REGULATION OF HUMAN RNA POLYMERASE II CTD MODIFICATIONS



Olga Kuznetsova

Brasenose College

University of Oxford

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy at the University of Oxford, Michaelmas Term 2015

Abstract

Transcription of human protein-coding genes and most small nuclear RNA genes is mediated by RNA Polymerase II (Pol II). During a cycle of transcription, Pol II recruits a variety of factors that facilitate transcription elongation, RNA processing and termination, through its long, unstructured C-terminal domain (CTD). The CTD in humans comprises 52 tandem heptapeptide repeats with the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$. Each amino acid of the heptapeptide can be chemically modified, which influences the recruitment of other protein factors to the transcription machinery. Not all enzymes that modify the CTD have been discovered. Recent studies have identified a novel CTD phosphatase: RPAP2 in humans and its yeast homologue Rtr1, which dephosphorylate phospho-Ser5 of the heptapeptide repeats. RPAP2 has been shown to stimulate 3' end cleavage of nascent snRNAs through recruitment of the Integrator complex, and unpublished work suggests the involvement of RPAP2 in regulating vertebrate developmental programs. However, the exact mechanisms that regulate the function of human RPAP2, and thus impact on CTD modification, are not well-understood.

This thesis presents a novel mechanism whereby RPAP2 recruits protein phosphatase 1 (PP1) to snRNA genes, where PP1 is postulated to activate P-TEFb to phosphorylate Ser2 of the CTD. At the same time, P-TEFb may have a role in activating the phosphatase activity of RPAP2. Furthermore, RPAP2 itself is shown to be recruited to a number of gene promoters by the RPRD1A protein, which also stimulates its phosphatase activity. RPAP2 was shown to have another role in regulating transcription termination: by recruiting the Integrator complex, which is shown here to mediate termination of snRNA genes, and by a so far unknown mechanism on a long protein-coding gene. An attempt was made to purify and crystallise the human RPAP2 to obtain a crystal structure, however the crystallisation trials were not successful. Finally, a correlation was found in human embryonic stem cells and induced pluripotent stem cells between low levels of RPAP2 and high levels of CTD Ser5P, suggesting a potential involvement of RPAP2 in regulating transcription at a key developmental stage.

The results presented here contribute to the understanding of human transcriptional mechanisms and the numerous interactions within the transcription machinery. In particular, the mechanism of terminating the transcription of snRNA genes is identified. An interesting possibility is the regulation of development and stem cell differentiation by RPAP2; however the exact pathways by which this occurs are yet to be discovered.

Acknowledgements

Firstly, I would like to express my gratitude to Prof Shona Murphy for accepting me into her group and for making this project possible. She has a profound understanding of both the joys and challenges of scientific research, and her support, patience, and guidance have been invaluable. Moreover, I am thankful to all the members of the "Big Mamas Lab", past and present: Justyna Zaborowska, Dawn O'Reilly, Clélia Laitem, Pilar Vazquez, Hadeel Al-Rawaf, Martin Dienstbier, Firdaus Isa, Alice Taylor, Alice Stanley, Joana Guiro, Isabel Instone, Harry Fischl and Johann Kufs, for their mentorship, company, laughter, and almost family-like support when experiments weren't working well.

Chapter 5 of the thesis would not have been possible without the help of Dr Sally Cowley and her colleagues in the James Martin Stem Cell Facility. I would particularly like to thank Cathy Browne for the practical help with growing and differentiating stem cells, and Sally Cowley for valuable discussions and for helping to keep my project on track. The idea to start the stem cell work in the first place originated following discussions with Dr Amber Mosley about the potential role of RPAP2 in development.

The work described in Chapter 3 was carried out in collaboration with the group of Prof Jack Greenblatt; in particular Dr Zuyao Ni, who invited me to take part in the project and provided siRNAs for the experiments. I would also like to thank Prof Dirk Eick for providing phospho-CTD antibodies, and Dr Sylvain Egloff for the CTD-RPAP2 plasmids, and for discussions and advice after I took over the RPAP2 project from him.

Prof Elspeth Garman gave me the possibility to carry out the structural biology work (Chapter 4 of the thesis) with support from members of her group. Special thanks extend to Dr Areej Abuhammad and Dr Edward Lowe for the technical support with growing crystals and for performing the diffraction experiments. Not only has Prof Garman assisted me with this project, treating me as a full member of her lab, but she has been a mentor to me since the summer of 2009, when I carried out my first summer project with her. To me, she is a perfect example of how to combine a successful scientific career, inspirational teaching, and a family and social life. Another great mentor is Dr Grant Churchill, with whom I carried out my Master's project. This experience was instrumental in my choice to pursue a PhD, and importantly, Grant and his lab members have shown me that science can, and should, be fun.

On a more personal note, I am very fortunate to have a number of amazing friends that have been with me through thick and thin and stood by me even when they understood nothing about my work.

And finally, none of this would have been possible without my family, who have given me so many opportunities in life, and through their experience of working in science were able to give me wise advice at every step of the way.

Table of contents

Abstract ..... **i**
Acknowledgements ..... **ii**
List of figures ..... **vii**
List of tables ..... **viii**
Abbreviations ..... **ix**
Chapter 1| Introduction **1**
1.1 Transcription of human genes **2**
1.1.1 RNA Polymerase II structure and function 2
1.1.2 Overview of the transcription cycle 3
1.1.3 RNA processing 7
1.1.4 β-actin gene structure and transcription 10
1.1.5 snRNA gene structure and transcription 11
1.1.6 Role of the Integrator complex 14
1.2 Overview of the Pol II CTD code **16**
1.2.1 Pol II CTD structure in yeast and human 16
1.2.2 Transition from initiation to elongation 18
1.2.3 Role of Ser7P 18
1.2.4 Roles of Thr4P and Tyr1P 20
1.2.5 Effects of proline isomerisation 21
1.2.6 Serine and threonine glycosylation 22
1.2.7 Gene-specific CTD modifications 23
1.3 Role of CTD-modifying enzymes in the transcription of human genes **23**
1.3.1 Overview of CTD-modifying enzymes 23
1.3.2 RPAP2 is a Ser5P phosphatase 28
1.3.3 Role of RPAP2 in snRNA 3’ end processing 29
1.3.4 RPAP2 and Rtr1 interaction partners 30
1.3.5 RPAP2 domains and modifications 33
1.3.6 Structural studies of RPAP2 35
1.4 CTD modifications in stem cells and during development **38**
1.4.1 Rtr1 mutations in zebrafish 38
1.4.2 Roles and regulation of HOX genes in embryonic development 39

1.4.3 Pol II CTD modifications on *HOX* genes in stem cells 41
1.5 Aims of the thesis 42
Chapter 2| Regulation of Pol II function by Pol II-associated factors 44
2.1 Introduction 45
2.2 Results 47
2.2.1 Recruitment of the PP1 phosphatase by RPAP2 47
2.2.2 Role of the Integrator complex in transcription termination 51
2.2.3 RPAP2 regulates transcription termination of a protein-coding gene 57
2.3 Conclusions 59
Chapter 3| Regulation of RPAP2 activity 61
3.1 Introduction 62
3.2 Results 64
3.2.1 The phosphatase activity of RPAP2 may be regulated by P-TEFb 64
3.2.2 RPRD1A is present on several human genes 66
3.2.3 Recruitment of RPAP2 to genes 68
3.2.4 RPRD1B does not impact on CTD Ser5 phosphorylation 70
3.3 Conclusions 72
Chapter 4| Structural and functional analysis of human RPAP2 74
4.1 Introduction 75
4.2 Results 76
4.2.1 Bioinformatic prediction of human RPAP2 domains and secondary structure 76
4.2.2 Expression and purification of GST-tagged human RPAP2 81
4.2.3 Crystallisation trials with RPAP2 RV 86
4.2.4 Functional analysis of RPAP2 RV 86
4.3 Conclusions 89
Chapter 5| Modification of the Pol II CTD during stem cell differentiation 91
5.1 Introduction 92
5.2 Results 94
5.2.1 CTD Ser5P is high in stem cells 94
5.2.2 Transfection of Myc-tagged RPAP2 into stem cells 97
5.3 Conclusions 98
Chapter 6| Discussion 100
6.1 Regulation of Pol II function by Pol II-associated factors (Chapter 2) 101

6.1.1 Recruitment of the PP1 phosphatase by RPAP2 101
6.1.2 Role of the Integrator complex in transcription termination 104
6.1.3 RPAP2 may regulate transcription termination of a protein-coding gene 105
6.1.4 Future directions 106
6.2 Regulation of RPAP2 activity (Chapter 3) 107
6.2.1 The phosphatase activity of RPAP2 may be regulated by P-TEFb 107
6.2.2 Role of the RPRD proteins in recruiting RPAP2 108
6.2.3 Future directions 110
6.3 Structural and functional analysis of human RPAP2 (Chapter 4) 111
6.3.1 Prediction of human RPAP2 secondary structure, and expression and purification 111
6.3.2 Functional analysis of RPAP2 RV 112
6.3.3 Future directions 112
6.4 Modification of the Pol II CTD during stem cell differentiation (Chapter 5) 113
6.4.1 Levels of Ser5P and CTD-modifying enzymes across cell types 113
6.4.2 Future directions 115
6.5 Conclusions 116
Chapter 7| Materials and Methods 117
7.1 Materials 118
7.1.1 Antibodies 118
7.1.2 Antibiotics 119
7.1.3 Buffers and solutions 119
7.1.4 Reagents 121
7.1.5 PCR Oligonucleotides: 122
Primers used in Chapter 2: 122
Primers used in Chapter 3: 123
7.1.6 Bacteria and cell lines 124
7.1.7 Bacterial and tissue culture media 124
7.2 Methods 126
DNA-related methods 126
7.2.1 Real-time quantitative Polymerase Chain Reaction (qPCR) 126
7.2.2 DNA sequencing 127
7.2.3 Protein expression 127

RNA-related methods 127
7.2.4 Purification of total RNA 127
7.2.5 RNAi-mediated knockdown 128
7.2.6 RNase protection assay 128
Protein-related methods 129
7.2.7 Cell Culture 129
7.2.8 Stem cell differentiation 130
7.2.9 Freezing 131
7.2.10 Transfection into HeLa cells and iPS cells 131
7.2.11 ChIP analysis 132
7.2.12 SDS-polyacrylamide gel electrophoresis (SDS-PAGE) 134
7.2.13 Coomassie staining 134
7.2.14 Western blot 134
7.2.15 GST fusion protein purification 135
7.2.16 Ion exchange protein purification 136
7.2.17 Differential scanning calorimetry 136
7.2.18 Crystallisation trials 137
7.2.19 Phosphatase assay 137
Bacterial methods 138
7.2.20 Transformation of competent bacterial cells 138

List of figures

Figure 1-1. Organisation of the eukaryotic pre-initiation complex. Adapted from (Krishnamurthy and Hampsey, 2009) 4
Figure 1-2. Model of the early elongation checkpoint. 6
Figure 1-3. Organisation of the human β-actin gene. 11
Figure 1-4. Structure of a U snRNA gene. U1 and U2 genes are used as an example. 13
Figure 1-5. CTD serine phosphorylation patterns during one round of a transcription cycle. 18
Figure 2-1. PPP1CA recruitment to the U2 snRNA is regulated by RPAP2. 49
Figure 2-2. Levels of Ser5P and Ser2P on the U2 gene after PP1 knockdown. 50
Figure 2-3. Effect of Integrator knockdown on the transcription of the β-actin gene. 53
Figure 2-4. Effect of Integrator knockdown on the transcription of the U1 and U2 genes. 54
Figure 2-5. RPA on the U2 gene following Integrator knockdown. 56
Figure 2-6. ChIP of Tyr1P on the KPNB1 gene after RPAP2 knockdown. 58
Figure 3-1. The effect of P-TEFb on the phosphatase activity of RPAP2. 65
Figure 3-2. ChIP of Pol II and RPRD1A on a range of mRNA and U2 snRNA genes. 68
Figure 3-3. RPRD1A recruitment to genes by RPAP2. 69
Figure 3-4. RPRD1B does not affect Ser5P levels. 71
Figure 4-1. Prediction of disorder from the human RPAP2 amino acid sequence. 78
Figure 4-2. Prediction of the human RPAP2 secondary structure by JPred. 80
Figure 4-3. Coomassie stain and Western blot of GST-RPAP2 RV purification. 82
Figure 4-4. Western blots of RPAP2 and GST after MonoQ purification. 84
Figure 4-5. Differential scanning calorimetry spectrum of RPAP2 RV. 85
Figure 4-6. X-ray diffraction pattern of the RPAP2 RV sample. 87
Figure 4-7. Phosphatase assay using recombinant RPAP2 RV on a GST-CTD substrate. 88
Figure 5-1. Levels of CTD modifications and CTD-modifying enzymes in ES cells. 95
Figure 5-2. Western blots of Pol II, CTD Ser5P, and RPAP2 across a range of stem and cancer cell lines. 96
Figure 5-3. Transfection of Myc-RPAP2 into iPS cells and HeLa cells. 97
Figure 6-1. Proposed sequential model for the function of PP1 on the U2 snRNA gene. 103

List of tables

Table 1-1. List of known CTD-modifying enzymes in mammals and yeast. 24

Abbreviations

3’ ss	3’ splice site
3C	chromatin conformation capture
5’ss	5’ splice site
bp	base pairs
BRE	TFIIB recognition element
BSA	bovine serum albumin
CBC	cap-binding complex
ChIP	chromatin immunoprecipitation
CID	CTD-interacting domain
CPSF	cleavage and polyadenylation specificity factor
CstF	cleavage stimulation factor
CTD	C-terminal domain
DNA	deoxyribonucleic acid
dNTP	deoxynucleotide triphosphate
DPE	downstream core promoter element
DRB	5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole
DSC	differential scanning calorimetry
DSE	downstream sequence element
DSIF	DRB sensitivity-inducing factor
DUF	domain of unknown function
EB	embryoid body
EDTA	ethylenediaminetetraacetic acid
EGF	epidermal growth factor
ESC	embryonic stem cell
GMP	guanosine monophosphate
GSH	reduced glutathione
GST	glutathione-S-transferase
GTF	general transcription factors
h	hour

HIV-1	human immunodeficiency virus type 1
HOX	homeobox
Int	Integrator subunit
iPSC	induced pluripotent stem cell
IPTG	isopropyl β-D-1-thiogalactopyranoside
kb	kilobases
kDa	kiloDaltons
KM	KM05283
M	molar
min	minutes
miRNA	microRNA
MLL	mixed lineage leukemia
mol	moles
mRNA	messenger RNA
MTE	motif ten element
MudPIT	multidimensional protein identification technology
NELF	negative elongation factor
NMR	nuclear magnetic resonance
nt	nucleotides
O/N	overnight
PAF_c	polymerase-association-factor complex
PcG	Polycomb group protein
PCR	polymerase chain reaction
PDB	Protein Data Bank
Pol I/II/III	RNA Polymerase I/II/III
PP1	protein phosphatase 1
PSE	promoter sequence element
P-TEFb	positive transcription elongation factor
qPCR	quantitative PCR
RA	read-around
RA	retinoic acid

RARE	retinoic acid response element
RNA	ribonucleic acid
RPA	RNase protection assay
RPAP2	RNA Polymerase II-associated protein 2
rRNA	ribosomal RNA
RT	read-through
RT	room temperature
Rtr1	regulator of transcription 1
SAGA	Spt-Ada-Gcn5-acetyltransferase
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
sec	seconds
SEC	Super Elongation Complex
SF1	splicing factor 1
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleic protein
TAF	TBP-associated factor
TBP	TATA-binding protein
TFII	transcription factor II
T_m	melting temperature
tRNA	transfer RNA
TSS	transcription start site
U2AF	U2 auxiliary factor

Chapter 1| Introduction

1.1 Transcription of human genes

1.1.1 RNA Polymerase II structure and function

All mammalian processes, from development to responses caused by environmental stimuli to metabolism, rely on the correct and timely expression of tens of thousands of genes. The first step of gene expression is transcription: the process of reading the DNA template and synthesising a complementary RNA copy. This is achieved by the action of an enzyme known as RNA polymerase. In eukaryotes, three RNA polymerases transcribe distinct sets of genes: RNA polymerase I (Pol I) transcribes ribosomal RNA (rRNA) genes, in particular, the 28S, 18S and 5.8S subunits from the 47S precursor (Russell and Zomerdijk, 2006). RNA polymerase II (Pol II) transcribes protein-coding genes, and also some small nuclear RNA (snRNA) genes and micro RNA (miRNA) (Egloff and Murphy, 2008, Lee et al., 2004). Finally, RNA polymerase III (Pol III) is responsible for the transcription of transfer RNAs (tRNAs), the 5S rRNA and some snRNAs (White, 2011, Roeder and Rutter, 1969). This thesis is concerned with transcription by RNA Pol II.

Human Pol II is a 12-subunit complex of over 0.5 MDa, composed of the proteins Rpb1-Rpb12 (Myer and Young, 1998). The structure of the ten-subunit core (lacking Rpb4 and Rpb7) of the budding yeast Pol II has been solved by X-ray crystallography to 2.8 Å resolution. It shows the main body of the enzyme formed by Rpb1 and Rpb2 and a mobile 'clamp', formed by the N- and C-terminal domains of Rpb1 and the C-terminal domain of Rpb2, that allows DNA entry and unwinding (Cramer et al., 2000, Cramer et al., 2001). The largest subunit, Rpb1, contains a long and unstructured C-terminal domain (CTD) that through its post-translational modifications recruits a wide variety of cofactors during the transcription cycle and regulates diverse processes (Egloff and Murphy, 2008). The Pol II

CTD will be described in more detail below. Generally, from here onwards any mention of yeast refers to *Saccharomyces cerevisiae*, unless stated otherwise.

Pol II on its own is unable to recognise promoter sequences, and promoter binding and transcription initiation is made possible by a host of accessory factors that form a complex together with Pol II (Matsui et al., 1980). These are known as the General Transcription Factors (GTFs), and comprise TFIIA, TFIID, TFIIB, TFIIF, TFIIE, and TFIIH. These factors together with Pol II make up the pre-initiation complex (Landeira et al.), which assembles at the promoter in a stepwise manner and unwinds the double-stranded DNA (see Figure 1-1) (Grunberg and Hahn, 2013, Myer and Young, 1998, Liu et al., 2013). A more detailed description of the transcription cycle is presented below.

1.1.2 Overview of the transcription cycle

The first stage of transcription in higher eukaryotes is usually the sequence-specific binding of transcription factors to an enhancer element, often located at a considerable distance from the gene it regulates. Recent technological advances including next-generation sequencing and chromatin conformation capture (3C) have revealed extensive chromatin looping that connects enhancers to gene promoters and allows assembly of GTFs (Daniel et al., 2014).

GTFs bind to conserved regions within the promoter, known as the core promoter elements (Shandilya and Roberts, 2012). These include the TATA box, Initiator, TFIIB recognition elements (BRE), the downstream core promoter element (DPE) in higher eukaryotes, and the motif ten element (MTE). The most ancient and evolutionarily-

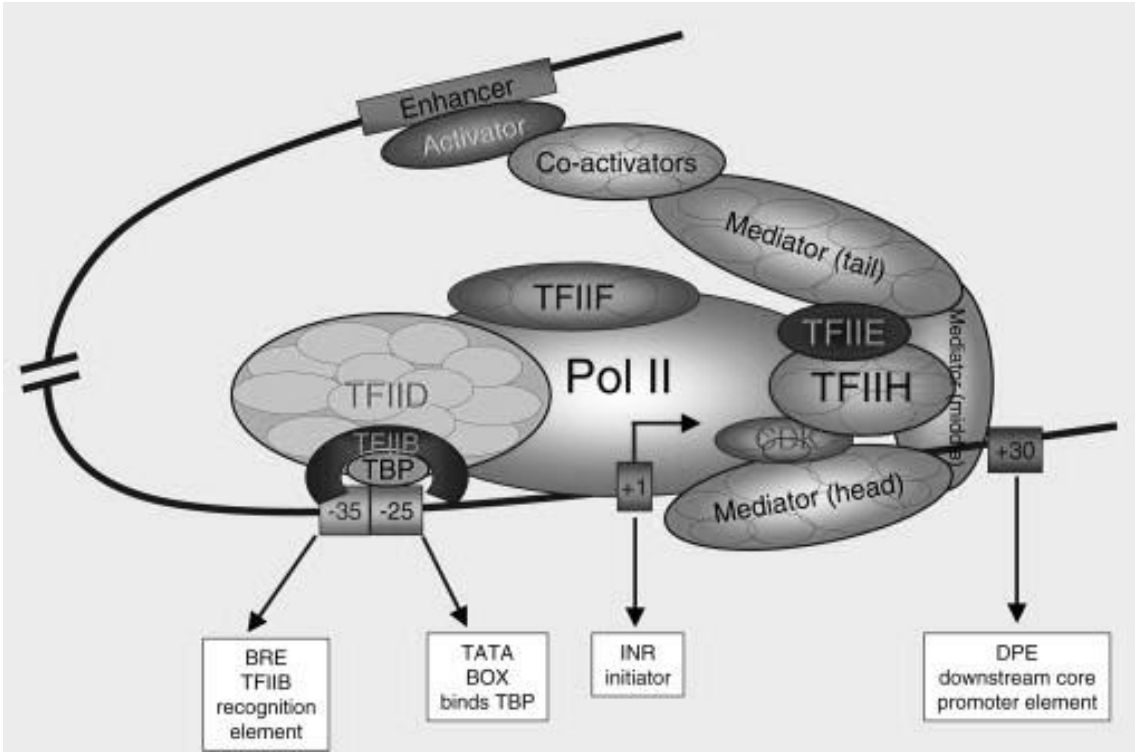


Figure 1-1. Organisation of the eukaryotic pre-initiation complex.
Adapted from (Krishnamurthy and Hampsey, 2009)

conserved promoter element is the TATA box, which has the sequence TATAWAAR and is located around 30 bp upstream of the transcription start site (TSS) (Juven-Gershon et al., 2008, Carninci et al., 2006). It is recognised by the TATA-binding protein (TBP) of the TFIID complex, while TFIIB binds flanking sequences known as BREs (Deng et al., 2009). TFIID comprises TBP and a set of TBP-associated factors (TAFs), and is the first GTF to bind to the promoter and stimulate PIC formation. The structure of human TFIID has recently been solved to 11.6 Å resolution by cryo-electron microscopy (Bieniossek et al., 2013). Interestingly, the majority of gene promoters in humans and *S. cerevisiae* lack a consensus TATA box, and are therefore referred to as 'TATA-less'**.** When TBP is part of the TFIID complex, it preferentially binds to these TATA-less promoters, despite its name

suggesting otherwise (Basehoar et al., 2004, Yang et al., 2007). The TATA-containing promoters are instead bound by TBP as part of the Spt-Ada-Gcn5-acetyltransferase (SAGA) complex, which contains several of the same subunits as TFIID and is in some cases functionally redundant (Bhaumik and Green, 2002, Lee et al., 2000, Rhee and Pugh, 2012). Thus, promoter recognition is driven by a variety of DNA elements, which may vary between genes and therefore require different GTFs for binding.

After the opening of the DNA double helix by TFIIB, the rest of the GTFs and the polymerase assemble. In particular, TFIIH is required to melt the double-stranded DNA at the promoter to facilitate transcription initiation (Laine and Egly, 2006). The GTF assembly is further stimulated by the 1.2 MDa multisubunit Mediator complex, which also has additional roles in transcription elongation (Kim et al., 1994, Chen et al., 2012, Takahashi et al., 2011). Transcription initiation usually first entails the production of short, 'abortive' transcripts up to 15 nucleotides long, until the polymerase clears the promoter and enters productive elongation, which is marked by a more stable association of Pol II with the template (Saunders et al., 2006, Kugel and Goodrich, 2000).

Before Pol II begins transcribing the body of the gene, it is subject to a promoter-proximal pause approximately 25 nt downstream of the TSS, which serves as an important checkpoint in transcription regulation (Jonkers et al., 2014). The pause is mediated by Pol II associating with two factors, negative elongation factor NELF and the DRB sensitivity-inducing factor DSIF (Wada et al., 1998a, Yamaguchi et al., 1999). DSIF is composed of the proteins Spt4 and Spt5 that have homologues in yeast, while NELF is a four-subunit complex of proteins that has no yeast homologues. During the pause, Spt5 interacts with the capping enzyme, which adds the 7-methylguanylate cap to the 5' end of the nascent

RNA (Wen and Shatkin, 1999, Lidschreiber et al., 2013). To release the pause, the Cdk9 subunit of the positive transcription elongation factor P-TEFb phosphorylates Spt5 and NELF, which causes the dissociation of NELF and converts DSIF into a positive elongation factor (see Figure 1-2). More recently however, it has been shown using the GRO-seq (global run-on sequencing) method and Cdk9 inhibitors that the early elongation checkpoint is associated with a Pol II pause only on a subset of genes. On other genes, the polymerase reads through the gene without pausing, but its elongation is nevertheless dependent on P-TEFb activity (Jonkers et al., 2014, Laitem et al., 2015). During

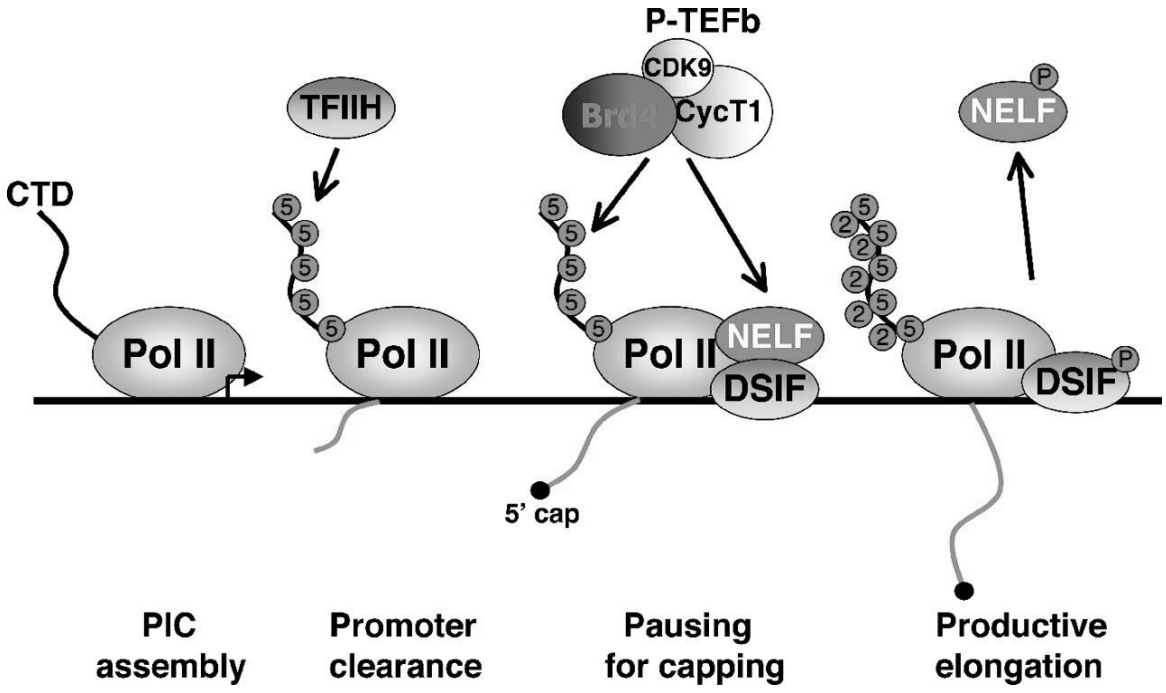


Figure 1-2. Model of the early elongation checkpoint.
The four stages of early transcription are depicted, from left to right: 1. transcription initiation, 2. phosphorylation of the Pol II CTD on Ser5 by TFIIH, 3. Pol II pause mediated by NELF and DSIF, and recruitment of P-TEFb and Brd4, and 4. release of the pause through phosphorylation of NELF, DSIF and Ser2 of the CTD. Adapted from (Zhou and Yik, 2006).

elongation, P-TEFb also phosphorylates Pol II on the CTD, further stimulating the recruitment of RNA processing and termination factors (Shandilya and Roberts, 2012, Lis et al., 2000, Chiba et al., 2010). Thus, sequential recruitment of positive and negative elongation factors regulates the early elongation checkpoint, Pol II pausing, and pause release.

Although previously it was thought that the majority of transcription regulation occurs at initiation, recent evidence suggests that elongation is also a tightly-regulated process that requires a multitude of cofactors. One example, described above, is the promoter-proximal pause, which serves as a 'checkpoint' before productive elongation. Once Pol II has overcome the pause, its progression through the gene body is regulated by a variety of factors, together making up the Super Elongation Complex (SEC). This is composed of P-TEFb, eleven-nineteen Lys-rich leukaemia protein (ELL), AF9, AFF4 and other proteins (Luo et al., 2012). The SEC is recruited to Pol II through the polymerase-association- factor complex (PAF_C) (He et al., 2011). At the end of the transcription cycle, the polymerase pauses and eventually dissociates from the template to be recycled for the next round of transcription. During transcription of eukaryotic protein-coding genes, this is coupled to 3' cleavage and polyadenylation of the nascent RNA, which will now be described.

1.1.3 RNA processing

As a pre-mRNA copy is synthesised by Pol II, it undergoes a process known as maturation, or RNA processing. Processing of a pre-mRNA results in the formation of a mature mRNA that is ready to be translated, and involves 5' capping, intron splicing, and 3' cleavage and polyadenylation (Bentley, 2014, Zhou et al., 2012). Processing is coupled to transcription, as components of the transcription machinery recruit processing factors to the nascent

RNA. Transcription and processing factors also regulate each other's activities (Lenasi et al., 2011, Martins et al., 2011, Hirose et al., 1999, Rasmussen and Lis, 1993, Moteki and Price, 2002).

As the nascent transcript of approximately 20-30 nucleotides in length emerges from the Pol II core, the polymerase recruits the capping complex. Both Spt5 of DSIF and the C-terminal domain of Pol II, with a phosphorylation on Ser5, interact with the capping apparatus. The polymerase undergoes a DSIF-dependent promoter-proximal pause that serves as a 'checkpoint' to allow capping to occur (Wen and Shatkin, 1999), although most of the evidence for this mechanism has come from studies in yeast and *Drosophila*, rather than humans. The capping complex consists of the capping enzyme, guanine-7-methyltransferase, and the cap binding complex (CBC) (Glover-Cutter et al., 2008, Shuman, 2001). The reaction entails addition of GMP to a 5' diphosphate of the RNA, followed by methylation of the guanosine cap to form 7-methylguanosine, m7G(5')ppp(5')X (Shatkin, 1976). The cap protects the nascent RNA from degradation and mediates cytoplasmic export and translation (Lewis and Izaurralde, 1997).

During transcription elongation, non-coding introns within genes are subject to an excision process known as splicing. This is achieved by the spliceosome: a large complex comprising five small nuclear RNAs (snRNAs): U1, U2, U3, U4 and U6, and a plethora of proteins, together forming small nuclear ribonucleoproteins, snRNPs (Hoskins and Moore, 2012, Chen and Manley, 2009). The snRNPs recognise several sequences in the the intron through base pairing, and the spliceosome assembles in a stepwise manner. In the early (E) complex, U1 snRNP binds the 5' splice site (5' ss), while splicing factor 1 (SF1) binds to the branch point and U2 auxiliary factor (U2AF) binds to the polypyrimidine tract

and 3’ splice site (Brow, 2002). U2 snRNP then utilises ATP to bind to the branch point in place of SF1 and U2AF, forming the A complex. U4, U6 and U5 snSNPs are then recruited through interactions with exonic sequences and the U2 snRNP to form the B complex. U1 and U4 subsequently dissociate, and the exon in excised in two steps: first the phosphodiester bond at the 5’ ss is broken and joins to the branch point to form a lariat structure. Finally, the 5’ exon attacks the 3’ ss, ligating the two exons together and removing the intron (Matera and Wang, 2014).

The spliceosome can also produce a number of different mRNAs from one pre-mRNA through alternative splicing. In humans, almost every gene is alternatively spliced, leading to a large expansion of the proteome relative to the number of genes (Nilsen and Graveley, 2010). Like capping, splicing occurs co-transcriptionally and is regulated by the C-terminal domain of Pol II, which recruits snRNPs and accessory proteins (Mortillaro et al., 1996, Yuryev et al., 1996). Splicing factors can also regulate transcription, as demonstrated by elongation defects and P-TEFb reduction on genes in SC35 mutant cells (Lin et al., 2008).

At the end of the transcription cycle, the mRNA is cleaved and polyadenylated. The cleavage site within the gene is denoted by a conserved A(A/T)TAAA, followed by a GT-rich downstream sequence element (DSE) (Gil and Proudfoot, 1984). The process starts with the recognition of the AAUAAA sequence by the Cleavage and Polyadenylation Specificity Factor (CPSF), and the DSE by the Cleavage Stimulation Factor (CstF). Cleavage requires two additional proteins: Cleavage Factors I and II (CFI and CFII), and is mediated by the 73 kDa subunit of CPSF (CPSF73) (Millevoi and Vagner, 2010, Shi et al., 2009). Cleavage occurs 10-35 bases downstream of the A(A/U)UAAA sequence in the nascent

RNA, and the free 3' end is modified by the addition of 200-300 adenosines, catalysed by the polyA polymerase (Edmonds et al., 1971). In some cases, like the beta-actin gene, transcription terminates when the exonuclease Xrn2, recruited by the cleavage and polyadenylation factors, degrades the downstream pre-mRNA and causes dissociation of Pol II from the template (West et al., 2004).

Although cleavage, polyadenylation and termination occur at the end of the transcription cycle, they are influenced by events occurring earlier. For example, recruitment of Pcf11 occurs through the Pol II C-terminal domain phosphorylated on Ser2 (Meinhart and Cramer, 2004). Often, genes have more than one polyA site, and the usage of one site over the others depends on the level of conservation, or 'strength' of the site, as well as on the protein factors that mediate cleavage and polyadenylation (Di Giammartino et al., 2011).

The transcriptional processes presented above hold true for most protein-coding genes transcribed by Pol II. A separate class of genes, the small nuclear RNA genes (snRNA) genes are structured differently to mRNA genes, and require a separate set of cofactors for transcription. In most of the experiments presented in this thesis, transcription of an mRNA gene – usually β-actin – was compared to that of an snRNA gene. The subsequent sections of this chapter will give an overview of two main genes that were used as models for experiments presented in this thesis.

1.1.4 β-actin gene structure and transcription

The human β-actin gene (*ACTB*) is one of the six isoforms within the actin family that are encoded on chromosome 7. It is a cytoskeletal protein implicated in organelle transport

within the cell, and in cell motility (Ng et al., 1985, Perrin and Ervasti, 2010). The β-actin gene is over 3 kb long and contains six exons (Figure 1-3) (Nakajima-Iijima et al., 1985).

The transcription of the β-actin gene has been well-described. The strong promoter contains both a CCAAT box and a TATA box (Nakajima-Iijima et al., 1985). Transcription terminates by Pol II pausing at an MAZ_4 site downstream of the polyA signal, and the action of Xrn2 to degrade the nascent mRNA after the cleavage event. The transcript is then polyadenylated (Gromak et al., 2006). In the present work, β-actin was used as a model to study transcription of protein-coding genes. The other model gene type analysed was an snRNA gene, described below.

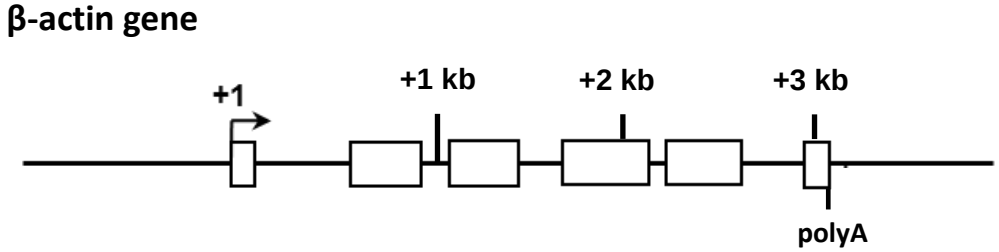


Figure 1-3. Organisation of the human β-actin gene.
The arrow denotes the TSS, while exons are represented by boxes. There is a cleavage and polyadenylation signal at the end of the last exon (pA).

1.1.5 snRNA gene structure and transcription

So far the discussion has focused on the transcription of mRNA genes and processing of the transcripts, however this is not the only gene class transcribed by Pol II. Small nuclear RNA genes (snRNA), except U6 and the component of the minor spliceosome, U6atac, are also transcribed by Pol II (Medlin et al., 2003, Egloff and Murphy, 2008, Turunen et al.,

2013). snRNAs are required for pre-mRNA splicing, and are concentrated within Cajal bodies in the nucleus: a sub-class of nuclear bodies that have a role in the assembly of the spliceosome and RNA processing (Dundr, 2012, Novotny et al., 2011, Lamond and Carmo-Fonseca, 1993, Matera and Ward, 1993).

The most extensively-studied snRNA genes are U1 and U2, and they will be used to illustrate the principles involved. In humans, U2 genes are arranged in 6.1 kb tandem repeats in chromosome 17, while U1 genes are also clustered in repeat units on chromosome 1 (Lindgren et al., 1985a, Lindgren et al., 1985b). Furthermore, U1 has a number of pseudogenes with varying degrees of homology to the 'true' U1 coding region and flanking sites (Lindgren et al., 1985b). Until recently, the pseudogenes were thought to be non-functional, however a recent study showed that they are in fact expressed, and inhibition of a specific U1 pseudogene (or variant) leads to large-scale transcriptome changes (O'Reilly et al., 2013).

In terms of gene structure, snRNA genes are much shorter than mRNA genes, and do not contain introns (Figure 1-4). The U1 and U2 promoters contain an essential proximal sequence element (PSE), and a distal sequence element (DSE) that serves as an enhancer, approximately 50-70 bp and 250 bp upstream of the TSS, respectively (reviewed in (Hernandez, 2001)). Recognition of the PSE requires a different complex to that which binds the TATA box on mRNA genes: a major component is PTF (also named $SNAP_c$) comprising five subunits, as well as components of the general transcription machinery: TBP, TFIIA, TFIIB, TFIIF, TFIIE and TFIIH and Pol II itself (Henry et al., 1998, Kuhlman et al., 1999) . Furthermore, the set of TAFs found associated with TBP differs from that found on mRNA genes, with TAF1, TAF4 and TAF10 absent from snRNA genes

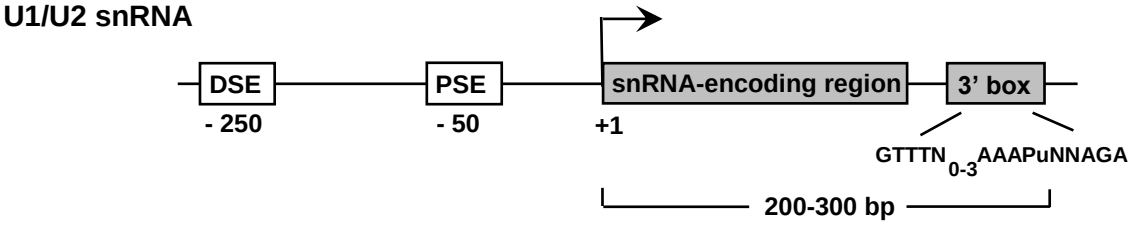


Figure 1-4. Structure of a U snRNA gene.
U1 and U2 genes are used as an example.

(Zaborowska et al., 2012). Interestingly, P-TEFb is not required for the elongation of transcription of snRNA genes, even though it is still necessary for the 3' end processing (Medlin et al., 2005). Differences in the general transcription machinery and elongation factors in the expression of mRNA and snRNA genes probably reflect differences in the structures of the genes, and thus different requirements for regulation of transcription and RNA processing.

Another feature of the U snRNA genes is a 3' processing element called the 3' box, which is distinct from the polyA signal in mRNA genes and is located 9-19 nucleotides downstream of the coding region (Hernandez, 1985, Yuo et al., 1985). Unlike mRNA genes, correct 3' end processing of snRNA genes requires compatible promoter elements, and promoters of mRNA genes cannot substitute for the snRNA PSE and DSE (Hernandez and Weiner, 1986, Hernandez and Lucito, 1988, de Vegvar et al., 1986). Both transcription of snRNA genes and 3' processing of the transcripts depend on the function of the Pol II C-terminal domain, and it was proposed that the C-terminal domain is required for 3' box recognition (Medlin et al., 2003, Jacobs et al., 2004). Despite this, 3'

processing is tolerant of 3’ box mutations, since it is also dependent on the presence of a 3’ stem-loop (Ezzeddine et al., 2011, Cuello et al., 1999). 3’ end processing requires both RNA and protein co-factors, however no binding of the mRNA cleavage and polyadenylation machinery has been observed (Uguen and Murphy, 2004). Recently, a new, snRNA-specific complex was described: Integrator, so far only found in metazoans, which directs site-specific pre-snRNA 3’ cleavage in a similar manner to that of CPSF (Egloff et al., 2010, Baillat et al., 2005). The function of Integrator will be discussed in more detail below.

Although the coding region of snRNA genes is only 200-300 base pairs, transcription of both U1 and U2 genes continues beyond the 3’ end of the gene, with nuclear run-on experiments showing that the full transcript is approximately 1 kb in length (Cuello et al., 1999, O'Reilly et al., 2013, James Faresse et al., 2012). After the pre-mRNA has been synthesized and processed, it is exported to the cytoplasm, where it is trimmed from the 3’ end and the 5’ cap is further methylated. The snRNA then assembles with Sm proteins, and this structure is re-imported back into the nucleus for further maturation (Matera and Wang, 2014).

1.1.6 Role of the Integrator complex

The first description of Integrator came from the discovery of two homologues of the human cleavage and polyadenylation factors CPSF73 and CPSF100, named RC68 and RC74, respectively (Dominski et al., 2005). The two proteins form a complex, with the C-terminus of RC68 mediating this interaction. Furthermore, RC68 contains the histidine motif typical of the β-lactamase superfamily and was proposed to possess cleavage

activity similar to CPSF73. Knockdown of RC68 resulted in a G_1 cell cycle arrest in HeLa cells.

