Expression of human snRNA genes from beginning to end

Sylvain Egloff, Dawn O'Reilly and Shona Murphy[1]
Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, U.K.

Abstract

In addition to protein-coding genes, mammalian pol II (RNA polymerase II) transcribes independent genes for some non-coding RNAs, including the spliceosomal U1 and U2 snRNAs (small nuclear RNAs). snRNA genes differ from protein-coding genes in several key respects and some of the mechanisms involved in expression are gene-type-specific. For example, snRNA gene promoters contain an essential PSE (proximal sequence element) unique to these genes, the RNA-encoding regions contain no introns, elongation of transcription is P-TEFb (positive transcription elongation factor b)-independent and RNA 3'-end formation is directed by a 3'-box rather than a cleavage and polyadenylation signal. However, the CTD (C-terminal domain) of pol II closely couples transcription with RNA 5' and 3' processing in expression of both gene types. Recently, it was shown that snRNA promoter-specific recognition of the 3'-box RNA processing signal requires a novel phosphorylation mark on the pol II CTD. This new mark plays a critical role in the recruitment of the snRNA gene-specific RNA-processing complex, Integrator. These new findings provide the first example of a phosphorylation mark on the CTD heptapeptide that can be read in a gene-type-specific manner, reinforcing the notion of a CTD code. Here, we review the control of expression of snRNA genes from initiation to termination of transcription.

Introduction

The vertebrate snRNA (small nuclear RNA) genes transcribed by pol II (RNA polymerase II) are ubiquitously and generally highly expressed, providing ideally a model system to elucidate both fundamental and gene-type-specific mechanisms underlying transcription and co-transcriptional RNA processing.

The structure of human snRNA genes transcribed by pol II

The human snRNA genes transcribed by pol II are exemplified by the best-characterized genes encoding the spliceosomal U1 and U2 snRNAs. These genes are repeated many times in the human genome and, in the case of the U2 snRNA genes, are located within 6.1 kb tandem repeats on chromosome 17 [1]. The U1 genes are more loosely clustered repeat units of up to 45 kb on chromosome 1 [2]. The genes are simple and compact with short TATA-less promoters minimally containing an enhancer-like DSE (distal sequence element), and an essential snRNA gene-specific PSE (proximal sequence element) (see [3]). snRNA

gene transcripts are not spliced and the 3'-ends are not polyadenylated, presumably to discourage association with the translation machinery. The snRNA gene-specific 3'-box, located 9–19 bp downstream of the RNA-encoding region, is required for correct 3'-end formation of snRNAs [4]. A diagram of the major *cis*-acting elements is shown in Figure 1. The primary transcript of the U1 genes extends to just beyond the 3'-box, while the primary transcript of the U2 genes extends for up to 1 kb [5,6] (Figure 1), most likely to be due to differences in transcription termination (see below).

snRNA gene promoters recruit gene-type-specific pre-initiation complexes (Figure 2)

Several factors that activate transcription of both mRNA and snRNA genes (e.g. Oct-1 and Staf) bind to sites in the DSE [3]. Oct-1 potentiates interaction of the multisubunit snRNA gene-specific transcription factor, PTF (PSE-binding transcription factor)/SNAPc (snRNA activator protein complex)/PBP (PSE-binding protein), with the PSE [3,7–9]. In common with mRNA genes, the pol II-dependent snRNA genes require the general factors TBP (TATA-box-binding protein), TFIIB (transcription factor IIB), TFIIA, TFIIE and TFIIF for transcription *in vitro* [3,10]. A more complete characterization of the transcription machinery has been hampered by the relatively low level of *in vitro* transcription from templates with a U1/U2 promoter. There is currently no direct evidence that TFIIH is required for transcription. However, phosphorylation of Ser[5] of

Figure 1 | **The structure of human snRNA genes transcribed by pol II**

The diagram shows the DSE and PSE *cis*-acting promoter elements and the 3' box *cis*-acting RNA-processing element of pol II-transcribed snRNA genes boxed, with their position relative to the transcription start site noted below. The start site of transcription is marked with an arrow above the line in this and subsequent Figures. The extent of the transcription unit of the U1 and U2 genes is indicated below the line by an arrow.
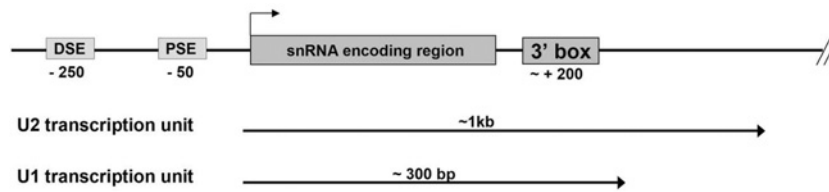

