1	Splicing Factor U2AF 35kDa Subunit	-1.231	5.62E-01
24	P26368	U2AF2	Splicing Factor U2AF 65 KD Subunit	-5.505	1.48E-01
	3' end processing and termination factors				
25	Q10570	CPSF1	Cleavage And Polyadenylation Specific Factor 1	-2.019	4.76E-01
26	Q9P2I0	CPSF2	Cleavage And Polyadenylation Specific Factor 2	-0.535	8.55E-01
27	G5E9W3	CPSF3	Cleavage And Polyadenylation Specific Factor 3	-2.999	2.14E-01
28	B7Z7B0	CPSF4	Cleavage And Polyadenylation Specific Factor 4	-0.116	9.69E-01
29	O43809	CPSF5	Cleavage And Polyadenylation Specific Factor 5	0.222	9.59E-01
30	F8WJN3	CPSF6	Cleavage And Polyadenylation Specific Factor 6	-1.317	6.89E-01
31	Q9H0D6	XRN2	5'-3' Exoribonuclease 2	-2.128	5.38E-01

Supplementary table 3: Peptide counts of proteins not interacting with the mutant YFFF in all 5 biological replicates

Peptide counts of proteins not interacting with the mutant YFFF													
			rWT						**YFFF**				
	Experiment number		**1**	**2**	**3**	**4**	**5**		**1**	**2**	**3**	**4**	**5***
	Uniprot ID	Gene Name	Peptide counts						Peptide counts				
1	Q9NRY2	INIP	2	5	3	5	8		0	0	0	0	0
2	Q9NPJ6	MED4	6	10	13	13	13		0	0	0	0	0
3	Q9UL03	INTS6	10	27	32	41	45		0	0	0	0	0
4	Q6P2C8	MED27	3	9	11	12	10		0	0	0	0	0

5	Q9BUE0	MED18	3	4	3	3	6		0	0	0	0	0
6	O95402	MED26	5	14	14	18	11		0	0	0	0	0
7	Q9BTT4	MED10	1	5	6	6	7		0	0	0	0	0
8	Q5T8T7	MED22	1	2	6	7	6		0	0	0	0	0
9	Q6P9B9	INTS5	3	15	13	17	26		0	0	0	0	0
10	Q68E01	INTS3	12	31	35	34	50		0	1	1	0	0
11	Q9NWA0	MED9	1	1	6	3	6		0	0	0	0	0
12	A0JLT2	MED19	1	3	3	3	5		0	0	0	0	0
13	Q96CB8	INTS12	1	6	9	15	13		0	0	0	0	0
14	Q9H0H0	INTS2	4	11	17	18	30		0	0	0	0	0
15	Q96G25	MED8	4	8	8	9	11		0	0	1	0	0
16	Q8N201	INTS1	16	43	48	60	74		0	0	1	0	0
17	Q9H944	MED20	1	2	4	7	6		0	0	1	0	0
18	Q96HW7	INTS4	2	12	15	31	41		0	0	0	0	0
19	Q9Y3C7	MED31	2	4	5	5	5		0	1	0	0	0
20	Q9H0M0	WWP1	3	12	18	20	27		0	0	0	0	0
21	Q9NVC6	MED17	3	6	19	17	21		0	0	1	0	0
22	Q9NV88	INTS9	2	5	7	13	21		0	0	0	0	0
23	Q96HR3	MED30	1	4	6	6	5		0	0	1	0	0
24	O60244	MED14	6	29	33	30	34		0	0	2	0	0
25	O00308	WWP2	5	12	13	27	24		0	0	0	0	0
26	Q15648	MED1	6	10	28	31	31		0	0	1	0	0
27	O75586	MED6	1	8	9	6	6		0	0	1	0	0
28	Q9BQ15	NABP2	0	5	1	4	7		0	0	0	0	0
29	Q9H204	MED28	0	1	3	3	4		0	0	0	0	0
30	O43513	MED7	0	1	3	7	8		0	0	0	0	0
31	Q9Y2Z0	SUGT1	1	1	2	3	4		1	0	0	0	0
32	Q96J02	ITCH	13	25	26	29	34		0	2	4	1	0
33	Q13503	MED21	1	2	1	4	3		0	0	1	0	0
34	Q75QN2	INTS8	0	6	6	18	23		0	0	0	0	0
35	Q9P086	MED11	0	1	2	4	3		0	0	0	0	0