At the same time, the full Integrator complex was purified as a component of a larger complex containing the protein named “deleted in split hand/foot 1” (DSS1), BRCA2 and Pol II (Baillat et al., 2005). Integrator contains at least 12 polypeptides in human, most of which do not have known homology with other proteins. However, RC68 and RC74 (Integrator subunits 11 and 9, respectively) are homologous with CPSF subunits, and Int6, Int7 and Int12 also contain known domains. Baillat and colleagues also demonstrated the role of Integrator in processing U1 and U2 pre-sRNAs, as RNAi knockdown of the catalytic subunit Int11 leads to a build-up of unprocessed pre-snRNA.

To associate with Pol II, Integrator requires a double mark on the Pol II C-terminal domain, with phosphorylations on Ser2 and Ser7 on two consecutive heptapeptide repeats (Egloff et al., 2010). In order for Int11 to cleave the 3’ end of the pre-snRNA, it requires a physical interaction with Int9. Although Int9 itself is devoid of catalytic activity, it appears to be an accessory protein that stimulates Int11 (Albrecht and Wagner, 2012). 3’ processing also requires association of Integrator with the mRNA elongation factors NELF and DSIF at the 3’ end of the gene (Yamamoto et al., 2014). In contrast, yeast does not possess Integrator homologues, and snRNA processing occurs by an independent pathway involving the RNase III homologue Rnt1 (Seipelt et al., 1999, Abou Elela and Ares, 1998, Chanfreau et al., 1997), while termination requires the Nrd1/Nab3/Sen1 complex (Steinmetz et al., 2001).

In addition to their well-described role in snRNA processing, several Integrator subunits have also been found in an mRNA processing complex (Shi et al., 2009). Recently, Gardini and colleagues have shown a cleavage-independent function for Integrator on mRNA genes. Integrator is recruited to epidermal growth factor (EGF)-responsive immediate early genes on which Pol II is paused near the promoter. Upon EGF stimulation, Integrator both activates transcription initiation and recruits P-TEFb as part of the super-elongation complex to release Pol II from the pause (Gardini et al., 2014). The Integrator complex therefore possesses several functions that are gene type-specific. Furthermore, Integrator is shown to associate with NELF, Spt5 and Pol II to mediate promoter-proximal pausing and pause release on some protein-coding genes, for example at the HIV-1 long terminal repeat (Stadelmayer et al., 2014).

It has been mentioned in several instances that Pol II regulates a variety of transcriptional processes, including Integrator recruitment, through its C-terminal domain. In the next section the C-terminal domain and its function will be discussed in more detail.

1.2 Overview of the Pol II CTD code

1.2.1 Pol II CTD structure in yeast and human

As mentioned previously, the large subunit of RNA Pol II, Rbp1, contains a long unstructured C-terminal domain, henceforth referred to as the CTD. The CTD comprises tandem repeats of a heptapeptide with the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$, which is proposed to have evolved by amplification of a short sequence. The number of repeats varies between organisms and reflects the organism's complexity, with the human CTD

containing 52 such repeats, while yeast contains only 26 (Chapman et al., 2008, Liu et al., 2010).

Although not necessary for gene transcription *in vitro*, deletion of the CTD in both yeast and mammals is lethal, highlighting its essential function in cell growth (Allison et al., 1988, Bartolomei et al., 1988). Moreover, the functional unit of the CTD is a diheptad, rather than a single heptad repeat (Stiller and Cook, 2004). Each residue of the repeat can be reversibly modified, most commonly by phosphorylation or proline isomerisation (Egloff et al., 2012a). The CTD undergoes phosphorylation at the beginning of the transcription cycle, converting the RNAP IIA form into RNAP IIO, followed by dephosphorylation at the end of the cycle. The phosphorylated CTD serves as a platform for the assembly of a variety of factors that mediate gene transcription and RNA processing (Hsin and Manley, 2012, Heidemann et al., 2013). This changing pattern of modifications during the transcription cycle led to the idea of a 'CTD code': that CTD modifications and combinations thereof give rise to distinct processes by engaging different parts of the transcription machinery (Buratowski, 2003).

Although the YSPTSPS sequence is the most commonly-found among CTD repeats, it displays some variation particularly towards the C-terminal end of the CTD in higher eukaryotes. The tyrosine residue at position 1 and the proline at position 6 are universally present in all repeats, however the other five residues can vary. In fact, in humans 31 CTD repeats are non-consensus, which allows for a wider variety of modifications (Chapman et al., 2005). The effects of the various modifications, both on consensus- and non-consensus repeats, will now be discussed.

1.2.2 Transition from initiation to elongation

At the beginning of the transcription cycle, Pol II binds to the promoter together with general transcription factors (GTFs); at that stage, the Pol II CTD is unphosphorylated (Egloff et al., 2012a). Early studies on yeast showed that soon after transcription initiation, the CTD is phosphorylated on Ser5 by the TFIIH kinase subunit Kin28, which recruits both the capping enzyme to protect the 5' end of the nascent RNA from degradation, and the Set1 methyltransferase to trimethylate lysine 4 on histone H3 (H3K4). The chromatin immunoprecipitation (ChIP) signal of Ser5P peaks at the promoter and then diminishes about 200 nucleotides into the gene body. Correspondingly, Ser2P levels rise in the coding region and peak at the 3' end of the gene (Komarnitsky et al., 2000). This pattern is also observed in mammalian cells (Egloff et al., 2012a), and it is now known that phosphorylation of Ser2 is necessary for splicing, termination and 3' end processing by recruiting the spliceosome, and the cleavage and polyadenylation factors such as Pcf11 (Gu et al., 2013, Davidson et al., 2014). Therefore, transcription elongation and termination are mediated by factors responding to the changing levels of Ser5P and Ser2P, which also allows for RNA processing.

1.2.3 Role of Ser7P

More recently, serine 7 of the YSPTSPS repeat was shown to also be subject to phosphorylation (Chapman et al., 2007). On a number of mammalian genes Ser7P peaks at the promoter, is present in the gene body, and peaks again at the 3' end of the gene. This was also confirmed in a genome-wide study in yeast, which demonstrated that high Ser7P levels are present on actively-transcribed genes, both coding and non-coding, and therefore may regulate transcription elongation (Tietjen et al., 2010). Therefore, Ser2P,

Ser5P and Ser7P have different profiles on genes (shown in Figure 1-5), which reflect their differing functions during a cycle of transcription. The exact role of Ser7 phosphorylation was demonstrated by mutating Ser7 to alanine, which led to a defect in snRNA 3' end processing (Egloff et al., 2007). The recruitment of this mutant Pol II to snRNA genes was not affected, suggesting that the 3' end processing depends on steps downstream of Pol II binding. Indeed, the defect results from the loss of the Integrator complex, which requires phosphorylated Ser7 for binding and which cleaves the 3' end of the nascent snRNA (Egloff et al., 2012b).

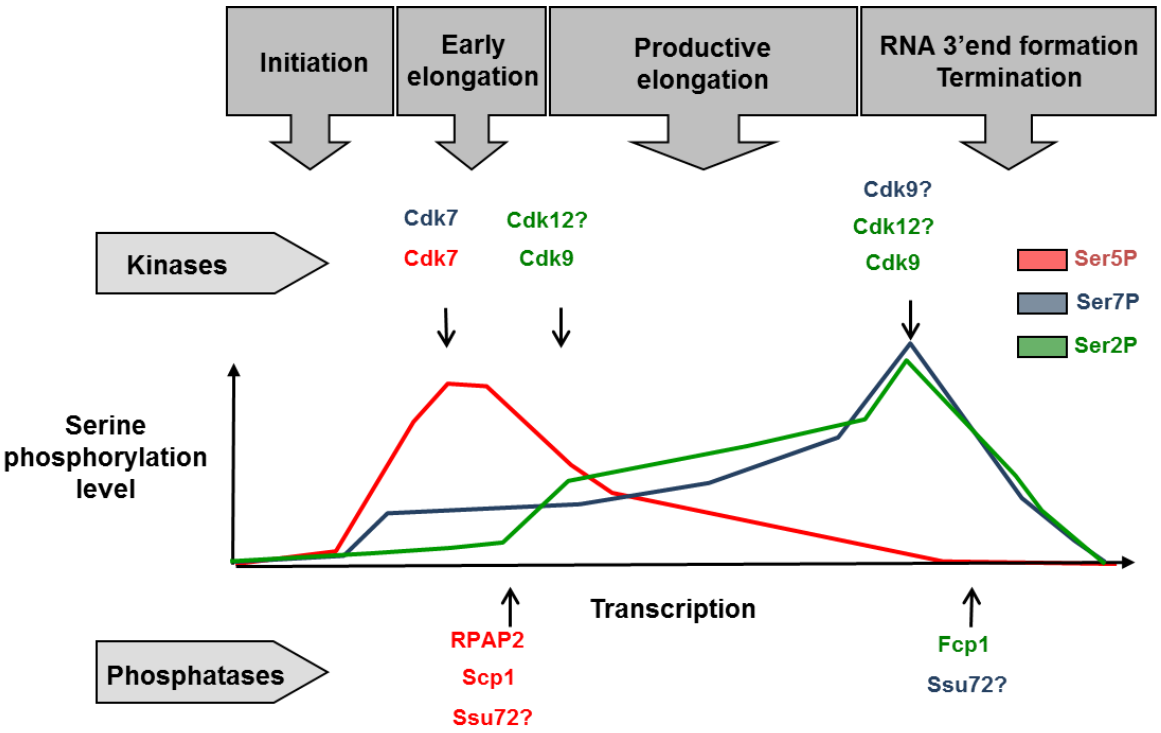


Figure 1-5. CTD serine phosphorylation patterns during one round of a transcription cycle. The levels of the phosphoserines are shown with the red, blue and green lines, while the relevant kinases and phosphatases are shown above and below (adapted from (Egloff et al., 2012a)).

1.2.4 Roles of Thr4P and Tyr1P

Phosphorylation of threonine 4 was first discovered in *S. pombe* on both the hypophosphorylated (Pol IIA) and the hyperphosphorylated (Pol IIO) forms of the polymerase, but the function of this modification was unknown at the time. The same study also observed phosphorylation of Tyr1 (Sakurai and Ishihama, 2002). More recently, analysis of Thr4P in a chicken cell line showed that Thr4 was essential for cell viability. When Thr4 is mutated to valine, the cells show no difference in mRNA transcription, polyadenylation and splicing, but there is a significant defect in histone mRNA 3' end processing (Hsin et al., 2011). Histone mRNA processing in vertebrates requires factors distinct from the usual mRNA 3' processing machinery (Marzluff et al., 2008), for example the stem-loop binding protein SLBP, which is lost from two histone genes in Thr4Val mutant cells. In yeast, the Thr4 mark has a different role: it is required for chromatin remodelling on a subset of genes involved in phosphate and galactose metabolism (Rosonina et al., 2014).

In mammalian cells Thr4 is phosphorylated by Plk3 and is associated with transcription elongation, with a Thr4P chromatin immunoprecipitation (ChIP) signal present genome-wide at low levels in gene bodies and peaking downstream of the polyA site (Hintermair et al., 2012). Mutation to alanine causes an accumulation of Pol II just after the TSS, suggesting an elongation defect. However, some genes are activated by the mutation, which may indicate that Thr4P also regulates transcription initiation.

The function of Tyr1 phosphorylation was first demonstrated in yeast. Tyr1P levels are found to increase downstream of the TSS, but drop before the polyA site (Mayer et al., 2012). The phospho group on Tyr1 impairs binding by CTD-interacting domain (CID)-

containing termination factors such as Nrd1, Pcf11 and Rtt103. Thus, Tyr1P acts as an anti-termination mark, and termination factors can only bind in regions where the mark is low, such as near the TSS for Nrd1 or near the polyA site for Pcf11 and Rtt103.

In human cells, the picture is quite different. Tyr1P peaks at gene promoters and co-localises with early transcription marks, such as Ser5P and Ser7P. Also unlike in yeast, vertebrate Tyr1P is associated with Pol II transcription in the antisense (AS) direction, and with Pol II-transcribed enhancers (Descostes et al., 2014, Hsin et al., 2014). However, promoter structure differs significantly between yeast and higher eukaryotes, and yeast do not have the same type of enhancers as metazoans do (Petrascheck et al., 2005), in line with the different functions of Tyr1P across species.

1.2.5 Effects of proline isomerisation

Although proline residues cannot be modified directly, the peptide bond between the proline and its preceding amino acid can exist in either the *cis* or *trans* conformation. Depending on this bond conformation, the CTD structure may either allow or sterically inhibit the binding of various proteins, and may thereby affect modification of other CTD residues.

Proline isomerisation between the *cis* and *trans* states is catalysed by Pin1 in mammals and Ess1 in yeast (Kubicek et al., 2012). There is now structural evidence for how the conformation of the peptidyl-prolyl bond affects the interaction with CTD-binding proteins. For example, the Ser5P phosphatase Scp1 can only bind to the CTD when both prolines are in the more energetically-favourable *trans* conformation (Zhang et al., 2006). Further studies have demonstrated that the phosphatase activity of Scp1 is not affected

by Pin1 in an *in vitro* assay, as ~80% of the prolines are naturally in the *trans* conformation, and the addition of Pin1 does not shift this ratio at a fast enough rate to affect Scp1 (Zhang et al., 2012b). Likewise, the cleavage factor Pcf11 only binds the CTD when both prolines of the heptad are in *trans* (Noble et al., 2005).

The affinities of CTD-binding proteins for different conformations of the proline residues can lead to competition. In yeast the termination factor Nrd1, which regulates transcription of short non-coding RNAs, preferentially binds to the CTD with phosphorylated Ser5, with the Ser5P-Pro6 bond in the *cis* conformation (Kubicek et al., 2012). As already mentioned, the polyadenylation factor Pcf11 needs a *trans* bond for binding. Ess1 can thereby activate the Ser5P phosphatase Ssu72, also requiring a *trans* bond for binding, to dephosphorylate Ser5P and therefore make Nrd1 dissociate from the CTD and allow binding of Pcf11 (Werner-Allen et al., 2011, Singh et al., 2009). Therefore, besides the direct modification of residues within the YSPTSPS heptad, isomerisation of peptidyl-prolyl bonds can have a large impact on transcription and RNA processing.

1.2.6 Serine and threonine glycosylation

In addition to phosphorylation and proline isomerisation, the CTD can be subject to glycosylation through an *O*-linked GlcNAc group on the serine and threonine residues. Glycosylation prevents phosphorylation of the same residues (Kelly et al., 1993). When the enzymes that catalyse glycosylation, OGT and OGA, are inhibited, pre-initiation complex assembly and Pol II recruitment to the promoter are disrupted, highlighting a role for CTD modifications at the very beginning of the transcription cycle (Ranuncolo et al., 2012).

1.2.7 Gene-specific CTD modifications

Certain CTD modifications have a gene-specific character, as already described for Thr4P on histone genes and Ser7P on snRNA genes. There are more such examples, especially when non-consensus CTD repeats are considered. The arginine of repeat 31 (YSPTSPR) can be methylated by CARM1, which represses the expression of some snRNAs and small nucleolar (Landeira et al.)RNAs (Sims et al., 2011, Egloff et al., 2012a).

Thus, the CTD code can regulate not just general transcription but that of specific gene classes, resulting in a more fine-tuned orchestration of gene expression. However, despite the recent advances in understanding the CTD code, many questions remain. For example, it is not clear how many repeats of the CTD need to bear a particular modification in order for it to have an effect. Also, the impact of combinations of modifications, especially on separate repeats, is not understood. The next section will focus on the enzymes that modify the CTD, and how they interact with other factors in a diverse range of roles.

1.3 Role of CTD-modifying enzymes in the transcription of human genes

1.3.1 Overview of CTD-modifying enzymes

In recent years, a large body of data has uncovered a multitude of CTD kinases, phosphatases and isomerases, along with other enzymes that modify residues in non-consensus repeats. Table 1 summarises the known CTD-modifying enzymes in humans, and in yeast where the human enzyme is unknown. Phosphorylation of Ser2 and Ser5 of the CTD are the most well-studied of all modifications, so more is known about the

Modification	Enzyme(s)		Function	References
Tyr1 phosphorylation	Kinase	c-Abl	Promotes elongation in yeast by preventing binding of Rtt103 and Pcf11 termination factors to the CTD	(Baskaran et al., 1997, Mayer et al., 2012)
	Phosphatase	Rtr1 in yeast		
Ser2 phosphorylation	Kinase	Cdk9, Cdk12/13 Brd4	Modulates mRNA 3' end processing and transcription termination by recruiting CPSF and CstF	(Washington et al., 2002, Lin et al., 2002, Shim et al., 2002, Jang et al., 2005, Yang et al., 2008, Devaiah et al., 2012, Bartkowiak et al., 2010, Blazek et al., 2011, Clemente-Blanco et al., 2011)
	Phosphatase	Fcp1, PP1		
Pro3 isomerisation	Pin1 in humans, Ess1 in yeast			(Zhang et al., 2012b)
Thr4 phosphorylation	Kinase	Cdk9	Histone mRNA 3' end processing	(Hsin et al., 2011)
	Phosphatase	?		
Ser5 phosphorylation	Kinase	Cdk7, Cdk8	Promoter escape, recruitment of the capping and splicing enzymes	(Washington et al., 2002, Ramanathan et al., 2001, Glover-Cutter et al., 2009, Yeo et al., 2003, Krishnamurthy et al., 2004, Clemente-Blanco et al., 2011, Egloff et al., 2012b)
	Phosphatase	RPAP2, Scp1, Ssu72, PP1		
Pro6 isomerisation	Pin1 in humans, Ess1 in yeast		*Cis* to *trans* bond conversion stimulates dephosphorylation	(Bataille et al., 2012, Zhang et al., 2012b)

			of Ser5P by Ssu72	
Ser7 phosphorylation	Kinase	Cdk7, Cdk9	Directs 3' end processing of snRNA genes by recruiting the Integrator complex	(Egloff et al., 2007, Bataille et al., 2012)
	Phosphatase	Ssu72 in yeast		
Arg7 methylation (in non-consensus repeats)	CARM1		Repression of snRNA and snoRNA transcription	(Sims et al., 2011)
Ser and Thr glycosylation	OGT, OGA		Pol II recruitment to promoters, and transcription initiation	(Kelly et al., 1993, Ranuncolo et al., 2012)

Table 1-1. List of known CTD-modifying enzymes in mammals and yeast.

kinases and phosphatases involved. Serine 2 is phosphorylated in mammals by Cdk9, which together with Cyclin T1, T2 or K form the positive transcription elongation factor P-TEFb (Marshall et al., 1996, Peterlin and Price, 2006). P-TEFb regulates the transition from abortive elongation to productive elongation after Pol II has escaped the promoter, and its activity is inhibited by 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB). P-TEFb achieves this by both phosphorylating the Pol II CTD, and two negative elongation factors: NELF and DRB-sensitivity inducing factor (DSIF), which causes the dissociation of NELF and the conversion of DSIF into a positive elongation factor (Kim and Sharp, 2001, Fujinaga et al., 2004).

The long-standing view that P-TEFb is the Ser2 kinase has been challenged in a recent paper by Czudnochowski *et al* (Czudnochowski et al., 2012). The authors used mass spectrometry and *in vitro* phosphorylation assays to show that P-TEFb in fact

phosphorylates Ser5 instead of Ser2, and requires a phospho-Ser7 mark in order to perform its kinase function. However, the indiscriminate approach of mass spectrometry and using anti-phospho-CTD antibodies that may cross-react with other modifications leaves room for further investigation and confirmation of these findings.

More recently, several other Ser2 kinases have been identified. These include Cdk12 and Cdk13, which are orthologues of the budding yeast CTD kinase Ctk1 (Bartkowiak et al., 2010, Blazek et al., 2011). Cdk12 in particular has a role in protecting cells against genomic instability: it phosphorylates the CTD predominantly on DNA damage-response genes such as BRCA1. Another newly-described kinase is the epigenetic regulator Brd4. Its structure is unlike that of a typical kinase, with eight regions that are homologous to kinase consensus motifs; however its CTD kinase activity has been demonstrated in *in vitro* assays and in cells (Devaiah et al., 2012). Moreover, it recruits P-TEFb to genes during interphase and stimulates CTD phosphorylation by P-TEFb, while another pool of Brd4 recruits P-TEFb at anaphase, acting as a 'mitotic bookmark' to allow rapid transcription reinitiation in the new cell cycle (Jang et al., 2005, Yang et al., 2008, Devaiah and Singer, 2013). Another important CTD kinase is Cdk7 of the TFIIH complex. Originally described as a Ser5 kinase at mRNA gene promoters (Roy et al., 1994, Akoulitchev et al., 1995), it was more recently shown to also phosphorylate Ser7 at both mRNA and snRNA genes, although Ser7 kinase activity requires a phospho-Ser5 mark in budding yeast (Glover-Cutter et al., 2009, Akhtar et al., 2009). Cdk7 regulates promoter-proximal pausing of Pol II, but also 3' pausing and transcription termination.

Having considered the kinases, the discussion moves to the CTD phosphatases: an important but less well-studied class of enzymes than the kinases. The Ser2P phosphatase

Fcp1 was initially discovered in a yeast two-hybrid screen as a TFIIF-interacting protein and a component of the Pol II holoenzyme. It was then shown to oppose the Ser2 kinase function of Ctk1 in yeast, and its knockdown increases the levels on Ser2P on genes (Archambault et al., 1997, Cho et al., 2001, Bataille et al., 2012). Fcp1 interacts with components of the spliceosome and is postulated to link transcription with splicing (Licciardo et al., 2003). Furthermore, it is important for recycling Pol II for the transcription of heat shock genes in *Drosophila* (Fuda et al., 2012).

The major Ser5P phosphatase in yeast and humans is Ssu72, which is associated with the cleavage and polyadenylation factor CPF and impacts on transcription elongation, termination and 3' end processing, in particular termination of snoRNAs in yeast (Dichtl et al., 2002, Krishnamurthy et al., 2004, Ganem et al., 2003, Hsin and Manley, 2012). Ssu72 also dephosphorylates Ser7P (Zhang et al., 2012a). In mammalian cells, the activity of Ssu72 is regulated by the cleavage and polyadenylation scaffold protein symplekin through a direct interaction (Xiang et al., 2010). As previously described, the phosphatase function of Ssu72 also depends on the CTD modification context: for example, Ssu72 has a preference for Ser5P with the neighbouring peptidyl-prolyl bond in *trans*, and its activity is reduced when Thr4 that neighbours the Ser5 is phosphorylated (Werner-Allen et al., 2011, Singh et al., 2009, Luo et al., 2013).

Besides Ssu72, several other Ser5P phosphatases are known. The family of three small C-terminal domain phosphatases (SCPs) preferentially dephosphorylate Ser5P and have a gene-type-specific role in silencing neuronal genes in non-neuronal tissues (Yeo et al., 2005, Yeo et al., 2003). Finally, the most recently-discovered Ser5P phosphatases are RPAP2 and its yeast homologue Rtr1, which will be the focus of the next section.

1.3.2 RPAP2 is a Ser5P phosphatase

The first evidence of the role of RPAP2 in transcription came from a proteomic analysis of the human transcription machinery interactome. Several proteins of the basal transcription machinery were tandem affinity purification (TAP)-tagged, purified, and subject to liquid chromatography-tandem mass spectrometry (LC-MS/MS). Over 800 interaction partners were found, including the open reading frame C1orf82, subsequently renamed to RNA polymerase II associated protein 2 (RPAP2) (Jeronimo et al., 2007). In addition to binding Pol II in the soluble nuclear fraction, RPAP2 was also found in complex with the Integrator complex, Mediator, the Prefoldin family of chaperones, and AAA+ ATPases.

In parallel, the yeast homologue of RPAP2, which is called regulator of transcription 1 (Rtr1), was identified through a screen of genes necessary for heat stress tolerance (Gibney et al., 2008). Rtr1 mostly localises in the cytoplasm under normal conditions, but shuttles into the nucleus in higher temperatures. Its highly-conserved N-terminal domain containing four cysteine residues confers temperature sensitivity, and like the human RPAP2, Rtr1 directly interacts with Pol II. Moreover, at both normal and high temperatures, Rtr1 mediates transcription of genes from the GAL1 promoter, suggesting a role in the transcription of protein-coding genes.

In subsequent studies, Rtr1 was shown to dephosphorylate the CTD Ser5P, as evidenced by a rise of phospho-Ser5 in *rtr1* deletion strains, and also by ChIP analysis and *in vitro* enzyme assays using recombinant Rtr1 (Kim et al., 2009, Mosley et al., 2009). These studies also confirmed the requirement of one of the conserved cysteine residues, Cys73, for the catalytic activity of Rtr1.

Consistent with the yeast studies, human RPAP2 was also shown to act as a CTD Ser5P phosphatase, with a particular role in snRNA gene transcription (Egloff et al., 2012b). In order for RPAP2 to bind the CTD on snRNA genes, the heptad repeats need to carry the Ser7P mark. However, this mark is not required for RPAP2 recruitment to protein-coding genes. Consistent with its role in Ser5P dephosphorylation, RPAP2 levels are highest next to gene promoters where the Ser5P mark begins to drop, and its levels drop throughout the gene body. The Pol II-binding activity is contained within the first 334 amino acids of RPAP2, and mutations of the conserved cysteines abolish the interaction. However, it was not known at this point whether the cysteines confer catalytic activity, or simply form a structure such as a zinc finger that allows protein-protein interactions.

Interestingly, other roles for RPAP2 besides CTD dephosphorylation have been revealed. Like its yeast homologue, human RPAP2 also shuttles between the cytosol and nucleus; however this shuttling serves to import Pol II subunits into the nucleus (Forget et al., 2013). The CTD is not required for this RPAP2-Pol II interaction. Thus, RPAP2 appears to be a multifunctional protein, although it is not yet known how these separate roles are coordinated and regulated.

1.3.3 Role of RPAP2 in snRNA 3' end processing

Although also present on mRNA genes, human RPAP2 is now known to have a gene-specific role in regulating snRNA 3' end processing. In fact, RPAP2 exists in at least two complexes: one containing the Pol II large subunit Rbp1, and another without Rbp1 but containing several subunits of the Integrator complex (Egloff et al., 2012b). These include Int4, Int5, Int6 and Int7, but do not include the catalytic subunit Int11 and its cofactor

Int9. This suggests that Integrator is not recruited to genes as a fully-assembled complex, and that different subunits can bind at different times.

As already mentioned, RPAP2 binds to snRNA genes when the CTD is phosphorylated on Ser7, a mark that is needed for transcription and processing of snRNAs (Egloff et al., 2007). In fact, when RPAP2 is knocked down by RNAi, the levels of correctly-processed pre-U2 snRNA drop in cells, while a larger proportion of the RNA is read through (i.e. the 3' box is not recognised). At the same time, there is a decrease in Integrator recruitment to the gene, including the catalytic subunit Int11. This suggests that RPAP2 recruits Integrator to snRNA genes, possibly in several stages with the catalytic subunits recruited later, and Integrator travels along the gene with the polymerase to direct 3' processing of the nascent RNA (see Figure 1-6).

1.3.4 RPAP2 and Rtr1 interaction partners

Knowing the interaction partners of a protein can often give insight into its function. The first study of the human transcription interaction network already demonstrated the association od RPAP2 with Mediator and Integrator, which suggests a role in transcription regulation, and specifically transcription of snRNA genes that require the Integrator complex for processing (Jeronimo et al., 2007). The full list of RPAP2 interacting partners, however, also contains the phosphatase PPP1CA, and with lower confidence scores, PPP1CB, PPM1A and PPM1B. PPP1CA and PPP1CB are catalytic subunits of the PP1 phosphatase (Aggen et al., 2000). The roles of PPM1A and PPM1B are to dephosphorylate IKKβ in the TNFα signalling pathway (Sun et al., 2009). The PP1 phosphatase dephosphorylates the CTD at Ser2P and Ser5P, as shown by *in vitro* assays, immunoprecipitation and mass spectrometry, and inhibition with okadaic acid and

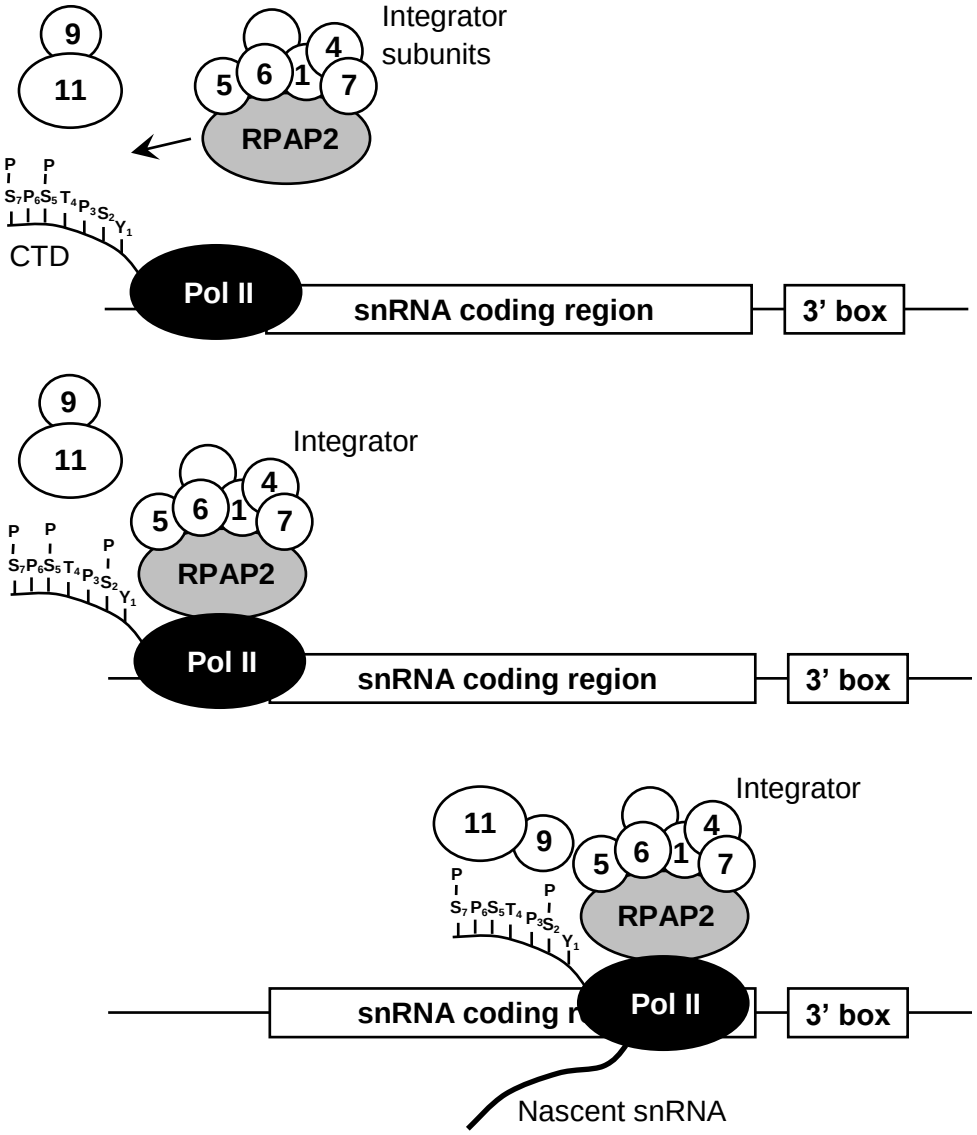


Figure 1-6. Model of Integrator recruitment to snRNA genes by RPAP2.
RPAP2 recognises the Ser7P mark on the CTD, and recruits several Integrator subunits (depicted as numbered circles) to the snRNA gene. After transcription initiation, the catalytic Integrator subunits are also recruited, and RPAP2 dephosphorylates Ser5P of the CTD to allow transcription elongation to proceed (adapted from (Egloff et al., 2012b)).

endogenous inhibitory protein NIPP1 (Jerebtsova et al., 2011, Washington et al., 2002). It is targeted to *Drosophila* genes by the nuclear targeting subunit dPNUTS, which is essential for development (Ciurciu et al., 2013). PP1 also activates transcription of the *p21* gene through histone H3S10 dephosphorylation, and activates Cdk9 by dephosphorylating its Ser175P (Valin et al., 2013, Ammosova et al., 2011). Given that PP1 binds to RPAP2 and possesses CTD phosphatase activity, it is possible that RPAP2 recruits PP1 to the CTD to dephosphorylate Ser5P, or that the two proteins interact to perform a different transcription regulatory role.

A more recent study of RPAP2 also used mass spectrometry to discover its interaction partners within the two complexes found in the nucleus (Egloff et al., 2012b). Along with a number of Pol II subunits and general transcription factors, the PPM1B interaction was also verified. The fact that RPAP2 binds several phosphatases is intriguing, as it may indicate that the phosphatase activity of RPAP2 depends on other factors.

The interactome of *S. cerevisiae* Rtr1 has been extensively analysed by multidimensional protein identification technology (MudPIT) (Hsu et al., 2014). In addition to the known interactions with Pol II subunits, Rtr1 was also shown to interact with a number of other factors, most notably Ctk1, the yeast homologue of the human Cdk12. Rtr1 is phosphorylated on Ser217, but not in yeast strains with a *ctk1* deletion; however the phosphorylation does not affect the phosphatase activity of Rrt1. Additionally, the binding of Rtr1 to Pol II is dependent on Ctk1, which may suggest that CTD Ser2 phosphorylation by Ctk1 is required for Rtr1 recruitment. Contrary to the human RPAP2 data, no phosphatases were found to interact with Rtr1.

In a separate study, human RPAP2 was shown to interact with three recently-described CTD-binding proteins, RPRD1A, RPRD1B and RPRD2 (Ni et al., 2011). They are related to the yeast termination factor Rtt103, and RPRD1A is known to be a cell cycle regulator and a regulator of Wnt signalling (Lunde et al., 2010, Wu et al., 2010, Liu et al., 2002). RPRD1A and RPRD1B preferentially bind to the phosphorylated CTD, and overexpression of both proteins reduces the levels of Ser5P at the promoter of the *LEO1* gene. This is consistent with the observation that knocking down RPAP2 causes an increase in Ser5P, if the RPRD proteins recruit RPAP2 (Egloff et al., 2012b). Indeed, so far the RPRD proteins have not been shown to possess phosphatase activity, so it is possible that RPRD1A and RPRD1B recruit RPAP2 to genes, and/or regulate its activity.

1.3.5 RPAP2 domains and modifications

RPAP2 is highly conserved in eukaryotes, but is not found in bacteria or archaea. The human RPAP2 protein is 612 amino acids long, while the *S. cerevisiae* homologue is only 226 amino acids. The highest degree of conservation between species lies within the N-terminal Domain of Unknown Function 408 (DUF408), spanning approximately 60 amino acids (Figure 1-7). In particular, the DUF408 contains four regularly-spaced cysteine residues (the fourth residue can also be a histidine in yeast) that form a metal-binding motif, until recently assumed to be a zinc finger (Egloff et al., 2012b). The DUF408 is required for binding both Pol II and Integrator, as its deletion abolishes the interaction with both complexes. However, RPAP2 does not contain any known phosphatase domain, so is either a novel type of phosphatase or is not a phosphatase at all, despite the *in vitro* assays and ChIP data suggesting otherwise.

The human RPAP2 also contains a coiled coil domain close to the N-terminus, and a putative transmembrane domain in the C-terminus, although membrane localisation has not been observed (http://www.uniprot.org/uniprot/Q8IXW5). Mass spectrometry data have uncovered a number of modification sites: mostly phosphorylation on serine, threonine and tyrosine residues, but also ubiquitylation of Lys259 and acetylation of Lys509. However, so far none of these modifications have been confirmed by site-specific methods (www.phosphosite.org/uniprotAccAction.do?id=Q8IXW5). Thus, very little is known about the structure and organisation of the human RPAP2, with most of the information having been obtained from sequence analysis and indirect mass spectrometry data, rather than from mutagenesis and other experimental evidence.

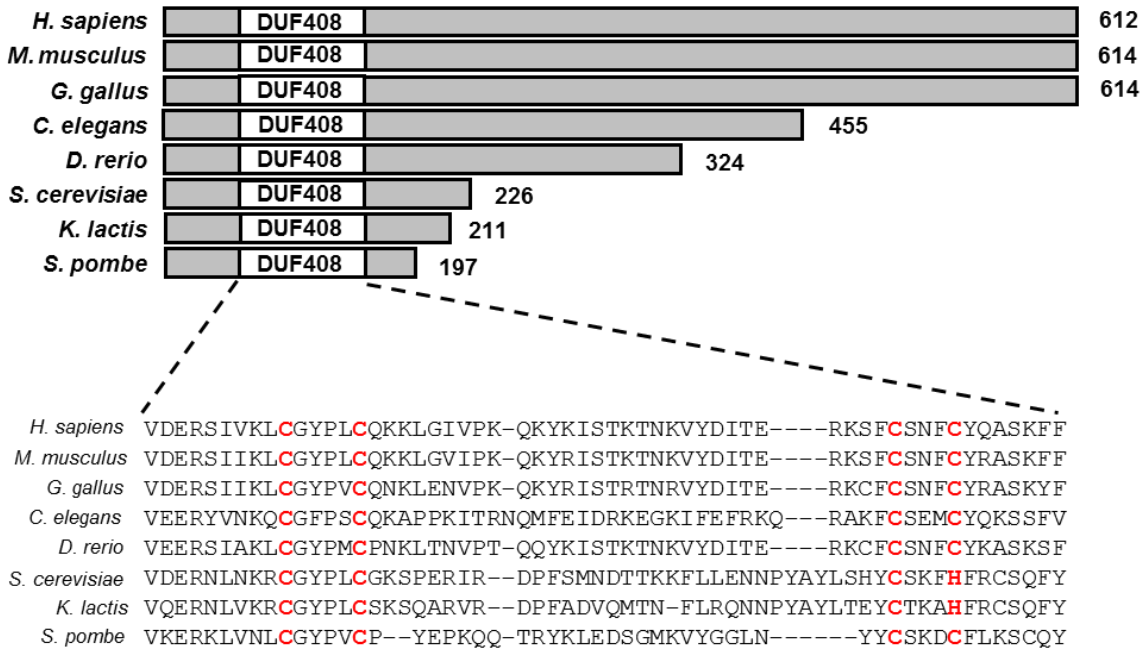


Figure 1-7. Conservation of RPAP2 across eukaryotes.
The sequence of the DUF408 in each species is displayed at the bottom, with the conserved cysteine/histidine motif highlighted in red (adapted from (Egloff et al., 2012b)). Two conserved glutamate residues are shown above the *H. sapiens* diagram

1.3.6 Structural studies of RPAP2

Given all the evidence so far, it was unclear whether RPAP2/Rtr1 is actually a CTD phosphatase, or whether it is a scaffold protein that simply recruits another phosphatase such as PP1 or PPM1A/PPM1B. The answer apparently emerged when the X-ray crystal structure of the yeast *Kluyveromyces lactis* Rtr1 was solved to 2.5 Å resolution (PDB 4FC8, (Xiang et al., 2012)). The structure showed five anti-parallel α-helices, with two long loops, although of the 211 amino acids, only residues 1-152 were visible (see Figure 1-8 A). The metal-binding site at the interface of the four cysteines is located near the surface of the protein (grey sphere in Figure 1-8 B), and there was nothing in the structure that resembled a phosphatase motif or active site. Moreover, *in vitro* enzyme assays of both the *K. lactis* Rtr1 and human RPAP2 (Xiang et al., 2012) did not detect any phosphatase activity, although the conditions used differed from previous studies, in that HeLa nuclear extract or purified Cdk7 were used to pre-phosphorylate the CTD substrate, as opposed to purified P-TEFb (Egloff et al., 2012b). Xiang and colleagues concluded that RPAP2 was not, in fact, a phosphatase.

This view was challenged when the structures of *K. lactis* and *S. cerevisiae* Rtr1 were solved to 2.1 Å (PDB 4M3O) and 4 Å resolutions, respectively (see Figure 1-9), and additional NMR data were collected to observe the interaction between the N-terminal and C-terminal regions of the protein (Hsu et al., 2014). The crystal structure, NMR data, and *in vitro* phosphatase assays showed that the C-terminal region of the protein, absent from the structure solved by Xiang *et al*., forms a loop that folds over the metal-binding motif and regulates the enzymatic activity of the N-terminal domain in a noncompetitive

A

```
S. S.                αA                  αB                  αC        αD
KlRtr1       1  MATIEEIKEVVLKPYTNHRQLTIREVETISINLIDLLITKDVKDARTMKYISRFLTKQDYADLVQERNLVKRCGYPLCSKSQ
ScRtr1       1  MATIEDIKETALIPFQKHRQLSMHEAEVITLEIIGLLCDSECKDEKTLKYLGRFLTPDMYQDLVDERNLNKRCGYPLCGKSP
AgRtr1       1  MATIEDIQKQVLVAYQQHRQLSLKEADLLTVELLELLSQTECTP-ETLKYVGRFFTKTTYSDLIDERNLNKLCGYPLCGQSQ
HsRPAP2     29  TSTLKQEDASKRKAELEAAVRKKIEFERKALHIVEQLLEENITEEFLMECG-RFITPAHYSDVVDERSIVKLCGYPLCQKKL

S. S.                                    αE
KlRtr1       83  ARVR-Dpfadvqmtn-flrQNNPYAYLTEYCTKAHFRCSQFYQFQLSDEALFARVGVHL------DDYEPPSEIQLLEevl
ScRtr1       83  ERIR-DPFSMNDTTKKFLLENNPYAYLSHYCSKFHFRCSQFYQVQLSDEALFARTGVHLFEDPEQDKHDIDFKVTLFEELL
AgRtr1       82  GRIR-DPYENRRVSN-FLTQNNPYKYLTSFCSKLHFRCSQFYQVQLSDEALFARVGVHL-------DDYEAGRVLLLEEAM
HsRPAP2     110  GIVPKQKYKISTKTN----KVYDITERKSFCSNFCYQASKFFEAQIPKTPVWVREEERH----------PDFQLLKEEQS

S. S.          αF
KlRtr1      156  ---aresdvkqmirgmtdlkladdd-sre----emerdisdmladikiveinepniigdlgrdr   211
ScRtr1      163  REKASEEDIKSLISGLKKLGLNPDSGTTEKDDTELEDDLSKWLAQIKIVENDNPSILGDFTRED   226
AgRtr1      154  ---AEERDLRTMIRGMNGLSLGENG-AGD----ELQKDLSEWLSEVRIVENERPPLLGDLMKE    208
HsRPAP2     176  GHSGEEVQLCSKAIKTSDIDNPSHFEKQYESSSSSTHSDSSSDNEQDFVSSILPGNRPNSTNIRP   240 (612)
```

B

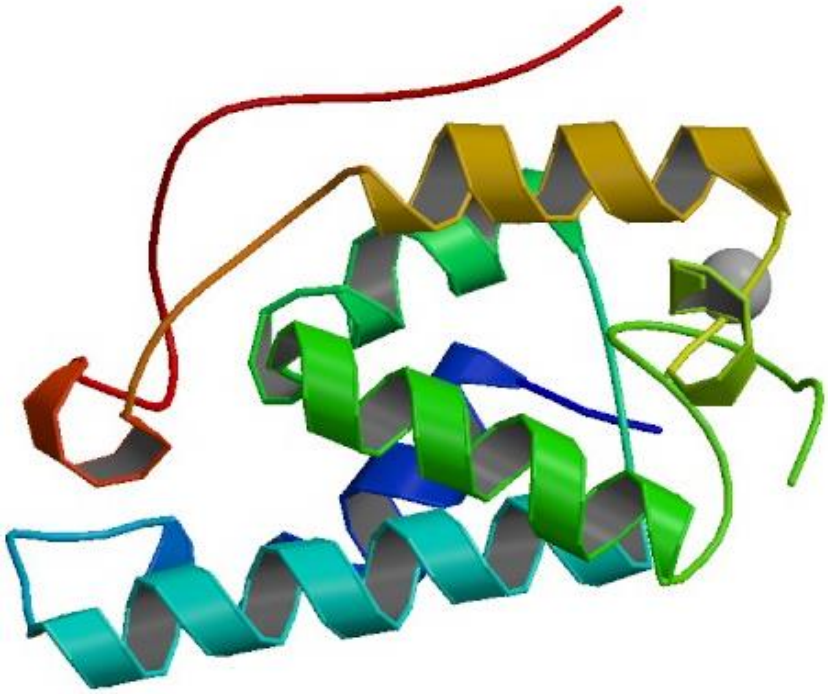


Figure 1-8. Diagram of the structure of the *K. lactis* Rtr1, solved to 2.5 Å resolution.