Figure 2 | **snRNA gene promoters recruit gene-type-specific pre-initiation complexes**

(**A**) The diagram shows the *cis*-acting promoter elements and *trans*-acting factors involved in formation of a pre-initiation complex on a typical protein-coding gene. These include a TATA box, at −25 relative to the transcription start site, USEs (upstream sequence elements), generally located within approx. 200 bp of the start site and an Enhancer, which can be far upstream or downstream of the start site. USEs and Enhancers bind a range of sequence-specific DNA-binding factors and the TATA box is recognized by the general factor TFIID. The pre-initiation complex minimally comprises the general transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (see [3]). Initiation may also require interaction of the pol II CTD with the Mediator complex [35]. (**B**) The diagram shows the *cis*-acting promoter elements and *trans*-acting factors involved in formation of a pre-initiation complex on a typical snRNA gene. These include a DSE that is recognized by factors such as Oct-1 and Staf1 that are shared with protein-coding genes and pol III-transcribed snRNA genes [3,7] and a PSE that is recognized by PTF, an snRNA gene-specific transcription factor. The pre-initiation complex minimally comprises TBP, TFIIA, TFIIB, TFIIE, TFIIF and possibly TFIIH [3,10]. TAF100 is found associated with these genes [13]. The involvement of other TAFs and TFIIH is yet to be demonstrated. Efficient initiation may also require interaction of the pol II CTD with the snRNA gene-specific Integrator complex [20].
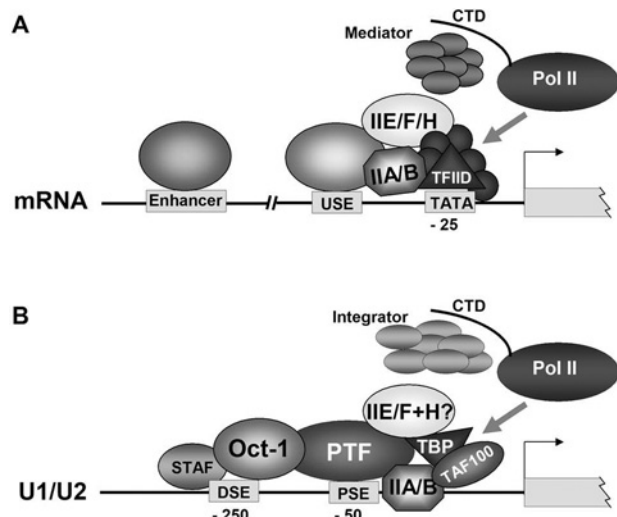
the $Y^1SPTSPS^7$ heptad repeat of the CTD (C-terminal domain) of pol II by the CDK7 (cyclin-dependent kinase 7) subunit of TFIIH probably activates co-transcriptional 5′ capping of all pol II transcripts (see below), suggesting that the kinase at least is associated with snRNA genes. In support of this, pol II transcribing snRNA genes is phosphorylated on Ser^5 of the CTD repeat [11]. TFIID, the TAFII (TBP/TBP-associated factor II) complex, which is involved in transcription of mRNA-coding genes, cannot substitute for TBP in transcription of the human U1 gene *in vitro*, and the form of TBP used is unclear [12]. ChIP (chromatin immunoprecipitation) analysis indicates that TAF-5 (TAFII100) is associated with snRNA genes [13] but it is not clear which, if any other TAFs are involved. Thus PTF/SNAPc/PBP is, so far, the only component of the pre-initiation complex that is specific to pol II-transcribed snRNA genes.

Interestingly, some snRNA genes, including those for U6 spliceosomal snRNA and 7SK RNA, are transcribed by pol III despite promoter structures very similar to those of the U1 and U2 genes [3]. Pol III-dependent snRNA genes are characterized by the presence of an additional TATA box at −25 relative to the site of transcription initiation, which favours recruitment of pol III-specific transcription factors [3].

The function of the 3′-box

The 3′-end formation of human snRNAs transcribed by pol II occurs in several steps initiated by recognition of the *cis*-acting 3′-box [4]. This 13–16-nt-long element directs the production of a 3′ extended pre-snRNA, which is a specific substrate for subsequent processing that generates the mature 3′-end after transport to the cytoplasm [14–16].