36	Q96P16	RPRD1A	1	15	23	12	15		0	8	8	0	0
37	Q96RN5	MED15	0	1	5	6	7		0	0	0	0	0
38	Q5TA45	CPSF3L	0	2	8	8	13		0	0	0	0	0
39	Q15369	TCEB1	1	1	2	5	3		0	0	0	3	0
40	O95104	SCAF4	0	2	11	7	5		0	0	0	0	0
41	Q9NX70	MED29	0	3	2	3	3		0	0	0	0	0
42	O75448	MED24	0	4	6	17	20		0	0	0	0	0
43	Q5VT52	RPRD2	0	8	18	15	7		0	0	0	0	0
44	Q6DN90	IQSEC1	0	6	12	6	11		0	0	0	0	0
45	Q9Y2X0	MED16	0	3	2	8	12		0	0	0	0	0
46	Q9NVH2	INTS7	0	6	6	23	30		0	0	0	1	0
47	A8MU58	AIMP2	0	1	2	2	2		0	0	0	0	0
48	Q5JSJ4	INTS6L	2	8	8	10	8		0	0	0	0	0
49	Q5TEJ8	THEMIS2	0	1	3	3	6		0	0	0	0	0
50	P30153	PPP2R1A	0	2	1	4	6		0	0	1	0	0
51	Q99590	SCAF11	0	3	5	3	3		0	0	0	0	0
52	Q13418	ILK	0	3	5	3	3		0	0	2	0	0
53	Q53G59	KLHL12	0	1	2	1	1		0	0	0	0	0
54	O00329	PIK3CD	0	2	3	4	2		0	0	0	0	0
55	Q13049	TRIM32	0	1	1	2	3		0	0	0	0	0
56	Q14145	KEAP1	0	2	8	2	3		0	0	0	0	0
57	Q13501	SQSTM1	0	1	3	0	3		0	0	0	1	0
58	Q14344	GNA13	0	1	2	1	1		0	0	0	0	0
59	H3BQA8	WDR61	1	1	0	1	1		0	0	0	0	0
60	O00505	KPNA3	1	3	4	4	4		0	1	1	0	0
61	Q14157	UBAP2L	0	2	2	1	1		0	0	0	0	0
62	Q15418	RPS6KA1	2	6	6	5	7		0	2	2	0	0
63	Q8ND56	LSM14A	0	2	4	0	1		0	0	0	0	0
64	Q9ULK4	MED23	0	1	1	9	12		0	0	0	0	0
65	Q13451	FKBP5	2	3	2	4	8		1	1	2	0	0
66	P04637	TP53	0	2	1	0	1		0	0	1	0	0

67	Q71RC2	LARP4	0	0	2	3	3		0	0	0	0	0
68	Q16576	RBBP7	0	3	4	4	2		0	1	3	2	0
69	P13807	GYS1	0	3	2	0	1		0	0	0	0	0

Supplementary table 4: Log2Fold change (YFFF/rWT) of proteins not interacting with Pol II of both, the rWT and the mutant YFFF.