A. sequence alignment of Rtr1 from several species, including human RPAP2. Yellow bars above the sequence indicate alpha-helices as observed in the *K. lactis* structure. The conserved cysteine motif is shown in orange letters, while the residues missing in the structures are shown in grey italics. B. crystal structure of the *K. lactis* Rtr1 (PDB code 4FC8). Both figures adapted from (Xiang et al., 2012).

fashion. Furthermore, divalent metal ions also inhibit the enzyme activity. Therefore when Rtr1 is purified with an EDTA wash step, it shows robust phosphatase activity towards a synthetic substrate, 6,8-difluoro-4-methylumbelliferyl phosphate (DiFMUP). The phosphatase activity depends on a conserved residue, Glu66, and mutation of this residue to alanine inhibits catalytic activity and results in a growth defect comparable with a deletion of the entire protein. Glu66 is located at the outer face of the protein, in what was determined to be its active site. In the study by Xiang *et al*., Rtr1 was purified in the presence of zinc without an ethylenediaminetetraacetic acid (EDTA) step, which may explain the lack of enzyme activity.

Hsu and colleagues therefore argue that Rtr1 is a structurally-novel type of phosphatase, with enzymatic activity confirmed by both structural evidence and mutagenesis. The structure of the human RPAP2, however, is still not known. Considering that it is significantly larger than the yeast homologues, it is unclear what function the less-conserved C-terminal part of the human protein serves, and whether the enzyme is also subject to auto-inhibition.

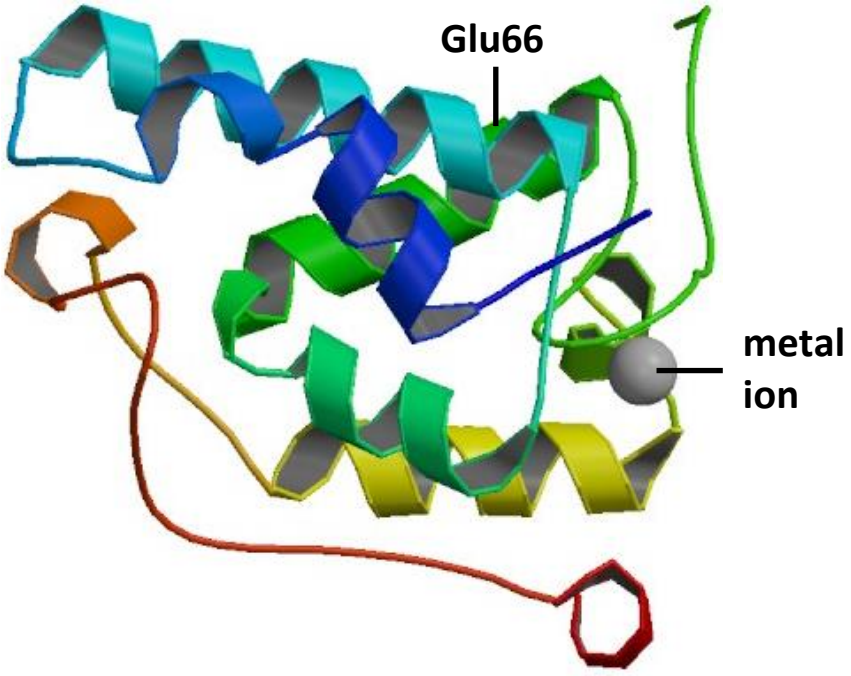


Figure 1-9. Diagram of the structure of the *K. lactis* Rtr1, solved to 2.1 Å resolution.
PDB code 4M3O. The Glu66 residue necessary for enzymatic activity and metal ion are labelled. Figure adapted from (Hsu et al., 2014).

1.4 CTD modifications in stem cells and during development

1.4.1 Rtr1 mutations in zebrafish

An important question is the role of CTD modifications in stem cells, and how their regulation changes during differentiation and development. All studies of RPAP2 and Rtr1 published so far have been conducted either in cancer cells such as HeLa, or in yeast. Recently however, it has emerged that Rtr1 may have an important role in embryonic development. The group of Dr Amber Mosley (Indiana University School of Medicine, Indianapolis, USA) knocked down Rtr1 in zebrafish for 48 hours using a morpholino oligonucleotide. Only four out of 54 embryos hatched, and showed a range of

developmental defects including curly tails, absence of eyes, and small heads (Dr Amber Mosley, personal communication).

Visually, the Rtr1 knockdown embryos are very similar to *pandora* and *foggy* zebrafish mutants, which are defective in the elongation factors Spt6 and Spt5, respectively (Keegan et al., 2002). Both mutants display similar embryonic defects, including curled tails, pericardial oedema and a reduced number of ventricular myocardial cells. The authors conclude that Spt5 and Spt6 are important for transcription regulation during zebrafish embryonic development. Given the similarity with the Rtr1 knockdown, it was postulated by Mosley and colleagues that Rtr1 is also important in embryonic development, potentially through its role in regulating CTD modifications. A possible target of Rtr1 in early zebrafish embryos is the *HOX* gene cluster, which is the master regulator of development along the head-to-tail axis.

1.4.2 Roles and regulation of HOX genes in embryonic development

Homeobox (HOX) proteins are helix-turn-helix transcription factors that regulate gene expression during embryonic development, although *HOX* expression continues throughout adult life (Lappin et al., 2006). The gene family is present in animals, plants and fungi, and has evolved through a series of duplications (Schilling and Knight, 2001). In mammals, there are 39 *HOX* genes in clusters on four different chromosomes, and the timing and location of their expression along the anteroposterior axis roughly follows the order of the genes along the chromosome; a concept known as 'colinearity' (Lewis, 1978, Duboule and Dolle, 1989).

HOX genes begin to be expressed during early gastrulation, a developmental stage marked by the formation of the three germ layers and the major body axis. Highlighting their importance in a wide range of developmental programmes, defects in expression of *HOX* genes result in a variety of disorders, ranging from deformities of the hands and feet to bronchopulmonary sequestration to cancers (Lappin et al., 2006). For example, translocations of the *HOXA9* and *HOXD13* genes are found in acute myeloid leukaemia (AML), while rearrangement of the mixed-lineage leukaemia gene *MLL1* leads to aberrant HOX gene expression (De Braekeleer et al., 2005, Pineault et al., 2003). Therefore, to avoid potentially severe developmental defects, the correct spatiotemporal expression of *HOX* genes requires stringent regulation.

In recent years, the study of human developmental programmes has been greatly aided by the use of embryonic stem cells (ESCs), initially derived from human blastocysts. Like cells within the early embryo, ESCs are pluripotent given the correct culture conditions, and have fewer technical and ethical limitations than if working with embryos *in vivo* (Vallier and Pedersen, 2005, Thomson et al., 1998).

HOX gene regulation has been studied in both human and mouse ESCs, as well as mouse embryos. In ESCs, as in the very early stages of embryonic development, *HOX* gene expression is silenced by its chromatin environment. The *HOX* clusters are spatially separated into inactive chromatin domains that are marked with the repressive histone 3 lysine 27 trimethyl mark (H3K27me3) that is added by the Polycomb group proteins (PcG). There are also low levels of the activating H3K4me3 mark, added by the Trithorax group proteins (TrxG) (Noordermeer et al., 2011, Soshnikova and Duboule, 2009, Schwartz and Pirrotta, 2007, Atkinson et al., 2008). This double mark is present in

'bivalent' chromatin, in which genes are transcribed at very low levels but are poised for active transcription. In fact, in mouse and human ESCs there are several large bivalent domains spread over all four *HOX* clusters (Bernstein et al., 2006, Mikkelsen et al., 2007, Guenther et al., 2007). In *Xenopus* or *Drosophila*, however, no bivalent domains are observed, suggesting that they only have a function in mammals (Gan et al., 2010, Akkers et al., 2009).

Activation of *HOX* gene expression during embryogenesis is stimulated by several signalling pathways, including retinoic acid (RA), Wnt, sonic hedgehog, and fibroblast growth factor. For example, RA binds to retinoic acid receptors located at RA-responsive DNA elements (RAREs) in the *HOX* gene clusters (Lengerke et al., 2007, Lengerke et al., 2008, Kashyap et al., 2011). As the embryo develops or the ESCs differentiate, the repressive chromatin at the *HOX* clusters opens up to allow active transcription. This is associated with rapid chromatin remodelling, with an increase in H3K4me3 within an hour, followed by an erasure of the H3K27me3 Polycomb mark (Kashyap et al., 2011). The clusters also physically move into a nuclear compartment with active transcription (Noordermeer et al., 2011). Therefore, transcription of *HOX* genes is regulated at the level of chromatin modifications, which determines the accessibility of the genes to the transcription machinery. However, another layer of regulation is at the level of Pol II itself and its CTD modifications.

1.4.3 Pol II CTD modifications on *HOX* genes in stem cells

Recent advances in genome-wide techniques such as chromatin immunoprecipitation coupled to sequencing (ChIP-seq), have given much insight into chromatin rearrangements and consequent changes in transcription during development. It is

known that in embryonic stem cells, Polycomb-regulated genes are bound by Pol II that is phosphorylated on Ser5 of the CTD (Landeira et al., 2010, Stock et al., 2007, Min et al., 2011). Specifically, it has been shown using a ChIP approach that in mouse ES cells, bivalent genes, including several *HOX* genes, are bound by Pol II carrying the Ser5P mark. The polymerase on these genes is poised for transcription, but in the absence of a Ser2P mark the genes are only transcribed at a low level (Stock et al., 2007, Kashyap et al., 2011). The Ser5P mark was not found on other untranscribed genes that do not have bivalent chromatin. To prevent premature transcription of bivalent genes, the chromatin is compacted by the histone H2A mono-ubiquitinating function of the Polycomb complex 1 (PRC1). However, inhibition of PRC1 by deleting one of its component proteins led to an upregulation of bivalent gene transcription without any major changes in the CTD modification profile. So far it remains unclear how Pol II can transcribe genes in ES cells without the Ser2P mark on the CTD, or without erasure of the Ser5P mark. Given that RPAP2 is a major Ser5P phosphatase, it is possible that regulation of RPAP2, either on the level of expression or enzymatic activity, contributes to the CTD profile observed in differentiating stem cells.

1.5 Aims of the thesis

The present thesis aims to investigate the modes of regulating Pol II CTD modifications in human cancer cells and stem cells. Although a number of CTD modifications and their downstream effects are known, in many cases it is still unclear how Pol II function is affected by factors that interact with the polymerase. For example, recent studies have shown that the Integrator complex drives snRNA maturation, but how Pol II terminates transcription of snRNA genes was unknown at the start of this research project.

Therefore, Chapter 2 of this thesis reports a closer examination the role of Integrator in transcription of the U1 and U2 snRNA genes. Furthermore, the CTD Ser5P phosphatase RPAP2, which recruits Integrator to snRNA genes, is shown to have a new role in recruiting the PP1 phosphatase to snRNA genes, and to have a putative role in regulating termination of transcription of protein-coding genes.

Chapters 3 and 4 present experiments that examine the structure and regulation of RPAP2 itself. A recombinant, truncated form of RPAP2 has been successfully purified, however crystallisation trials have not yielded protein crystals. The function of RPAP2 on human genes is shown to be affected both by the positive elongation factor P-TEFb, and by the RPRD complex, which recruits RPAP2 to genes and stimulates its Ser5P phosphatase activity.

Chapter 5 describes investigations into CTD modifications in human embryonic stem cells and induced pluripotent stem cells, and shows a correlation between Ser5P and RPAP2 levels, which may give an insight into transcription regulation and developmental programmes.

The Discussion (Chapter 6) sums up the findings of this work and points to some possible future directions. Chapter 7 describes the materials and methods used for the experimental work.

Chapter 2| Regulation of Pol II function by Pol II-associated factors

2.1 Introduction

Transcription of all mRNA, microRNA and most sn/snoRNA genes is mediated by RNA Polymerase II (Pol II). It is a large complex comprising 12 subunits in human, with the largest subunit Rbp1 containing a long, unstructured C-terminal domain (CTD) (Myer and Young, 1998). Post-translational modifications of the CTD heptapeptide repeats, which make up the "CTD code", enable the recruitment of various co-factors that regulate all aspects of Pol II function ((Egloff et al., 2012a) and Section 1.2 of the Introduction). Phosphorylation of the CTD Ser5 is associated with transcription initiation, while phosphorylation of Ser2 by the P-TEFb complex marks the transition into elongation, as P-TEFb also phosphorylates the negative elongation factors NELF and DSIF. Phosphorylation of Tyr1 inhibits transcription termination in yeast, and dephosphorylation is needed for termination to proceed (Komarnitsky et al., 2000, Egloff et al., 2012a, Kim and Sharp, 2001, Ni et al., 2008, Peterlin and Price, 2006, Wada et al., 1998b, Mayer et al., 2012). Combinations of modifications in the same repeat can have differing effects: for example, isomerisation of proline 6 to the *cis* conformation by Pin1 allows Ssu72 to dephosphorylate Ser5, while the phosphatase function is inhibited by the same proline in *trans* (Werner-Allen et al., 2011, Zhang et al., 2012b). Considering the huge number of possible combinations in the CTD code, it is still unclear how some of the modifications interact to regulate Pol II function.

Modification of the residues in the CTD heptapeptide repeats is mediated by a plethora of kinases, phosphatases, acetylases, methylases and isomerases (Egloff et al., 2012a, Sims et al., 2011, Zhang et al., 2012b). Some enzymes have specificity for more than one residue, for example Rtr1, the yeast homologue of the Ser5P phosphatase RPAP2, has

recently been shown to dephosphorylate yeast Tyr1 *in vitro (Hsu et al., 2014)*. Likewise, one residue can be modified by different enzymes: for example Ser2 is usually phosphorylated by Cdk9 of the P-TEFb complex, but can also be phosphorylated by Cdk12 on long genes, particularly those implicated in the DNA damage response (Blazek et al., 2011). In fact, recent studies highlight the gene-specific nature of CTD modifications (Egloff et al., 2012a).

To add yet another layer of complexity, the CTD-modifying enzymes themselves are subject to diverse regulation and association with interacting factors. It has been shown that recruitment of the Integrator complex by RPAP2 to snRNA genes leads to 3' end cleavage of the nascent RNA. Recruitment of a subcomplex of Integrator requires a specific CTD signature, with both Ser2 and Ser7 phosphorylated (Egloff et al., 2012b). It was postulated that recruitment of RPAP2 allows the assembly of the full Integrator complex, and subsequent 3' end processing of the pre-snRNA. However, the factors that regulate RPAP2 itself are not known.

The present chapter concerns the elucidation of the mechanisms by which regulation of Pol II-associated factors impact on the CTD code and transcription. In particular, it focuses on the roles of the Ser5 phosphatase RPAP2 and its association with other factors. Considering that the role of RPAP2 as a Ser5 phosphatase has been disputed, the possibility of the existence of a separate Ser5 phosphatase recruited by RPAP2 is first examined. A strong candidate was PP1, previously shown to dephosphorylate phospho-Ser5 *in vitro*, although it is also known to have other roles such as activation of P-TEFb (Ammosova et al., 2011, Washington et al., 2002). The interplay between RPAP2 and PP1 was studied by knocking down either factor and performing ChIP to investigate

recruitment and Ser5 phosphatase activity. Another line of study reported here investigates the RPAP2-interacting Integrator complex. Although the Integrator complex is known to mediate snRNA 3' end processing, the mechanisms of human snRNA gene transcription termination were unknown until now, so this study looks at the involvement of Integrator in the process. Furthermore, the yeast homologue of RPAP2 has recently been more directly implicated in mRNA transcription termination, through a phosphatase activity on the CTD Tyr1 in yeast. Thus this work also investigated the role of RPAP2 in dephosphorylation of Tyr1 in human cells using ChIP of the phospho-tyrosine mark on Pol II transcribing an mRNA gene.

2.2 Results

2.2.1 Recruitment of the PP1 phosphatase by RPAP2

The study by Xiang and colleagues (Xiang et al., 2012), in which the authors solved the crystal structure of *K. lactis* Rtr1 to 2.5 Å resolution and found neither an active site nor phosphatase activity, cast some doubts on whether RPAP2/Rtr1 really is a CTD Ser5P phosphatase. It is possible that RPAP2 only recruits the true phosphatase to the genes. Indeed, a proteomic analysis of the human transcription interaction network by Jeronimo *et al.* (Jeronimo et al., 2007) uncovered PPP1CA and PPP1CB, two catalytic subunits of protein phosphatase 1 (PP1), as high-confidence RPAP2 interaction partners. Involved in a large number of cellular processes, PP1 has also been found in a complex with Pol II, and was shown *in vitro* and in cultured cells to dephosphorylate both Ser2 and Ser5, although this was never confirmed by ChIP on genes (Jerebtsova et al., 2011, Washington et al., 2002). Moreover, PP1 has other targets beyond the Pol II CTD: recently it was demonstrated that by removing the inhibitory phosphate on Ser175 of Cdk9, PP1

stimulates its activity (Jerebtsova et al., 2011), although there is now conflicting evidence for the role of Ser175 in the regulation of P-TEFb (Mbonye et al., 2013). Therefore, the possibility that RPAP2 recruits PP1 to genes, where it serves as the CTD Ser5P phosphatase, is examined in the work discussed below.

For the following experiments, two catalytic subunits of PP1, PPP1CA and PPP1CB, were knocked down in HeLa cells using siRNAs (referred to as "PP1 knockdown" or "siPP1", as loss of these subunits was postulated to abolish enzyme activity). Only the PPP1CA antibody was used for the chromatin immunoprecipitation experiments, as the PPP1CB antibody was not suitable for this purpose. ChIP of PPP1CA was performed and its profile on the U2 snRNA and the β-actin genes after RPAP2 and PP1 knockdown by siRNA was analysed (Figure 2-1 A). PPP1CA was found on the U2 gene, peaking towards the 3' end, while the levels on β-actin were at background levels for all primer pairs examined. Due to a lack of primer pairs in the 3' end of the β-actin gene that were suitable for qPCR, it was not possible to determine whether PP1 binds at the β-actin 3' end, or if it is entirely absent from this gene. Therefore, only the U2 gene was considered for subsequent ChIP experiments. After knockdown of RPAP2, Pol II levels decreased throughout the U2 gene body, as previously demonstrated (Egloff et al., 2012b). No such decrease was observed after PP1 knockdown (Figure 2-1 B), suggesting that PP1 itself does not affect transcription elongation.

Following this, PPP1CA levels on U2 after RPAP2 knockdown were analysed. If RPAP2 recruits PP1 to genes, PP1 levels should drop in the absence of RPAP2. Indeed, PPP1CA was diminished after RPAP2 knockdown to levels comparable with a knockdown of PP1 itself, suggesting that RPAP2 is recruiting it to the gene, directly or indirectly.

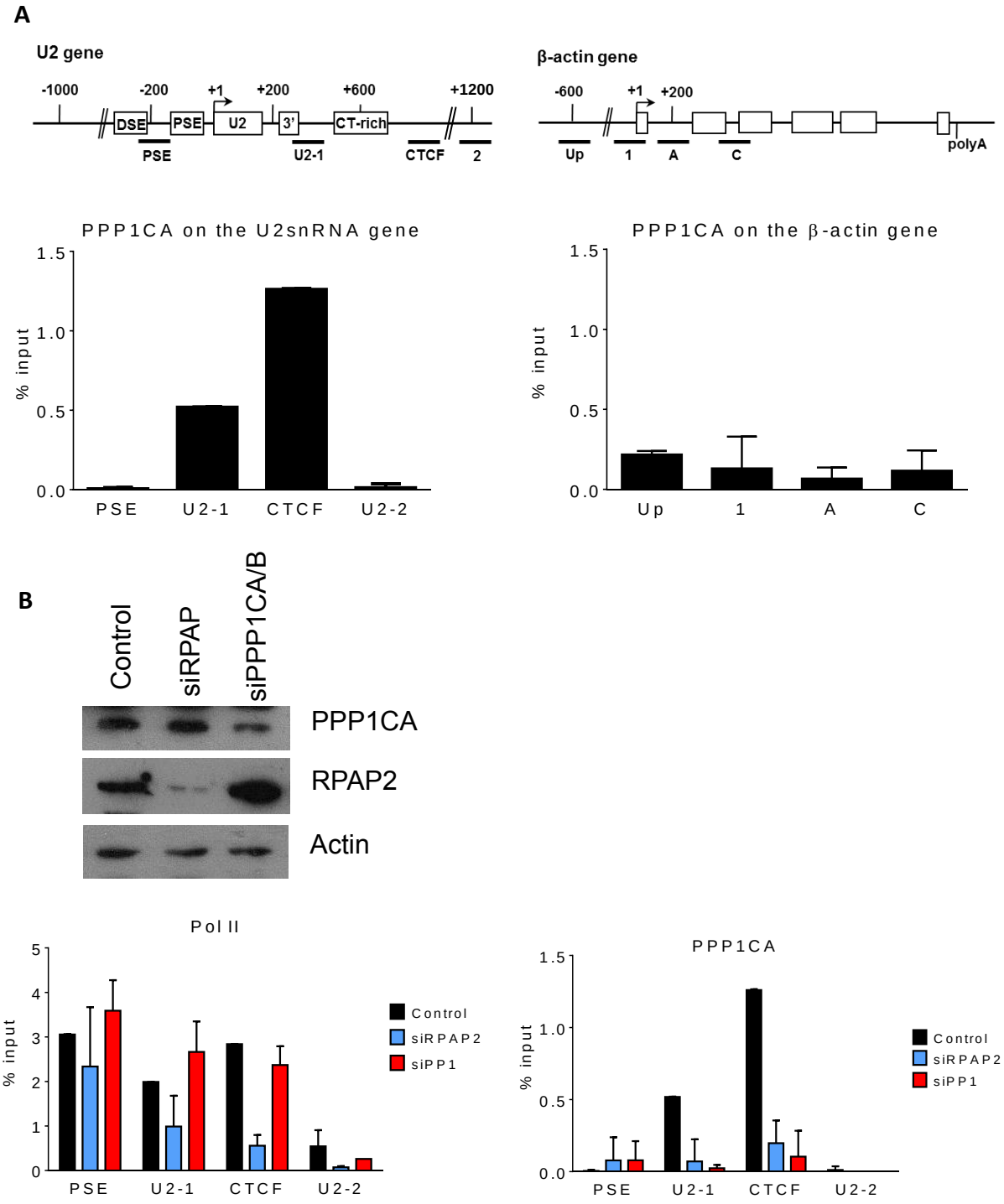


Figure 2-1. PPP1CA recruitment to the U2 snRNA is regulated by RPAP2.

A. ChIP-qPCR of PPP1CA on the U2 snRNA and β-actin genes, using four primer pairs for each gene: at the promoter (PSE), in the gene body (U2-1) and downstream of the open reading frame (CTCF and 2) for the U2 gene, and upstream of the gene (Up), at the TSS (1) and within the gene body (A and C) for the β-actin gene. PPP1CA occupancy is expressed as % input on the bar graphs. B. Western blot demonstrating PPP1CA and RPAP2 knockdown using specific siRNAs, and ChIP profiles of Pol II and PPP1CA on the U2 gene after knockdown of RPAP2 and two PP1 catalytic subunits (PPP1CA and PPP1CB), using the same primer pairs as in A. In all graphs, data are from at least three biological repeats, and error bars indicate the standard deviation.

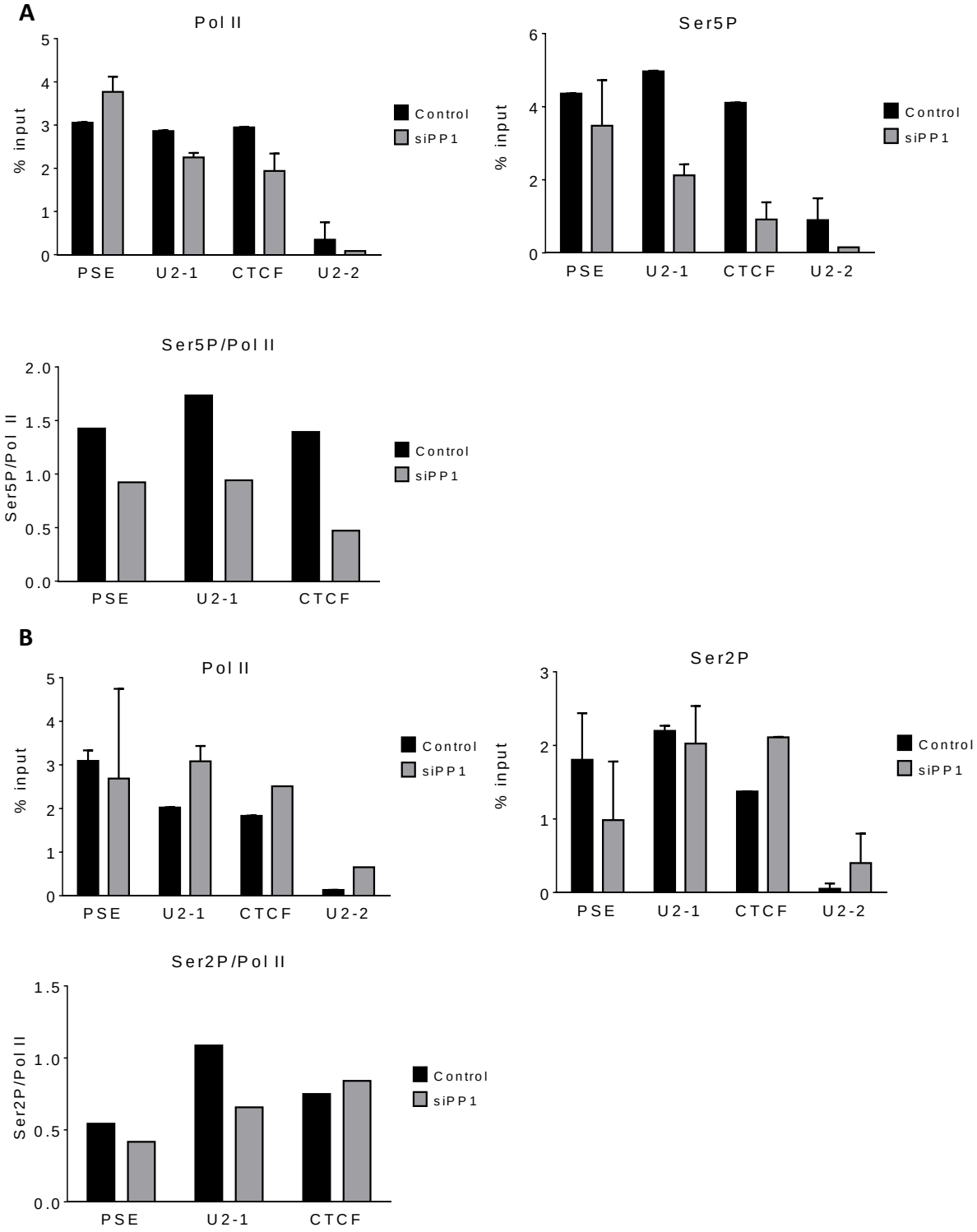


Figure 2-2. Levels of Ser5P and Ser2P on the U2 gene after PP1 knockdown.

A. ChIP-qPCR showing the levels of phospho-Ser5 and B. phospho-Ser2 on the U2 snRNA gene after RPAP2 and PPP1CA/1CB knockdown, using the same sets of primers as in Figure 2-1. Data are an average of three biological repeats for Ser5P and two repeats for Ser2P, and error bars indicate the standard deviation.

Finally, it remained to be confirmed whether PP1 has any phospho-CTD phosphatase activity on the U2 gene. ChIP analysis of Ser5P on U2 showed a decrease after PP1 knockdown, particularly towards the end of the gene. This could be due to the rise in RPAP2 following PP1 knockdown, as seen from the Western blot. Moreover, there was decrease in Ser2 phosphorylation after PP1 knockdown (Figure 2-2). Therefore, it appears that although PP1 is recruited to the U2 snRNA gene by RPAP2, it does not act as a Ser2P or Ser5P phosphatase.

2.2.2 Role of the Integrator complex in transcription termination

So far, the work in this chapter has focused on transcription elongation, and the transition from initiation to elongation regulated by RPAP2. Attention now turns to the termination pathway. At the end of the transcription cycle, the nascent RNA is cleaved, and Pol II disengages from the template (Shandilya and Roberts, 2012). The 3' end of the RNA is then formed by polyadenylation for mRNA, or 3' end processing for snRNA (Albrecht and Wagner, 2012, Bentley, 2014, Egloff and Murphy, 2008, Egloff et al., 2008, Gromak et al., 2006, Yuo et al., 1985). These processes have been extensively studied in budding yeast, *S. cerevisiae*. Termination of transcription of mRNA genes is tightly coupled to 3' end processing by the cleavage and polyadenylation complex, while for termination of sn- or snoRNA gene transcription, the Nrd1-Nab3-Sen1 complex is recruited. However, factors that regulate transcription termination of human genes are less well-understood. While human mRNA genes terminate by a similar pathway to yeast (Egloff et al., 2008, Lunde et al., 2010, Shandilya and Roberts, 2012), there is no equivalent of an Nrd1 complex for snRNA genes. Egloff and others have shown that the Integrator complex, comprising at least 12 subunits in humans, is recruited to snRNA

genes by RPAP2 and cleaves the 3’ end of the nascent snRNA with its catalytic subunits, Int9 and Int11 (Albrecht and Wagner, 2012, Baillat et al., 2005, Egloff et al., 2012b).

Thus, the role of RPAP2 and the Integrator complex in termination of transcription of human genes was further investigated. The work in this section is included in a publication in Nucleic Acids Research ((O'Reilly et al., 2014), and see Appendix 1). First, the effect of several Integrator subunits on the transcription of the β-actin mRNA gene and the U1 and U2 snRNA genes was examined. The two catalytic Integrator subunits (Int9 and Int11), as well as Int5, were separately knocked down by siRNA, and then a ChIP of Pol II was performed on several genes. Figure 2-3 A shows a Western blot of the three Integrator subunits, confirming that all the knockdowns were successful. To assess the difference in the Pol II profile on the β-actin gene, two primer pairs were used: one near the transcription start site (1A), and the other at the site of the Pol II pause after the Poly A site (Pause). After knockdown of all three subunits, there was no difference in Pol II levels, suggesting that Integrator does not regulate transcription of the β-actin gene (Figure 2-3 B).

Then, the U1 and U2 snRNA genes were analysed. After Integrator knockdown and ChIP, five primer pairs were used for each gene in qPCR, spanning the gene body and also regions upstream and downstream (Figure 2-4). For both genes, there is a drop of Pol II in the gene body following Int9 and Int11 knockdown, indicating a transcription defect. Furthermore, after knockdown there is a rise of Pol II on both genes after the 3’ end. This suggests that after loss of the catalytic Integrator subunits, the polymerase reads through past the 3’ end instead of terminating. The same effects are observed to a lesser extent

on U2 but not U1 following Int5 knockdown. Currently, the function of Int5 is not known, but from these results it appears to be also involved in transcription regulation. The ChIP

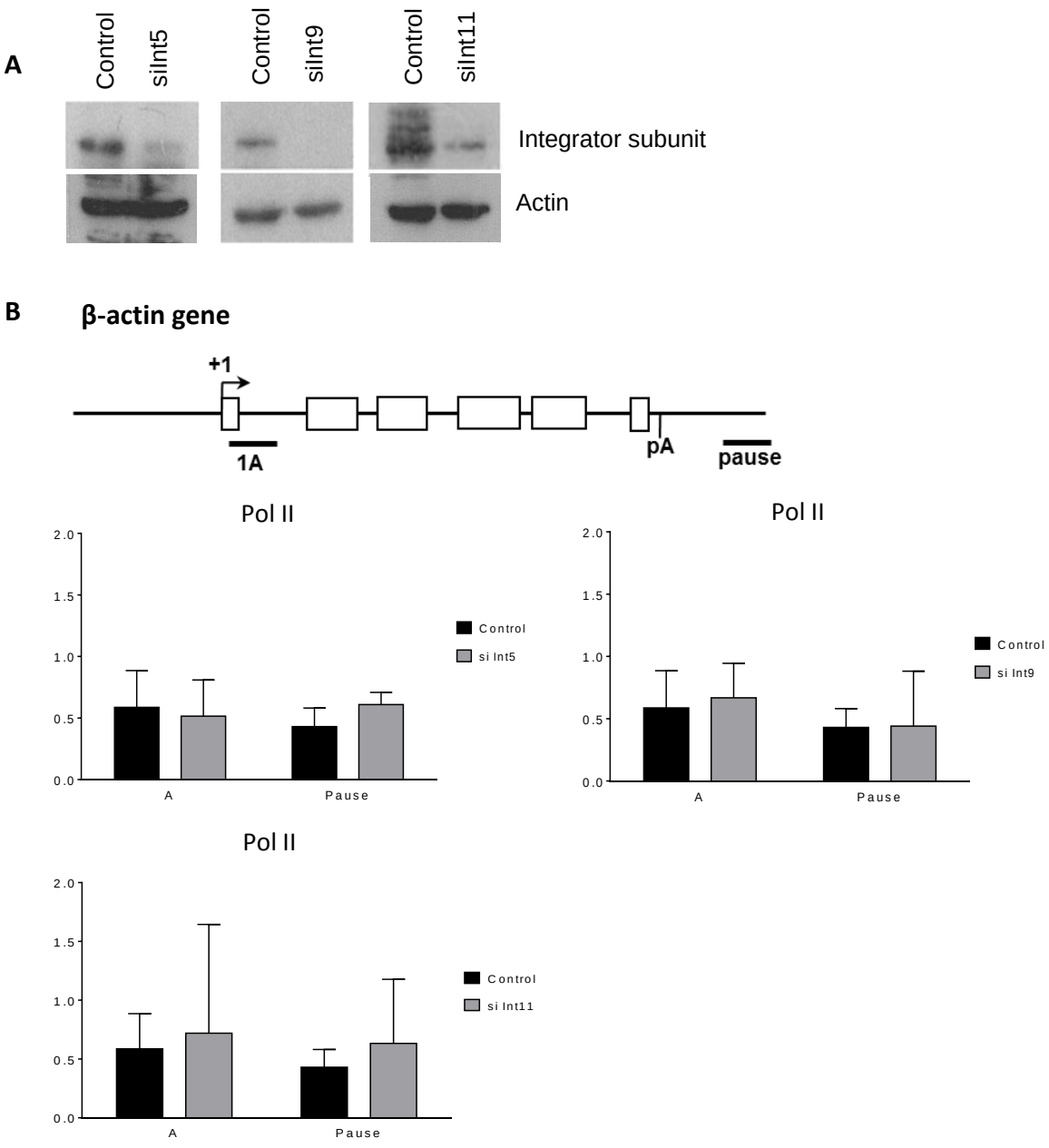


Figure 2-3. Effect of Integrator knockdown on the transcription of the β-actin gene.
A. Knockdown of three Integrator subunits (Int5, Int9 and Int11) in HeLa cells with specific siRNAs. Actin is used as a loading control. B. ChIP profile of Pol II on the β-actin gene after Integrator subunit knockdown, using two primer pairs: at the beginning of the open reading frame (1A) and at the Pol II pause site downstream of the gene (pause). Data are from at least three separate experiments, and error bars indicate the standard deviation.

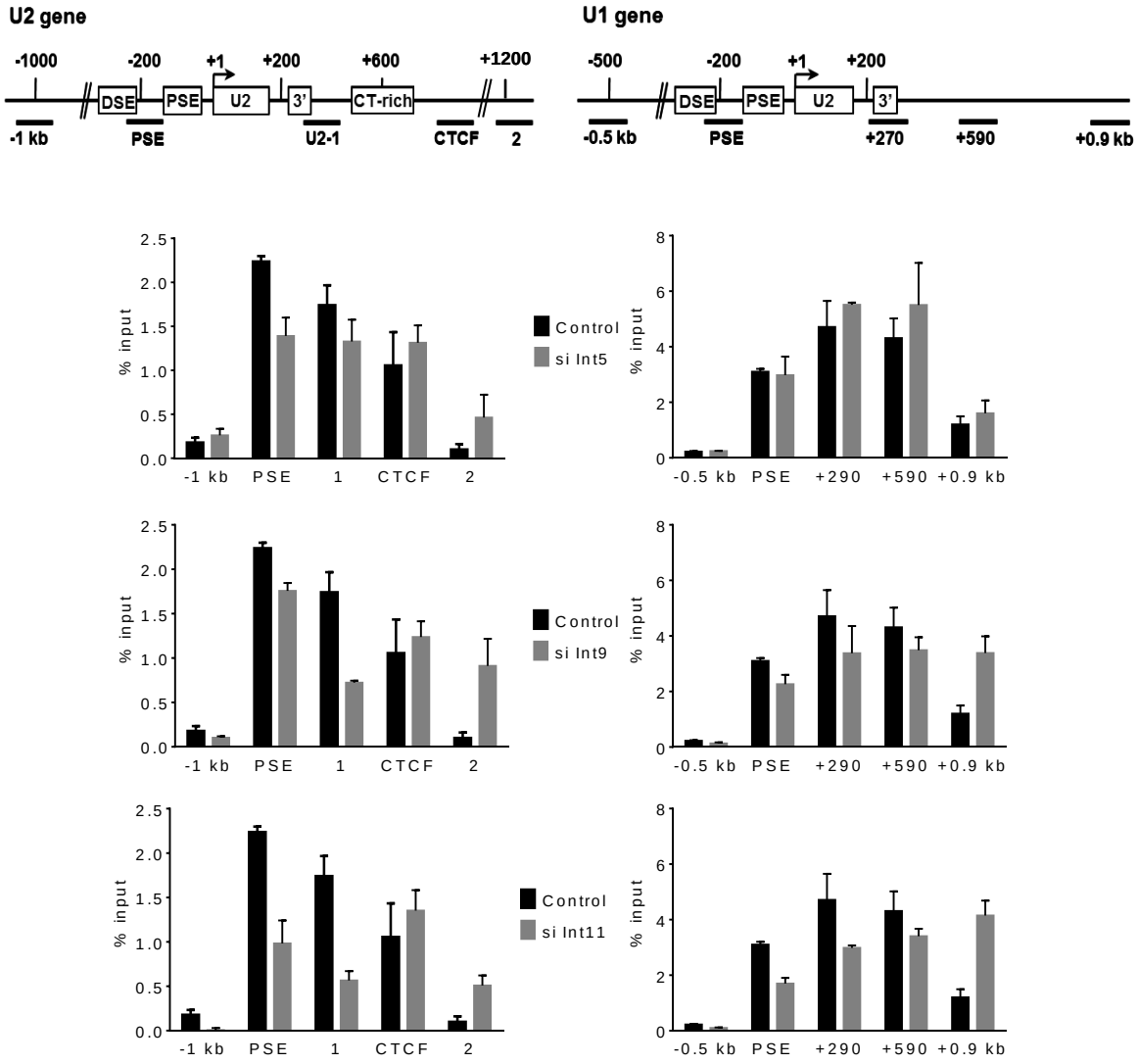


Figure 2-4. Effect of Integrator knockdown on the transcription of the U1 and U2 genes.
ChIP of Pol II on the U2 and U1 snRNA genes after separate knockdown of three Integrator subunits: Int5, Int9 and Int11. The primer pairs used for both genes are shown on the schematic. The data are from at least three biological repeats; error bars indicate the standard deviation.

results therefore suggest that Int9 and Int11, and to a lesser extent Int5, are responsible for regulating transcription termination of human snRNA genes.

To further investigate the role of Integrator in regulation of snRNA transcription, Int5, Int9 and Int11 were knocked down and total RNA was purified from cells. RNase protection analysis (RPA) was then performed using a radioactively-labelled probe spanning the U2 gene and 3' box. All RNAs synthesised from the U2 gene that are present in the cell lysate bind to the complementary probe. RNase is then used to degrade all single-stranded RNA, however probe-target duplexes cannot be degraded, and reflect the size and relative abundance of U2 RNAs in each sample (Figure 2-5 A). If the RNA has been successfully cleaved, it spans only the U2 coding region (labelled pre-U2). Defects in cleavage lead to a read-through (RT) past the 3' box, up to the mismatch at the end of the probe. If the transcription fails to terminate, the polymerase reads into the next U2 repeat (labelled read-around, RA).