The demonstration that transcription of the human U2 gene continues for approx. 800 nt beyond the 3′-box *in vivo* [5] strongly suggested that this element directs RNA processing rather than transcription termination. In addition, the 3′-box can direct accurate 3′-end formation of RNA substrates *in vitro* [17], providing further evidence that it is a *bone fide* RNA-processing element. Both protein and RNA components are required for *in vitro* processing and at

Figure 3 | Integrating transcription of snRNA genes with RNA 3′-end formation

The model summarizes how transcription may be linked to RNA processing. The snRNA gene-specific pre-initiation complex containing PTF and/or other yet to be identified snRNA gene-specific transcription factors recruits pol II, P-TEFb and a Ser[7] kinase. Phosphorylation of the pol II on Ser[2] by P-Tefb, Ser[5] by the CDK7 subunit of TFIIH and Ser[7] by a yet to be identified kinase recruits capping enzymes and the Integrator complex. After transcription of the 3′ box, Integrator associates with the RNA and, together with additional factors including an RNA component, processes the RNA to produce pre-snRNA, which is later trimmed. Processing leads to termination of transcription, which may occur by mechanisms similar to those causing termination of transcription of protein-coding genes [33]. For example, changes in interactions between the pol II CTD and processing factors and/or dephosphorylation of the CTD and/or attack of the now uncapped 5′ end of the downstream transcript by exonucleases such as Xrn2 may prompt termination of transcription. In the case of the U1 snRNA genes, proteins binding just downstream of the 3′ box may cause transcription to terminate more abruptly than in the U2 snRNA genes.
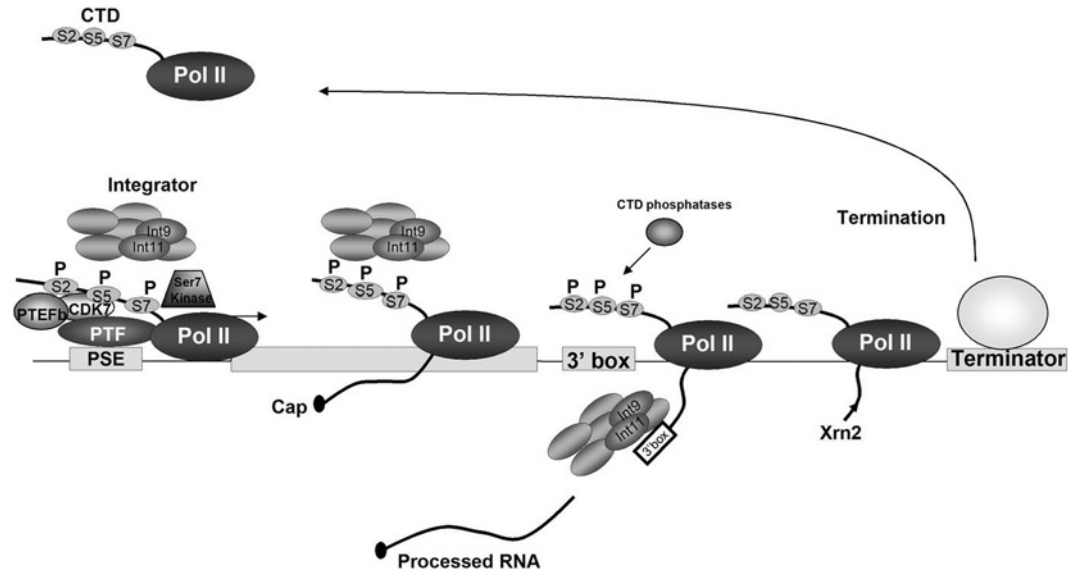
least one heat-labile component is involved [18]. In common with *in vitro* polyadenylation, processing is activated by a 7-methyl-G cap on substrate RNA [18]. There is no evidence for the involvement of mRNA-processing factors in 3′-box recognition *in vitro* [17,18]. More recently, a large complex termed Integrator has been shown to play a role in pre-snRNA 3′-end formation *in vivo* [19,20]. This complex contains homologues of the CPSF-100 (cleavage and polyadenylation specificity factor-100) and CPSF-73 subunits of the CPSF, termed respectively either RC74 and RC68 [21] or Int9 and Int11 [19]. ChIP analysis indicates that Integrator subunits are associated with U1 and U2 genes *in vivo* and RNAi (RNA interference)-mediated knockdown of Int9 or Int11 causes a defect in pre-snRNA production [19]. There is now good evidence that CPSF-73 is the cleavage activity for 3′-end formation of polyadenylated RNAs [22]. RC68/Int9 is therefore a good candidate for the endonuclease implicated in 3′-box-directed processing *in vitro* [17]. In line with the findings from analysis of *in vitro* 3′-box-directed processing, no factors involved in mRNA 3′-end formation have been found associated with the Integrator complex [19]. Taken together, these findings indicate that the 3′-box is an RNA-processing element analogous to the polyadenylation signal commonly found in protein-coding genes, recognized by a distinct set of processing factors. Interestingly, two 3′-box-dependent processing products are produced *in vitro*: one corresponding to the *in vivo* pre-snRNA and one slightly longer at the 3′-end [17]. Production of pre-snRNA may therefore involve endonucleolytic cleavage by RC68/Int9 followed by limited exonucleolytic trimming.