	Log2Fold change (YFFF/rWT) of proteins not interacting with Pol II of both, the rWT and the mutant YFFF				
	Uniprot ID	**Gene Name**	**Description**	**Log2Fold Change (YFFF/rWT)**	**p-value**
1	Q9NRY2	INIP	INTS3 and NABP interacting protein	-14.548	5.588E-08
2	Q9NPJ6	MED4	Mediator Complex Subunit 4	-13.671	1.447E-09
3	Q9UL03	INTS6	Integrator Complex Subunit 6	-13.490	3.673E-09
4	Q6P2C8	MED27	Mediator Complex Subunit 27	-13.159	2.356E-09
5	Q9BUE0	MED18	Mediator Complex Subunit 18	-12.981	1.433E-12
6	O95402	MED26	Mediator Complex Subunit 26	-12.838	1.41E-11
7	Q9BTT4	MED10	Mediator Complex Subunit 10	-12.827	2.053E-08
8	Q5T8T7	MED22	Mediator Complex Subunit 22	-12.637	3.805E-08
9	Q6P9B9	INTS5	Integrator Complex Subunit 5	-12.259	7.930E-08
10	Q68E01	INTS3	Integrator Complex Subunit 3	-12.146	2.420E-05
11	Q9NWA0	MED9	Mediator Complex Subunit 9	-11.932	1.696E-06
12	A0JLT2	MED19	Mediator Complex Subunit 19	-11.858	4.219E-08
13	Q96CB8	INTS12	Integrator Complex Subunit 12	-11.533	2.203E-05
14	Q9H0H0	INTS2	Integrator Complex Subunit 2	-11.520	2.799E-08
15	Q96G25	MED8	Mediator Complex Subunit 8	-11.496	2.292E-04
16	Q8N201	INTS1	Integrator Complex Subunit 1	-11.496	2.449E-06
17	Q9H944	MED20	Mediator Complex Subunit 20	-11.388	2.521E-04
18	Q96HW7	INTS4	Integrator Complex Subunit 4	-11.258	5.507E-06
19	Q9Y3C7	MED31	Mediator Complex Subunit 31	-11.180	5.368E-04
20	Q9H0M0	WWP1	WW Domain containing E3 Ubiquitin Protein Ligase 1	-11.161	3.837E-07
21	Q9NVC6	MED17	Mediator Complex Subunit 17	-11.092	1.153E-04

22	Q9NV88	INTS9	Integrator Complex Subunit 9	-11.079	9.394E-07
23	Q96HR3	MED30	Mediator Complex Subunit 30	-11.069	2.195E-04
24	O60244	MED14	Mediator Complex Subunit 14	-10.979	1.094E-04
25	O00308	WWP2	WW Domain containing E3 Ubiquitin Protein Ligase 2	-10.976	5.113E-07
26	Q15648	MED1	Mediator Complex Subunit 1	-10.736	9.263E-06
27	O75586	MED6	Mediator Complex Subunit 6	-10.708	6.511E-04
28	Q9BQ15	NABP2	Nucleic acid binding protein 2	-10.628	4.230E-03
29	Q9H204	MED28	Mediator Complex Subunit 28	-9.820	4.124E-03
30	O43513	MED7	Mediator Complex Subunit 7	-9.771	3.981E-03
31	Q9Y2Z0	SUGT1	SGT1 Homolog, MIS12 Kinetochore Complex Assembly Cochaperone	-9.705	3.864E-11
32	Q96J02	ITCH	Itchy E3 Ubiquitin Protein Ligase	-9.434	2.206E-03
33	Q13503	MED21	Mediator Complex Subunit 21	-9.335	2.326E-03
34	Q75QN2	INTS8	Integrator Complex Subunit 8	-9.223	4.178E-03
35	Q9P086	MED11	Mediator Complex Subunit 11	-8.966	4.348E-03
36	Q96P16	RPRD1A	Regulation of nuclear pre-mRNA domain containing 1A (CTD phosphatase)	-8.924	1.335E-02
37	Q96RN5	MED15	Mediator Complex Subunit 15	-8.887	4.168E-03
38	Q5TA45	CPSF3L	Cleavage and Polyadenylation specificity factor 3-like (Integrator Complex Subunit 11)	-8.729	4.677E-03
39	Q15369	TCEB1	Transcription elongation factor B subunit 1	-8.481	5.145E-03
40	O95104	SCAF4	SR-related CTD associated factor 4	-8.349	4.985E-03
41	Q9NX70	MED29	Mediator Complex Subunit 29	-8.304	4.207E-03
42	O75448	MED24	Mediator Complex Subunit 24	-7.916	4.396E-03
43	Q5VT52	RPRD2	Regulation of nuclear pre-mRNA domain containing 2 (CTD phosphatase)	-7.903	4.556E-03
44	Q6DN90	IQSEC1	IQ motif and Sec7 Domain 1	-7.801	4.395E-03
45	Q9Y2X0	MED16	Mediator Complex Subunit 16	-7.639	4.243E-03
46	Q9NVH2	INTS7	Integrator Complex Subunit 7	-7.255	3.588E-02