As seen from the autoradiograph, shown in Figure 2-5, knockdown of Int9 and Int11 results in a severe processing defect, as evidenced by a rise in RT, and knockdown of all three subunits leads to a termination defect, with an increase in RA. These data, together with the ChIP data, point to a role of Integrator not only in snRNA 3' end processing but also in termination.

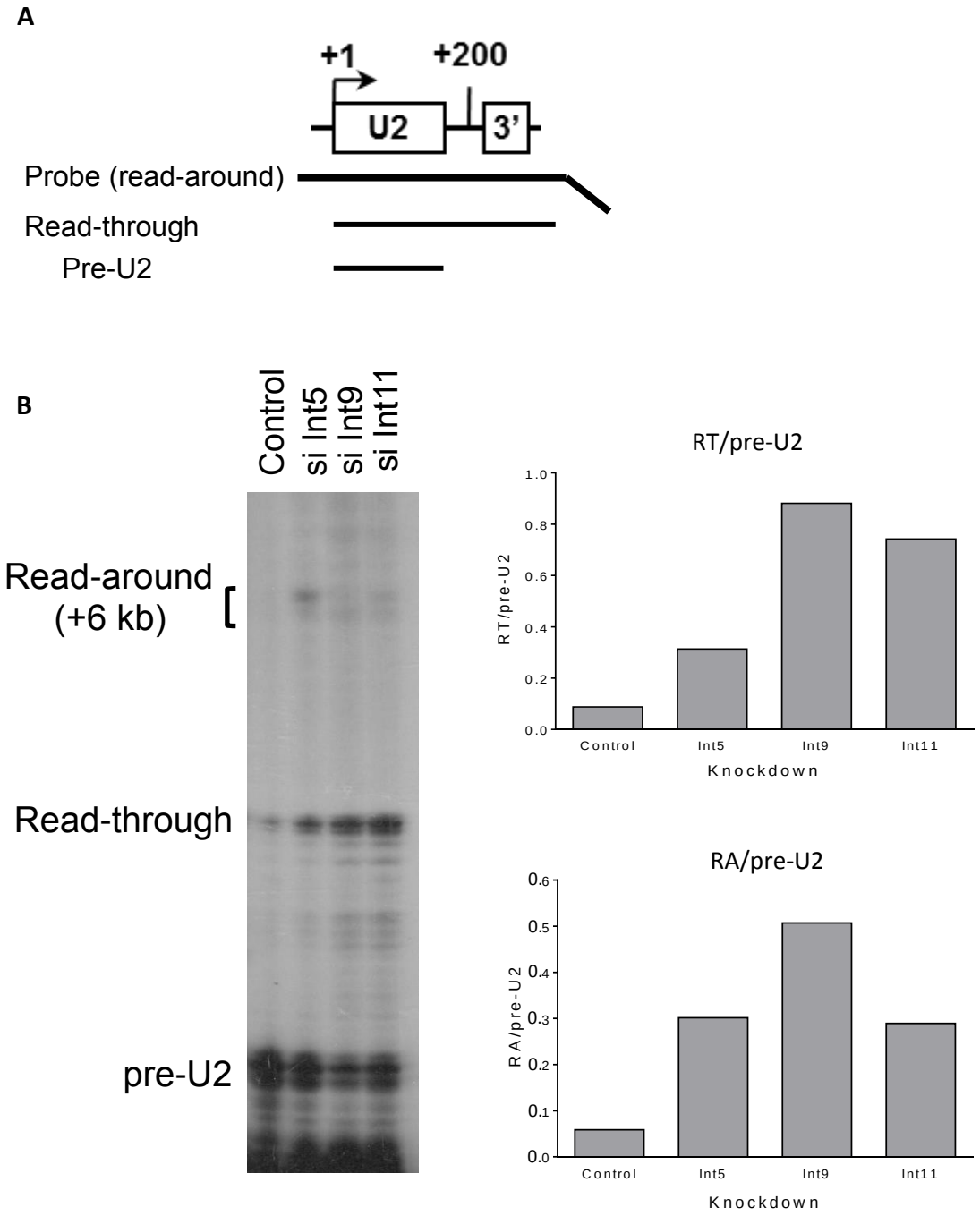


Figure 2-5. RPA on the U2 gene following Integrator knockdown.

A. Radioactive probe used for the RNase protection assay (RPA), spanning the entire region of the U2 gene and 3' box, with a mismatch at the end. Undigested fragments corresponding to read-through and pre-U2 are shown below the probe. B. RNase protection assay of RNA extracts from HeLa cells following knockdown of three Integrator subunits: Int5, Int9 and Int11. The RNase protection was performed by Prof Shona Murphy and normalised to pre-U2 RNA. Termination and processing defects were observed by calculating the Read-around/pre-U2 and Read-through/pre-U2 ratios, respectively.

2.2.3 RPAP2 regulates transcription termination of a protein-coding gene

The role of RPAP2 in promoting transcription elongation by dephosphorylating Ser5P has been shown by several groups, both in yeast and in human. This role is generally accepted, even in light of some conflicting evidence (Xiang et al., 2012). However, the involvement of RPAP2 in regulating transcription may go beyond Ser5. In a recent paper (Hsu et al., 2014), yeast Rtr1 was shown to dephosphorylate CTD phospho-Tyr1 *in vitro*. It is known that in yeast, phosphorylated Tyr1 inhibits the recruitment of termination factors Nrd1, Pcf11 and Rtt103 (Mayer et al., 2012). In human and chicken cells, Tyr1P has a different role in promoting antisense transcription, and is also associated with enhancers (Descostes et al., 2014, Hsin et al., 2014). The work in the present chapter therefore set out to investigate whether RPAP2 also acts as a Tyr1P phosphatase in human cells, and if so, what effect knockdown of RPAP2 would have on transcription.

This hypothesis was tested by inhibiting RPAP2 expression in HeLa cells by siRNA, and performing ChIP of Pol II and Tyr1P. To analyse the Pol II and phospho-CTD profile, a long mRNA gene, KPNB1, for which the Tyr1P profile is known (Laitem et al., 2015) was chosen. Tyr1P peaks at the promoter, then drops throughout the gene body, and peaks again near the polyA site. If RPAP2 is a Tyr1P phosphatase in human as well as yeast, knockdown would result in a rise of phospho-tyrosine.

As seen from the qPCR data, RPAP2 knockdown does not, in fact, lead to an accumulation of Tyr1P but rather a reduction at the beginning and end of the gene (Figure 2-6). Furthermore, the peak of Pol II at the 3’ end of the gene is shifted, suggesting an earlier Pol II pause and possible premature termination. Overall, the data do not support the hypothesis that RPAP2 is a Tyr1P phosphatase in human cells, at least on mRNA genes.

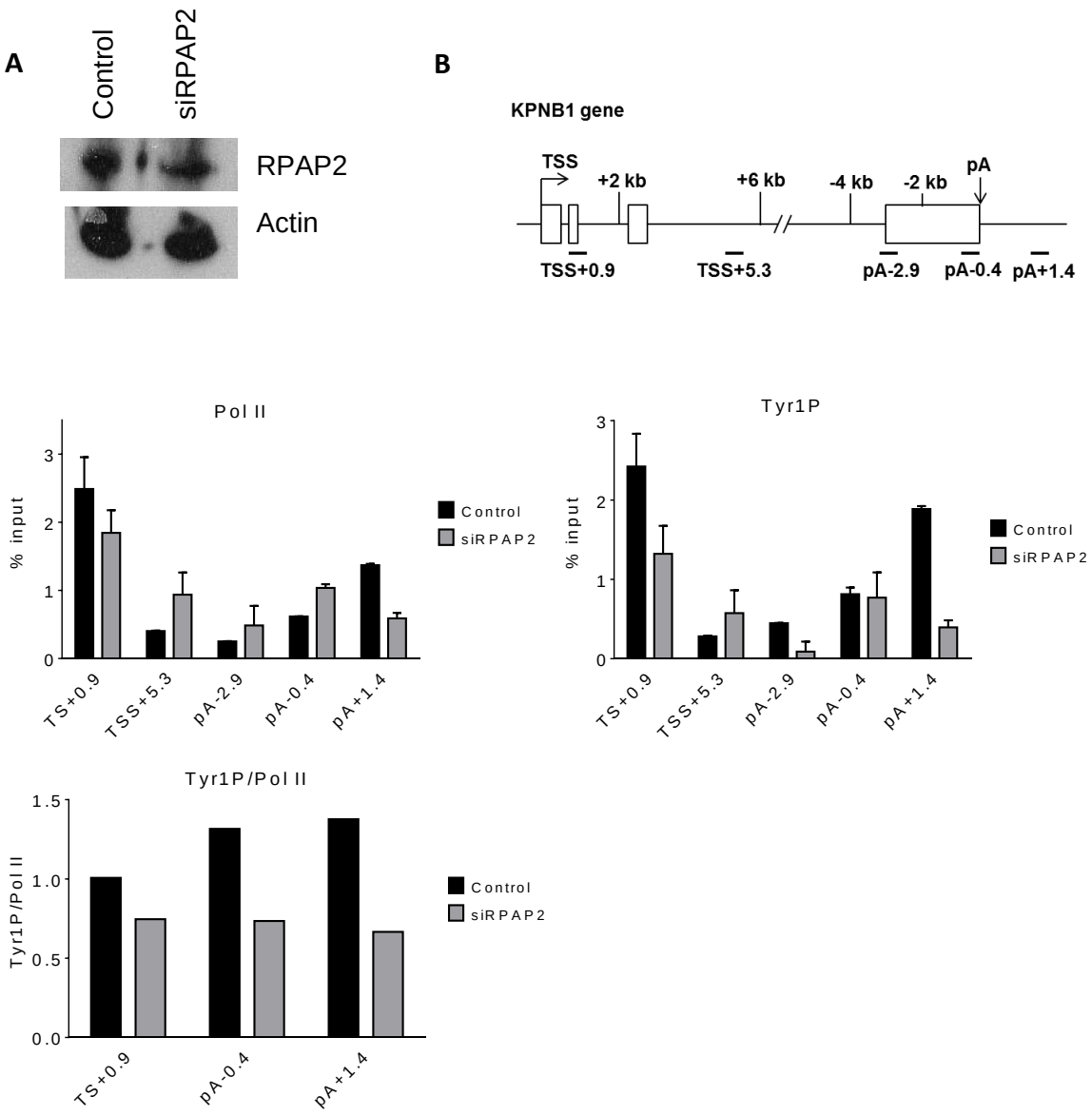


Figure 2-6. ChIP of Tyr1P on the KPNB1 gene after RPAP2 knockdown.

A. Western blot of RPAP2 knockdown in HeLa cells. B. ChIP of Pol II and Tyr1P on the KPNB1 gene after RPAP2 knockdown in HeLa cells, using five primer pairs spanning the whole length of the gene. The data are from two biological repeats, and error bars represent the standard deviation. On the Tyr1/Pol II graph, only the values for the primer pairs at the beginning and at the end of the gene were related, as in the other samples the Pol II levels were too low to give meaningful values.

2.3 Conclusions

This work reported in this chapter examined the effects of several Pol II-binding proteins on Pol II CTD modifications and transcription.

First, it was investigated whether RPAP2 recruits a phosphatase to the Pol II CTD, where it would dephosphorylate Ser5, with a focus on PP1 as a likely candidate. The data show that, while PP1 requires RPAP2 for recruitment to the U2 snRNA gene, it does not function as a Ser5P or Ser2P phosphatase. Furthermore, ChIP analysis reveals that a knockdown of PP1 in fact causes a decrease in Ser5P and Ser2P levels. The previous *in vitro* study that indicated that PP1 has phosphatase activity is not supported by these data; however the results here suggest that PP1 may stimulate the kinase activity of P-TEFb. Moreover, PP1 appears to be a negative regulator of RPAP2, as evidenced by the rise of RPAP2 in the Western blot and a decrease in CTD Ser5P on the U2 gene upon PP1 knockdown.

Next, the function of the Integrator complex on RNA processing and termination of transcription of Pol II-transcribed genes was analysed. Knockdown of the catalytic Integrator subunits followed by ChIP of Pol II shows a transcription defect in the gene bodies of the U1 and U2 snRNA genes, and a termination defect whereby the polymerase continues transcribing past the known transcription unit. These data are in agreement with RNase protection assay results, which show both a 3' end processing and termination defect after Int9 and Int11 knockdown, as indicated by a relative increase of read-through and read-around RNA. This suggests that Integrator is necessary not only for snRNA 3' end processing but also for termination, coupling the two processes. This is the first insight into the termination mechanism on human snRNA genes, and it differs

markedly from the yeast Nrd1-dependent mechanism of termination. Integrator knockdown does not affect Pol II levels on the β-actin gene, consistent with the known mechanism of Pol II pausing followed by Xrn2-mediated termination (Gromak et al., 2006).

Knowing that RPAP2 regulates transcription termination of snRNA genes through recruitment of Integrator, it was examined whether RPAP2 might also regulate transcription termination of mRNA genes. One way in which it could do so is to dephosphorylate phospho-Tyr1 of the CTD, a termination-inhibitory mark, as has been already shown in yeast (Hsu et al., 2014). However, knockdown of RPAP2 followed by ChIP of Tyr1P on KPNB1, a long mRNA gene, showed no increase in tyrosine phosphorylation, suggesting that in humans RPAP2 may not function as a Tyr1 phosphatase. In addition, a transcription defect was observed following RPAP2 knockdown, with a peak of Pol II towards the end of the gene but before the known pause site around 1.4 kb downstream of the PolyA site. This could be indicative of premature termination, which would suggest a role for RPAP2 in regulating mRNA gene termination independent of Tyr1P phosphatase activity.

Chapter 3| Regulation of RPAP2 activity

3.1 Introduction

During a cycle of transcription, the Pol II CTD is modified by a number of enzymes, and the modifications of residues within its YSPTSPS heptapeptide repeats govern the progression of the polymerase along the gene. Serine 5 of the repeat is phosphorylated at the beginning of the transcription cycle by Cdk7 of the TFIIH complex to allow 5' processing factors, for example capping enzymes, to bind (Akhtar et al., 2009, Akoulitchev et al., 1995, Bentley, 2014, Lenasi et al., 2011, Moteki and Price, 2002, Shatkin, 1976). Phospho-Ser5 (Ser5P) can be dephosphorylated by several phosphatases, including Scp1, Ssu72, and the recently-characterised RPAP2 (Rtr1 in yeast) (Yeo et al., 2003, Ganem et al., 2003, Krishnamurthy et al., 2004, Mosley et al., 2009, Egloff et al., 2012b).

A potential role for RPAP2 in regulating Pol II activity was first suggested by the results of an affinity purification-mass spectrometry analysis of the human transcription interaction network. RPAP2, at the time known only by its open reading frame C1orf82, was discovered as a Pol II-interacting protein (Jeronimo et al., 2007). Subsequent studies on both the human RPAP2 and the yeast Rtr1 uncovered several functions of the protein, including dephosphorylation of the CTD Ser5P in humans and recruiting the Integrator complex to enable 3' end processing of nascent pre-snRNAs. The phosphatase activity was confirmed in both *in vitro* assays and ChIP of Ser5P on several human genes. When the crystal structure of the yeast *K. lactis* Rtr1 was solved to 2.5 Å resolution, however, it resulted in controversy. The structure did not have an obvious active site, and *in vitro* phosphatase assays using both yeast and human RPAP2 gave negative results (Xiang et al., 2012). This led to questions about the function of RPAP2, and whether it is simply a scaffold for the 'true' Ser5P phosphatase. However, this view was countered by a follow-

up study, where the protein preparation was washed with EDTA before performing the phosphatase assay, hence removing inhibitory divalent metal ions, which yielded an active protein. Furthermore, the phosphatase activity of the N-terminal domain (DUF) of the protein was shown in a new 2.1 Å crystal structure to be regulated by a C-terminal loop, which lies at the interface with the DUF, forming a crevice where the substrate may bind (Hsu et al., 2014). Hence, Rtr1 and by extension RPAP2 appear to possess phosphatase activity, which is regulated by the protein's structure as well as by external factors.

A recent analysis of the TAP- and Flag-tagged *S. cerevisiae* Rtr1 interactome revealed that Rtr1 interacts mainly with Pol II subunits, along with several other proteins of the transcription machinery, including Leo1 of the PAF complex and Ctk1, the yeast homologue of Cdk12. In fact, the interaction between Rtr1 and Pol II is decreased when Ctk1 is deleted, leading to the suggestion that CTD Ser2 phosphorylation is necessary for Rtr1 recruitment. Moreover, Ser217 of Rtr1 is phosphorylated in a Ctk1-dependent manner, although direct phosphorylation by Ctk1 has not yet been observed (Smith-Kinnaman et al., 2014).

The experiments described in this chapter investigated the regulation of the phosphatase activity of RPAP2 by other factors. In particular, it was interesting to understand what role P-TEFb plays in regulating RPAP2, particularly in light of the necessity of P-TEFb in the *in vitro* phosphatase assay (Egloff et al., 2012b). Furthermore, it is known that proteins of the RPRD family, namely RPRD1A and RPRD1B, are involved in regulating CTD Ser5 phosphorylation (Ni et al., 2011). The present study examined the interaction between

RPRD proteins and RPAP2, and the hypothesis that an RPRD complex recruits RPAP2 to genes in order to stimulate its phosphatase activity.

3.2 Results

3.2.1 The phosphatase activity of RPAP2 may be regulated by P-TEFb

In their effort to characterise RPAP2 function (Egloff et al., 2012b), the authors performed an enzymatic assay to examine phosphatase activity. Prior to dephosphorylation by recombinant RPAP2, the CTD substrate was pre-phosphorylated by either HeLa nuclear extract or purified P-TEFb. However, phosphatase activity was only observed when purified P-TEFb was used, and not nuclear extract (Dr Sylvain Egloff, unpublished observations). It was therefore postulated that P-TEFb may positively regulate RPAP2 activity, and this regulation is lost in nuclear extract due to the presence of other factors or additional CTD modifications. To test this hypothesis, the kinase subunit of P-TEFb, Cdk9, was inhibited by treating HeLa cells with the Cdk9 inhibitor KM05283 (from here on referred to as KM) at a final concentration of 100 μM for 30 minutes. A Western blot of CTD Ser2P shows a reduction after KM treatment, demonstrating that the inhibition was successful and Cdk9 was no longer able to phosphorylate the CTD. No change was observed in the total levels of Pol II or CTD Ser7P, and there was a slight drop in the levels of Ser5P (Figure 3-1 A).

To examine whether P-TEFb regulates the phosphatase activity of RPAP2, chromatin immunoprecipitation and qPCR of Pol II and Ser5P were performed. It is known that on snRNA genes, unlike protein-coding genes, P-TEFb is not necessary for transcription

elongation (Medlin et al., 2005), and therefore its inhibition would not interfere with Pol II progression along the gene. For this reason, the U2 snRNA gene was chosen to examine

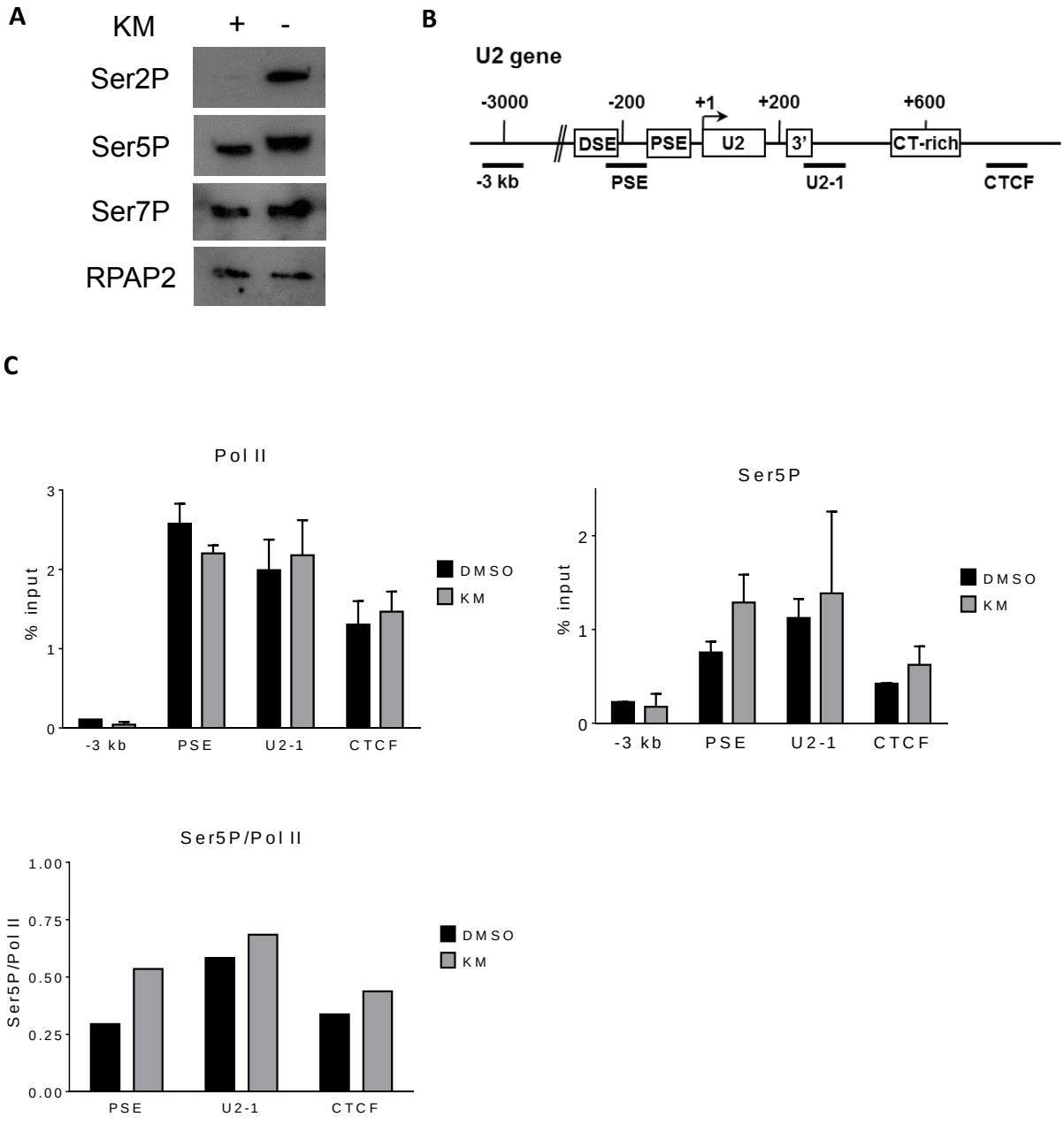


Figure 3-1. The effect of P-TEFb on the phosphatase activity of RPAP2.

A. Western blot of CTD phospho-serines and RPAP2 in HeLa cell extracts following KM treatment. A reduction in phospho-Ser2 confirms that the treatment was successful. B. Schematic of the U2 snRNA gene, showing the location of the primer pairs used. C. ChIP of Ser5P and Pol II on the U2 snRNA gene following KM treatment, using the same primer pairs as shown above. The data are from three biological repeats and error bars indicate the standard deviation.

phospho-Ser5 levels as opposed to an mRNA gene like β-actin. If P-TEFb activates the phosphatase activity of RPAP2, inhibition of Cdk9 is expected to lead to an increase of Ser5P.

The distribution of Pol II and phospho-serines on the U2 snRNA gene is known from previous studies (Egloff et al., 2012a). There is a peak of Pol II occupancy at the DSE, approximately 200 base pairs upstream of the TSS, and its levels gradually drop through the gene body and downstream past the 3' box element. As shown in Figure 3-1 B and C, the Pol II profile on U2 is as expected, and did not change following KM treatment. However, after normalisation for Pol II levels, there was a modest increase in Ser5P especially at the beginning of the gene, indicating a potential regulation.

3.2.2 RPRD1A is present on several human genes

Given that CTD modification requires precise timing to correctly co-ordinate different stages of the transcription cycle, the CTD-modifying enzymes must be subject to regulation themselves to ensure that they are only active at the appropriate time. However, it is still not known which factors regulate the activity of RPAP2, the Ser5P phosphatase. Recently, several proteins have been identified that impact on CTD phospho-Ser5 levels. Of these, RPRD1A and RPRD1B, with RPRD1A known for its roles in cell cycle control (Liu et al., 2002, Wu et al., 2010), have been demonstrated to form a complex that binds directly to Pol II. Phosphorylation of CTD Ser5 and Ser7 at the LEO1 gene promoter is decreased when RPRD1A or RPRD1B are overexpressed (Ni et al., 2011). In collaboration with the group of Prof Jack Greenblatt (Department of Molecular Genetics, University of Toronto, Canada), it was examined whether the RPRD complex

regulates the phosphatase activity of RPAP2. This section of the thesis has been published ((Ni et al., 2014) and see Appendix 2).

To examine the function of the RPRD complex, it was necessary to determine whether it is present on other human genes. A ChIP of the RPRD1A subunit was performed in HeLa cells, and its enrichment analysed on a panel of genes by qPCR. In parallel, Pol II levels on the same genes were also analysed to ensure that they are transcribed above background. Figure 3-2 shows the Pol II and RPRD1A occupancy on the promoters of the RPRD1A gene itself, on the gene for the Mediator subunit MED12, genes coding for components of the Paf elongation complex: LEO1 and CTR9, and on the U2 snRNA gene. The region 300 bp downstream of the TSS of the β-actin gene was also included, as the primers for the β-actin promoter were not good enough to give a reliable signal, while the U2 snRNA gene intergenic region (primer U2.3) was used as a negative control.

All of the genes analysed were transcribed by Pol II, with the highest levels of Pol II at β-actin and the U2 snRNA gene promoters, but RPRD1A was only present above background on LEO1, RPRD1A, MED12 and CTR9, but absent from the β-actin and the U2 snRNA gene promoters.

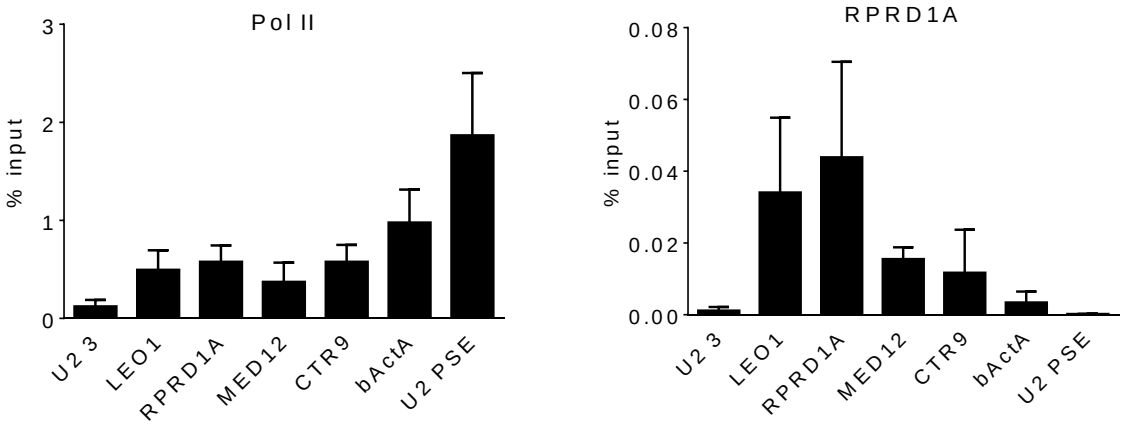


Figure 3-2. ChIP of Pol II and RPRD1A on a range of mRNA and U2 snRNA genes.
For LEO1, RPRD1A, MED12 and CTR9, primer pairs at promoters were used. For β-actin, the primer pairs were 300 bp downstream of the promoter. A U2 intergenic region (primer pair U2.3) was used as a negative control. The results are an average of three biological repeats, with error bars indicating the standard deviation.

3.2.3 Recruitment of RPAP2 to genes

Next, the involvement of RPRD1A in recruiting RPAP2 to *LEO1*, *RPRD1A*, *MED12* and *CTR9* was investigated. RPRD1A was knocked down in HeLa cells using siRNA, with a Western blot confirming that the knockdown was successful (Figure 3-3 A). The enrichment of RPAP2 on the genes was then analysed before and after knockdown. For all four genes, RPAP2 was present at the promoter above background, however in the absence of RPRD1A, RPAP2 dropped to background levels. This strongly suggests that RPRD1A plays a role in recruiting RPAP2 to genes.

It was then postulated that a loss of RPAP2 would lead to a rise in CTD Ser5P at the gene promoters. A ChIP of Pol II showed a small drop in polymerase levels after RPRD1A knockdown, analogous to the reduction seen after RPAP2 knockdown (see results reported in Chapter 2). However, there was a rise in Ser5P on all of the promoters following RPRD1A knockdown (see Ser5P/Pol II graph in Figure 3-3 B), indicating that at

least on some genes, RPRD1A affects the activity of RPAP2 and consequently CTD phosphorylation.

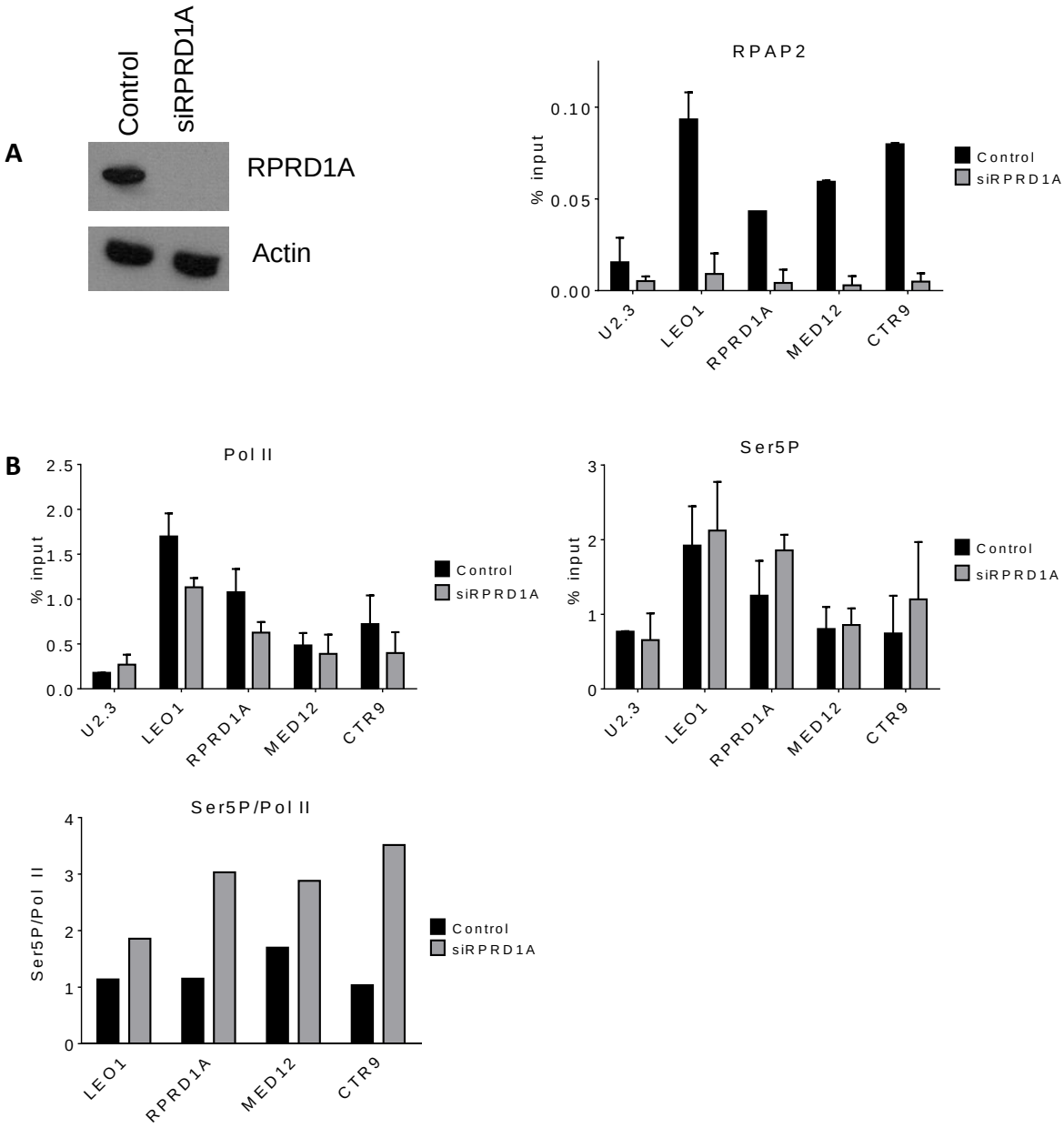


Figure 3-3. RPRD1A recruitment to genes by RPAP2.

A. Western blot of an RPRD1A knockdown in HeLa cells, showing a reduction in RPRD1A levels, and ChIP of RPAP2 on the promoters of the *LEO1, RPRD1A*, *MED12* and *CTR9* gene promoters after RPRD1A knockdown. The intergenic U2.3 primer pair was used as a negative control. B. ChIP of CTD Ser5P after knocking down RPRD1A on a range of the same selection of genes as above. In all cases, the results are an average of three biological repeats, and error bars indicate the standard deviation.

3.2.4 RPRD1B does not impact on CTD Ser5 phosphorylation

It has been shown by Ni *et al.* in their structural analysis (Ni et al., 2014) that RPRD1A forms homodimers and can also heterodimerise with RPRD1B through their coiled-coil domains to form a binding scaffold for RPAP2 on the CTD. Given that RPRD1B can be present in a complex with RPRD1A, it was postulated that RPRD1B may have the same effect on CTD Ser5 phosphorylation as does RPRD1A. To test this, RPRD1B was knocked down with siRNA in HeLa cells and Pol II and Ser5P enrichment were analysed by ChIP on the same gene promoters as for the RPRD1A knockdown, and on an intergenic region of the U2 snRNA gene as a negative control.

After knocking down RPRD1B, a Western blot was performed (Figure 3-4) to confirm that the knockdown was successful. As seen in the Pol II ChIP, polymerase levels drop after RPRD1B knockdown. However, unlike with RPRD1A, there was no significant difference in phospho-Ser5 levels after RPRD1B knockdown on LEO1, RPRD1A and CTR9 promoters, and there was a decrease in phospho-Ser5 on MED12 (Figure 3-4, Ser5P/Pol II graph).

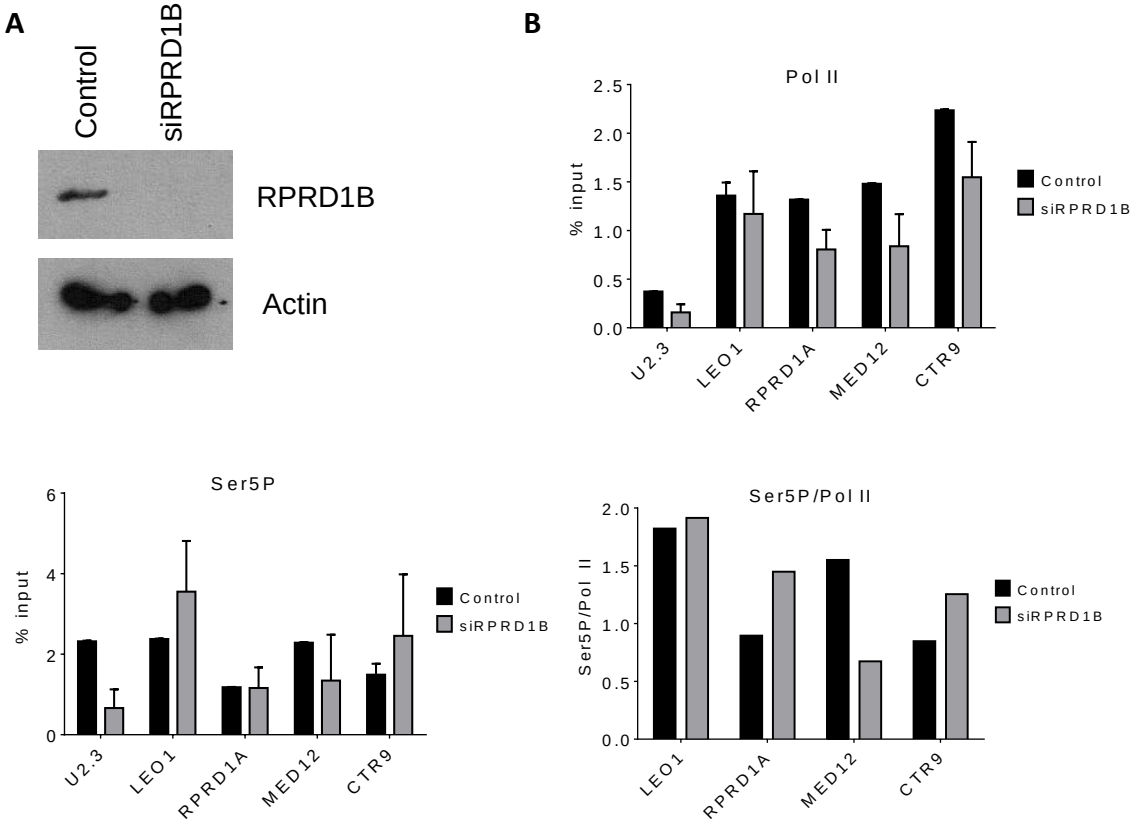


Figure 3-4. RPRD1B does not affect Ser5P levels.

A. Western blot of a knockdown of RPRD1B in HeLa cells, showing a reduction in RPRD1B levels post-knockdown. B. ChIP of total Pol II and CTD Ser5P after knocking down RPRD1B on the LEO1, RPRD1A, MED12 and CTR9 gene promoters. The intergenic U2.3 primer pair was used as a negative control. In all cases, the results are an average of at least three biological repeats, and error bars indicate the standard deviation.

3.3 Conclusions

The experiments described in this chapter investigated the role of several proteins in regulating the phosphatase activity of RPAP2.

In previous work on characterising the activity of RPAP2 *in vitro*, phosphatase activity was only observed when the substrate was phosphorylated with purified P-TEFb, and not HeLa nuclear extract (Dr Sylvain Egloff, unpublished observations). It was therefore postulated that P-TEFb may stimulate RPAP2's activity. To test this hypothesis, the Cdk9 inhibitor KM was used, and Ser5P levels were analysed by ChIP on the U2 gene, as inhibition of P-TEFb does not impact on Pol II elongation on snRNA genes. There was a modest increase in Ser5P following Cdk9 inhibition, suggesting that at least on the U2 snRNA gene, the activity of RPAP2 may be regulated by P-TEFb.

Another possible regulator of RPAP2 is the RPRD1 family of proteins. RPRD1A and RPRD1B are known CTD-interacting proteins, and have been shown to affect Ser5 and Ser7 phosphorylation. A collaboration with another group sought to determine whether the RPRD1 proteins recruit RPAP2 to genes and regulate its phosphatase activity.

The results show that RPRD1A is present at the promoters of several protein-coding genes, however not on the U2 snRNA gene. The levels of RPRD1A on the β-actin gene were low, however they may be higher at the promoter. This was not possible to test due to the lack of suitable qPCR primers. RPAP2 is present on the same genes as RPRD1A except β-actin and U2 snRNA, but its levels drop to background when RPRD1A is knocked down. In the same study in which the above data were published (Ni et al., 2014), Ni *et al*. demonstrated that RPRD1A and RPRD1B bind directly to RPAP2. Taken together, these

results support the idea that RPRD1A recruits RPAP2 to human genes. It can be concluded that during transcription of snRNA genes and protein-coding genes, two different mechanisms are used to recruit RPAP2 – phosphorylation of Ser7 of the CTD for snRNA genes, and RPRD1A and RPRD1B for mRNA genes.

Moreover, knockdown of RPRD1A leads to a rise of Ser5P on some of these genes, most likely through an inhibition of RPAP2 recruitment. On the other hand, knock down of RPRD1B does not have such an effect on Ser5P. It may seem surprising that the loss of RPRD1A can lead to a rise in Ser5P, while loss of RPRD1B has no effect, especially if the two subunits are present in the same complex. However there can, in theory, be three different types of complex: RPRD1A homodimers, RPRD1B homodimers, and RPRD1A-RPRD1B heterodimers. The binding affinities for non-canonical CTD repeats are different for the three types of complex (Ni et al., 2014), and it is not known on which CTD repeats RPRD1A mediates RPAP2 recruitment. It is also not known which combinations of RPRD1A and RPRD1B dimers are present on which genes, and whether there is redundancy in the oligomerisation. It is therefore possible that on the genes analysed in this section, RPRD1A is the essential subunit, while loss of RPRD1B can be compensated for by RPRD1A homodimerisation.

Chapter 4| Structural and functional analysis of human RPAP2

4.1 Introduction

Having uncovered some aspects of human RPAP2 function, attention now turns to its possible structure. From sequence analysis alone, RPAP2 and its yeast homologue Rtr1 do not have any known phosphatase motifs and, despite *in vitro* phosphatase activity, there has been controversy as to whether or not RPAP2 dephosphorylates the Pol II CTD.

The first crystal structure of Rtr1 from *Kluyveromyces lactis* was solved by Xiang and colleagues to 2.5 Å resolution ((Xiang et al., 2012) and Figure 1-8 B in the Introduction). The structure was shown to be mostly alpha-helical and to contain a metal-binding site, however it lacked a number of C-terminal residues and did not display an obvious active site. Consequently, the authors concluded that Rtr1 is not, in fact, a phosphatase. In a more recent publication, Hsu and colleagues analysed both the *K. lactis* and *S. cerevisiae* Rtr1, and demonstrated that the protein is an active phosphatase, with activity depending on an EDTA wash step during the purification process (Hsu et al., 2014). The 2.1 Å *K. lactis* Rtr1 crystal structure (Figure 1-9 in the Introduction) revealed an auto-inhibitory region in the C-terminal portion of the protein, which was absent in the structure solved by Xiang *et al*. There now appears to be sufficient evidence for the phosphatase activity of yeast Rtr1, based on the results of *in vitro* enzymatic assays, ChIP, and the most recent crystal structure.

At 612 amino acids, the human RPAP2 is substantially larger than the yeast protein (226 amino acids in *S. cerevisiae*, 211 in *K. lactis*). Although there is a high degree of conservation between different species in the first 150 amino acids, the rest of the human protein is far less conserved, and its function is unclear (Figure 1-7 in the Introduction). Previous studies by Egloff and colleagues (Egloff et al., 2012b)

demonstrated phosphatase activity that localises to the N-terminal half of the human protein. It is therefore predicted that the N-terminal domain of the human RPAP2 would have a similar structure to the yeast protein. What is not clear at present is how RPAP2 binds the CTD, for which a crystal structure of RPAP2 in complex with a CTD substrate would be needed.