Integrating the transcription of snRNA genes with RNA 3′-end formation (Figure 3)

In vivo, the 3′ box is only recognized in transcripts from an snRNA promoter [23,24], emphasizing the close link between transcription and RNA processing. Recent work has demonstrated that the CTD of the large subunit of pol II and its phosphorylation are required for 3′-box-dependent RNA 3′-end formation *in vivo* [6,25], indicating that processing occurs co-transcriptionally. The CTD plays a major role in co-ordinating transcription with co-transcriptional RNA processing in expression of protein-coding genes (recently reviewed in [26,27]. In mammalian pol II, this structure comprises 52 repeats with the consensus Y[1]SPTSPS[7] and its deletion can cause a major defect in capping, splicing and polyadenylation of pre-mRNAs. The CTD undergoes reversible phosphorylation *in vivo* during transcription. Phosphorylation of Ser[5] at initiation by the

CDK7 subunit of TFIIH allows the CTD to interact with and activate factors that cap the 5′-end of the nascent RNA. Subsequent phosphorylation of Ser[2] by P-TEFb (positive transcription elongation factor b), comprising CDK9 and a cyclin T subunit, increases the ability of the CTD to activate splicing and cleavage/polyadenylation. In yeast, Ser[2] phosphorylation potentiates interaction with the polyadenylation factor Pcf11. Activation of polyadenylation in mammalian *in vitro* systems by phospho-CTD and inhibition of polyadenylation in *Drosophila* by CDK9 inhibitors suggest that a similar mechanism operates in higher eukaryotes. Ser[2] phosphorylation is also required to overcome an early block to transcription caused by the N-TEFs (negative elongation factors), DSIF [DRB (5,6-dichloro-1β-D-ribofuranosylbenzimidazole) sensitivity-inducing factor] and NELF (negative elongation factor), and the conversion of pol II into a processively elongating form during transcription of protein-coding genes [28].

The CTD of pol II is also phosphorylated during transcription of snRNA genes [11], although the transcription cycle differs from that of protein-coding genes in some respects [29]. ChIP studies indicate that P-TEFb is associated with U2 snRNA genes and that the CTD of pol II transcribing U2 genes is phosphorylated on Ser[2] in addition to Ser[5] [11]. Inhibitors of P-TEFb abolish recognition of the 3′-box *in vivo* [6,11], and mutation of Ser[2] to alanine residue, which cannot be phosphorylated, reduces 3′-box-dependent processing [11,20]. Taken together, these results suggest that phosphorylation of Ser[2] by CDK9 plays a key role in 3′ box-dependent processing *in vivo*. However, Ser[2] phosphorylation by P-TEFb does not appear to be necessary for elongation of transcription of snRNA genes [11]. Interestingly, repeats 1–25 of the wild-type CTD, which are largely consensus, activate recognition of the 3′ box more effectively than the repeats between 27 and 52, which generally lack a serine residue in position 7 [6]. In contrast, both halves of the CTD are equally effective in activating polyadenylation [30]. In accordance with these findings, selective mutation of Ser[7] to alanine affects recognition of the 3′ box but does not affect polyadenylation of transcripts from several protein-coding genes [20,31]. Interestingly, mutation of Ser[7] to alanine residue also abrogates recruitment of Integrator to the U1 and U2 snRNA genes. It has now been demonstrated that Ser[7] is phosphorylated during transcription of snRNA genes and that Integrator binds specifically to the phosphorylated CTD *in vitro* only if there is a serine residue in position 7 [20], Thus, as shown for activation of capping and polyadenylation, CTD phosphorylation potentiates interactions with a 3′ box-specific processing factor, resulting in increased recruitment and possibly also allosteric activation. These studies on snRNA 3′-end formation have therefore highlighted new aspects of the role of the pol II CTD in RNA processing. Since Ser[2] phosphorylation is also implicated in recognition of the 3′ box, it is possible that phosphorylation of both Ser[2] and Ser[7] is required for efficient recruitment of the Integrator complex.