47	A8MU58	AIMP2	Aminoacyl tRNA Synthetase Complex-interacting Multifuntional protein 2	-7.053	4.231E-03
48	Q5JSJ4	INTS6L	Integrator Complex Subunit 6 Like	-6.739	4.555E-03
49	Q5TEJ8	THEMIS2	Thymocyte selection associated family member 2	-6.695	4.375E-03
50	P30153	PPP2R1A	Protein phosphatase 2 regulatory subunit A, alpha	-6.577	7.267E-03
51	Q99590	SCAF11	SR-related CTD associated factor 11	-6.371	4.849E-03
52	Q13418	ILK	Integrin linked kinase	-6.116	4.462E-02
53	Q53G59	KLHL12	Kelch like family member 12	-6.051	4.223E-03
54	O00329	PIK3CD	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Delta	-5.961	4.118E-03
55	Q13049	TRIM32	Tripartite Motif Containing 32	-5.938	7.280E-03
56	Q14145	KEAP1	Kelch Like ECH Associated Protein 1	-5.919	2.511E-02
57	Q13501	SQSTM1	Sequestosome 1	-5.918	4.041E-02
58	Q14344	GNA13	G Protein Subunit Alpha 13	-5.908	6.915E-03
59	H3BQA8	WDR61	WD Repeat Domain 61	-5.784	5.656E-03
60	O00505	KPNA3	Karyopherin Subunit Alpha 3	-5.738	3.355E-02
61	Q14157	UBAP2L	Ubiquitin associated protein 2 like	-5.721	6.337E-03
62	Q15418	RPS6KA1	Ribosomal Protein S6 kinase A1	-5.715	3.025E-02
63	Q8ND56	LSM14A	LSM14A mRNA processing body assembly factor	-5.583	4.459E-02
64	Q9ULK4	MED23	Mediator Complex Subunit 23	-5.570	8.126E-03
65	Q13451	FKBP5	FK506 Binding protein 5	-5.353	1.853E-02
66	P04637	TP53	Tumor protein p53	-5.263	4.374E-02
67	Q71RC2	LARP4	La Ribonucleoprotein Domain Family Member 4	-5.112	4.040E-02
68	Q16576	RBBP7	Retinoblastoma Binding Protein 7	-5.100	4.517E-02
69	P13807	GYS1	Glycogen Synthase 1	-5.082	4.381E-02

Supplementary table 5: Log2Fold change (S2AAA/rWT) of proteins interacting with Pol II of both, rWT and the S2AAA mutant.

	Log2Fold change (S2AAA/rWT) of proteins interacting with Pol II of rWT and the S2AAA mutant				
	Uniprot ID	**Gene Name**	**Description**	**Log2Fold Change (S2AAA/rWT)**	**p-value**
	Polymerase Subunits				
1	P24928	RPB1	RNA Polymerase II subunit A	1.429	0.415
2	P30876	RPB2	RNA Polymerase II subunit B	1.380	0.584
3	P19387	RPB3	RNA Polymerase II subunit C	2.232	0.479
4	O15514	RPB4	RNA Polymerase II subunit D	1.009	0.859
5	P19388	RPB5	RNA Polymerase II subunit E	2.257	0.418
6	U3KPY1	RPB6	RNA Polymerase II subunit F	-4.169	0.374
7	P62487	RPB7	RNA Polymerase II subunit G	3.394	0.427
8	P52434	RPB8	RNA Polymerase II subunit H	2.407	0.220
9	P36954	RPB9	RNA Polymerase II subunit I	4.350	0.340
10	P62875	RPB10	RNA Polymerase II subunit L	7.651	0.111
11	P52435	RPB11	RNA Polymerase II subunit J	5.215	0.217
12	P53803	RPB12	RNA Polymerase II subunit K	5.023	0.267
	Integrator Complex				
13	Q8N201	INTS1	Integrator Complex Subunit 1	-1.025	0.662
14	Q9H0H0	INTS2	Integrator Complex Subunit 2	-1.192	0.690
15	Q68E01	INTS3	Integrator Complex Subunit 3	-1.337	0.566
16	Q96HW7	INTS4	Integrator Complex Subunit 4	-1.453	0.650
17	Q6P9B9	INTS5	Integrator Complex Subunit 5	-2.179	0.562
18	Q9UL03	INTS6	Integrator Complex Subunit 6	-1.605	0.554
19	Q9NVH2	INTS7	Integrator Complex Subunit 7	-2.794	0.443
20	Q75QN2	INTS8	Integrator Complex Subunit 8	-5.354	0.126
21	Q9NV88	INTS9	Integrator Complex Subunit 9	-3.784	0.251
22	Q5TA45	CPSF3L	Integrator Complex Subunit 11	-1.492	0.626
23	Q96CB8	INTS12	Integrator Complex Subunit 12	-0.082	0.985
	Mediator complex				