The experiments described in this chapter were aimed at investigating the possible structure and activity of human RPAP2. First, bioinformatic analyses were used to predict the domain architecture and secondary structure of RPAP2, beyond the first 200 amino acids that have already been studied by x-ray crystallography in Rtr1. Then, a short RPAP2 construct comprising only the first predicted globular domain was expressed in bacteria and purified using GST affinity purification and ion exchange chromatography. Unfortunately, an attempt to crystallise the protein was unsuccessful, as the resulting crystals appeared to originate from a small molecule rather than protein. Finally, a phosphatase assay was performed to determine whether the recombinant truncated RPAP2 had phosphatase activity.

4.2 Results

4.2.1 Bioinformatic prediction of human RPAP2 domains and secondary structure

In order to understand which regions of a protein form globular domains that would more likely form crystals, domain and secondary structure prediction algorithms can be employed. For this study, three platforms were used to perform such predictions and in every case, the input was the primary sequence of RPAP2.

Firstly, GlobPlot (http://globplot.embl.de) was utilised. GlobPlot uses the calculated propensities of each amino acid to be either in a disordered or globular region of a protein, and from this calculates the propensity of a particular region of a given protein to be ordered or disordered based on its amino acid content (Linding et al., 2003). The results for RPAP2 are presented in Figure 4-1 A. The propensity sums for three large regions of the protein have negative slopes, suggesting the presence of three globular domains, with disordered regions present for the first ~20 amino acids, amino acids 200-235, and the region between amino acids 400 and 500.

To confirm the results obtained with GlobPlot, the IUPred algorithm was also used (http://iupred.enzim.hu/). In contrast to GlobPlot, IUPred relies on the principle that ordered proteins or protein domains are stabilised by an extensive network of inter-residue interactions, while disordered proteins or regions are not. The propensity of each amino acid to form interactions with neighbouring residues is calculated, and ordered regions elucidated. According to the prediction (Figure 4-1 B), RPAP2 contains three globular domains, with disordered regions in the first ~40 amino acids, then around residue 200-320, and finally between residues 400-490.

Although the two programs used different parameters to calculate disorder, the predictions were similar, giving more confidence in the result. Overall, it appears that human RPAP2 contains three ordered domains of similar sizes, each making up approximately a sixth of the protein. The next stage was to estimate the secondary structure of the protein, and this was carried out using the JPred server (http://www.compbio.dundee.ac.uk/www-jpred/). JPred compares the protein sequence both to sequences of proteins with known structures within the PDB, and to sequence

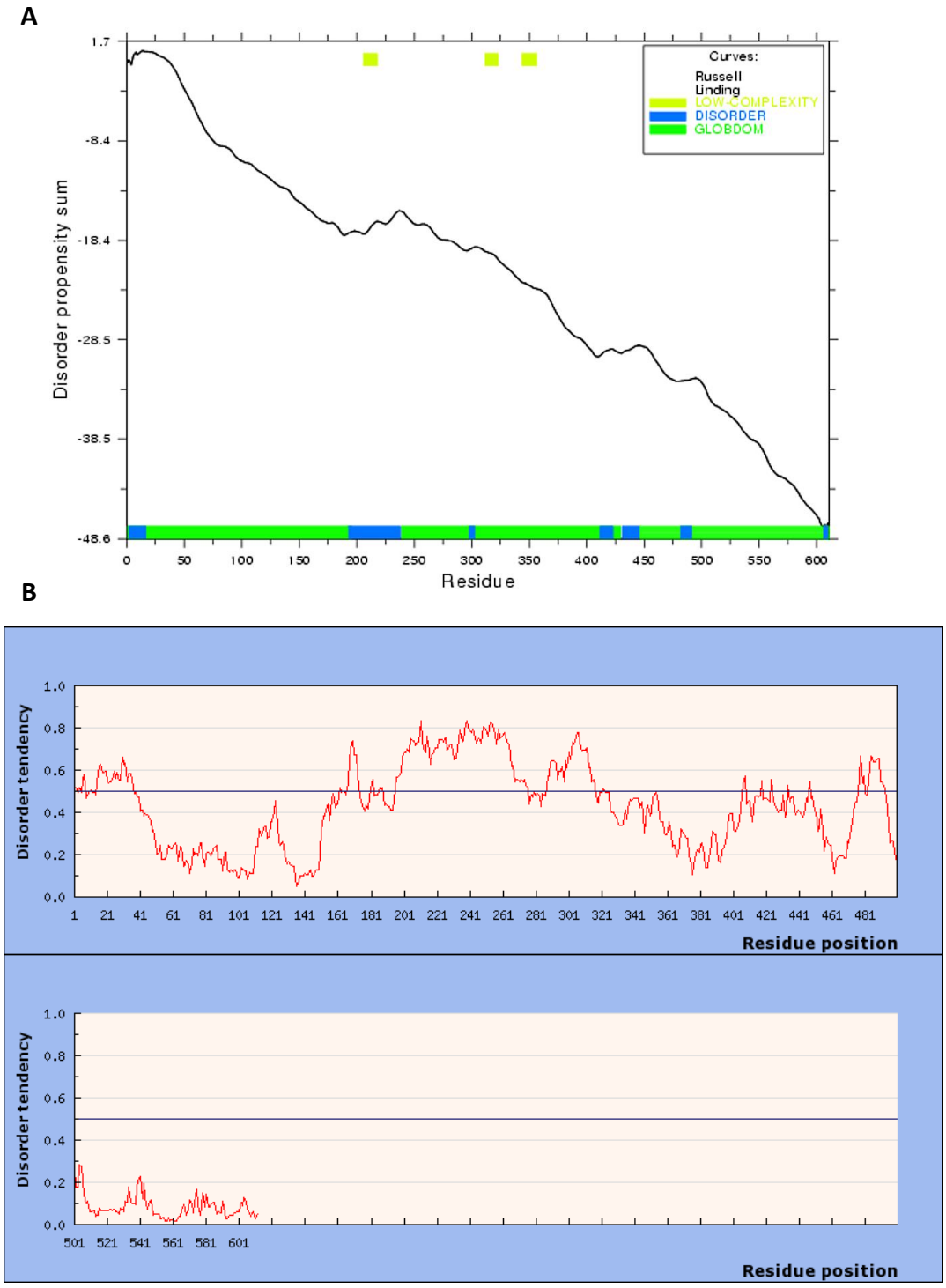


Figure 4-1. Prediction of disorder from the human RPAP2 amino acid sequence.
Prediction from Globplot. The *x*-axis denotes the residue number, while the *y*-axis shows the disorder propensity sum. The regions at which the line has a positive slope are predicted to be disordered (also shown in blue), while those where the line has a negative slope are predicted to be globular domains. B. Prediction from IUPred. Here, the *y*-axis shows the disorder tendency, with values above 0.5 implying that the region is disordered, and values below 0.5 indicating a globular domain.

clusters in UniRef90. It subsequently uses position-specific scoring matrices and hidden Markov models to predict secondary structure (Cole et al., 2008).

The secondary structure is predicted (Figure 4-2) to be mostly composed of alpha-helices, which are present in the regions between amino acids 40-100, 330-460, and 500-612. Several short beta-strands are also predicted to be present throughout the length of the protein. This secondary structure distribution correlates with the globular domain prediction from GlobPlot and IUPred. In short, human RPAP2 is likely to be composed of three globular regions, each exhibiting mostly alpha-helical structure.

The predictions described so far were carried out before the publication of the *K. lactis* Rtr1 structure by Xiang *et al.* (Xiang et al., 2012). Because the N-terminal portion of the protein is very well-conserved between species, it is now possible to compare the predicted structure of human RPAP2 to the known structure of yeast Rtr1. Looking at Figure 1-8 A in the Introduction, the alpha-helices lie in the regions corresponding to the human RPAP2 residues 32-40 (αA), 51-67 (αB), 73-80 (αC), 84-96 (αD), 138-150 (αE), and 179-191 (αF). In contrast, JPred predicts helices in the residues 37-65 (αA), 72-79 (αB), 89-97 (αC), and 137-150 (αD). The first helix predicted by JPred is shown as two separate helices in the *K. lactis* structure, but otherwise the location and length of the helices are well-predicted. The beta-strands predicted by JPred are, in fact, long loops in the *K. lactis* structure. Overall, all the predictions carried out appear consistent with each other and with the published yeast structure.

```
1    MADFAGPSSAGRKAGAPRCSRKAAGTKQTSTLKQEDASKRKAELEAAVRKKIEFERKALH
     -----------------------------------HHHHHHHHHHHHHHHHHHHHHHHH

61   IVEQLLEENITEEFLMECGRFITPAHYSDVVDERSIVKLCGYPLCQKKLGIVPKQKYKIS
     HHHHH------HHHHHHHH--------HHHHHHHHHH------------------EEEE

121  TKTNKVYDITERKSFCSNFCYQASKFFEAQIPKTPVWVREEERHPDFQLLKEEQSGHSGE
     E-----E--HHHH---HHHHHHHHHHHHHH----------------EEE-----------

181  EVQLCSKAIKTSDIDNPSHFEKQYESSSSSTHSDSSSDNEQDFVSSILPGNRPNSTNIRP
     -HHHHH------------------------------------------------------

241  QLHQKSIMKKKAGHKANSKHKDKEQTVVDVTEQLGDCKLDSQEKDATCELPLQKVNTQSS
     -------------------------EEE--------------------------------

301  SNSTLPERLKASENSESEYSRSEITLVGISKKSAEHFKRKFAKSNQVSRSVSSSVQVCPE
     ------------------------------------------------------------

361  VGKRNLLKVLKETLIEWKTEETLRFLYGQNYASVCLKPEASLVKEELDEDDIISDPDSHF
     -----HHHHHHHHHHHHHHHHHHEEE---------------------------HHHHH

421  PAWRESQNSLDESLPFRGSGTAIKPLPSYENLKKETEKLNLRIREFYRGRYVLGEETTKS
     HHHHHH-----------------------HHHHHHHHHH--HEE----EEEE-------

481  QDSEEHDSTFPLIDSSSQNQIRKRIVLEKLSKVLPGLLVPLQITLGDIYTQLKNLVRTFR
     ----------------HHHHHHHHHHH-HHHHHHHHHHH------HHHHHHHHHHHH--

541  LTNRNIIHKPAEWTLIAMVLLSLLTPILGIQKHSQEGMVFTRFLDTLLEELHLKNEDLES
     -----EEE----HHHHHHHHHHHH----HHHHHHH---HHHHHHHHHHHH----HHHHHH

601  LTIIFRTSCLPE
     HHHH--------
```

Figure 4-2. Prediction of the human RPAP2 secondary structure by JPred.
The primary sequence of the full-length human RPAP2 is shown, with an indication of the residue number. Under the sequence, the prediction of secondary structure is shown. Alpha-helices are shown as pink 'H', while beta-sheets are shown as yellow 'E'. Unstructured regions are indicated by dashes.

4.2.2 Expression and purification of GST-tagged human RPAP2

To obtain a crystal structure, a recombinant protein is typically expressed in bacteria and then purified to obtain a highly-concentrated (around 10 mg/ml), homogenous sample. A truncated construct of RPAP2 in a pGEX-6P-1 vector was chosen for expression. The construct contains a GST tag at the N-terminal end, which can be cleaved using PreScission protease. The construct, named RPAP2 RV, corresponds to amino acids 1-194 and had been cloned by Dr Sylvain Egloff by cutting the full-length cDNA at the EcoRV site 582 base pairs from the first codon (Dr Sylvain Egloff, unpublished). It contains the DUF408 domain, which had previously been shown to contain a metal-binding site and to be required for CTD binding (Egloff et al., 2012b), and comprises the first globular domain of the protein, as predicted by the bioinformatics analyses. For these reasons, the construct was considered to be a good candidate for purification.

The plasmid DNA was transformed into Rosetta™ *E. coli* cells, and when the bacteria reached the logarithmic growth phase, protein expression was induced with IPTG. After harvesting, the bacteria were sonicated to release the protein, which was separated from the bacterial debris by ultracentrifugation and then incubated with glutathione-sepharose beads to bind the GST tag and separate the recombinant RPAP2 from the rest of the bacterial proteins. The beads were washed with several concentrations of salt to remove non-specifically bound proteins, and finally GST-RPAP2 was eluted with 25 mM glutathione (GSH). The different stages of purification are shown on the Coomassie gel and RPAP2 Western blot in Figure 4-3.

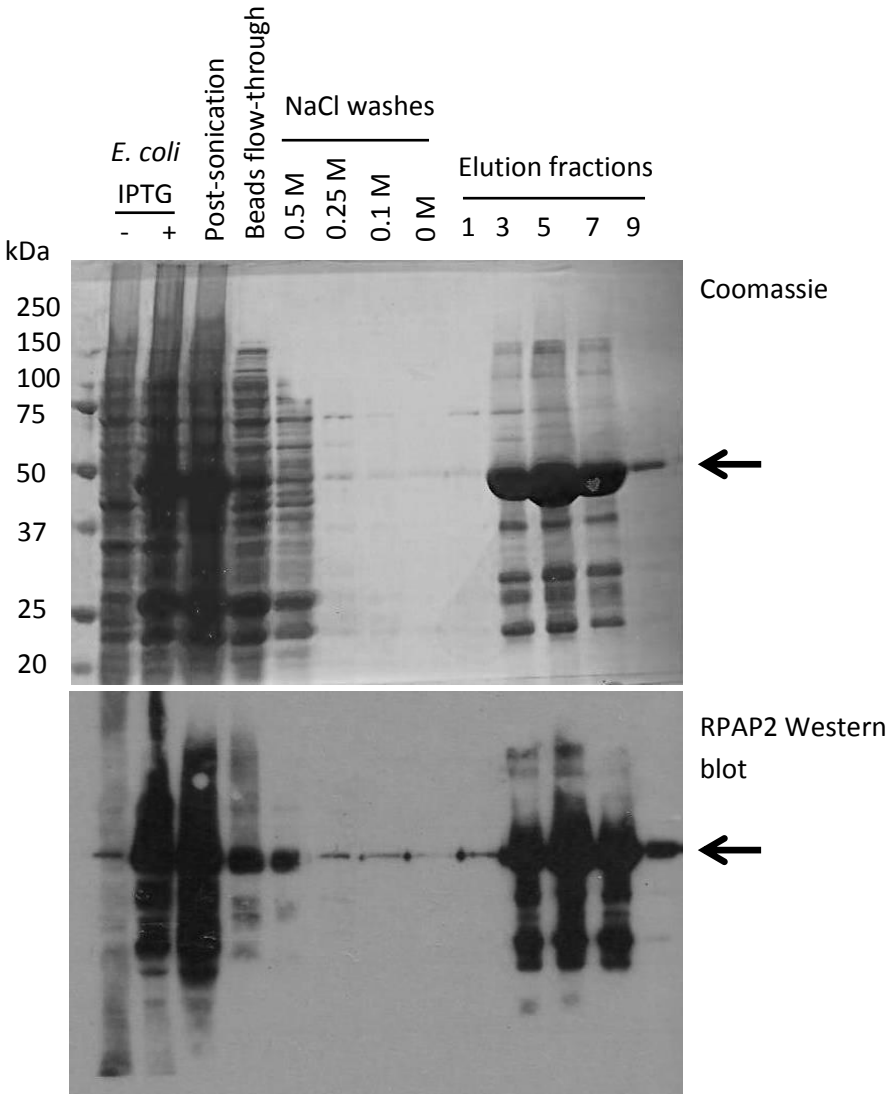


Figure 4-3. Coomassie stain and Western blot of GST-RPAP2 RV purification.

Top: Coomassie stain gel of the different stages of purification. Bottom: RPAP2 Western blot of the same samples. On both gels, the black arrow points to the GST-RPAP2 protein at approximately 50 kDa. The lanes, from left to right, show: Rosetta™ *E. coli* transformed but pre-induction; Rosetta™ post-induction with IPTG; bacterial lysate post-sonication; flow-through of non-bound protein after incubation with glutathione-sepharose beads; four salt washes of the beads at decreasing concentrations of NaCl; and five non-consecutive elution fractions with 25 mM GSH.

There is a strong induction of protein expression after addition of IPTG to the bacteria, and the protein was retained in the soluble fraction after sonication. After incubation with glutathione-sepharose beads, most of the protein was eluted in fractions 3-7, and was then pooled and dialysed to remove contaminating GSH.

Despite a strong expression and successful first purification, at approximately 200 mg protein purified from a 500 ml bacterial culture, contaminants remained in the sample (see “elution fractions” lanes in Figure 4-3). Many of these bands, particularly ones of lower molecular weight, may correspond to degradation products, since they were detected with the RPAP2 antibody. The resulting sample was not of sufficient purity for crystallisation trials, so a second step of purification was introduced. However, prior to this, GST-RPAP2 RV was incubated overnight with PreScission protease to cleave the GST tag. The sample was then once again incubated with glutathione-sepharose beads to bind the free GST, and the flow-through collected and concentrated. This flow-through was expected to contain mostly RPAP2 RV, potentially with some contamination from GST and other proteins. RPAP2 RV and free GST are of almost identical molecular weight (both approximately 25 kDa), so it was not possible to separate the two proteins based on size. Instead, they were separated using ion exchange chromatography.

Ion exchange chromatography is used to separate proteins based on their charges. At a fixed buffer pH, different proteins will carry a different net charge depending on their isoelectric point (pI). The pI of free GST is 5.84, while the predicted pI of RPAP2 RV is 8.88 (calculated using the ProtParam tool on ExPASy). The protein sample to be purified was dialysed into a buffer of pH 8.0, at which RPAP2 is expected to have a net positive charge, and GST a net negative charge. The sample was passed through a MonoQ column (Figure

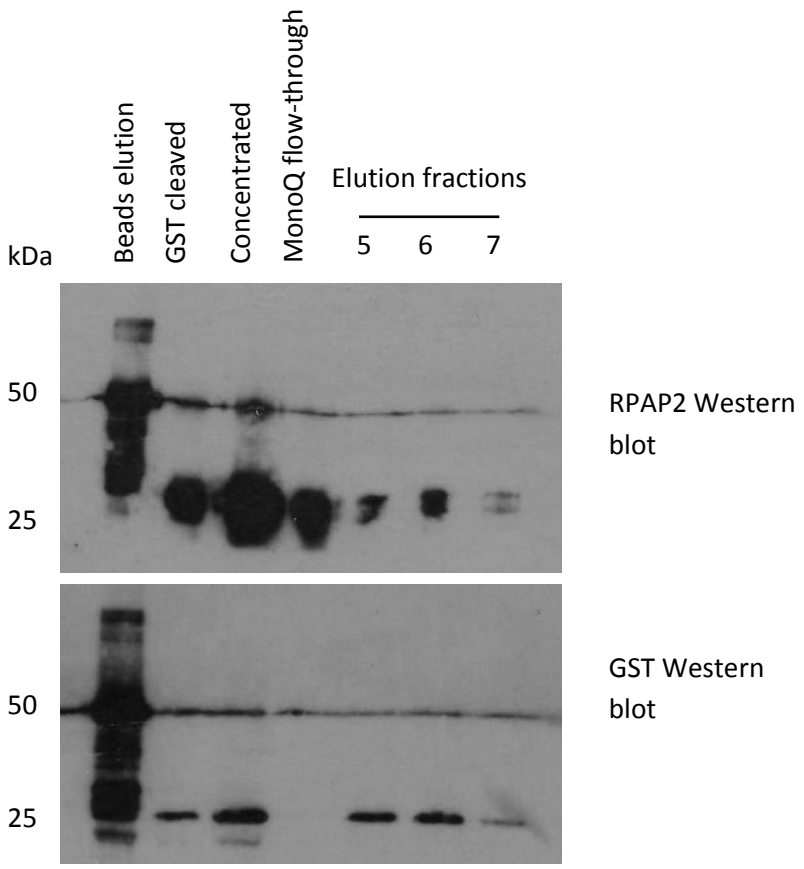


Figure 4-4. Western blots of RPAP2 and GST after MonoQ purification.
Top: Western blot of several stages of protein purification detected with an RPAP2 antibody. Bottom: the same samples detected with a GST antibody. The lanes, from right to left, are: eluate of GST-RPAP2 from glutathione-sepharose beads; the eluate after treatment with PreScission protease to cleave the GST tag; the same sample concentrated; flow-through of unbound protein from the MonoQ column; three elution fractions (numbered 5,6,7) from the MonoQ column. The bands at 25 kDa correspond to the cleaved RPAP2 RV and GST.

4-4). As an anion exchanger, the MonoQ was expected to bind GST but not RPAP2. Indeed, RPAP2 was collected in the flow-through without binding to the column, while no GST was detected in this fraction.

Before putting the RPAP2 RV sample, concentrated to 10 mg/ml, into crystallisation trials, its thermal stability was measured by differential scanning calorimetry (DSC). The sample

is subjected to a temperature gradient, and its heat capacity is measured. This allows the melting temperature of the protein, T_m to be determined and therefore predict its stability during crystallisation. The DSC measurements were carried out by Dr David Staunton at the Biophysical Instrument Facility, Department of Biochemistry, University of Oxford. The spectrum is shown in Figure 4-5. The protein appears stable over a wide temperature range up to 75°C, after which it denatures. Although the peak was not exceptionally well-defined, the protein was deemed of sufficient quality for crystallisation trials.

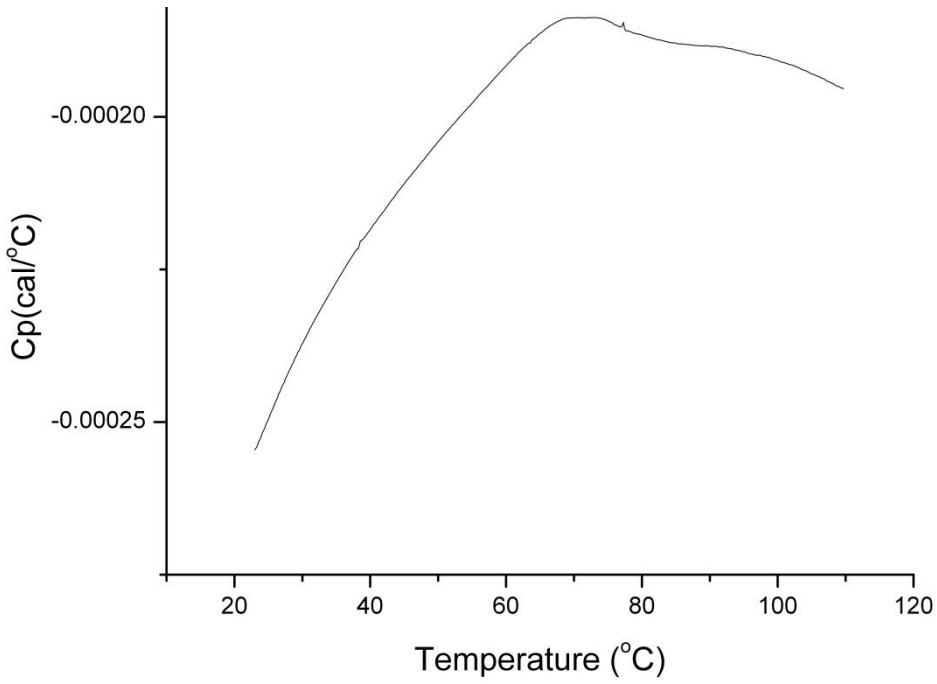


Figure 4-5. Differential scanning calorimetry spectrum of RPAP2 RV.
The *x*-axis shows the temperature range to which the protein sample was subjected; the *y*-axis shows the heat capacity. The T_m of the protein is at the peak of the curve, approximately 75°C.

4.2.3 Crystallisation trials with RPAP2 RV

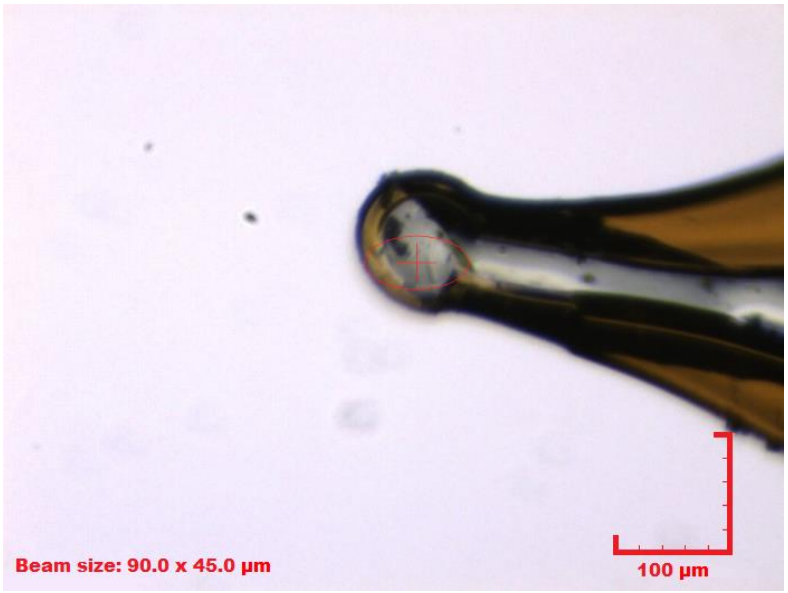
After the RPAP2 RV protein had been purified and tested for thermal stability, it was concentrated to 10 mg/ml and put into crystallisation trials. JCSG+, PACT, and MORPHEUS plates were set up. After 4 weeks, crystals of approximately 40 µm in length were found in two wells of the PACT screen.

One crystal was harvested, cryocooled in 30% ethylene glycol and 70% reservoir buffer, and analysed by X-ray diffraction at the I04 beamline of the Diamond Light Source synchrotron with an X-ray wavelength of 0.98 Å and rotation angle of 0.5 degrees. Images were collected on a Pilatus 6M detector, with a detector distance of 402 mm. The crystal preparation and data collection were carried out by Dr Edward Lowe, Department of Biochemistry, Oxford, UK. The diffraction pattern shown in Figure 4-6 is of poor quality, and the reflections more likely correspond to a small molecule than a protein crystal. The most probable explanation for this is contamination with the glutathione that was used to elute GST-RPAP2 from the glutathione-sepharose beads. Although the protein was subsequently dialysed to remove contaminants, some may have remained.

4.2.4 Functional analysis of RPAP2 RV

To test whether the purified RPAP2 RV is an active phosphatase, a phosphatase assay was carried out as described in (Egloff et al., 2012b). Egloff and colleagues determined that the smallest RPAP2 construct that had phosphatase activity contained amino acids 1-334, while the 1-167 construct lacked phosphatase activity. RPAP2 RV is larger than the 1-167 protein and therefore is more likely to contain the residues necessary for activity.

A



B

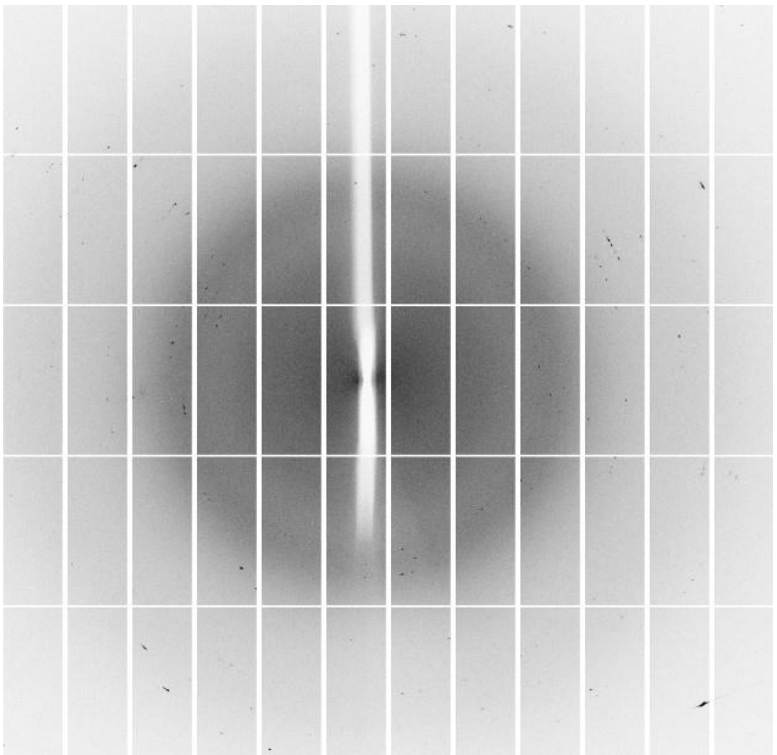


Figure 4-6. X-ray diffraction pattern of the RPAP2 RV sample.

A. Image of a crystal in the cryoloop. B. Diffraction pattern of the crystal shown in A.

A recombinant GST-CTD containing 25 heptapeptide repeats was phosphorylated with recombinant Cdk9-Cyclin T1 using ATP. After this, the ATP was removed in a desalting column, and the substrate was incubated with increasing quantities of RPAP2 RV from 0 to 10 pmoles. The sample was run on an SDS-PAGE gel for a Western blot (Figure 4-7 A). None of the phosphoserine levels change with an increase of RPAP2, suggesting that the purified RPAP2 was not active. A positive control using the non-specific lambda phosphatase (λ PP) was introduced (Figure 4-7 B) to show that the assay was working under these conditions. 0.5 pmoles of λ PP was sufficient to remove the phosphate from Ser5P of the CTD. This confirms that the lack of change in Ser5P levels in Figure 4-7 A was due to the inactivity of RPAP2 in these particular conditions.

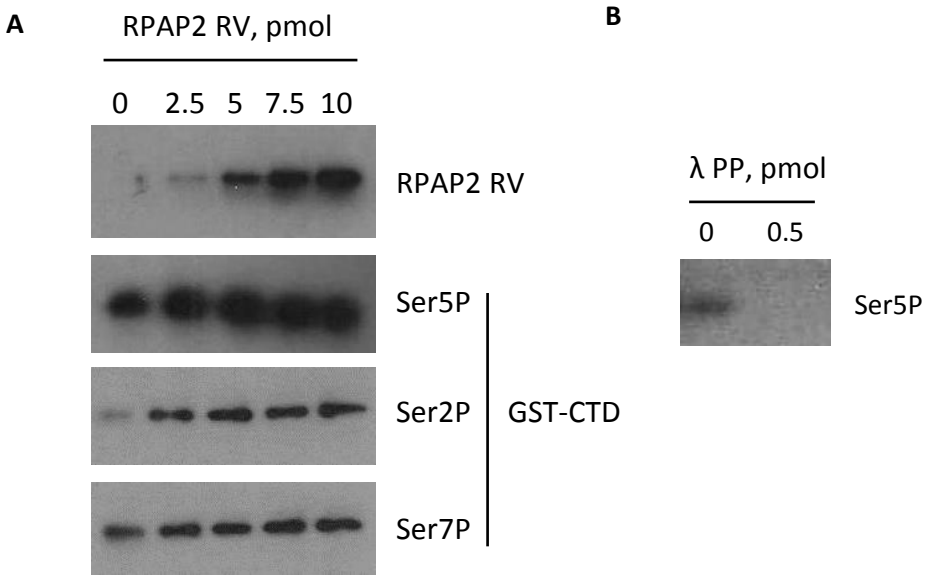


Figure 4-7. Phosphatase assay using recombinant RPAP2 RV on a GST-CTD substrate.
A. Western blots showing the effect on phospho-serines of a GST-CTD substrate with increasing quantities of recombinant RPAP2 RV. The lanes are, in order: Control (no RPAP2 added); 0 pmol, 2.5 pmol, 5 pmol, 7.5 pmol and 10 pmol RPAP2 RV per reaction. B. Positive control of CTD Ser5P dephosphorylation with 0.5 pmol lambda phosphatase (marked as λ PP).

4.3 Conclusions

The experiments presented here describe the structure prediction and purification of the human RPAP2 protein. Predicting the tertiary structure of RPAP2 would have resulted in a low-confidence prediction, given that the only structures available of known RPAP2 homologues are of the yeast Rtr1, which is a third of the size of RPAP2. However, domain architecture and secondary structure predictions were possible. According to these predictions, the human RPAP2 contains three globular domains with unstructured regions between them. The secondary structure was predicted to be mostly alpha-helical, which corresponded well to the known structure of the yeast Rtr1. However, the JPred algorithm also predicted some small beta-strands, which were absent in the yeast structure.

With the domain predictions in mind, the first domain of the human RPAP2 (named RPAP2 RV) was expressed in *E. coli* with a GST tag, and purified using a two-step method. The purification, including a GST tag cleavage, yielded a protein that was deemed free from contaminants, and was put in crystallisation trials. However, the only crystals that were observed came from a small molecule rather than protein. It is possible that the RPAP2 protein had degraded during the purification, or that the conditions tested in the crystal trials were not those that would produce nucleation. To ensure a suitable product, the purification strategy would need to be changed to remove all impurities, including small molecules such as glutathione, and more biophysical tests should be carried out to determine the stability and secondary structure of the purified protein.

When crystallising a protein, it is necessary to ensure that it retains its activity after purification. To test for this, a phosphatase assay was carried out with the RPAP2 RV

protein, using a pre-phosphorylated GST-CTD substrate. However, increasing amounts of RPAP2 in the reaction did not reduce Ser5P on the substrate. There may be several reasons for this: firstly, RPAP2 may have degraded during the purification. This could be a contributing factor in the lack of RPAP2 crystals in the crystallisation trials, and the weak DSC signal. Additionally, the protein was purified without an EDTA wash step, which could have resulted in contamination with an inhibitory metal ((Hsu et al., 2014) and section 1.3.6 in Introduction). The purification, crystallisation and functional analyses were carried out before the publication of the new *K. lactis* crystal structure by Hsu *et al.*, and the requirement for particular purification conditions was not known at the time. It is therefore possible that purified RPAP2 would retain activity if additional purification steps were used, or if the construct were longer, such as the 1-334 construct described by Egloff *et al.* that had shown phosphatase activity in a previous publication (Egloff et al., 2012b). However, a longer construct would not necessarily be suitable for crystallisation, so a range of constructs would need to be synthesised and tested for both activity and structure.

Chapter 5| Modification of the Pol II CTD during stem cell differentiation

5.1 Introduction

As known from the literature and as described in previous chapters, the Pol II CTD Ser5P phosphatase RPAP2 has roles in regulating transcription and RNA processing, as well as Pol II import into the nucleus. Most studies on RPAP2 so far have been conducted in cancer cells such as HeLa, or on its homologue Rtr1 in yeast. However, some recent findings have indicated that RPAP2 may also have a function in development.

In an unpublished study from the laboratory of Dr Amber Mosley, Indiana University School of Medicine, Rtr1 expression was knocked down in zebrafish embryos using a morpholino oligonucleotide. After a 48 hour treatment with 3.0 ng morpholino oligonucleotide per embryo, the zebrafish displayed a range of phenotypes such as curly tails, defects in eye development, pericardial oedema, and small heads. Moreover, 93% of all treated embryos did not hatch. A similar zebrafish phenotype has been previously observed in the *fog* and *pan* mutants – lacking the transcription elongation factors Spt5 and Spt6, respectively. The developing embryos also displayed defects in circulation, as well as curly tails (Keegan et al., 2002). In both cases, the phenotypes involve defects in body axis patterning and cardiac and circulation development, and likely arise from dysregulation of homeobox (*HOX*) genes.

The highly-conserved *HOX* genes encode a family of helix-turn-helix transcription factors that regulate development along the anterior-posterior axis, but continue to be expressed throughout the organism's life. During embryonic development, they are expressed in a precise order, with the 3' end genes within the cluster expressed earlier and towards the anterior end of the embryo, while the 5' genes are expressed later and at the posterior end (Noordermeer et al., 2011). Mutation or faulty expression of *HOX*

genes leads to a number of diseases in humans, such as defects in limb formation and aggressive leukemias (De Braekeleer et al., 2005, Lappin et al., 2006, Pineault et al., 2003).

Given the importance of correct timing in *HOX* gene activation, stringent mechanisms have evolved to regulate the *HOX* cluster expression during development. Regulation is effected both through *cis* elements, such as enhancers containing retinoic acid response elements, and through epigenetic marks (Atkinson et al., 2008, Kashyap et al., 2011). In mouse embryonic stem cells (ESCs), which mimic early embryonic development, the *HOX* loci are marked with both histone H3 lysine 27 trimethylation (H3K27me3) and low levels of histone H3 lysine 4 trimethylation (H3K4me3): a combination known as bivalent chromatin (Kashyap et al., 2011). The genes are thereby repressed, but poised for expression. Upon stimulation, for instance by retinoic acid, the H3K27me3 mark is erased while H3K4me3 levels increase, and *HOX* gene transcription is upregulated.

Genome-wide studies in mouse and human ESCs reveal that promoters of genes with the bivalent chromatin mark contain low levels of Pol II phosphorylated on Ser5 of the CTD (Stock et al., 2007, Landeira et al., 2010, Min et al., 2011). Other studies observed Ser5P specifically on *HOX* genes (Stock et al., 2007, Kashyap et al., 2011). Considering the potential role of RPAP2 in zebrafish development, it was postulated that RPAP2 may regulate *HOX* expression by dephosphorylating Pol II Ser5P and thereby activating gene transcription. RPAP2 itself would need to be stringently regulated to ensure the correct spatio-temporal order of *HOX* expression. The present chapter conducts the first known analysis of RPAP2 in stem cells, however due to time constraints the regulation of *HOX* genes was not studied.

5.2 Results

5.2.1 CTD Ser5P is high in stem cells

It appears that RPAP2 plays a role in embryonic development, as downregulation of the gene causes severe developmental defects in zebrafish. However, so far there have been no studies on the involvement of RPAP2 in developmental regulation in higher organisms. In mammalian embryonic stem cells, developmental regulator genes are marked by Pol II phosphorylated on Ser5. It was therefore interesting to investigate whether RPAP2 has a role in regulating mammalian development through CTD dephosphorylation.

To begin testing this hypothesis, the total levels of Pol II, CTD Ser5P, and CTD-modifying enzymes were examined in human stem cells. The HUES-2 cell line was grown on Matrigel without feeder cells, and then differentiated into embryoid bodies (EBs) by growing the cells in Aggrewells. The same number of HUES-2 cells, cells within EBs and HeLa cells was then used for Western blot analysis (Figure 5-1 A). Actin was used as a loading control, however because the levels of housekeeping genes might differ between stem cells and differentiated cells or cancer cells, total histone H3 was also analysed. In this case, no variation was observed in the levels of actin or histone H3 across the different cell types.

As seen from the Western blots, the levels of total Pol II and all the CTD-modifying enzymes are low in ES cells and rise in EBs. However, RPAP2 is absent from ES cells and very low in EBs (see Figure 5-2), with high levels present only in HeLa cells. Strikingly, Ser5P is highest in ES cells and drops approximately 5-fold in EBs and HeLa cells when related to total Pol II (Figure 5-1 A and B). The absence of Ser5P phosphatases, particularly RPAP2, could explain the high levels of Ser5P in stem cells.

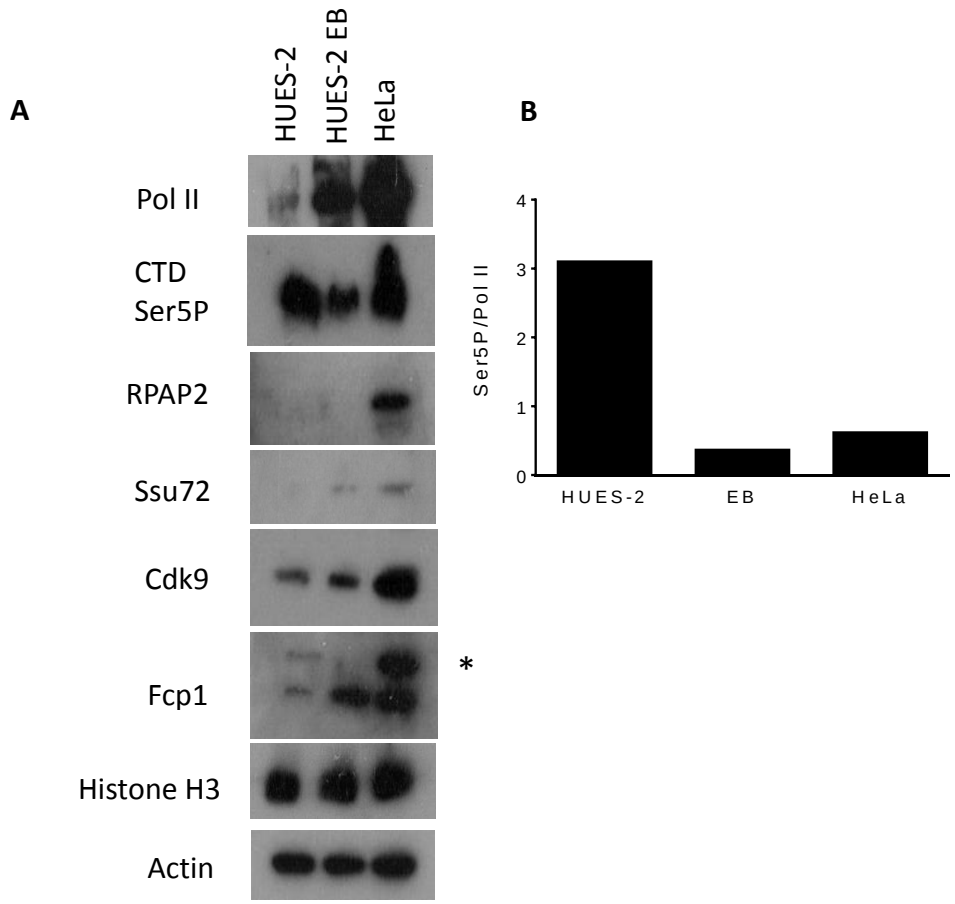


Figure 5-1. Levels of CTD modifications and CTD-modifying enzymes in ES cells.
A. Western blot of Pol II, CTD phospho-Ser5 and a number of CTD-modifying enzymes in human embryonic stem cells (HUES-2), embryoid bodies (EB) differentiated from the stem cells, and HeLa cells. Total histone H3 and actin are used as loading controls. B. Quantification of the ratio between Ser5P and Pol II across HUES-2, embryoid bodies and HeLa cells. * the top band in the Fcp1 blot may correspond to a phosphorylated form of the protein.