It was previously shown that the protein-coding gene-specific transcription factor TFIID mediates interaction between the polyadenylation factor CPSF and the CTD [32]. Likewise, the snRNA promoter-specific factor PTF may act in concert with phosphorylation of Ser[7] of the CTD to specifically recruit Integrator and other factors involved in 3′ box-dependent processing and/or recruit the Ser[7] kinase to snRNA genes (see Figure 3). This would provide a simple explanation for the link between a PSE-containing promoter and recognition of the 3′ box *in vivo*.

Interestingly, in addition to affecting 3′ processing, mutation of Ser[7] to alanine residue causes a defect in transcription without affecting pol II recruitment [20]. Integrator recruitment may therefore also be required for efficient initiation to occur.

The role of the 3′ box in transcription termination

Mutation of the 3′ box affects termination of transcription of snRNA genes [5]. Since polyadenylation signals are also required for transcription termination (recently reviewed in [33]), the molecular mechanisms operating in snRNA and protein-coding genes may be analogous even though a different complement of processing factors is involved. The Torpedo Model of termination suggests that degradation of the nascent transcript from the unprotected 5′-end produced by cleavage at the polyadenylation site destabilizes the polymerase. In support of this model, the 5′–3′ exonuclease Xrn2 is required for efficient termination of transcription of protein-coding genes in human cells [34]. Recognition of the polyadenylation signal can also result in changes in the interaction of factors with the pol II CTD and/or CTD phosphorylation, all of which may contribute to termination (Allosteric Model, see [33]). Likewise, the 3′ box may promote termination as a result of RNA cleavage and/or changes in the association of factors with the CTD or changes in CTD phosphorylation.

Transcription of the U2 snRNA gene extends for approx. 800 bp after the 3′ box, while transcription of the U1 gene terminates abruptly shortly after the 3′ box [5]. *In vivo* footprinting suggests that a factor or complex binds to the DNA just downstream of the U1 3′ box [5]. The footprinted region effectively terminates transcription when it is placed downstream of the U2 3′ box [5], implicating factors binding to this region in the termination process. There are no sequence similarities to the 'U1 terminator' 800 bp downstream of the U2 3′ box, suggesting that different strategies are used to terminate transcription of these two highly related snRNA genes.

There has been much progress towards understanding the control of expression of the pol II-transcribed mammalian snRNA genes in recent years. However, there are still many outstanding questions. A more complete understanding of the mechanisms involved in transcription and co-transcriptional RNA processing awaits the identification of

all the factors involved in pre-initiation complex assembly, poll II recruitment and 3′ box-directed processing.

References

1 Lindgren, V., Ares, M., Weiner, A.M. and Franke, U. (1985) Human genes for U2 small nuclear RNA map to a major adenovirus 12 modification site on chromosome 17. Nature **314**, 115–116

2 Lindgren, V., Bernstein, L.B., Weiner, A.M. and Franke, U. (1985) 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**, 2172–2180

3 Hernandez, N. (2001) Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. J. Biol. Chem. **276**, 26733–26736

4 Hernandez, N. (1992) Transcription of vertebrate snRNA genes and related genes. in Transcriptional Regulation, Vol. 1 (McKnight, S.L. and Yamamoto, K.R., eds.), pp. 281–313, Cold Spring Harbor Laboratory Press, Cold Spring Harbor

5 Cuello, P., Boyd, D.C., Dye, M.J., Proudfoot, N.J. and Murphy, S. (1999) Transcription of the human U2 snRNA genes continues beyond the 3′ box *in vivo*. EMBO J. **18**, 2867–2877

6 Medlin, J.E., Uguen, P., Taylor, A., Bentley, D.L. and 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**, 925–934

7 Murphy, S., (1997) Differential *in vivo* activation of the class II and class III snRNA genes by the POU-specific domain of Oct-1. Nucleic Acids Res. **25**, 2068–2076