24	Q15648	MED1	Mediator Complex Subunit 1	-1.562	0.662
25	Q9NPJ6	MED4	Mediator Complex Subunit 4	-6.496	0.115
26	O75586	MED6	Mediator Complex Subunit 6	-2.498	0.514
27	O43513	MED7	Mediator Complex Subunit 7	-3.205	0.374
28	Q96G25	MED8	Mediator Complex Subunit 8	-4.227	0.387
29	Q9BTT4	MED10	Mediator Complex Subunit 10	-3.271	0.431
30	Q9P086	MED11	Mediator Complex Subunit 11	-5.924	0.138
31	O60244	MED14	Mediator Complex Subunit 14	-3.984	0.342
32	Q96RN5	MED15	Mediator Complex Subunit 15	-5.369	0.117
33	Q9Y2X0	MED16	Mediator Complex Subunit 16	-4.729	0.117
34	Q9NVC6	MED17	Mediator Complex Subunit 17	-5.181	0.137
35	Q9BUE0	MED18	Mediator Complex Subunit 18	-3.517	0.374
36	A0JLT2	MED19	Mediator Complex Subunit 19	-2.609	0.374
37	Q9H944	MED20	Mediator Complex Subunit 20	-1.875	0.689
38	Q9ULK4	MED23	Mediator Complex Subunit 23	-3.432	0.187
39	O75448	MED24	Mediator Complex Subunit 24	-3.665	0.299
40	O95402	MED26	Mediator Complex Subunit 26	-7.038	0.062
41	Q9H204	MED28	Mediator Complex Subunit 28	-6.301	0.139
42	Q9NX70	MED29	Mediator Complex Subunit 29	-3.028	0.374
43	Q96HR3	MED30	Mediator Complex Subunit 30	-0.086	0.986
44	Q9Y3C7	MED31	Mediator Complex Subunit 31	-3.914	0.374

Supplementary table 6: Conditions for chromatin immunoprecipitation experiments.

Protein	Antibody	Antibody quantity	Number of cells/ChIP	Protein G beads/ChIP
Pol II	ab9110	10 µg	25 x 10^6	100 µl
H3K4me3	ab8580	2 µg	5 x 10^6	20 µl
H3K27ac	ab4729	2 µg	5 x 10^6	20 µl

7. List of Abbreviations

A/Ala	alanine
CAN	Acetonitrile
AS	Antisense
Bp	basepair
ChIP-seq	Chromatin immunoprecipitation sequencing
CID	CTD-interacting domain
CTD	carboxy-terminal domain
CUT	Cryptic unstable transcript
DGE	Differential Gene Expression
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
DTT	Dithiothreitol
E/Glu	Glutamate
EBNA	Epstein-Barr virus nuclear antigen
EBV	Epstein Barr Virus
EC	Elongation complex
F/Phe	phenylalanine
FBS	Fetal bovine serum
GTF	General transcription factor
HA	haemagglutinin
Hg	Human genome
IAA	Indole-3-acetic acid
IGB	Integrated Genome Browser
IP	Immunoprecipitation
kDa	kilodalton
LC-MS	Liquid chromatography-mass spectrometry
mRNA	messenger RNA
MS	mass spectrometry
ncRNA	non-coding RNA
N_D	Number of dead cells
N_L	Number of living cells
Nt	nucleotide
N_V	Percentage of viable cells
PBS	phosphate buffer saline
PC/PCA	Principal Component/Analysis