To examine whether the CTD modification profile seen above is not restricted to the HUES-2 cell line but is common for stem cells in general, the study was extended to include a human iPS cell line, OX1-18, and embryoid bodies derived from it. The dedifferentiated HEK293 cell line was also added to the panel. Whole-cell lysates of all the cell lines were analysed by Western blot for total Pol II, CTD Ser5P and RPAP2 (Figure 5-2). As seen before, Ser5P is higher in ES cells than in EBs, and is also higher in iPS cells

than iPS-derived EBs. Correspondingly, RPAP2 is completely absent both from ES and iPS cells, while being present at low levels in EBs and rising to much higher levels in dedifferentiated cell lines. Despite the general trend of high Ser5P in stem cells, the phosphorylation profile differs between ES cells and iPS cells. In HUES-2 and HUES-2 EBs, the polymerase is predominantly in a lower M_w form, while the blot of Ser5P in iPS cells and EBs shows two bands, which suggests the presence of additional CTD marks, possibly Ser2P (Figure 5-2, Ser5P Western). The results above indicate that high CTD Ser5P levels are a general feature of stem cells, although the profile of other CTD marks may vary between stem cell lines. There is also a potential role for RPAP2 in regulating differentiation by dephosphorylating Ser5P on developmental regulator genes.

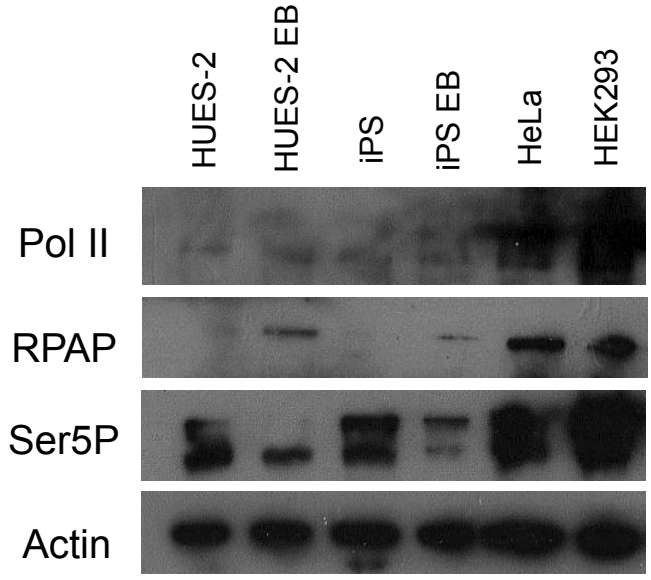


Figure 5-2. Western blots of Pol II, CTD Ser5P, and RPAP2 across a range of stem and cancer cell lines.

The lanes, from left to right, are: HUES-2, HUES-2-derived EBs, human iPS cells, iPS-derived EBs, HeLa, and HEK293 cells. Actin was used as a loading control.

5.2.2 Transfection of Myc-tagged RPAP2 into stem cells

Because the absence of RPAP2 in stem cells correlates with high CTD Ser5P, and its rise during differentiation into embryoid bodies correlates with a drop in Ser5P, it was necessary to determine if it is RPAP2 that is responsible for regulating Ser5P levels. Potentially, another phosphatase such as Ssu72 could be contributing or solely responsible for these observations. To test this, Myc-tagged RPAP2 was transfected into OX1-18 iPSCs using the Lipofectamine2000 transfection reagent. Unfortunately, under a number of conditions the transfection was not successful, as shown by the Myc Western blot in Figure 5-3 A. Only the endogenous Myc band is visible, with no additional bands after transfection. Conversely, Myc-RPAP2 was successfully transfected into HeLa cells (Figure 5-3 B) at levels comparable to endogenous Myc, but lower than endogenous RPAP2. Transfection experiments will nonetheless continue with more optimisation.

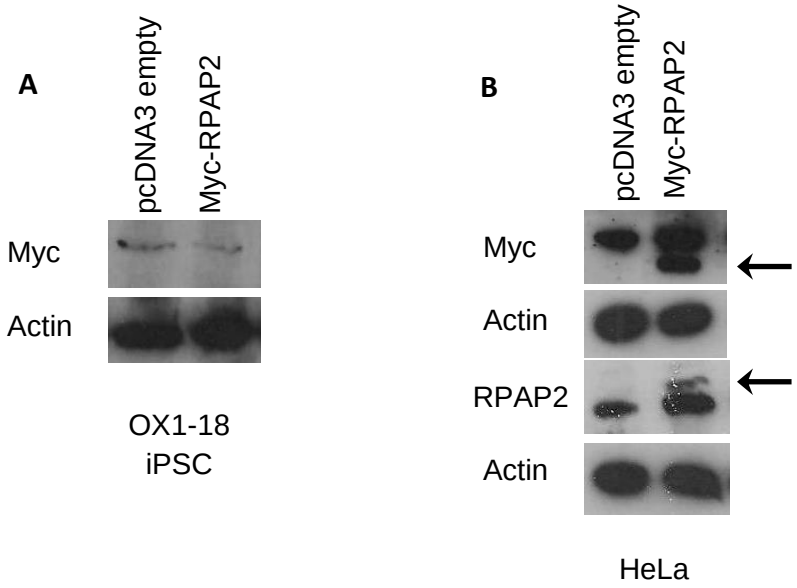


Figure 5-3. Transfection of Myc-RPAP2 into iPS cells and HeLa cells.
A. Western blot of Myc in OX1-18 iPS cells before and after transfection with Myc-RPAP2. The band in the Myc blot shows the endogenous protein at ~100 kDa. B. Western blots of Myc and RPAP2 in HeLa cells before and after transfection with Myc-RPAP2. The bands present in both the control and transfected cells in the Myc and RPAP2 blots correspond to endogenous Myc and RPAP2; the arrows point to the Myc-tagged RPAP2.

5.3 Conclusions

Recently, RPAP2 has been proposed to regulate embryonic development based on knockdown studies in zebrafish. From the resultant phenotypes, it was suggested that RPAP2 may affect the expression of the developmental 'master regulator' *HOX* genes. In this chapter, human embryonic stem cells were used as a model for early embryonic development to investigate the role of RPAP2 in Pol II function and in differentiation.

A Western blot analysis of the total levels of RPAP2 in human embryonic stem cells (HUES-2) and embryoid bodies showed that RPAP2 was very low, below the limit of detection. Correspondingly, the relative levels of CTD Ser5P were far higher in ESCs than in EBs or in HeLa cells. Although the occupancy of RPAP2 and Ser5P on individual genes was not investigated, it is possible that the low levels of RPAP2 are responsible for Ser5P not being dephosphorylated. Another Ser5P phosphatase, Ssu72, is also low in ESCs and EBs, which could provide another explanation for the high levels of Ser5P. It is possible that the absence of both phosphatases contributes towards the high levels of total Ser5P in ESCs, but that further in differentiation, e.g. in EBs, they regulate distinct sets of genes.

The study was then expanded to include induced pluripotent stem cells (iPSCs) of the OX1-18 line as another model of early development. As with ESCs, iPSCs displayed low levels of RPAP2 and higher Ser5P than embryoid bodies derived from the iPSCs. This is in line with the hypothesis that RPAP2 is responsible for dephosphorylating Ser5P during differentiation. The exact CTD modification profile appears different in iPSCs and ESCs, as evidenced by the relative strength of bands in the Ser5P blot, and it is possible that this is a general feature of CTD marks across different cell types.

The present project aimed to differentiate iPSCs into monocytes using haematopoietic growth factors; however this was not successful due to difficulties in making embryoid bodies as a result of faulty reagents, and then time constraints. Furthermore, the project planned to confirm the role of RPAP2 in CTD code regulation by transfecting Myc-tagged RPAP2 into iPSCs and observing changes in Ser5P levels, as well as markers of stem cell differentiation. Again, this was not possible because of difficulties in transfecting stem cells. These studies will continue, as it is an important experiment to help determine the involvement of RPAP2 in development.

Chapter 6| Discussion

6.1 Regulation of Pol II function by Pol II-associated factors (Chapter 2)

6.1.1 Recruitment of the PP1 phosphatase by RPAP2

The progression of Pol II along DNA during transcription is regulated by a number of proteins that associate with the polymerase. These can include proteins that modify Pol II itself, such as the various kinases, phosphatases and other enzymes that modify the C-terminal domain of Pol II, or proteins that recruit co-factors that direct RNA processing, chromatin modification and Pol II pausing and termination. A large network of protein-protein interactions fine-tune the different stages of a transcription cycle, in ways that are not all understood.

One protein shown previously to regulate Pol II is the human CTD Ser5P phosphatase RPAP2, or Rtr1 in yeast. In a recent proteomic analysis, human RPAP2 was shown to interact with the PP1 phosphatase. Among its many cellular functions, PP1 is known to dephosphorylate the Pol II CTD on Ser2 and Ser5 residues *in vitro*, although whether it performs the same function on genes was not known. Chapter 2 of this thesis set out to analyse the interplay between RPAP2 and PP1, and whether PP1 functions as a CTD phosphatase on human genes.

PP1 was shown by ChIP to be present on the U2 snRNA gene, peaking towards the 3' end of the gene, but it was not found at the beginning of the β-actin gene (Chapter2, Figure 2-1). In the experiments described in this chapter, only the U2 gene was analysed. Knockdown of RPAP2 by siRNA led to a dramatic reduction of PP1 levels on U2, similar to those observed after knockdown of the PP1 catalytic subunits. This could partly be

explained by a reduction in Pol II on the gene after RPAP2 knockdown, but the levels of PP1 fell lower than could be accounted for by the reduction of Pol II (Figure 2-1). When the levels of Ser2P and Ser5P were analysed after PP1 knockdown, instead of seeing a rise as would be expected for a phosphatase, Ser5P and Ser2P both decreased (Figure 2-2). The results suggest that at least on the U2 gene, PP1 does not act as a CTD Ser5P or Ser2P phosphatase. This is at odds with the previously-published *in vitro* data, however the substrate specificity of phosphatases *in vitro* may not reflect their actual cellular targets. It is likely, therefore, that PP1 is recruited by RPAP2 to the U2 gene to execute a different function.

The yeast homologue of a PP1 catalytic subunit, called Glc7, has been shown to regulate small nucleolar RNA (snoRNA) gene termination through its association with the Nrd1 complex (Nedea et al., 2008). Although there are no Nrd1 homologues in vertebrates, it may still be possible for PP1 to direct termination of snRNA and snoRNA genes through a different mechanism. However, the Pol II profile on the U2 gene after PP1 knockdown did not show a significant termination defect (Figure 2-1). PP1 has been shown to positively regulate P-TEFb by dephosphorylating Thr186 of Cdk9 and thereby aiding release from the 7SK RNA/HEXIM complex, and also by dephosphorylating Ser175 of Cdk9 (Chen et al., 2008, Ammosova et al., 2011). It would be expected, therefore, that knocking down PP1 would result in a fall of Ser2P on genes, which was indeed observed at the beginning of the U2 gene. Cdk9 is known to be present at the beginning of the U2 gene, and its levels rise further in the gene body (Egloff et al., 2009). A model is therefore proposed here whereby RPAP2 recruits PP1 to U2, which in turn stimulates Cdk9 activity to

phosphorylate Ser2 of the CTD (Figure 6-1). Cdk9, in turn, activates RPAP2, as evidenced by results from Chapter 3.

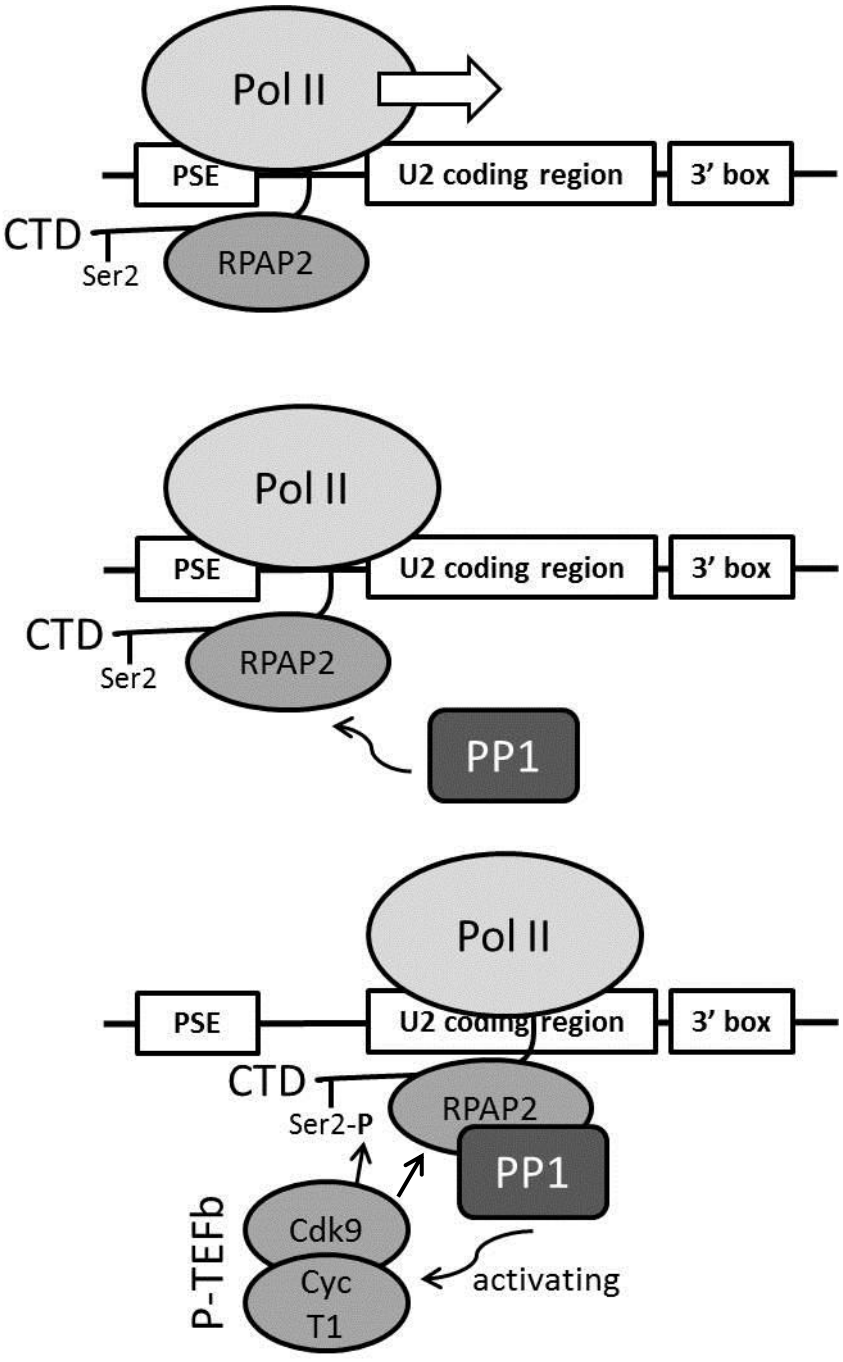


Figure 6-1. Proposed sequential model for the function of PP1 on the U2 snRNA gene.
RPAP2 recruits PP1 to the Pol II CTD, where it activates P-TEFb to phosphorylate Ser2 of the CTD.

Thus, RPAP2 has the potential to regulate the transition of Pol II into productive elongation after the promoter-proximal pause by linking CTD Ser5P dephosphorylation to P-TEFb activation and Ser2 phosphorylation.

6.1.2 Role of the Integrator complex in transcription termination

Another factor that regulates Pol II activity and interacts with RPAP2 is the Integrator complex, which directs 3' cleavage of snRNA genes but does not have a known function on mRNA genes, despite being present in a complex on some of them (Baillat et al., 2005). Previously, Integrator has been shown to be recruited to the Pol II CTD by the Ser2P-Ser7P double mark, and to require RPAP2 for recruitment. Although the 3' end processing mechanisms on snRNA genes have been studied extensively, the exact mechanism of transcription termination was not known. Yeast, which does not have the Integrator complex or Integrator homologues, directs termination *via* the Nrd1-Nab3-Sen1 pathway ((Steinmetz et al., 2001) and "Role of the Integrator complex" section in the Introduction), but Nrd1 is absent in humans and does not appear to have a role in snRNA transcription. Therefore, the work here investigated the relationship between human snRNA gene processing and termination, and the involvement of Integrator.

Transcription termination was first analysed by a Pol II ChIP on the U1 and U2 snRNA genes, and the β-actin mRNA gene. Individually knocking down three subunits of the Integrator complex, including the catalytic Int9 and Int11, did not change the Pol II profile on the β-actin gene (Chapter 2, Figure 2-3). However, knockdown of Int9 and Int11 on both U1 and U2 resulted in a reduction of Pol II within the gene body, and high levels towards the 3' end of both genes (Figure 2-4). This is indicative of an elongation defect and a termination defect, as Pol II does not leave the template and continues transcribing

further than normal past the 3’ box. A less pronounced termination defect was observed after knocking down the Int5 subunit, which does not itself possess catalytic activity, but which may stabilise the complex and contribute to efficient transcription termination and snRNA processing.

The processing and termination defects following Integrator subunit knockdown were further confirmed in an RNase protection assay. Knocking down Int5, Int9 or Int11 resulted in an increase of unprocessed pre-snRNAs (seen on the blot as read-through), which have not been cleaved after the 3’ box, and also an increase in long transcripts that had read around the entire U2 unit and into the next U2 gene repeat, without terminating (Figure 2-5). In both cases, the defect was most pronounced after the Int9 knockdown, consistent with a stronger termination defect seen by ChIP in the siInt9 sample. Although Int9 itself does not catalyse pre-snRNA cleavage, it is an essential co-factor for Int11-directed cleavage (Albrecht and Wagner, 2012), and from the results presented here it appears to also have the biggest impact on termination.

Overall, the results demonstrate that the Integrator complex couples snRNA 3’ end processing to termination, in a different mechanism to the Xrn2-mediated termination of mRNA genes, or Nrd1-dependent termination of yeast snRNA genes (O'Reilly et al., 2014).

6.1.3 RPAP2 may regulate transcription termination of a protein-coding gene

Having considered snRNA transcription termination, the focus of this study now shifts to termination of mRNA genes. It has been demonstrated that phosphorylated Tyr1 of the CTD acts as an anti-termination mark in yeast, preventing the binding of the termination factors such as Nrd1 (Mayer et al., 2012). Furthermore, *in vitro* assays by Hsu *et al.*

suggested that yeast Rtr1 is responsible for dephosphorylating Tyr1P, therefore promoting termination (Hsu et al., 2014). In human cells however, Tyr1P is not involved in termination; instead it is correlated with antisense transcription, and Pol II carrying the Tyr1P mark is enriched at gene enhancers (Descostes et al., 2014). The study here investigated the role of RPAP2 in Tyr1P dephosphorylation in human cells, and its effect on transcription.

ChIP analysis of the Tyr1P mark on the long mRNA gene KPNB1 shows that after RPAP2 knockdown, the Tyr1P mark falls, rather than rises, which suggests that RPAP2 does not act as a Tyr1P phosphatase at least on this gene (Chapter 2, Figure 2-6). Furthermore, after RPAP2 knockdown the levels of Pol II increase slightly in the gene body. The 3' peak of Pol II, usually associated with transcription termination, is shifted from 1.4 kb past the polyA site to 400 bp before the polyA site. It therefore appears that a knockdown of RPAP2 may cause premature termination on mRNA genes, although this would need to be confirmed by nuclear run-on experiments.

6.1.4 Future directions

Although much is known about mechanisms of Pol II-directed transcription, the interplay between different protein factors that interact with the polymerase and modulate its activity has in many cases not been elucidated. From the results presented in Chapter 2, it was postulated that the CTD Ser5P phosphatase RPAP2 regulates the activity of P-TEFb through recruitment of the PP1 phosphatase on snRNA genes. Potentially, this mechanism also operates on mRNA genes, where the timely placement of the Ser2P mark ensures correct transcription termination. However, this link needs to be investigated in more detail. The next step should be to analyse whether PP1 is bound to KPNB1 and

other mRNA genes, and whether it is recruited by RPAP2 as it is to the U2 gene. Furthermore, the link between PP1 recruitment and P-TEFb activation should be confirmed using specific phospho-Cdk9 antibodies in a ChIP analysis of P-TEFb on relevant genes.

Regarding Integrator, it would be necessary to determine the exact mechanism by which Integrator directs termination of snRNA genes. Termination of mRNA genes proceeds *via* the polymerase pausing after the polyA site, and the 'torpedo' model, whereby the Xrn2 5'–3' exonuclease degrades the cleaved transcript and physically dissociates Pol II from the template (Gromak et al., 2006, Shandilya and Roberts, 2012, Brannan et al., 2012). In contrast, yeast sn/snoRNA genes terminate through the action of the Nrd1 complex (Nedea et al., 2008). A thorough analysis of the molecular mechanisms in eukaryotic snRNA termination would help to outline the role of Integrator.

6.2 Regulation of RPAP2 activity (Chapter 3)

6.2.1 The phosphatase activity of RPAP2 may be regulated by P-TEFb

After some general mechanisms of human transcription regulation through modification of Pol II had been described, Chapter 3 focused on the regulation of RPAP2 itself. Despite having no homology to any known phosphatases, RPAP2 and Rtr1 had been shown in several studies to dephosphorylate the Pol II CTD Ser5P residues both *in vitro* and in cells by ChIP (Egloff et al., 2012b, Hsu et al., 2014, Mosley et al., 2009). However, recombinant human RPAP2 was only able to dephosphorylate its substrate when the CTD was pre-phosphorylated with purified P-TEFb and not with nuclear extract (Dr Sylvain Egloff, unpublished observations). The first part of Chapter 3 therefore investigated whether P-

TEFb may be necessary for regulation of RPAP2, and addressed the question with a ChIP analysis of Ser5P on the U2 snRNA gene after P-TEFb inhibition by a chemical inhibitor, KM. Indeed, there is a rise in the levels of Ser5P after P-TEFb inhibition, which indicates that P-TEFb is likely to be at least in part responsible for RPAP2 activation (Chapter 3, Figure 3-1). The limitation of this approach is that KM has several other targets besides P-TEFb, such as the Cdk1 kinase that regulates mitotic entry (Mancebo et al., 1997, Lee and Nurse, 1988). However, the alternative of knocking down the Cdk9 component of P-TEFb would require a two-day treatment with the siRNA, which would result in large-scale changes in gene expression that may confound the observations.

Another reason for the observations made by Egloff *et al*. could be that while purified P-TEFb only phosphorylated the three serines of the CTD, HeLa nuclear extract may have carried out additional modifications. While it is known that RPAP2 requires the Ser7P mark for binding (Egloff et al., 2012b), other as-yet unknown marks or combinations thereof could be inhibitory and prevent RPAP2 binding to the CTD.

6.2.2 Role of the RPRD proteins in recruiting RPAP2

There is now evidence that another complex recruits RPAP2 to genes and modulates its activity. The RPRD proteins, which comprise RPRD1A, RPRD1B and RPRD2, contain CTD-interaction domains that allow them to bind to Pol II and regulate CTD Ser5P levels. Overexpression of RPRD1A and RPRD1B resulted in a fall of both Ser5P and Ser7P on the *LEO1* gene promoter, although a direct effect on RPAP2 has not been reported until now (Ni et al., 2011).

The research presented here shows that RPRD1A interacts with a variety of protein-coding gene promoters, but not with that of the U2 snRNA gene or the coding region of β-actin (Chapter 3, Figure 3-2). The RPRD1B antibodies available were not suitable for ChIP experiments, so whether RPRD1B is also present is not known. On the promoters where RPRD1A is found, it is responsible for the recruitment of RPAP2, as evidenced by a drop of RPAP2 to background levels after an RPRD1A knockdown (Figure 3-3). The effect of RPRD1B on RPAP2 recruitment is currently unknown. However, knockdown of RPRD1A, but not RPRD1B, leads to an increase of Ser5P on these promoters (Figures 3-3 and 3-4). This would suggest separate and distinct actions of RPRD1A and RPRD1B. The two RPRD proteins have been shown to form both homodimers and heterodimers that act as a scaffold on the CTD for assembly of other factors, such as RPAP2 (Ni et al., 2014). However, the distribution of these dimers on different genes has not been investigated, and it is therefore not known whether the genes analysed here are bound by RPRD1A homodimers or RPRD1A-RPRD1B heterodimers. Several possibilities could explain the difference in Ser5P levels between the two knockdowns. Perhaps the four genes examined only bind RPRD1A homodimers, and therefore RPRD1B knockdown has no effect, or alternatively RPRD1A-RPRD1B heterodimers are present, but only the RPRD1A subunit is responsible for RPAP2 activation.

The results also suggest a differential mode of recruitment of RPAP2 to mRNA and snRNA genes. It has been shown previously that on snRNA genes, RPAP2 requires the phosphorylation of Ser7 of the CTD (Egloff et al., 2012b), however on mRNA genes Ser7 phosphorylation does not impact the early stages of transcription (Egloff, 2012, Egloff et

al., 2007), and RPRD1A and RPRD1B are now shown to be responsible for RPAP2 recruitment.

6.2.3 Future directions

It is already known that RPAP2 requires the Ser7P mark on the CTD for binding and dephosphorylation of Ser5P. However, the context of other modifications may also be important in allowing or blocking RPAP2 binding to its substrate. The role of other modifications could be investigated by constructing CTD peptides with a range of combinations of modifications and performing pulldowns or affinity assays with RPAP2 to assess binding ability.

One other possible explanation for the fact that RPAP2 functions as a phosphatase *in vitro* in the presence of purified P-TEFb is that during the purification process, RPRD proteins could have co-purified with the P-TEFb and activated RPAP2. Therefore, the presence of RPRD1A and RPRD1B in the P-TEFb purification could be examined by Western blot or mass spectrometry. Furthermore, it would be interesting to investigate why RPRD1B knockdown does not affect Ser5P levels on the genes examined. Finally, although RPRD2 (the longer RPRD1A/B homologue) is known to bind to the CTD, nothing is known about its effect on CTD phosphorylation. Its presence on individual genes and changes in Ser5P after RPRD2 knockdown could be analysed by ChIP as has been done for the other two subunits.

6.3 Structural and functional analysis of human RPAP2 (Chapter 4)

6.3.1 Prediction of human RPAP2 secondary structure, and expression and purification

Considering that RPAP2 is a novel type of phosphatase, determining its structure is of great interest and importance. The structure of the yeast homologue Rtr1 is already known to 2.1 Å resolution (Hsu et al., 2014, Xiang et al., 2012), however the human RPAP2 is much larger, and the architecture and features of its C-terminal portion are not currently known. Furthermore, the exact mode of RPAP2 binding to the CTD is not clear. Therefore, the possible structure of the human RPAP2 was investigated for the purpose of comparing the human structure to the yeast one, and with the eventual goal of crystallising the protein in complex with a CTD substrate.

Several software programs were used to predict some structural elements of RPAP2 from its primary sequence. Both GlobProt and IUPred predicted three globular domains, approximately spanning residues 30-200, 300-400, and 500-612 (Chapter 4, Figure 4-1). The domains are separated by long unstructured regions, the functions of which are not known. The secondary structure prediction from JPred suggested the presence of mostly alpha-helices, which correlated almost exactly with the known secondary structure of Rtr1 (Figure 4-2).

Towards the structural studies, the first domain of the human RPAP2 was expressed and purified. At the time of the structural experiments the study by Hsu *et al.* had not yet been published, so the importance of EDTA treatment of the sample was not known. Omitting this step from the purification procedure therefore resulted in an inactive

enzyme. However, the two-step purification protocol resulted in a pure protein, which demonstrated thermal stability up to 75°C in a DSC scan (Figures 4-3, 4-4 and 4-5). Nevertheless, protein crystals did not form in the JCSG, MORPHEUS, and PACT screens. This could be due to the protein degrading or incorrectly folding during the process of purification, or that some contaminants present in the sample prohibited crystallisation. In fact, crystals of what was most likely a small molecule were observed in two wells of the PACT screen, suggesting the presence of impurities, most probably glutathione (Figure 4-6). In order to obtain crystals of RPAP2, a different strategy would need to be employed, testing several constructs and performing more rigorous wash steps.

6.3.2 Functional analysis of RPAP2 RV

After purification, the RPAP2 RV protein was tested for phosphatase activity, but no change was observed in CTD Ser5P levels after addition of RPAP2 (Figure 4-7). With the knowledge gained from the Hsu *et al.* publication (Hsu et al., 2014), it is now clear why the protein did not perform its function. Most importantly, the purification protocol did not include an EDTA wash step, which enhances phosphatase activity. However, the GST-CTD substrate was also pre-phosphorylated with recombinant P-TEFb, instead of the complex purified from cells as was carried out by Egloff and colleagues (Egloff et al., 2012b). This may not be of great importance; however it is also plausible that the purified P-TEFb was bound to other proteins, such as the RPRD proteins, that stimulated the activity of RPAP2.

6.3.3 Future directions

The crystal trials with RPAP2 RV were not successful, and there are a number of options available to improve the chances of successful crystallisation. Firstly, the length and

features of the construct are important, since long unstructured regions can reduce the likelihood of crystal formation (Uversky and Dunker, 2010). It would therefore be necessary to make several truncated RPAP2 constructs and test each of them for stability and capacity to crystallise. The purification procedure can also be improved to better remove impurities, for example through more extensive dialysis, as well as to fully separate RPAP2 from the cleaved GST. Alternatively, it may be possible to express and purify RPAP2 with the GST tag, as the tag may stabilise the protein and prevent degradation.

If the improvements outlined above were to result in a successful crystal structure, the next step would be to co-crystallise RPAP2 with a CTD diheptapeptide phosphorylated on Ser7, as this modification is required for RPAP2 binding (Egloff et al., 2012b).

6.4 Modification of the Pol II CTD during stem cell differentiation (Chapter 5)

6.4.1 Levels of Ser5P and CTD-modifying enzymes across cell types

So far, there has been little analysis of enzymes that modify the Pol II CTD in stem cells and during development. Several genome-wide approaches, including the ENCODE project, have identified the distribution of CTD phospho-serines in stem cell lines, and there is information about CTD phosphorylation patterns on subsets of genes; in particular on genes that regulate developmental programs. For example, the *HOX* genes that regulate the anteroposterior body axis patterning are bound by Pol II that is phosphorylated on Ser5P, and marked by both permissive and restrictive 'bivalent' chromatin (Bernstein et al., 2006, Mikkelsen et al., 2007, Stock et al., 2007). The

polymerase is poised for transcription, but is not actively transcribing in the absence of appropriate differentiation signals. A Ser5P phosphatase therefore has the potential to have a crucial role in activating the transcription of *HOX* genes and therefore stem cell differentiation.

The RPAP2 Ser5P phosphatase, thus far investigated mostly in cancer cells and yeast, was recently demonstrated to regulate development in zebrafish. An RPAP2 knockdown in early embryos leads to severe developmental defects that suggest potential *HOX* gene mis-regulation. Embryonic stem cells and induced pluripotent stem cells (ESCs and iPSCs) were used to analyse the role of RPAP2 in early embryonic development and its effect on Ser5P.

In the present work, whole-cell Western blots in a panel of stem cell lines and embryoid bodies showed several trends (Chapter 5, Figures 5-1 and 5-2). Firstly, the levels of Pol II and CTD-modifying enzymes are low in stem cells and rise in embryoid bodies. The most marked is the absence of RPAP2 from ES and iPS cells, and its reappearance at high levels in HeLa and HEK293 cells. Another feature is the high levels of Ser5P in ES and iPS cells, compared to that in embryoid bodies and dedifferentiated cells. The low levels of RPAP2 and high levels of Ser5P correlate, but causation has not yet been demonstrated. Technical limitations prevented further differentiation of the stem cell lines into monocytes and macrophages. It is therefore not possible to comment on the Ser5P and RPAP2 levels in a differentiated cell line.

One limitation of the study is that Western blot analysis does not reveal how much of the Ser5P Pol II is DNA-bound, or the gene-type distribution of the polymerase at the

different stages of differentiation. This could only be analysed by looking at RPAP2 and Ser5P genome-wide. Furthermore, the high Ser5P levels in ES and iPS cells could instead be due to the lack of Ssu72, another Ser5P phosphatase. Ssu72 is present at promoters and 5' untranslated regions on a range of genes in ESCs, colocalising with CTD Ser5P (Chen et al., 2014). The list of genes includes *NANOG*, involved in ESC self-renewal. A way to analyse the contribution of Ssu72 and RPAP2 would be to overexpress either protein in stem cells and observe the changes in Ser5P, both by Western blot in cell lysate and by ChIP on genes. Transfection of Myc-RPAP2 into iPS cells was attempted but not successful (Figure 5-3), so more transfection conditions need to be tested.

6.4.2 Future directions

So far, only an inverse-correlative link between RPAP2 and Ser5P has been established in stem cell lines. Although it appears that RPAP2 may regulate CTD phosphorylation during stem cell differentiation and development, this needs to be confirmed with gene-specific approaches. For this, tagged RPAP2 and Ssu72 would first need to be ectopically overexpressed in ES or iPS cells, and their effect on Ser5P levels examined. At the same time, changes in expression of *HOX* genes and other developmental regulators could be analysed by ChIP. To assess whether RPAP2 expression is activated immediately upon differentiation, the cells could be treated with retinoic acid to induce differentiation, and levels of RPAP2 analysed at various time points after treatment.

To look directly at RPAP2 localisation in the genome of differentiating stem cells, genome-wide ChIP-seq could be employed. Because the commercially-available RPAP2 antibodies are not suitable for ChIP-seq, a stem cell line stably expressing tagged RPAP2 is

required. This approach would allow analysis of RPAP2 binding to *HOX* genes and other genes carrying the bivalent chromatin mark.

6.5 Conclusions

In conclusion, the studies presented in this thesis demonstrate several novel modes of regulating eukaryotic transcription by RNA polymerase II. In particular, the CTD phosphatase RPAP2 was shown to impact on transcription regulation through a recruitment of the PP1 phosphatase to snRNA genes, where it may stimulate the kinase activity of P-TEFb, and by regulating the termination of a long mRNA gene. Furthermore, the mechanism of snRNA gene termination was elucidated, which involves the Integrator complex that also directs snRNA 3' end cleavage.

RPAP2 itself was found to be regulated by two recently-described proteins of the RPRD family, which recruit it to a number of mRNA genes and stimulate its phosphatase activity. Finally, a potential involvement of RPAP2 in stem cell differentiation and development was outlined, but further experiments are needed to confirm its exact role.

Chapter 7| Materials and Methods

7.1 Materials

7.1.1 Antibodies

Antibodies were stored at 4^{o}C or at -20^{o}C

Antibodies	**Company**
Anti-RNA Polymerase II (sc-N20; rabbit polyclonal)	Santa Cruz
Anti-RNA polymerase CTD phosphorylated on Serine 2 (3E10), 5 (3EB) and 7 (4E12) (all rat monoclonal)	Gift from Prof Dirk Eick (Chapman et al., 2007)
Anti-RNA polymerase CTD phosphorylated on Serine 2 (ab5095)	Abcam
Anti-RNA polymerase CTD phosphorylated on Serine 2 (H5, ab24758; mouse monoclonal)	Abcam
Anti-RNA polymerase CTD phosphorylated on Serine 5 (ab5131; rabbit polyclonal)	Abcam
Anti-RNA polymerase CTD phosphorylated on Tyrosine 1 (61383; mouse monoclonal)	Active Motif
Anti-actin (sc-1615; goat polyclonal)	Santa Cruz
Anti-RPAP2 (17401-1-AP; rabbit polyclonal)	Proteintech
Anti-Histone H3 (ab1791; rabbit polyclonal)	Abcam
Anti-Cdk9 (C-20; rabbit polyclonal)	Santa Cruz
Anti-Int5 (A301-268A; rabbit polyclonal)	Bethyl
Anti-Int9 (A300-412A; rabbit polyclonal)	Bethyl
Anti-Int11 (ab75276; rabbit polyclonal)	Abcam
Anti-RPRD1A (P15RS, C-18; rabbit polyclonal)	Santa Cruz
Anti-RPRD1B (A303-782A; rabbit polyclonal)	Bethyl
Anti-PPP1CA (A300-904A; rabbit polyclonal)	Bethyl
Anti-GST (27-4577-01; goat polyclonal)	GE Healthcare
Anti-Myc epitope (ab1932; goat polyclonal)	Abcam
Goat anti-mouse HRP (1858413)	Pierce

Donkey anti-goat HRP (sc-2020)	Santa Cruz
Goat anti-rabbit HRP (12-348)	Millipore

7.1.2 Antibiotics

Ampicillin (100 µg/ml; Sigma Aldrich)

7.1.3 Buffers and solutions

- Phosphate buffered saline (PBS) - 140 mM NaCl, 3.4 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4

- Tris buffered saline - 150 mM NaCl, 40 mM Tris-HCl pH 7.4

- TBE buffer - 10 mM Tris, 250 mM orthoborate, 2 mM EDTA

- ChIP lysis buffer - 10 mM Tris-HCl pH 8.0, 0.25% Triton X-100, 10 mM EDTA, 0.5 mM EGTA, Protease inhibitor cocktail

- ChIP wash buffer - 10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, Protease inhibitor cocktail

- ChIP sonication buffer - 10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1mM EDTA, 0.5 mM EGTA, 0.1% SDS, Protease inhibitor cocktail

- RIPA buffer - 10 mM Tris-HCl pH 8.0, 150 mM NaCl, 1mM EDTA, 0.5 mM EGTA, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate

- High salt wash buffer - 10 mM Tris-HCl pH 8.0, 500 mM NaCl, 1mM EDTA, 0.5 mM EGTA, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate

- LiCl wash buffer - 10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA, 1 mM EGTA, 1% NP-40, 1% sodium deoxycholate

- Elution buffer - 100 mM NaHCO3, 1% SDS, 10 mM DTT

- Proteinase K buffer (5×) - 50 mM Tris-HCl pH 7.5, 25 mM EDTA, 1.2% SDS

- EB buffer – 10 mM Tris-HCl, pH 8.1

- GST purification buffer – 20 mM Tris-HCl pH 8, 100 mM NaCl

- PreScission protease buffer – 50 mM Tris pH 7.0, 150 mM NaCl, 1 mM DTT

- SDS running buffer – 25 mM Tris, 250 mM glycine, 1% SDS

- Coomassie staining solution – 0.25% Coomassie brilliant blue, 50 % methanol, 10% acetic acid

- De-staining solution – 10 % acetic acid, 20% methanol

- Bacterial lysis buffer – 50 mM glucose, 25 mM Tris, 10 mM EDTA

- Bacterial lysis solution – 1% SDS, 0.2 M NaOH

- Neutralisation buffer – 3 M potassium acetate, 5 M acetic acid

- Kinase buffer – 150 mM NaCl, 40 mM HEPES pH 7.6, 10 mM $MgCl_2$, 5 mM DTT, 500 µM ATP

- Phosphatase buffer – 50 mM Tris-HCl pH 6.5, 10 mM $MgCl_2$, 20 mM KCl, 5 mM DTT

- RNA lysis buffer – 140 mM NaCl, 1.5 mM $MgCl_2$, 10 mM Tris pH 7.5, 1% NP-40

- HBS (2x) – 20 mM Tris pH 7.5, 1 M NaCl, 20 mM $MgCl_2$

- PK solution – 0.4 mg/ml proteinase K, 0.4% SDS

7.1.4 Reagents

Chemical reagents were purchased from VWR International or Sigma, apart from the reagents listed below:

- RNase-DNase-free water (Qiagen)

- LipofectAMINE 2000 (Invitrogen)

- Dimethylsulfoxide (DMSO; Merck)

- Precision plus protein Dual Colour Standards (Bio-RAD)

- Dithiothreitol (DTT; Roche)

- GlutaMAX (Gibco)

- 2-mercaptoethanol (Gibco)

- dNTPs (Invitrogen)

- BSA (10 mg/ml; New England Biolabs)

- Protease inhibitor cocktail (Glover-Cutter et al.)

- Qiaquick PCR purification kit (Qiagen)

- QIAGEN Plasmid Maxi Kit (Qiagen)

- QuantiTect SYBR Green PCR master mix (Qiagen)

- DNase 1 (Promega)

- Proteinase K (Promega)

7.1.5 PCR Oligonucleotides:

All oligonucleotides were made by Sigma Aldrich or MWG. The sequences of all primers are listed 5′ to 3′ below.