8 Murphy, S., Yoon, J.B., Gerster, T. and Roeder, R.G. (1992) Oct-1 and Oct-2 potentiate functional interactions of a transcription factor with the proximal sequence element of small nuclear RNA genes. Mol. Cell. Biol. **12**, 3247–3261

9 Yoon, J.B., Murphy, S., Bai, L., Wang, Z. and Roeder, R.G. (1995) Proximal sequence element-binding transcription factor (PTF) is a multisubunit complex required for transcription of both RNA polymerase II- and RNA polymerase III-dependent small nuclear RNA genes. Mol. Cell. Biol. **15**, 2019–2027

10 Kuhlman, T.C., Cho, H., Reinberg, D. and 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**, 2130–2141

11 Medlin, J., Scurry, A., Taylor, A., Zhang, F., Peterlin, B.M. and Murphy, S. (2005) P-TEFb is not an essential elongation factor for the intronless human U2 snRNA and histone H2b genes. EMBO J. **24**, 4154–4165

12 Bernues, J., Simmen, K.A., Lewis, J.D., Gunderson, S.I., Polycarpou-Schwarz, M., Moncollin, V., Egly, J.M. and Mattaj, I.W. (1993) Common and unique transcription factor requirements of human U1 and U6 snRNA genes. EMBO J. **12**, 3573–3585

13 Christova, R. and Oelgeschlager, T. (2002) Association of human TFIID–promoter complexes with silenced mitotic chromatin *in vivo*. Nat. Cell Biol. **4**, 79–82

14 Kiss, T. (2004) Biogenesis of small nuclear RNPs. J. Cell Sci. **117**, 5949–5951

15 Mougin, A., Torterotot, F., Branlant, C., Jacobson, M.R., Huang, Q. and Pederson, T. (2002) A 3′-terminal minihelix in the precursor of human spliceosomal U2 small nuclear RNA. J. Biol. Chem. **277**, 23137–23142

16 Huang, Q., Jacobson, M.R. and Pederson, T. (1997) 3′ processing of human pre-U2 small nuclear RNA: a base-pairing interaction between the 3′ extension of the precursor and an internal region. Mol. Cell. Biol. **17**, 7178–7185

17 Uguen, P. and Murphy, S. (2003) The 3′ ends of human pre-snRNAs are produced by RNA polymerase II CTD-dependent RNA processing. EMBO J. **22**, 4544–4554

18 Uguen, P. and 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**, 2987–2994

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

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

21 Dominski, Z., Yang, X.C., Purdy, M., Wagner, E.J. and 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**, 1489–1500

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

23 de Vegvar, H.E., Lund, E. and Dahlberg, J.E. (1986) 3′ End formation of U1 snRNA precursors is coupled to transcription from snRNA promoters. Cell **47**, 259–266

24 Hernandez, N. and Weiner, A.M. (1986) Formation of the 3′ end of U1 snRNA requires compatible snRNA promoter elements. Cell **47**, 249–258

25 Jacobs, E.Y., Ogiwara, I. and 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**, 846–855

26 Phatnani, H.P. and Greenleaf, A.L. (2006) Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev. **20**, 2922–2936

27 Egloff, S. and Murphy, S. (2008) Cracking the RNA polymerase II CTD code. Trends Genet. **24**, 280–288

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

29 Egloff, S. and Murphy, S. (2008) Role of the C-terminal domain of RNA polymerase II in expression of small nuclear RNA genes. Biochem. Soc. Trans. **36**, 537–539

30 Fong, N., Bird, G., Vigneron, M. and Bentley, D.L. (2003) A 10 residue motif at the C-terminus of the RNA pol II CTD is required for transcription, splicing and 3′ end processing. EMBO J. **22**, 4274–4282

31 Chapman, R.D., Heidemann, M., Albert, T.K., Mailhammer, R., Flatley, A., Meisterernst, M., Kremmer, E. and Eick, D. (2007) Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. Science **318**, 1780–1782

32 Dantonel, J.C., Murthy, K.G., Manley, J.L. and Tora, L. (1997) Transcription factor TFIID recruits factor CPSF for formation of 3′ end of mRNA. Nature **389**, 399–402

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

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

35 Malik, S. and Roeder, R.G. (2005) Dynamic regulation of pol II transcription by the mammalian Mediator complex. Trends Biochem. Sci. **30**, 256–263