PIC	pre-initiation complex
Pol I/II/III	DNA dependent RNA-Polymerase I/II/III
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RNAP	RNA Polymerase
rRNA	ribosomal RNA
RT	read-through
rWT	recombinant wild type
S	sense
S. cerevisiae	Saccharomyces cerevisiae
S. pombe	Saccharomyces pombe
S/Ser	serine
SDS	sodium dodecylsulphate
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
T/Thr	threonine
TFA	Trifluoroacetic acid
tRNA	Transfer RNA
TSS	transcription start site
Y/Tyr	tyrosine

8. Appendix

A. Publications

- **Shah, N.,** Maqbool, M.A., Yahia, Y., Aabidine, A.Z.E., Forne, I., Decker, T.M., Martin, D., Schuller, R., Krebs, S., Blum, H., Imhof, A., Eick, D., and Andrau, J.C. (2017). Tyrosine-1 of RNA polymerase II CTD controls global termination of gene transcription in mammals (In preparation).

- Schuller, R., Forne, I., Straub, T., Schreieck, A., Texier, Y., **Shah, N.,** Decker, T.M., Cramer, P., Imhof, A., and Eick, D. (2016). Heptad-Specific Phosphorylation of RNA Polymerase II CTD. Molecular cell *61*, 305-314.

- Decker, T.M., Kluge, M., Krebs, S., **Shah, N.,** Blum, H., Friedel, C.C., and Eick, D. (2017). Transcriptome analysis of dominant-negative Brd4 mutants identifies Brd4-specific target genes of BET inhibitor JQ1. Scientific reports (In press).

B. Oral and Poster presentation

2014	Poster: 11th EMBL conference: Transcription and Chromatin, Heidelberg, Germany
2015	Poster: Munich Interact, Germany
2015	Talk: 6th Munich Chromatin day, Germany
2016	Poster: 12th EMBL conference: Transcription and Chromatin, Heidelberg, Germany
2016	Poster: Munich Chromatin dynamic symposium, Germany
2017	Poster: Chromatin and epigenetics: from mechanism to function, Munich, Germany

C. Acknowledgements

Here I would like to extend my thanks to all the people who supported me during my time as a PhD student.

Foremost I would like to thank my supervisor Prof. Dr. Dirk Eick for accepting me as his student and for his continuous support, teaching, patience and inspiration. His immense knowledge and guidance has helped me during all the time of my research.

Current and past members of the Eick lab: Dr. Corinna Hintermeir, Dr. Michaela Rohrmoser, Dr. Kirsten Voss, Dr. Kaspar Burger, Dr. Yves Texier, Dr. Martin Heidemann, Dr. Markus Kellner and Anita Gruber-Eber for the great lab atmosphere, scientific discussions and all the fun we had in and outside the lab. I would specially like to thank Tim Decker, who over the years became my great friend. It has always been a pleasure discussing about different aspects of life, culture, sports, science, and of course the game of thrones with you. I would also like to thank 'My dear friend', Dr. Roland Schüller for introducing me to this project.

My scientific collaborators, Dr. Jean-Christophe Andrau, Dr. Muhammad Ahmad Maqbool, Yousra Yahia, Dr. Ignasi Forne and Prof. Dr. Axel Imhof for the successful collaboration and fruitful discussions during this project. I would like to extend my special acknowledgement to Yousra for making my stay in Montpellier a pleasant one.

My thesis committee members, Prof. Dr. Wolfgang Hammerschmidt and Prof. Dr. Axel Imhof for all their invaluable scientific advices and suggestions during my research.

To my friends in Germany and back home in India, thank you for your thoughts, good wishes, emails, phone calls and being there whenever I needed a friend.

I would like to express my deepest gratitude to my parents and my brother for the unconditional love and support they bestowed upon me. I would also like to acknowledge my In-laws for their faith in me. Last but not least, the best thing to happen in my life in the last 4 years is finding my best friend, soulmate and wife, Jaini. Thank you for being a pillar of strength for me and for all your love, support and encouragement.