Primers used in Chapter 2:

Primer	Sequence (5′ to 3′)
U2 PSE F	ATGAGAGTGGGACGGTGA
U2 PSE R	CACTTGATCTTAGCCAAAA
U2-1 F	ACGAGTCCTGTGACGCGCCGG
U2-1 R	CTCCGGGTGGGTCCCATTCCT
U2-CTCF F	CCTCCCCGCCTCTCCCTCGCTC
U2-CTCF R	GGACAAATAGCCAACGCATGCGG
U2-2 F	CGGTGTTTGGTTTTCGCTTCCTGTG
U2-2 R	GGACACGTATGTTTACTGCAGCAC
U2 -1 kb F	TGATATAAACAACGGGGGCGCGATG
U2 -1 kb R	GCTGTGTAGGGAAGCTGCCTCTCTG
βAct Up F	GCTGCGGCTGGGTAGGTTTG
βAct Up R	CACTTAGAAGTCGCAGGACC
βAct 1 F	CCAATCAGCGTGCGCCGTTCCGA
βAct 1 R	GGTGTGGACGGGCGGCGGATC
βAct A F	GGGCAACCGGCGGGGTCTTT
βAct A R	ACGCAGTTAGCGCCCAAAGG
βAct C F	TGGGCCACTTAATCATTCAAC
βAct C R	CCTCACTTCCAGACTGACAGC
βAct 1A F	CCGCAGTGCAGACTTCCGAGCCG
βAct 1A R	CTCCCGCGGGGCAGCCTCGGCC
βAct pause F	GGGACTATTTGGGGGTGTCT
βAct pause R	TCCCATAGGTGAAGGCAAAG
U1 -0.5 kb F	GCGGAATGCTGCACTACAAGG
U1 -0.5 kb R	CCTTTCGGTGGCCAGATGGAGC
U1 PSE F	GGAAAGGGCTCGGGAGTGCGCG
U1 PSE R	CAGGTAAGTATGAGATCTTCGGG

U1 +290 F	ACTTTCTGGAGTTTCAAAAACAG
U1 +290 R	CTTCTTATTCCCCAGGCCTCCAC
U1 +590 F	CCTACCCCCCAAGTCACCTCTTAC
U1 +590 R	CGCCCCATCTGTGTAGTTTCT
U1 +0.9 kb F	GATGTCTCACTCCCACAGGC
U1 +0.9 kb R	GAGGGACTGGGGAGGACACTGG
KPNB1 TSS+0.9 F	AGGGAAGGGAAGTTAGGTTGCG
KPNB1 TSS+0.9 R	GAGCTTCCACATGGCCCCTA
KPNB1 TSS+5.3 F	GCAAAGCAGGATACTAAGTGAATCCGA
KPNB1 TSS+5.3 R	GAAGCCACAATTGACCTAGAGC
KPNB1 pA-2.9 F	GAGGAGTGTGCACGGATGCTGAA
KPNB1 pA-2.9 R	CCAAGATGGCCGATGTTATGG
KPNB1 pA-0.4 F	TAGTTACCGTCTGCTTGGGAAGATG
KPNB1 pA-0.4 R	CCTCTGACAGCAAGTCCAACATT
KPNB1 pA+1.4 F	GACTCATCACACCAAGGTCAC
KPNB1 pA+1.4 R	GATAGTGCTGGGAAGGAAATGG

Primers used in Chapter 3:

Primer	**Sequence (5’ to 3’)**
U2 -3 kb F	GAACTGACCACATTTCCAGGGCTCC
U2 -3 kb R	AATCAGGATATTCACGTGGACTACT
U2 PSE F	ATGAGAGTGGGACGGTGA
U2 PSE R	CACTTGATCTTAGCCAAAA
U2-1 F	ACGAGTCCTGTGACGCGCCGG
U2-1 R	CTCCGGGTGGGTCCCATTCCT
U2 CTCF F	CCTCCCCGCCTCTCCCTCGCTC
U2 CTCF R	GGACAAATAGCCAACGCATGCGG
U2-3 F	CAGCTGTGGCTGGACAGGTTGGACC
U2-3 R	CAGGGTTCCTGCACATTCACAGTTA
LEO1 F	ATACCCAGGCGAGAACAGGT
LEO1 R	GACAGTGTCGCAAAGATTCG
RPRD1A F	TGCTTCTTTGACAGAGTTTCCA
RPRD1A R	TGATGCCTGGGTTACTCTCA
MED12 F	TCTAGGCCAAAACCGAGCTA
MED12 R	GATGCAAACCAGGGAAGAAG
CTR9 F	GGAGAGTCAGACGCCAGATG
CTR9 R	GAAGCTTTTGTTGGCGTGTC
βActA F	GGGCAACCGGCGGGGTCTTT
βActA R	ACGCAGTTAGCGCCCAAAGG

7.1.6 Bacteria and cell lines

- *Escherichia coli* NM554 competent cells were used to produce plasmid DNA, while *Escherichia coli* Rosetta™ cells were used to produce recombinant protein.

- HeLa human cervical cancer adherent cells were used for ChIP, transient transfection and knockdown experiments.

- 293 human embryonic kidney cells were used for Western blots.

- Mouse embryonic fibroblasts were used for culturing human embryonic stem cells before conversion onto feeder-free culture.

- Human embryonic stem cells (HUES-2) and embryoid bodies were used for Western blots.

- OX1-18 human induced pluripotent stem cells were used for transient transfection, Western blots and differentiation into embryoid bodies.

7.1.7 Bacterial and tissue culture media

- NM554 cells were cultured in 2-YT medium at 37^{o}C, shaking overnight.

- Rosetta cells were cultured in 2-YT medium at 37^{o}C pre-induction, then moved to 21^{o}C overnight after induction.

- HeLa cells and 293 human embryonic kidney cells were grown in complete media consisting of Dulbecco's modified Eagle's medium, supplemented with 2 mM L-glutamine (Invitrogen), 10% heat inactivated foetal bovine serum (Invitrogen) and 100 Units/ml penicillin/streptomycin (Gibco).

- Mouse embryonic fibroblasts were grown on 0.1% gelatin in advanced DMEM (Gibco/Life Technologies) supplemented with 10% FCS (Biosera), 1% GlutaMAX (Gibco), 50 mM 2-mercaptoethanol and 100 Units/ml penicillin/streptomycin (PAA).

- Human embryonic stem cells were grown on MEF feeders in knockout DMEM (Gibco) supplemented with 20% knockout serum replacement (Gibco), 1% GlutaMAX, 1% non-essential amino acids (PAA), 0.1% 2-mercaptoethanol, 100 Units/ml penicillin/streptomycin and 10ng/ml basic fibroblast growth factor (R&D).

- Once converted to feeder-free media, all stem cells were grown on Matrigel (BD Biosciences) in mTeSR™1 media (StemCell Technologies) supplemented with 10µM ROCK inhibitor (Cal Biochem/Merck) and 100 Units/ml penicillin/streptomycin.

- To differentiate iPS cells into monocytes, EBs were produced by growth in mTeSR supplemented with 10 µM ROCK inhibitor, 50 ng/ml BMP4, 50 ng/ml VEGF and 20 ng/ml SCF. Once formed, the EBs were cultured in X-VIVO media supplemented with 1% GlutaMax, 0.1% 2-mercaptoethanol, 100 ng/ml M-CSF and 5 ng/ml interleukin-3.

7.2 Methods

DNA-related methods

7.2.1 Real-time quantitative Polymerase Chain Reaction (qPCR)

Real time quantitative PCR (qPCR) was used to measure the accumulation of PCR product during the exponential phase and to quantify the gene transcript levels *in vivo*. A Rotorgene 6000 thermal cycler (Corbett) was used for all qPCR experiments. The reaction contained 5 µl of 2x QuantiTect SYBR Green PCR Kit (Qiagen), 3 µl water, 10 picomoles of each primer and 1 µl of template DNA (usually 1/30th of ChIP DNA). The thermal cycling programme used was: 95^{o}C for 15 min, followed by 40 cycles at 94^{o}C for 15 sec, 57^{o}C for 15 sec and 72^{o}C for 30 sec. To confirm that a single PCR product was generated in the reaction, the last cycle ended with a 10 minute melt from 50^{o}C to 99^{o}C with a 1^{o}C increase every 5 seconds. Each amplicon should yield a clear peak corresponding to its melting temperature, and allows the detection of contaminating dsDNA such as primer dimers. At the end of the programme, a threshold value was set for all samples and standards over the linear range of the PCR reaction. The Rotorgene software converted the cycle number (Ct) for each standard to copies/µl and generated a standard curve. The Ct values for each sample were quantified relative to this standard. For ChIP analysis, a standard curve was produced for each primer pair using serial dilutions of the Input DNA sample (undiluted, 1:10, 1:50, 1:250 and 1:750 dilutions). Signals from all other samples were related to this standard curve, resulting in a "% of input" value for each sample. To adjust for background, the IgG % input was subtracted to give the final value. To allow normalisation of transcript levels taking into account the difference in DNA content

between inputs over several repeats, the concentration (in copies/µl) of the specific amplicon was divided by the total input DNA for each immunoprecipitation, giving the final value plotted on the graphs.

7.2.2 DNA sequencing

DNA sequencing was carried out to confirm that all constructs used were correct. The sequencing reaction used mini-prep or maxi-prep (Qiagen)-purified DNA with the appropriate primers. All sequencing reactions were carried out by Source BioScience UK limited, Department of Biochemistry, University of Oxford, UK.

7.2.3 Protein expression

To express GST-RPAP2, Rosetta™ *E. coli* cells were transformed with 1-2 µg of plasmid DNA and plated on an agar plate with ampicillin at 37^{O}C overnight. 100 ml 2xYT media with ampicillin was then inoculated with several colonies and grown while shaking at 37^{O}C until the log phase. Then, the culture was induced with 0.1 mM IPTG to start protein expression, and moved to 21^{O}C overnight.

RNA-related methods

7.2.4 Purification of total RNA

HeLa cells grown in 10 cm round dishes were washed with PBS, scraped into a 1.5 ml tube and centrifuged for 1 min at 12,000 g. The cells were then resuspended in 100 µl RNA lysis buffer and mixed with 100 µl of 2x high salt buffer and 1 µl DNase (Promega). This was incubated for 5 mins at 30°C, then mixed with 100 µl freshly-made PK solution and incubated for a further 30 mins at 37°C. Following this, RNA was extracted with 300 µl

phenol/chloroform and precipitated with 750 µl water. To remove DNA contaminants, 100 µl DNase buffer and 10 units of DNase were added and incubated for 1 h at 37°C. The purified RNA was extracted again with 100 µl phenol/chloroform and precipitated with 10 µl of 3 M sodium acetate and 250 µl ethanol. The RNA was resuspended in R-loop buffer (50 mM PIPES, pH 7.4, 1 mM NaEDTA, 500 mM NaCl) ready to use for RNase protection assays.

7.2.5 RNAi-mediated knockdown

RNAi duplexes were ordered from Dharmacon and resuspended to a concentration of 20 pM. HeLa cells were plated into 6-well Costar plates in complete DMEM media and grown until 60-70% confluent. Then the media was changed to DMEM without antibiotics to allow RNA uptake. 3 µl of siRNA and 3 µl LipofectAMINE2000 were diluted in 250 µl of Opti-MEM 1 (Invitrogen) reduced serum medium and incubated separately for 5 min at RT. siRNA and lipofectAMINE2000 were combined and incubated for a further 15 minutes at RT. 500 µl of this complex was added to each well of the 6-well plate. Media was changed after 6 hours. The next day, cells were split into 10 cm round dishes by pooling two wells of a 6-well plate together. Transfection was repeated as above with 3 µl siRNA in 250 µl Opti-MEM and 3 µl of LipofectAMINE2000 in 250 µl Opti-MEM.

7.2.6 RNase protection assay

RNase protection assays were carried out by Prof Shona Murphy as described in (Cuello et al., 1999). Briefly: 10 to 20 µg of RNA in 30 µl of R-loop buffer were incubated with 1 µl of riboprobe (prepared by Prof Shona Murphy) for 10 min at 80^{o}C and hybridized at 56°C overnight. After adding 300 µl of RNase protection solution (300 mM NaCl, 10 mM Tris -

HCL pH 7.4, 5 mM EDTA, 40 µg RNaseA and 1100 U RNaseT1) the samples were incubated at 18^{o}C for 2 h. After adding 30 µl of Proteinase K mix (1% SDS, 3.5 mg/ml of protease K), the mixture was than incubated for 30 min at 37^{o}C, phenol/chloroform purified, ethanol precipitated and analyzed on a 6% polyacrylamide-urea gel.

Protein-related methods

7.2.7 Cell Culture

All cell culture was performed in a category 2 laminar flow hood. HeLa cells and HEK 293 cells were grown at 37^{o}C in 5% CO_2 in sterile Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% heat-inactivated foetal bovine serum, 2 mM L-glutamine and 1x penicillin/streptomycin. Typically, adherent HeLa cells in 75 cm^2 flasks were lifted off by incubation with 2 ml of PBS with 1 mM EDTA for 5 min. This was followed by gentle tapping of the side of the flask. Cells were resuspended in 5 ml of complete DMEM media and passaged into new flasks at a 1:5 dilution.

ES cells were initially grown on mouse embryonic fibroblast (MEF) feeders. 0.6x10^6 mitomycin C-treated MEFs were plated in MEF medium (advanced DMEM supplemented with 10% FCS, 1% GlutaMAX and 50 mM 2-mercaptoethanol) in 6-well plates pre-treated with 0.1% gelatin for 1 h. The next day, the well was rinsed with KnockOut DMEM (KO-DMEM, Gibco) to remove traces of FCS, and the medium changed to hES medium (KO-DMEM supplemented with 20% KnockOut Serum Replacement, 1% GlutaMAX, 1% non-essential amino acids, 0.1% 2-mercaptoethanol, and 10 ng/ml Basic Fibroblast Growth Factor). Using a dissecting microscope, 10-11 good, undifferentiated hES colonies from a separate plate were cut into approximately 9 pieces with a sterile needle, detached from

plate and transferred into the MEFi well. The new colonies were fed by 50% media change every day for approximately 5-7 days of culture, until they were ready to split again. Before conversion to Matrigel, the colonies were lifted from the plate with TrypLE and passaged onto MEFi feeders as individual cells, not colonies.

To adapt hES or iPS cells to feeder-free Matrigel growth, a well of a 6-well plate was coated with Matrigel (BD Biosciences) diluted 1:12 in KO-DMEM and incubated at 37°C for 1-48 hours. A well of just-confluent hES or iPS cells was washed with PBS, and the cells are lifted with TrypLE. The lifted cells were diluted 1:10 in PBS, centrifuged for 5 min at 400 g, and resuspended to a concentration of 2.5 x 10^5 cells/ml in mTeSR with 10% penicillin/streptomycin. To ensure the cells adhered to the bottom of the well, Rock inhibitor was also added to a concentration of 10 µM. Matrigel was then aspirated from the well, the cells in mTeSR added at approximately 5 x 10^5 cells/well and incubated at 37°C overnight. The medium was changed every day until the cells were confluent and ready for passaging.

7.2.8 Stem cell differentiation

To differentiate hES or iPS cells into embryoid bodies (EBs), Aggrewells (Stemcell Technologies) were used. One well of an Aggrewell 800 plate can make 300 EBs, while one well of Aggrewell 400 can make 1200 EBs. Each well of an Aggrewell plate to be used was washed with PBS, filled with 1 ml of mTeSR medium, and spun at 3000 g for 1-2 min to remove any microscopic air bubbles. hES or iPS cells were washed with PBS and trypsinised with TrypLE. After dilution 1:10 with PBS and centrifugation, the cells were resuspended in mTeSR with 10 µM Rock inhibitor and added to the Aggrewells at a

quantity of 3.6 x 10^6 cells for Aggrewell 400 or 5 x 10^6 cells for Aggrewell 800. The plate was then centrifuged at 100 g for 3 minutes to allow the cells to fill the wells, and incubated at 37°C. The cells were fed by a 75% media change every day until EBs formed, after 3-4 days. EBs were harvested by pipetting PBS up and down with a serological pipette to lift the EBs from the wells, and then straining the liquid through a 40 µm cell strainer. The EBs harvested on the strainer were then eluted with more PBS and used for further differentiation or Western blot analysis.

7.2.9 Freezing

Cells to be stored at -80^{o}C were lifted from a 75 cm^2 flask with PBS supplemented with 1 mM EDTA and resuspended in 1 ml of complete DMEM media supplemented with 20% foetal bovine serum and 10% DMSO. The cells were then transferred into a cryotube and placed in a precooled Cryobomb (NALGENE) filled with isopropanol in a -80^{o}C freezer.

7.2.10 Transfection into HeLa cells and iPS cells

DNA and siRNA were transfected using Lipofectamine 2000 according to the manufacturer's instructions (Invitrogen). The day before transfection, HeLa cells were plated in complete DMEM media, and iPS cells in mTeSR media, in 6-well Costar dishes at a density of 2 × 10^5 cells per well to achieve 80-90% confluence on the day of transfection. Just before the transfection, the medium was changed to DMEM without penicillin/streptomycin for HeLa cells, or mTeSR without penicillin/streptomycin for iPS cells. Each DNA or siRNA to be transfected was diluted in 250 µl of Opti-MEM (GIBCO) reduced serum medium and incubated for 5 min at RT. LipofectAMINE 2000 was also diluted in Opti-MEM reduced serum medium and incubated for 5 min at RT. Typically, 1

µl of Lipofectamine 2000 was used per µg of DNA or siRNA in 250 µl of Opti-MEM. The DNA and Lipofectamine 2000 were combined and incubated for a further 15-20 min at RT. 500 µl of complexes was added, by slow dispensing, to the corresponding well. The media were changed after 6 h and cells were harvested after 24 h.

7.2.11 ChIP analysis

Chromatin immunoprecipitation was carried out as described in (Oshima et al., 2004), with some minor modifications. In short: HeLa cells were grown in 10 cm dishes in 10 ml of DMEM complete media. 1% formaldehyde was added to fix proteins to nucleic acids. After 10 min, the reaction was stopped by adding glycine to a final concentration of 125 mM. After washing with cold PBS, the cells were lysed by incubation on ice for 10 min in 1 ml cell lysis buffer (10 mM Tris-HCl pH 8.0, 0.25% Triton X-100, 10 mM EDTA, 1x protease inhibitor cocktail and 1x phosphatase inhibitor) and the cytoplasmic fraction was separated from the nucleus by spinning at 5000 x g for 5 min. The nuclei were then washed using 1 ml ChIP wash buffer (10 mM Tris-HCl pH 8.0, 200 mM NaCl, 1 mM EDTA, 1x protease inhibitor cocktail and 1x phosphatase inhibitor). The nuclear lysate was then resuspended in 200 µl sonication buffer (10 mM Tris-HCl pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.1% SDS, 1x protease inhibitor cocktail and 1x phosphatase inhibitor), incubated rotating at 4°C for 10 mins, and sonicated using a Bioruptor sonicator (Diagenode) for 2x 15 mins at high intensity, in cycles of 30 sec on, 30 sec off. The lysate was then spun at 20,000 g for 15 min at 4°C, and the supernatant collected and pre-cleared for 30 min at 4°C with 10 µl Dynabeads Protein G (Novex by Life Technologies). The supernatant was recovered, split into 50 µl fractions, and incubated with appropriate amounts of antibody at 4°C O/N.

At the same time, Dynabeads were blocked to prepare for immunoprecipitation. 15 µl Dynabeads for IP were washed with 100 µl RIPA buffer (10 mM Tris-HCl pH 8.0, 150 mM NaCl, 1 mM EDTA, 0.1% SDS, 1% Triton X-100 and 0.1% sodium deoxycholate) and incubated with 15 µl RIPA buffer containing 4 mg/ml herring sperm DNA and 4 mg/ml BSA at 4°C O/N.

On the second day, blocking solution was removed from the beads and the sonicated nuclear extract with antibody was added for 1 h at 4°C. The supernatant of the rabbit IgG control was kept as total input. The beads were washed 3 times with 300 µl of RIPA buffer, 3 times with high salt wash buffer (10 mM Tris-HCl pH 8.0, 500 mM NaCl, 1 mM EDTA, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate), twice with LiCl wash buffer (10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA, 1% NP-40, 1% sodium deoxycholate) and twice with TE buffer. The immunoprecipitate was then eluted in 2x15 min steps with 50 µl elution buffer (100 mM $NaHCO_3$, 1% SDS and 10 mM DTT), while 1/10 of the total input chromatin was also eluted. Each sample was incubated for 4-5 h at 65°C with 0.5 µl RNaseA and a final concentration of 200 mM NaCl to reverse the cross-linking and degrade the RNA. The DNA was then precipitated with 250 µl EtOH at -20°C O/N.

On the third day, the precipitated DNA was centrifuged at 20,000 g for 20 mins, and the supernatant discarded. Each pellet was resuspended in 100 µl TE buffer with 25 µl 5x proteinase K buffer (50 mM Tris-HCl pH 7.5, 25 mM EDTA and 1.25% SDS) and incubated for 1 h at 45°C. The DNA was then purified using the Qiagen PCR Cleanup kit and eluted in 30 µl water for qPCR analysis.

7.2.12 SDS-polyacrylamide gel electrophoresis (SDS-PAGE)

The resolving gels were prepared with final concentrations of: 375 mM Tris-HCl pH 8.8, 0.1% SDS, 8% to 12% acrylamide/bisacrylamide 37.5:1 (Biorad). The stacking gels were prepared with final concentrations of: 125 mM Tris-HCl pH 6.8, 0.1% SDS, 6% acrylamide/bisacrylamide 37.5. The gels were polymerised by adding ammonium persulfate and TEMED to final concentrations of 0.033% and 0.2%, respectively, and were allowed to set for at least 10 min. Protein samples were boiled for 5 min at 96^{o}C in 2x loading buffer containing DTT before loading on the gel. All gels were run in SDS running buffer.

7.2.13 Coomassie staining

Coomassie staining was carried out by adding a solution of Coomassie Brilliant blue in 40% methanol and 10% acetic acid to SDS-PAGE gels followed by an incubation of at least 1 h. Gels were destained by several incubations with a 20% methanol/10% acetic acid solution.

7.2.14 Western blot

Cells were collected in PBS and centrifuged at 5000 g for 3 min at 4^{o}C. The pellet was lysed using 50 µl of 1X loading buffer with 100 mM DTT and denatured by boiling at 90^{o}C for 5 minutes. Proteins separated by SDS-PAGE were transferred to a nitrocellulose or PVDF membrane (Hyperbond-C Extra, GE Healthcare and Amersham Hybond PVDF, GE Healthcare respectively) using a semi-dry blotter (Transblot, BioRad). The transfer was carried out at a constant voltage of 10 V for 45 min. The membranes were blocked in 5% milk and 0.1% Tween in PBS for 1 h at RT. The primary antibody was diluted at an

appropriate concentration in fresh blocking solution, and added to the membrane for 1 h at RT or O/N at 4°C. After 3 washes of 10 min each in 5% milk with 0.1% Tween in PBS, the membrane was incubated with HRP-conjugated secondary antibody in 5% milk and 0.1% Tween in PBS for 1 h at RT. After repeating the 3 washes as previously described, the bands were detected by enhanced chemiluminescence (VisualizerTM Westernblot detection kit, Upstate). Chemiluminescence was detected with BioMax LS Film (Kodak) and developed on a Kodak X-Omat 2000 processor.

7.2.15 GST fusion protein purification

After harvesting the bacteria described in the "Protein expression" section of the Materials and Methods by ultracentrifugation, the bacteria were resuspended in the GST purification buffer containing 0.5 mM DTT, 0.2 mM PMSF and 1x protease inhibitor (Glover-Cutter et al.), and sonicated on ice using a Soniprep 150 sonicator (MSE) for several cycles of 10 sec on, 1 min off until the suspension cleared. The bacterial lysate was centrifuged at 10,000 g for 1 hour at 4°C to separate the protein from the cell debris. The lysate was then incubated O/N with an appropriate quantity of 50% Glutathione Sepharose™ 4B bead slurry (GE Healthcare). During the incubation, the GST-tagged proteins were expected to bind to the beads, while the other bacterial proteins remain free in suspension. The following day, the sonicant and beads were transferred to a plastic column with a filter, and washed in four steps with GST purification buffer containing: 0.5 M NaCl, 0.25 M NaCl, 0.1 M NaCl, and 0 M NaCl in order to remove non-specifically-bound proteins from the beads. The GST-tagged protein was eluted in 1 ml fractions using 25 mM reduced glutathione (GSH) adjusted to pH 7. Small aliquots of the elutate were collected and analysed using SDS-PAGE.

The fractions containing large amounts of protein were then pooled and dialysed into the PreScission protease buffer to remove contaminating GSH. PreScission protease was then added to the sample, which was incubated at 4°C O/N to cleave the GST tag. Following cleavage, the sample was incubated once again with Glutathione Sepharose beads. At this stage, the cleaved GST would bind to the beads, while the RPAP2 would remain in solution. RPAP2 was collected as a flow-through from the beads and concentrated in a VIVASPIN20 concentrator (Sartorius Stedim Biotech) with a cut-off molecular weight 10 kDa at 1,500 g at 4°C.

7.2.16 Ion exchange protein purification

Proteins were separated by charge on a MonoQ 5/50 column (GE Healthcare) using an ÄKTA FPLC system (GE Healthcare) as per the manufacturer's instructions, at a flow rate of 0.3 ml/min, collecting 0.5 ml fractions. GST purification buffer was used at pH 8.0. The RPAP2 not bound to the column was collected in the flow-through before fraction collection began. Data were analysed using the UNICORN software (Amersham Biosciences Ltd).

7.2.17 Differential scanning calorimetry

To analyse the thermal stability of the purified RPAP2, 12 µM protein in the GST purification buffer was submitted to the biophysical facility at the Department of Biochemistry, University of Oxford, where DSC analysis was carried out by Dr David Staunton.

7.2.18 Crystallisation trials

The purified RPAP2 sample was concentrated to 10 mg/ml. Before the crystallisation trials, the sample was subject to a Pre-Crystallisation Test (Hampton Research) that determined that the protein concentration is correct. 400 nl protein was transferred using a mosquito LCP robot (TTP Labtech) to each well of three 96-well plates: JCSG, PACT, and MORPHEUS (Molecular Dimensions). After 4 weeks, small crystals were observed in wells A2 and A3 of the PACT screen, however these were too small to test for diffraction. The crystal from well A2 was mixed with the reservoir solution and a Seed Bead (Hampton Research), and vortexed until the crystal was broken up into small fragments. A new PACT screen was then set up with the seeds. The following day, a crystal was observed in well B11 of the new PACT screen.

The diffraction experiment was carried out by Dr Edward Lowe at the Department of Biochemistry, University of Oxford. The crystal was fished from the drop with a 100 µm diameter loop and dipped into a 30% ethylene glycol-70% reservoir cryoprotectant buffer. The crystal was then cryocooled in liquid nitrogen and taken to the synchrotron at 77K. Data were collected at the I04 beamline of the Diamond Light Source. The X-ray wavelength was 0.98 Å with a rotation axis of 0.5°. A Pilatus 6M detector at a distance of 402 mm from the sample was used to collect images.

7.2.19 Phosphatase assay

The phosphatase assay was carried out as described previously (Egloff et al., 2012b). Recombinant GST-CTD was incubated with recombinant Cdk9 and Cyclin T1 in kinase buffer for 1 h at 30°C. To stop the reaction and remove ATP, the sample was then spun

through a Bio-Spin 6 Tris column (BioRad). Phosphatase buffer and appropriate amounts of recombinant RPAP2 or bacterial lambda phosphatase were then added and incubated at 30°C for 1 h.

Bacterial methods

7.2.20 Transformation of competent bacterial cells

Competent NM554 bacterial cells were thawed on ice and gently resuspended. 1-2 µg of DNA was added to the cells and mixed gently, incubated on ice for 10 min and heat shocked at 37^{o}C for 5 min in a dry heater. The whole cell mixture was plated onto L-Agar plates with appropriate antibiotics and incubated at 37^{o}C overnight.

References

ABOU ELELA, S. & ARES, M., JR. 1998. Depletion of yeast RNase III blocks correct U2 3' end formation and results in polyadenylated but functional U2 snRNA. *EMBO J,* 17**,** 3738-46.

AGGEN, J. B., NAIRN, A. C. & CHAMBERLIN, R. 2000. Regulation of protein phosphatase-1. *Chem Biol,* 7**,** R13-23.

AKHTAR, M. S., HEIDEMANN, M., TIETJEN, J. R., ZHANG, D. W., CHAPMAN, R. D., EICK, D. & ANSARI, A. Z. 2009. TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. *Mol Cell,* 34**,** 387-93.

AKKERS, R. C., VAN HEERINGEN, S. J., JACOBI, U. G., JANSSEN-MEGENS, E. M., FRANCOIJS, K. J., STUNNENBERG, H. G. & VEENSTRA, G. J. 2009. A hierarchy of H3K4me3 and H3K27me3 acquisition in spatial gene regulation in Xenopus embryos. *Dev Cell,* 17**,** 425-34.

AKOULITCHEV, S., MAKELA, T. P., WEINBERG, R. A. & REINBERG, D. 1995. Requirement for TFIIH kinase activity in transcription by RNA polymerase II. *Nature,* 377**,** 557-60.

ALBRECHT, T. R. & WAGNER, E. J. 2012. snRNA 3' end formation requires heterodimeric association of integrator subunits. *Mol Cell Biol,* 32**,** 1112-23.

ALLISON, L. A., WONG, J. K., FITZPATRICK, V. D., MOYLE, M. & 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. *Mol Cell Biol,* 8**,** 321-9.

AMMOSOVA, T., OBUKHOV, Y., KOTELKIN, A., BREUER, D., BEULLENS, M., GORDEUK, V. R., BOLLEN, M. & NEKHAI, S. 2011. Protein phosphatase-1 activates CDK9 by dephosphorylating Ser175. *PLoS One,* 6**,** e18985.

ARCHAMBAULT, J., CHAMBERS, R. S., KOBOR, M. S., HO, Y., CARTIER, M., BOLOTIN, D., ANDREWS, B., KANE, C. M. & GREENBLATT, J. 1997. An essential component of a C-terminal domain phosphatase that interacts with transcription factor IIF in Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A,* 94**,** 14300-5.

ATKINSON, S. P., KOCH, C. M., CLELLAND, G. K., WILLCOX, S., FOWLER, J. C., STEWART, R., LAKO, M., DUNHAM, I. & ARMSTRONG, L. 2008. Epigenetic marking prepares the human HOXA cluster for activation during differentiation of pluripotent cells. *Stem Cells,* 26**,** 1174-85.

BAILLAT, D., HAKIMI, M. A., NAAR, A. M., SHILATIFARD, A., COOCH, N. & SHIEKHATTAR, R. 2005. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell,* 123**,** 265-76.

BARTKOWIAK, B., LIU, P., PHATNANI, H. P., FUDA, N. J., COOPER, J. J., PRICE, D. H., ADELMAN, K., LIS, J. T. & GREENLEAF, A. L. 2010. CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev,* 24**,** 2303-16.

BARTOLOMEI, M. S., HALDEN, N. F., CULLEN, C. R. & CORDEN, J. L. 1988. Genetic analysis of the repetitive carboxyl-terminal domain of the largest subunit of mouse RNA polymerase II. *Mol Cell Biol,* 8**,** 330-9.

BASEHOAR, A. D., ZANTON, S. J. & PUGH, B. F. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell,* 116**,** 699-709.

BASKARAN, R., CHIANG, G. G., MYSLIWIEC, T., KRUH, G. D. & WANG, J. Y. 1997. Tyrosine phosphorylation of RNA polymerase II carboxyl-terminal domain by the Abl-related gene product. *J Biol Chem,* 272**,** 18905-9.

BATAILLE, A. R., JERONIMO, C., JACQUES, P. E., LARAMEE, L., FORTIN, M. E., FOREST, A., BERGERON, M., HANES, S. D. & ROBERT, F. 2012. A universal RNA polymerase II CTD cycle is orchestrated by complex interplays between kinase, phosphatase, and isomerase enzymes along genes. *Mol Cell,* 45**,** 158-70.

BENTLEY, D. L. 2014. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet,* 15**,** 163-75.

BERNSTEIN, B. E., MIKKELSEN, T. S., XIE, X., KAMAL, M., HUEBERT, D. J., CUFF, J., FRY, B., MEISSNER, A., WERNIG, M., PLATH, K., JAENISCH, R., WAGSCHAL, A., FEIL, R., SCHREIBER, S. L. & LANDER, E. S. 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell,* 125**,** 315-26.

BHAUMIK, S. R. & GREEN, M. R. 2002. Differential requirement of SAGA components for recruitment of TATA-box-binding protein to promoters in vivo. *Mol Cell Biol,* 22**,** 7365-71.

BIENIOSSEK, C., PAPAI, G., SCHAFFITZEL, C., GARZONI, F., CHAILLET, M., SCHEER, E., PAPADOPOULOS, P., TORA, L., SCHULTZ, P. & BERGER, I. 2013. The architecture of human general transcription factor TFIID core complex. *Nature,* 493**,** 699-702.

BLAZEK, D., KOHOUTEK, J., BARTHOLOMEEUSEN, K., JOHANSEN, E., HULINKOVA, P., LUO, Z., CIMERMANCIC, P., ULE, J. & PETERLIN, B. M. 2011. The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev,* 25**,** 2158-72.

BRANNAN, K., KIM, H., ERICKSON, B., GLOVER-CUTTER, K., KIM, S., FONG, N., KIEMELE, L., HANSEN, K., DAVIS, R., LYKKE-ANDERSEN, J. & BENTLEY, D. L. 2012. mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell,* 46**,** 311-24.

BROW, D. A. 2002. Allosteric cascade of spliceosome activation. *Annu Rev Genet,* 36**,** 333-60.

BURATOWSKI, S. 2003. The CTD code. *Nat Struct Biol,* 10**,** 679-80.

CARNINCI, P., SANDELIN, A., LENHARD, B., KATAYAMA, S., SHIMOKAWA, K., PONJAVIC, J., SEMPLE, C. A., TAYLOR, M. S., ENGSTROM, P. G., FRITH, M. C., FORREST, A. R., ALKEMA, W. B., TAN, S. L., PLESSY, C., KODZIUS, R., RAVASI, T., KASUKAWA, T., FUKUDA, S., KANAMORI-KATAYAMA, M., KITAZUME, Y., KAWAJI, H., KAI, C., NAKAMURA, M., KONNO, H., NAKANO, K., MOTTAGUI-TABAR, S., ARNER, P., CHESI, A., GUSTINCICH, S., PERSICHETTI, F., SUZUKI, H., GRIMMOND, S. M., WELLS, C. A., ORLANDO, V., WAHLESTEDT, C., LIU, E. T., HARBERS, M., KAWAI, J., BAJIC, V. B., HUME, D. A. & HAYASHIZAKI, Y. 2006. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet,* 38**,** 626-35.

CHANFREAU, G., ELELA, S. A., ARES, M., JR. & GUTHRIE, C. 1997. Alternative 3'-end processing of U5 snRNA by RNase III. *Genes Dev,* 11**,** 2741-51.

CHAPMAN, R. D., CONRAD, M. & EICK, D. 2005. Role of the mammalian RNA polymerase II C-terminal domain (CTD) nonconsensus repeats in CTD stability and cell proliferation. *Mol Cell Biol,* 25**,** 7665-74.

CHAPMAN, R. D., HEIDEMANN, M., ALBERT, T. K., MAILHAMMER, R., FLATLEY, A., MEISTERERNST, M., KREMMER, E. & EICK, D. 2007. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. *Science,* 318**,** 1780-2.

CHAPMAN, R. D., HEIDEMANN, M., HINTERMAIR, C. & EICK, D. 2008. Molecular evolution of the RNA polymerase II CTD. *Trends Genet,* 24**,** 289-96.

CHEN, M. & MANLEY, J. L. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol,* 10**,** 741-54.

CHEN, R., LIU, M., LI, H., XUE, Y., RAMEY, W. N., HE, N., AI, N., LUO, H., ZHU, Y., ZHOU, N. & ZHOU, Q. 2008. PP2B and PP1alpha cooperatively disrupt 7SK snRNP to release P-TEFb for transcription in response to Ca2+ signaling. *Genes Dev,* 22**,** 1356-68.

CHEN, X. F., LEHMANN, L., LIN, J. J., VASHISHT, A., SCHMIDT, R., FERRARI, R., HUANG, C., MCKEE, R., MOSLEY, A., PLATH, K., KURDISTANI, S. K., WOHLSCHLEGEL, J. & CAREY, M. 2012. Mediator and SAGA have distinct roles in Pol II preinitiation complex assembly and function. *Cell Rep,* 2**,** 1061-7.

CHEN, Y., ZHANG, L., ESTARAS, C., CHOI, S. H., MORENO, L., JR., KARN, J., MORESCO, J. J., YATES, J. R., 3RD & JONES, K. A. 2014. A gene-specific role for the Ssu72 RNAPII CTD phosphatase in HIV-1 Tat transactivation. *Genes Dev,* 28**,** 2261-75.

CHIBA, K., YAMAMOTO, J., YAMAGUCHI, Y. & HANDA, H. 2010. Promoter-proximal pausing and its release: molecular mechanisms and physiological functions. *Exp Cell Res,* 316**,** 2723-30.

CHO, E. J., KOBOR, M. S., KIM, M., GREENBLATT, J. & BURATOWSKI, S. 2001. Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. *Genes Dev,* 15**,** 3319-29.

CIURCIU, A., DUNCALF, L., JONCHERE, V., LANSDALE, N., VASIEVA, O., GLENDAY, P., RUDENKO, A., VISSI, E., COBBE, N., ALPHEY, L. & BENNETT, D. 2013. PNUTS/PP1 regulates RNAPII-mediated gene expression and is necessary for developmental growth. *PLoS Genet,* 9**,** e1003885.

CLEMENTE-BLANCO, A., SEN, N., MAYAN-SANTOS, M., SACRISTAN, M. P., GRAHAM, B., JARMUZ, A., GIESS, A., WEBB, E., GAME, L., EICK, D., BUENO, A., MERKENSCHLAGER, M. & ARAGON, L. 2011. Cdc14 phosphatase promotes segregation of telomeres through repression of RNA polymerase II transcription. *Nat Cell Biol,* 13**,** 1450-6.

COLE, C., BARBER, J. D. & BARTON, G. J. 2008. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res,* 36**,** W197-201.

CRAMER, P., BUSHNELL, D. A., FU, J., GNATT, A. L., MAIER-DAVIS, B., THOMPSON, N. E., BURGESS, R. R., EDWARDS, A. M., DAVID, P. R. & KORNBERG, R. D. 2000. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science,* 288**,** 640-9.

CRAMER, P., BUSHNELL, D. A. & KORNBERG, R. D. 2001. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science,* 292**,** 1863-76.

CUELLO, P., BOYD, D. C., DYE, M. J., PROUDFOOT, N. J. & MURPHY, S. 1999. Transcription of the human U2 snRNA genes continues beyond the 3' box in vivo. *EMBO J,* 18**,** 2867-77.

CZUDNOCHOWSKI, N., BOSKEN, C. A. & GEYER, M. 2012. Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. *Nat Commun,* 3**,** 842.

DANIEL, B., NAGY, G. & NAGY, L. 2014. The intriguing complexities of mammalian gene regulation: how to link enhancers to regulated genes. Are we there yet? *FEBS Lett,* 588**,** 2379-91.

DAVIDSON, L., MUNIZ, L. & WEST, S. 2014. 3' end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. *Genes Dev,* 28**,** 342-56.

DE BRAEKELEER, M., MOREL, F., LE BRIS, M. J., HERRY, A. & DOUET-GUILBERT, N. 2005. The MLL gene and translocations involving chromosomal band 11q23 in acute leukemia. *Anticancer Res,* 25**,** 1931-44.

DE VEGVAR, H. E., LUND, E. & DAHLBERG, J. E. 1986. 3' end formation of U1 snRNA precursors is coupled to transcription from snRNA promoters. *Cell,* 47**,** 259-66.

DENG, W., MALECOVA, B., OELGESCHLAGER, T. & ROBERTS, S. G. 2009. TFIIB recognition elements control the TFIIA-NC2 axis in transcriptional regulation. *Mol Cell Biol,* 29**,** 1389-400.

DESCOSTES, N., HEIDEMANN, M., SPINELLI, L., SCHULLER, R., MAQBOOL, M. A., FENOUIL, R., KOCH, F., INNOCENTI, C., GUT, M., GUT, I., EICK, D. & ANDRAU, J. C. 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 & SINGER, D. S. 2012. BRD4 is an atypical kinase that phosphorylates serine2 of the RNA polymerase II carboxy-terminal domain. *Proc Natl Acad Sci U S A,* 109**,** 6927-32.

DEVAIAH, B. N. & SINGER, D. S. 2013. Two faces of brd4: mitotic bookmark and transcriptional lynchpin. *Transcription,* 4**,** 13-7.

DI GIAMMARTINO, D. C., NISHIDA, K. & MANLEY, J. L. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol Cell,* 43**,** 853-66.

DICHTL, B., BLANK, D., OHNACKER, M., FRIEDLEIN, A., ROEDER, D., LANGEN, H. & KELLER, W. 2002. A role for SSU72 in balancing RNA polymerase II transcription elongation and termination. *Mol Cell,* 10**,** 1139-50.

DOMINSKI, Z., YANG, X. C., PURDY, M., WAGNER, E. J. & MARZLUFF, W. F. 2005. A CPSF-73 homologue is required for cell cycle progression but not cell growth and interacts with a protein having features of CPSF-100. *Mol Cell Biol,* 25**,** 1489-500.

DUBOULE, D. & DOLLE, P. 1989. The structural and functional organization of the murine HOX gene family resembles that of Drosophila homeotic genes. *EMBO J,* 8**,** 1497-505.

DUNDR, M. 2012. Nuclear bodies: multifunctional companions of the genome. *Curr Opin Cell Biol,* 24**,** 415-22.

EDMONDS, M., VAUGHAN, M. H., JR. & NAKAZATO, H. 1971. Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. *Proc Natl Acad Sci U S A,* 68**,** 1336-40.

EGLOFF, S. 2012. Role of Ser7 phosphorylation of the CTD during transcription of snRNA genes. *RNA Biol,* 9**,** 1033-8.

EGLOFF, S., AL-RAWAF, H., O'REILLY, D. & MURPHY, S. 2009. Chromatin structure is implicated in "late" elongation checkpoints on the U2 snRNA and beta-actin genes. *Mol Cell Biol,* 29**,** 4002-13.

EGLOFF, S., DIENSTBIER, M. & MURPHY, S. 2012a. Updating the RNA polymerase CTD code: adding gene-specific layers. *Trends Genet,* 28**,** 333-41.

EGLOFF, S. & MURPHY, S. 2008. Role of the C-terminal domain of RNA polymerase II in expression of small nuclear RNA genes. *Biochem Soc Trans,* 36**,** 537-9.

EGLOFF, S., O'REILLY, D., CHAPMAN, R. D., TAYLOR, A., TANZHAUS, K., PITTS, L., EICK, D. & MURPHY, S. 2007. Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science,* 318**,** 1777-9.

EGLOFF, S., O'REILLY, D. & MURPHY, S. 2008. Expression of human snRNA genes from beginning to end. *Biochem Soc Trans,* 36**,** 590-4.

EGLOFF, S., SZCZEPANIAK, S. A., DIENSTBIER, M., TAYLOR, A., KNIGHT, S. & MURPHY, S. 2010. The integrator complex recognizes a new double mark on the RNA polymerase II carboxyl-terminal domain. *J Biol Chem,* 285**,** 20564-9.

EGLOFF, S., ZABOROWSKA, J., LAITEM, C., KISS, T. & MURPHY, S. 2012b. Ser7 phosphorylation of the CTD recruits the RPAP2 Ser5 phosphatase to snRNA genes. *Mol Cell,* 45**,** 111-22.

EZZEDDINE, N., CHEN, J., WALTENSPIEL, B., BURCH, B., ALBRECHT, T., ZHUO, M., WARREN, W. D., MARZLUFF, W. F. & WAGNER, E. J. 2011. A subset of Drosophila integrator proteins is essential for efficient U7 snRNA and spliceosomal snRNA 3'-end formation. *Mol Cell Biol,* 31**,** 328-41.

FORGET, D., LACOMBE, A. A., CLOUTIER, P., LAVALLEE-ADAM, M., BLANCHETTE, M. & COULOMBE, B. 2013. Nuclear import of RNA polymerase II is coupled with nucleocytoplasmic shuttling of the RNA polymerase II-associated protein 2. *Nucleic Acids Res,* 41**,** 6881-91.

FUDA, N. J., BUCKLEY, M. S., WEI, W., CORE, L. J., WATERS, C. T., REINBERG, D. & LIS, J. T. 2012. Fcp1 dephosphorylation of the RNA polymerase II C-terminal domain is required for efficient transcription of heat shock genes. *Mol Cell Biol,* 32**,** 3428-37.

FUJINAGA, K., IRWIN, D., HUANG, Y., TAUBE, R., KUROSU, T. & PETERLIN, B. M. 2004. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element. *Mol Cell Biol,* 24**,** 787-95.

GAN, Q., SCHONES, D. E., HO EUN, S., WEI, G., CUI, K., ZHAO, K. & CHEN, X. 2010. Monovalent and unpoised status of most genes in undifferentiated cell-enriched Drosophila testis. *Genome Biol,* 11**,** R42.

GANEM, C., DEVAUX, F., TORCHET, C., JACQ, C., QUEVILLON-CHERUEL, S., LABESSE, G., FACCA, C. & FAYE, G. 2003. Ssu72 is a phosphatase essential for transcription termination of snoRNAs and specific mRNAs in yeast. *EMBO J,* 22**,** 1588-98.

GARDINI, A., BAILLAT, D., CESARONI, M., HU, D., MARINIS, J. M., WAGNER, E. J., LAZAR, M. A., SHILATIFARD, A. & SHIEKHATTAR, R. 2014. Integrator regulates transcriptional initiation and pause release following activation. *Mol Cell,* 56**,** 128-39.

GIBNEY, P. A., FRIES, T., BAILER, S. M. & MORANO, K. A. 2008. Rtr1 is the Saccharomyces cerevisiae homolog of a novel family of RNA polymerase II-binding proteins. *Eukaryot Cell,* 7**,** 938-48.

GIL, A. & PROUDFOOT, N. J. 1984. A sequence downstream of AAUAAA is required for rabbit beta-globin mRNA 3'-end formation. *Nature,* 312**,** 473-4.

GLOVER-CUTTER, K., KIM, S., ESPINOSA, J. & BENTLEY, D. L. 2008. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat Struct Mol Biol,* 15**,** 71-8.

GLOVER-CUTTER, K., LAROCHELLE, S., ERICKSON, B., ZHANG, C., SHOKAT, K., FISHER, R. P. & 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. *Mol Cell Biol,* 29**,** 5455-64.

GROMAK, N., WEST, S. & PROUDFOOT, N. J. 2006. Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol Cell Biol,* 26**,** 3986-96.

GRUNBERG, S. & HAHN, S. 2013. Structural insights into transcription initiation by RNA polymerase II. *Trends Biochem Sci,* 38**,** 603-11.

GU, B., EICK, D. & BENSAUDE, O. 2013. CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. *Nucleic Acids Res,* 41**,** 1591-603.

GUENTHER, M. G., LEVINE, S. S., BOYER, L. A., JAENISCH, R. & YOUNG, R. A. 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell,* 130**,** 77-88.

HE, N., CHAN, C. K., SOBHIAN, B., CHOU, S., XUE, Y., LIU, M., ALBER, T., BENKIRANE, M. & ZHOU, Q. 2011. Human Polymerase-Associated Factor complex (PAFc) connects the Super Elongation Complex (SEC) to RNA polymerase II on chromatin. *Proc Natl Acad Sci U S A,* 108**,** E636-45.

HEIDEMANN, M., HINTERMAIR, C., VOSS, K. & EICK, D. 2013. Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochim Biophys Acta,* 1829**,** 55-62.

HENRY, R. W., MITTAL, V., MA, B., KOBAYASHI, R. & HERNANDEZ, N. 1998. SNAP19 mediates the assembly of a functional core promoter complex (SNAPc) shared by RNA polymerases II and III. *Genes Dev,* 12**,** 2664-72.

HERNANDEZ, N. 1985. Formation of the 3' end of U1 snRNA is directed by a conserved sequence located downstream of the coding region. *EMBO J,* 4**,** 1827-37.

HERNANDEZ, N. 2001. Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. *J Biol Chem,* 276**,** 26733-6.

HERNANDEZ, N. & LUCITO, R. 1988. Elements required for transcription initiation of the human U2 snRNA gene coincide with elements required for snRNA 3' end formation. *EMBO J,* 7**,** 3125-34.

HERNANDEZ, N. & WEINER, A. M. 1986. Formation of the 3' end of U1 snRNA requires compatible snRNA promoter elements. *Cell,* 47**,** 249-58.

HINTERMAIR, C., HEIDEMANN, M., KOCH, F., DESCOSTES, N., GUT, M., GUT, I., FENOUIL, R., FERRIER, P., FLATLEY, A., KREMMER, E., CHAPMAN, R. D., ANDRAU, J. C. & EICK, D. 2012. Threonine-4 of mammalian RNA polymerase II CTD is targeted by Polo-like kinase 3 and required for transcriptional elongation. *EMBO J,* 31**,** 2784-97.

HIROSE, Y., TACKE, R. & MANLEY, J. L. 1999. Phosphorylated RNA polymerase II stimulates pre-mRNA splicing. *Genes Dev,* 13**,** 1234-9.

HOSKINS, A. A. & MOORE, M. J. 2012. The spliceosome: a flexible, reversible macromolecular machine. *Trends Biochem Sci,* 37**,** 179-88.

HSIN, J. P., LI, W., HOQUE, M., TIAN, B. & MANLEY, J. L. 2014. RNAP II CTD tyrosine 1 performs diverse functions in vertebrate cells. *Elife,* 3**,** e02112.

HSIN, J. P. & MANLEY, J. L. 2012. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev,* 26**,** 2119-37.

HSIN, J. P., SHETH, A. & MANLEY, J. L. 2011. RNAP II CTD phosphorylated on threonine-4 is required for histone mRNA 3' end processing. *Science,* 334**,** 683-6.

HSU, P. L., YANG, F., SMITH-KINNAMAN, W., YANG, W., SONG, J. E., MOSLEY, A. L. & VARANI, G. 2014. Rtr1 is a dual specificity phosphatase that dephosphorylates Tyr1 and Ser5 on the RNA polymerase II CTD. *J Mol Biol,* 426**,** 2970-81.

JACOBS, E. Y., OGIWARA, I. & WEINER, A. M. 2004. Role of the C-terminal domain of RNA polymerase II in U2 snRNA transcription and 3' processing. *Mol Cell Biol,* 24**,** 846-55.

JAMES FARESSE, N., CANELLA, D., PRAZ, V., MICHAUD, J., ROMASCANO, D. & HERNANDEZ, N. 2012. Genomic study of RNA polymerase II and III SNAPc-bound promoters reveals a gene transcribed by both enzymes and a broad use of common activators. *PLoS Genet,* 8**,** e1003028.

JANG, M. K., MOCHIZUKI, K., ZHOU, M., JEONG, H. S., BRADY, J. N. & OZATO, K. 2005. The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. *Mol Cell,* 19**,** 523-34.

JEREBTSOVA, M., KLOTCHENKO, S. A., ARTAMONOVA, T. O., AMMOSOVA, T., WASHINGTON, K., EGOROV, V. V., SHALDZHYAN, A. A., SERGEEVA, M. V., ZATULOVSKIY, E. A., TEMKINA, O. A., PETUKHOV, M. G., VASIN, A. V., KHODORKOVSKII, M. A., ORLOV, Y. N. & NEKHAI, S. 2011. Mass spectrometry and biochemical analysis of RNA polymerase II: targeting by protein phosphatase-1. *Mol Cell Biochem,* 347**,** 79-87.

JERONIMO, C., FORGET, D., BOUCHARD, A., LI, Q., CHUA, G., POITRAS, C., THERIEN, C., BERGERON, D., BOURASSA, S., GREENBLATT, J., CHABOT, B., POIRIER, G. G., HUGHES, T. R., BLANCHETTE, M., PRICE, D. H. & COULOMBE, B. 2007. Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. *Mol Cell,* 27**,** 262-74.

JONKERS, I., KWAK, H. & LIS, J. T. 2014. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife,* 3**,** e02407.

JUVEN-GERSHON, T., HSU, J. Y., THEISEN, J. W. & KADONAGA, J. T. 2008. The RNA polymerase II core promoter - the gateway to transcription. *Curr Opin Cell Biol,* 20**,** 253-9.

KASHYAP, V., GUDAS, L. J., BRENET, F., FUNK, P., VIALE, A. & SCANDURA, J. M. 2011. Epigenomic reorganization of the clustered Hox genes in embryonic stem cells induced by retinoic acid. *J Biol Chem,* 286**,** 3250-60.

KEEGAN, B. R., FELDMAN, J. L., LEE, D. H., KOOS, D. S., HO, R. K., STAINIER, D. Y. & YELON, D. 2002. The elongation factors Pandora/Spt6 and Foggy/Spt5 promote transcription in the zebrafish embryo. *Development,* 129**,** 1623-32.

KELLY, W. G., DAHMUS, M. E. & HART, G. W. 1993. RNA polymerase II is a glycoprotein. Modification of the COOH-terminal domain by O-GlcNAc. *J Biol Chem,* 268**,** 10416-24.

KIM, J. B. & 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. *J Biol Chem,* 276**,** 12317-23.

KIM, M., SUH, H., CHO, E. J. & BURATOWSKI, S. 2009. Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2, 5, and 7. *J Biol Chem,* 284**,** 26421-6.

KIM, Y. J., BJORKLUND, S., LI, Y., SAYRE, M. H. & 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.

KOMARNITSKY, P., CHO, E. J. & BURATOWSKI, S. 2000. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev,* 14**,** 2452-60.

KRISHNAMURTHY, S. & HAMPSEY, M. 2009. Eukaryotic transcription initiation. *Curr Biol,* 19**,** R153-6.

KRISHNAMURTHY, S., HE, X., REYES-REYES, M., MOORE, C. & HAMPSEY, M. 2004. Ssu72 Is an RNA polymerase II CTD phosphatase. *Mol Cell,* 14**,** 387-94.

KUBICEK, K., CERNA, H., HOLUB, P., PASULKA, J., HROSSOVA, D., LOEHR, F., HOFR, C., VANACOVA, S. & STEFL, R. 2012. Serine phosphorylation and proline isomerization in RNAP II CTD control recruitment of Nrd1. *Genes Dev,* 26**,** 1891-6.

KUGEL, J. F. & GOODRICH, J. A. 2000. A kinetic model for the early steps of RNA synthesis by human RNA polymerase II. *J Biol Chem,* 275**,** 40483-91.

KUHLMAN, T. C., CHO, H., REINBERG, D. & HERNANDEZ, N. 1999. The general transcription factors IIA, IIB, IIF, and IIE are required for RNA polymerase II transcription from the human U1 small nuclear RNA promoter. *Mol Cell Biol,* 19**,** 2130-41.

LAINE, J. P. & EGLY, J. M. 2006. When transcription and repair meet: a complex system. *Trends Genet,* 22**,** 430-6.

LAITEM, C., ZABOROWSKA, J., ISA, N. F., KUFS, J., DIENSTBIER, M. & MURPHY, S. 2015. CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. *Nat Struct Mol Biol*.

LAMOND, A. I. & CARMO-FONSECA, M. 1993. Localisation of splicing snRNPs in mammalian cells. *Mol Biol Rep,* 18**,** 127-33.

LANDEIRA, D., SAUER, S., POOT, R., DVORKINA, M., MAZZARELLA, L., JORGENSEN, H. F., PEREIRA, C. F., LELEU, M., PICCOLO, F. M., SPIVAKOV, M., BROOKES, E., POMBO, A., FISHER, C., SKARNES, W. C., SNOEK, T., BEZSTAROSTI, K., DEMMERS, J., KLOSE, R. J., CASANOVA, M., TAVARES, L., BROCKDORFF, N., MERKENSCHLAGER, M. & FISHER, A. G. 2010. Jarid2 is a PRC2 component in embryonic stem cells required for multi-lineage differentiation and recruitment of PRC1 and RNA Polymerase II to developmental regulators. *Nat Cell Biol,* 12**,** 618-24.

LAPPIN, T. R., GRIER, D. G., THOMPSON, A. & HALLIDAY, H. L. 2006. HOX genes: seductive science, mysterious mechanisms. *Ulster Med J,* 75**,** 23-31.

LEE, M. & NURSE, P. 1988. Cell cycle control genes in fission yeast and mammalian cells. *Trends Genet,* 4**,** 287-90.

LEE, T. I., CAUSTON, H. C., HOLSTEGE, F. C., SHEN, W. C., HANNETT, N., JENNINGS, E. G., WINSTON, F., GREEN, M. R. & YOUNG, R. A. 2000. Redundant roles for the TFIID and SAGA complexes in global transcription. *Nature,* 405**,** 701-4.

LEE, Y., KIM, M., HAN, J., YEOM, K. H., LEE, S., BAEK, S. H. & KIM, V. N. 2004. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J,* 23**,** 4051-60.

LENASI, T., PETERLIN, B. M. & BARBORIC, M. 2011. Cap-binding protein complex links pre-mRNA capping to transcription elongation and alternative splicing through positive transcription elongation factor b (P-TEFb). *J Biol Chem,* 286**,** 22758-68.

LENGERKE, C., MCKINNEY-FREEMAN, S., NAVEIRAS, O., YATES, F., WANG, Y., BANSAL, D. & DALEY, G. Q. 2007. The cdx-hox pathway in hematopoietic stem cell formation from embryonic stem cells. *Ann N Y Acad Sci,* 1106**,** 197-208.

LENGERKE, C., SCHMITT, S., BOWMAN, T. V., JANG, I. H., MAOUCHE-CHRETIEN, L., MCKINNEY-FREEMAN, S., DAVIDSON, A. J., HAMMERSCHMIDT, M., RENTZSCH, F., GREEN, J. B., ZON,

L. I. & DALEY, G. Q. 2008. BMP and Wnt specify hematopoietic fate by activation of the Cdx-Hox pathway. *Cell Stem Cell,* 2**,** 72-82.

LEWIS, E. B. 1978. A gene complex controlling segmentation in Drosophila. *Nature,* 276**,** 565-70.

LEWIS, J. D. & IZAURRALDE, E. 1997. The role of the cap structure in RNA processing and nuclear export. *Eur J Biochem,* 247**,** 461-9.

LICCIARDO, P., AMENTE, S., RUGGIERO, L., MONTI, M., PUCCI, P., LANIA, L. & MAJELLO, B. 2003. The FCP1 phosphatase interacts with RNA polymerase II and with MEP50 a component of the methylosome complex involved in the assembly of snRNP. *Nucleic Acids Res,* 31**,** 999-1005.

LIDSCHREIBER, M., LEIKE, K. & CRAMER, P. 2013. Cap completion and C-terminal repeat domain kinase recruitment underlie the initiation-elongation transition of RNA polymerase II. *Mol Cell Biol,* 33**,** 3805-16.

LIN, P. S., DUBOIS, M. F. & DAHMUS, M. E. 2002. TFIIF-associating carboxyl-terminal domain phosphatase dephosphorylates phosphoserines 2 and 5 of RNA polymerase II. *J Biol Chem,* 277**,** 45949-56.

LIN, S., COUTINHO-MANSFIELD, G., WANG, D., PANDIT, S. & FU, X. D. 2008. The splicing factor SC35 has an active role in transcriptional elongation. *Nat Struct Mol Biol,* 15**,** 819-26.

LINDGREN, V., ARES, M., JR., WEINER, A. M. & FRANCKE, U. 1985a. Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. *Nature,* 314**,** 115-6.

LINDGREN, V., BERNSTEIN, L. B., WEINER, A. M. & FRANCKE, U. 1985b. Human U1 small nuclear RNA pseudogenes do not map to the site of the U1 genes in 1p36 but are clustered in 1q12-q22. *Mol Cell Biol,* 5**,** 2172-80.

LINDING, R., RUSSELL, R. B., NEDUVA, V. & GIBSON, T. J. 2003. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res,* 31**,** 3701-8.

LIS, J. T., MASON, P., PENG, J., PRICE, D. H. & WERNER, J. 2000. P-TEFb kinase recruitment and function at heat shock loci. *Genes Dev,* 14**,** 792-803.

LIU, J., LIU, H., ZHANG, X., GAO, P., WANG, J. & HU, Z. 2002. Identification and characterization of P15RS, a novel P15(INK4b) related gene on G1/S progression. *Biochem Biophys Res Commun,* 299**,** 880-5.

LIU, P., KENNEY, J. M., STILLER, J. W. & GREENLEAF, A. L. 2010. Genetic organization, length conservation, and evolution of RNA polymerase II carboxyl-terminal domain. *Mol Biol Evol,* 27**,** 2628-41.

LIU, X., BUSHNELL, D. A. & KORNBERG, R. D. 2013. RNA polymerase II transcription: structure and mechanism. *Biochim Biophys Acta,* 1829**,** 2-8.

LUNDE, B. M., REICHOW, S. L., KIM, M., SUH, H., LEEPER, T. C., YANG, F., MUTSCHLER, H., BURATOWSKI, S., MEINHART, A. & VARANI, G. 2010. Cooperative interaction of transcription termination factors with the RNA polymerase II C-terminal domain. *Nat Struct Mol Biol,* 17**,** 1195-201.

LUO, Y., YOGESHA, S. D., CANNON, J. R., YAN, W., ELLINGTON, A. D., BRODBELT, J. S. & ZHANG, Y. 2013. novel modifications on C-terminal domain of RNA polymerase II can fine-tune the phosphatase activity of Ssu72. *ACS Chem Biol,* 8**,** 2042-52.

LUO, Z., LIN, C., GUEST, E., GARRETT, A. S., MOHAGHEGH, N., SWANSON, S., MARSHALL, S., FLORENS, L., WASHBURN, M. P. & SHILATIFARD, A. 2012. The super elongation complex family of RNA polymerase II elongation factors: gene target specificity and transcriptional output. *Mol Cell Biol,* 32**,** 2608-17.

MANCEBO, H. S., LEE, G., FLYGARE, J., TOMASSINI, J., LUU, P., ZHU, Y., PENG, J., BLAU, C., HAZUDA, D., PRICE, D. & FLORES, O. 1997. P-TEFb kinase is required for HIV Tat transcriptional activation in vivo and in vitro. *Genes Dev,* 11**,** 2633-44.

MARSHALL, N. F., PENG, J., XIE, Z. & PRICE, D. H. 1996. Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. *J Biol Chem,* 271**,** 27176-83.

MARTINS, S. B., RINO, J., CARVALHO, T., CARVALHO, C., YOSHIDA, M., KLOSE, J. M., DE ALMEIDA, S. F. & CARMO-FONSECA, M. 2011. Spliceosome assembly is coupled to RNA polymerase II dynamics at the 3' end of human genes. *Nat Struct Mol Biol,* 18**,** 1115-23.

MARZLUFF, W. F., WAGNER, E. J. & DURONIO, R. J. 2008. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat Rev Genet,* 9**,** 843-54.

MATERA, A. G. & WANG, Z. 2014. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol,* 15**,** 108-21.

MATERA, A. G. & WARD, D. C. 1993. Nucleoplasmic organization of small nuclear ribonucleoproteins in cultured human cells. *J Cell Biol,* 121**,** 715-27.

MATSUI, T., SEGALL, J., WEIL, P. A. & ROEDER, R. G. 1980. Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. *J Biol Chem,* 255**,** 11992-6.

MAYER, A., HEIDEMANN, M., LIDSCHREIBER, M., SCHREIECK, A., SUN, M., HINTERMAIR, C., KREMMER, E., EICK, D. & CRAMER, P. 2012. CTD tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science,* 336**,** 1723-5.

MBONYE, U. R., GOKULRANGAN, G., DATT, M., DOBROWOLSKI, C., COOPER, M., CHANCE, M. R. & KARN, J. 2013. Phosphorylation of CDK9 at Ser175 enhances HIV transcription and is a marker of activated P-TEFb in CD4(+) T lymphocytes. *PLoS Pathog,* 9**,** e1003338.

MEDLIN, J., SCURRY, A., TAYLOR, A., ZHANG, F., PETERLIN, B. M. & MURPHY, S. 2005. P-TEFb is not an essential elongation factor for the intronless human U2 snRNA and histone H2b genes. *EMBO J,* 24**,** 4154-65.

MEDLIN, J. E., UGUEN, P., TAYLOR, A., BENTLEY, D. L. & MURPHY, S. 2003. The C-terminal domain of pol II and a DRB-sensitive kinase are required for 3' processing of U2 snRNA. *EMBO J,* 22**,** 925-34.

MEINHART, A. & CRAMER, P. 2004. Recognition of RNA polymerase II carboxy-terminal domain by 3'-RNA-processing factors. *Nature,* 430**,** 223-6.

MIKKELSEN, T. S., KU, M., JAFFE, D. B., ISSAC, B., LIEBERMAN, E., GIANNOUKOS, G., ALVAREZ, P., BROCKMAN, W., KIM, T. K., KOCHE, R. P., LEE, W., MENDENHALL, E., O'DONOVAN, A., PRESSER, A., RUSS, C., XIE, X., MEISSNER, A., WERNIG, M., JAENISCH, R., NUSBAUM, C., LANDER, E. S. & BERNSTEIN, B. E. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature,* 448**,** 553-60.

MILLEVOI, S. & VAGNER, S. 2010. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Res,* 38**,** 2757-74.

MIN, I. M., WATERFALL, J. J., CORE, L. J., MUNROE, R. J., SCHIMENTI, J. & LIS, J. T. 2011. Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev,* 25**,** 742-54.

MORTILLARO, M. J., BLENCOWE, B. J., WEI, X., NAKAYASU, H., DU, L., WARREN, S. L., SHARP, P. A. & BEREZNEY, R. 1996. A hyperphosphorylated form of the large subunit of RNA polymerase II is associated with splicing complexes and the nuclear matrix. *Proc Natl Acad Sci U S A,* 93**,** 8253-7.

MOSLEY, A. L., PATTENDEN, S. G., CAREY, M., VENKATESH, S., GILMORE, J. M., FLORENS, L., WORKMAN, J. L. & WASHBURN, M. P. 2009. Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. *Mol Cell,* 34**,** 168-78.

MOTEKI, S. & PRICE, D. 2002. Functional coupling of capping and transcription of mRNA. *Mol Cell,* 10**,** 599-609.

MYER, V. E. & YOUNG, R. A. 1998. RNA polymerase II holoenzymes and subcomplexes. *J Biol Chem,* 273**,** 27757-60.

NAKAJIMA-IIJIMA, S., HAMADA, H., REDDY, P. & KAKUNAGA, T. 1985. Molecular structure of the human cytoplasmic beta-actin gene: interspecies homology of sequences in the introns. *Proc Natl Acad Sci U S A,* 82**,** 6133-7.

NEDEA, E., NALBANT, D., XIA, D., THEOHARIS, N. T., SUTER, B., RICHARDSON, C. J., TATCHELL, K., KISLINGER, T., GREENBLATT, J. F. & NAGY, P. L. 2008. The Glc7 phosphatase subunit of the cleavage and polyadenylation factor is essential for transcription termination on snoRNA genes. *Mol Cell,* 29**,** 577-87.

NG, S. Y., GUNNING, P., EDDY, R., PONTE, P., LEAVITT, J., SHOWS, T. & KEDES, L. 1985. Evolution of the functional human beta-actin gene and its multi-pseudogene family: conservation of noncoding regions and chromosomal dispersion of pseudogenes. *Mol Cell Biol,* 5**,** 2720-32.

NI, Z., OLSEN, J. B., GUO, X., ZHONG, G., RUAN, E. D., MARCON, E., YOUNG, P., GUO, H., LI, J., MOFFAT, J., EMILI, A. & GREENBLATT, J. F. 2011. Control of the RNA polymerase II phosphorylation state in promoter regions by CTD interaction domain-containing proteins RPRD1A and RPRD1B. *Transcription,* 2**,** 237-42.

NI, Z., SAUNDERS, A., FUDA, N. J., YAO, J., SUAREZ, J. R., WEBB, W. W. & LIS, J. T. 2008. P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo. *Mol Cell Biol,* 28**,** 1161-70.

NI, Z., XU, C., GUO, X., HUNTER, G. O., KUZNETSOVA, O. V., TEMPEL, W., MARCON, E., ZHONG, G., GUO, H., KUO, W. H., LI, J., YOUNG, P., OLSEN, J. B., WAN, C., LOPPNAU, P., EL BAKKOURI, M., SENISTERRA, G. A., HE, H., HUANG, H., SIDHU, S. S., EMILI, A., MURPHY, S., MOSLEY, A. L., ARROWSMITH, C. H., MIN, J. & GREENBLATT, J. F. 2014. RPRD1A and RPRD1B are human RNA polymerase II C-terminal domain scaffolds for Ser5 dephosphorylation. *Nat Struct Mol Biol,* 21**,** 686-95.

NILSEN, T. W. & GRAVELEY, B. R. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature,* 463**,** 457-63.

NOBLE, C. G., HOLLINGWORTH, D., MARTIN, S. R., ENNIS-ADENIRAN, V., SMERDON, S. J., KELLY, G., TAYLOR, I. A. & RAMOS, A. 2005. Key features of the interaction between Pcf11 CID and RNA polymerase II CTD. *Nat Struct Mol Biol,* 12**,** 144-51.

NOORDERMEER, D., LELEU, M., SPLINTER, E., ROUGEMONT, J., DE LAAT, W. & DUBOULE, D. 2011. The dynamic architecture of Hox gene clusters. *Science,* 334**,** 222-5.

NOVOTNY, I., BLAZIKOVA, M., STANEK, D., HERMAN, P. & MALINSKY, J. 2011. In vivo kinetics of U4/U6.U5 tri-snRNP formation in Cajal bodies. *Mol Biol Cell,* 22**,** 513-23.

O'REILLY, D., DIENSTBIER, M., COWLEY, S. A., VAZQUEZ, P., DROZDZ, M., TAYLOR, S., JAMES, W. S. & MURPHY, S. 2013. Differentially expressed, variant U1 snRNAs regulate gene expression in human cells. *Genome Res,* 23**,** 281-91.

O'REILLY, D., KUZNETSOVA, O. V., LAITEM, C., ZABOROWSKA, J., DIENSTBIER, M. & MURPHY, S. 2014. Human snRNA genes use polyadenylation factors to promote efficient transcription termination. *Nucleic Acids Res,* 42**,** 264-75.

OSHIMA, S., NAKAMURA, T., NAMIKI, S., OKADA, E., TSUCHIYA, K., OKAMOTO, R., YAMAZAKI, M., YOKOTA, T., AIDA, M., YAMAGUCHI, Y., KANAI, T., HANDA, H. & WATANABE, M. 2004. Interferon regulatory factor 1 (IRF-1) and IRF-2 distinctively up-regulate gene expression and production of interleukin-7 in human intestinal epithelial cells. *Mol Cell Biol,* 24**,** 6298-310.

PERRIN, B. J. & ERVASTI, J. M. 2010. The actin gene family: function follows isoform. *Cytoskeleton (Hoboken),* 67**,** 630-4.

PETERLIN, B. M. & PRICE, D. H. 2006. Controlling the elongation phase of transcription with P-TEFb. *Mol Cell,* 23**,** 297-305.

PETRASCHECK, M., ESCHER, D., MAHMOUDI, T., VERRIJZER, C. P., SCHAFFNER, W. & BARBERIS, A. 2005. DNA looping induced by a transcriptional enhancer in vivo. *Nucleic Acids Res,* 33**,** 3743-50.

PINEAULT, N., BUSKE, C., FEURING-BUSKE, M., ABRAMOVICH, C., ROSTEN, P., HOGGE, D. E., APLAN, P. D. & HUMPHRIES, R. K. 2003. Induction of acute myeloid leukemia in mice by the human leukemia-specific fusion gene NUP98-HOXD13 in concert with Meis1. *Blood,* 101**,** 4529-38.

RAMANATHAN, Y., RAJPARA, S. M., REZA, S. M., LEES, E., SHUMAN, S., MATHEWS, M. B. & PE'ERY, T. 2001. Three RNA polymerase II carboxyl-terminal domain kinases display distinct substrate preferences. *J Biol Chem,* 276**,** 10913-20.

RANUNCOLO, S. M., GHOSH, S., HANOVER, J. A., HART, G. W. & LEWIS, B. A. 2012. Evidence of the involvement of O-GlcNAc-modified human RNA polymerase II CTD in transcription in vitro and in vivo. *J Biol Chem,* 287**,** 23549-61.

RASMUSSEN, E. B. & LIS, J. T. 1993. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci U S A,* 90**,** 7923-7.

RHEE, H. S. & PUGH, B. F. 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature,* 483**,** 295-301.

ROEDER, R. G. & RUTTER, W. J. 1969. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature,* 224**,** 234-7.

ROSONINA, E., YURKO, N., LI, W., HOQUE, M., TIAN, B. & MANLEY, J. L. 2014. Threonine-4 of the budding yeast RNAP II CTD couples transcription with Htz1-mediated chromatin remodeling. *Proc Natl Acad Sci U S A,* 111**,** 11924-31.

ROY, R., SCHAEFFER, L., HUMBERT, S., VERMEULEN, W., WEEDA, G. & EGLY, J. M. 1994. The DNA-dependent ATPase activity associated with the class II basic transcription factor BTF2/TFIIH. *J Biol Chem,* 269**,** 9826-32.

RUSSELL, J. & ZOMERDIJK, J. C. 2006. The RNA polymerase I transcription machinery. *Biochem Soc Symp***,** 203-16.

SAKURAI, H. & ISHIHAMA, A. 2002. Level of the RNA polymerase II in the fission yeast stays constant but phosphorylation of its carboxyl terminal domain varies depending on the phase and rate of cell growth. *Genes Cells,* 7**,** 273-84.

SAUNDERS, A., CORE, L. J. & LIS, J. T. 2006. Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol,* 7**,** 557-67.

SCHILLING, T. F. & KNIGHT, R. D. 2001. Origins of anteroposterior patterning and Hox gene regulation during chordate evolution. *Philos Trans R Soc Lond B Biol Sci,* 356**,** 1599-613.

SCHWARTZ, Y. B. & PIRROTTA, V. 2007. Polycomb silencing mechanisms and the management of genomic programmes. *Nat Rev Genet,* 8**,** 9-22.

SEIPELT, R. L., ZHENG, B., ASURU, A. & RYMOND, B. C. 1999. U1 snRNA is cleaved by RNase III and processed through an Sm site-dependent pathway. *Nucleic Acids Res,* 27**,** 587-95.

SHANDILYA, J. & ROBERTS, S. G. 2012. The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling. *Biochim Biophys Acta,* 1819**,** 391-400.

SHATKIN, A. J. 1976. Capping of eucaryotic mRNAs. *Cell,* 9**,** 645-53.

SHI, Y., DI GIAMMARTINO, D. C., TAYLOR, D., SARKESHIK, A., RICE, W. J., YATES, J. R., 3RD, FRANK, J. & MANLEY, J. L. 2009. Molecular architecture of the human pre-mRNA 3' processing complex. *Mol Cell,* 33**,** 365-76.

SHIM, E. Y., WALKER, A. K., SHI, Y. & BLACKWELL, T. K. 2002. CDK-9/cyclin T (P-TEFb) is required in two postinitiation pathways for transcription in the C. elegans embryo. *Genes Dev,* 16**,** 2135-46.

SHUMAN, S. 2001. Structure, mechanism, and evolution of the mRNA capping apparatus. *Prog Nucleic Acid Res Mol Biol,* 66**,** 1-40.

SIMS, R. J., 3RD, ROJAS, L. A., BECK, D., BONASIO, R., SCHULLER, R., DRURY, W. J., 3RD, EICK, D. & REINBERG, D. 2011. The C-terminal domain of RNA polymerase II is modified by site-specific methylation. *Science,* 332**,** 99-103.

SINGH, N., MA, Z., GEMMILL, T., WU, X., DEFIGLIO, H., ROSSETTINI, A., RABELER, C., BEANE, O., MORSE, R. H., PALUMBO, M. J. & HANES, S. D. 2009. The Ess1 prolyl isomerase is required for transcription termination of small noncoding RNAs via the Nrd1 pathway. *Mol Cell,* 36**,** 255-66.

SMITH-KINNAMAN, W. R., BERNA, M. J., HUNTER, G. O., TRUE, J. D., HSU, P., CABELLO, G. I., FOX, M. J., VARANI, G. & MOSLEY, A. L. 2014. The interactome of the atypical phosphatase Rtr1 in Saccharomyces cerevisiae. *Mol Biosyst,* 10**,** 1730-41.

SOSHNIKOVA, N. & DUBOULE, D. 2009. Epigenetic temporal control of mouse Hox genes in vivo. *Science,* 324**,** 1320-3.

STADELMAYER, B., MICAS, G., GAMOT, A., MARTIN, P., MALIRAT, N., KOVAL, S., RAFFEL, R., SOBHIAN, B., SEVERAC, D., RIALLE, S., PARRINELLO, H., CUVIER, O. & BENKIRANE, M. 2014. Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nat Commun,* 5**,** 5531.

STEINMETZ, E. J., CONRAD, N. K., BROW, D. A. & CORDEN, J. L. 2001. RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature,* 413**,** 327-31.

STILLER, J. W. & COOK, M. S. 2004. Functional unit of the RNA polymerase II C-terminal domain lies within heptapeptide pairs. *Eukaryot Cell,* 3**,** 735-40.

STOCK, J. K., GIADROSSI, S., CASANOVA, M., BROOKES, E., VIDAL, M., KOSEKI, H., BROCKDORFF, N., FISHER, A. G. & POMBO, A. 2007. Ring1-mediated ubiquitination of H2A restrains poised RNA polymerase II at bivalent genes in mouse ES cells. *Nat Cell Biol,* 9**,** 1428-35.

SUN, W., YU, Y., DOTTI, G., SHEN, T., TAN, X., SAVOLDO, B., PASS, A. K., CHU, M., ZHANG, D., LU, X., FU, S., LIN, X. & YANG, J. 2009. PPM1A and PPM1B act as IKKbeta phosphatases to terminate TNFalpha-induced IKKbeta-NF-kappaB activation. *Cell Signal,* 21**,** 95-102.

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., FLORENS, L., SEIDEL, C. W., LIN, C., SMITH, E. R., SHILATIFARD, A., CONAWAY, R. C. & CONAWAY, J. W. 2011. Human mediator subunit MED26 functions as a docking site for transcription elongation factors. *Cell,* 146**,** 92-104.

THOMSON, J. A., ITSKOVITZ-ELDOR, J., SHAPIRO, S. S., WAKNITZ, M. A., SWIERGIEL, J. J., MARSHALL, V. S. & JONES, J. M. 1998. Embryonic stem cell lines derived from human blastocysts. *Science,* 282**,** 1145-7.

TIETJEN, J. R., ZHANG, D. W., RODRIGUEZ-MOLINA, J. B., WHITE, B. E., AKHTAR, M. S., HEIDEMANN, M., LI, X., CHAPMAN, R. D., SHOKAT, K., KELES, S., EICK, D. & ANSARI, A. Z. 2010. Chemical-genomic dissection of the CTD code. *Nat Struct Mol Biol,* 17**,** 1154-61.

TURUNEN, J. J., NIEMELA, E. H., VERMA, B. & FRILANDER, M. J. 2013. The significant other: splicing by the minor spliceosome. *Wiley Interdiscip Rev RNA,* 4**,** 61-76.

UGUEN, P. & MURPHY, S. 2004. 3'-box-dependent processing of human pre-U1 snRNA requires a combination of RNA and protein co-factors. *Nucleic Acids Res,* 32**,** 2987-94.

UVERSKY, V. N. & DUNKER, A. K. 2010. Understanding protein non-folding. *Biochim Biophys Acta,* 1804**,** 1231-64.

VALIN, A., OUYANG, J. & GILL, G. 2013. Transcription factor Sp3 represses expression of p21CIP(1) via inhibition of productive elongation by RNA polymerase II. *Mol Cell Biol,* 33**,** 1582-93.

VALLIER, L. & PEDERSEN, R. A. 2005. Human embryonic stem cells: an in vitro model to study mechanisms controlling pluripotency in early mammalian development. *Stem Cell Rev,* 1**,** 119-30.

WADA, T., TAKAGI, T., YAMAGUCHI, Y., FERDOUS, A., IMAI, T., HIROSE, S., SUGIMOTO, S., YANO, K., HARTZOG, G. A., WINSTON, F., BURATOWSKI, S. & HANDA, H. 1998a. DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev,* 12**,** 343-56.

WADA, T., TAKAGI, T., YAMAGUCHI, Y., WATANABE, D. & HANDA, H. 1998b. Evidence that P-TEFb alleviates the negative effect of DSIF on RNA polymerase II-dependent transcription in vitro. *EMBO J,* 17**,** 7395-403.

WASHINGTON, K., AMMOSOVA, T., BEULLENS, M., JEREBTSOVA, M., KUMAR, A., BOLLEN, M. & NEKHAI, S. 2002. Protein phosphatase-1 dephosphorylates the C-terminal domain of RNA polymerase-II. *J Biol Chem,* 277**,** 40442-8.

WEN, Y. & SHATKIN, A. J. 1999. Transcription elongation factor hSPT5 stimulates mRNA capping. *Genes Dev,* 13**,** 1774-9.

WERNER-ALLEN, J. W., LEE, C. J., LIU, P., NICELY, N. I., WANG, S., GREENLEAF, A. L. & ZHOU, P. 2011. cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72. *J Biol Chem,* 286**,** 5717-26.

WEST, S., GROMAK, N. & PROUDFOOT, N. J. 2004. Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature,* 432**,** 522-5.

WHITE, R. J. 2011. Transcription by RNA polymerase III: more complex than we thought. *Nat Rev Genet,* 12**,** 459-63.

WU, Y., ZHANG, Y., ZHANG, H., YANG, X., WANG, Y., REN, F., LIU, H., ZHAI, Y., JIA, B., YU, J. & CHANG, Z. 2010. p15RS attenuates Wnt/{beta}-catenin signaling by disrupting {beta}-catenin.TCF4 Interaction. *J Biol Chem,* 285**,** 34621-31.

XIANG, K., MANLEY, J. L. & TONG, L. 2012. The yeast regulator of transcription protein Rtr1 lacks an active site and phosphatase activity. *Nat Commun,* 3**,** 946.

XIANG, K., NAGAIKE, T., XIANG, S., KILIC, T., BEH, M. M., MANLEY, J. L. & TONG, L. 2010. Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. *Nature,* 467**,** 729-33.

YAMAGUCHI, Y., TAKAGI, T., WADA, T., YANO, K., FURUYA, A., SUGIMOTO, S., HASEGAWA, J. & HANDA, H. 1999. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell,* 97**,** 41-51.

YAMAMOTO, J., HAGIWARA, Y., CHIBA, K., ISOBE, T., NARITA, T., HANDA, H. & YAMAGUCHI, Y. 2014. DSIF and NELF interact with Integrator to specify the correct post-transcriptional fate of snRNA genes. *Nat Commun,* 5**,** 4263.

YANG, C., BOLOTIN, E., JIANG, T., SLADEK, F. M. & MARTINEZ, E. 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene,* 389**,** 52-65.

YANG, Z., HE, N. & ZHOU, Q. 2008. Brd4 recruits P-TEFb to chromosomes at late mitosis to promote G1 gene expression and cell cycle progression. *Mol Cell Biol,* 28**,** 967-76.

YEO, M., LEE, S. K., LEE, B., RUIZ, E. C., PFAFF, S. L. & GILL, G. N. 2005. Small CTD phosphatases function in silencing neuronal gene expression. *Science,* 307**,** 596-600.

YEO, M., LIN, P. S., DAHMUS, M. E. & GILL, G. N. 2003. A novel RNA polymerase II C-terminal domain phosphatase that preferentially dephosphorylates serine 5. *J Biol Chem,* 278**,** 26078-85.

YUO, C. Y., ARES, M., JR. & WEINER, A. M. 1985. Sequences required for 3' end formation of human U2 small nuclear RNA. *Cell,* 42**,** 193-202.

YURYEV, A., PATTURAJAN, M., LITINGTUNG, Y., JOSHI, R. V., GENTILE, C., GEBARA, M. & CORDEN, J. L. 1996. The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins. *Proc Natl Acad Sci U S A,* 93**,** 6975-80.

ZABOROWSKA, J., TAYLOR, A. & MURPHY, S. 2012. A novel TBP-TAF complex on RNA polymerase II-transcribed snRNA genes. *Transcription,* 3**,** 92-104.

ZHANG, D. W., MOSLEY, A. L., RAMISETTY, S. R., RODRIGUEZ-MOLINA, J. B., WASHBURN, M. P. & ANSARI, A. Z. 2012a. Ssu72 phosphatase-dependent erasure of phospho-Ser7 marks on the RNA polymerase II C-terminal domain is essential for viability and transcription termination. *J Biol Chem,* 287**,** 8541-51.

ZHANG, M., WANG, X. J., CHEN, X., BOWMAN, M. E., LUO, Y., NOEL, J. P., ELLINGTON, A. D., ETZKORN, F. A. & ZHANG, Y. 2012b. Structural and kinetic analysis of prolyl-isomerization/phosphorylation cross-talk in the CTD code. *ACS Chem Biol,* 7**,** 1462-70.

ZHANG, Y., KIM, Y., GENOUD, N., GAO, J., KELLY, J. W., PFAFF, S. L., GILL, G. N., DIXON, J. E. & NOEL, J. P. 2006. Determinants for dephosphorylation of the RNA polymerase II C-terminal domain by Scp1. *Mol Cell,* 24**,** 759-70.

ZHOU, Q., LI, T. & PRICE, D. H. 2012. RNA polymerase II elongation control. *Annu Rev Biochem,* 81**,** 119-43.

ZHOU, Q. & YIK, J. H. 2006. The Yin and Yang of P-TEFb regulation: implications for human immunodeficiency virus gene expression and global control of cell growth and differentiation. *Microbiol Mol Biol Rev,* 70**,** 646-59.