mRNA 3' END FORMATION AND RNA POLYMERASE II TERMINATION IN *CAENORHABDITIS ELEGANS* OPERONS

by

ALFONSO GARRIDO-LECCA

B.A., Texas A&M University, 2001

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Molecular, Cellular and Developmental Biology

2012

This thesis entitled:
mRNA 3' end Formation and RNA Polymerase II Termination in *C. Elegans* Operons
written by Alfonso Garrido-Lecca
has been approved for the
Department of Molecular, Cellular and Developmental Biology

Joaquin Espinosa, Chair

Thomas Blumenthal, Advisor

David Bentley

Min Han

Ravinder Singh

Date:

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Garrido-Lecca, Alfonso (Ph.D., Molecular, Cellular, and Developmental Biology)

mRNA 3' end Formation and RNA Polymerase II Termination in *C. Elegans* Operons

Thesis directed by Professor Thomas Blumenthal

In most organisms, 3’ end formation of the pre-mRNA and transcription termination are tightly coupled, making it impossible to study these two processes independently from each other. *C. elegans*, however, contains polycistronic transcripts (operons) that naturally separate 3’ end processing from transcription termination, allowing me to ask questions that cannot be answered in other systems. I have used ChIP experiments in operons to study 3’ end formation and transcription termination in a unique context. I found that within operons Ser-5 and Ser-2 phosphorylation of RNAPII CTD colocalized with the expected sites of pre-mRNA processing; Ser-5p was associated with sites of co-transcriptional capping, while Ser-2p was associated with 3’ end formation sites. Moreover, I globally mapped the CstF-64 localization of all genes in the worm genome. I found that CstF-64 binds all 3’ ends of genes, even those in which termination does not occur. Interestingly, CstF-64 colocalized with Ser-2p at 3’ ends of genes, indicating that in *C. elegans* the CstF trimeric complex might be recruited by Ser-2 phosphorylation. I also present evidence that RNAPII at 3’ ends of internal genes in operons is paused, colocalizing with Ser-2p and CstF-64, similar to the pattern seen for terminal 3’ ends. These results indicate that 3’ ends marked by Ser-2p, bound by CstF-64 and containing paused RNAPII are not sufficient to cause transcription termination. Finally, I investigated the 3’ end formation mechanism at the 3’ end of internal genes in SL1-type operons. I found no evidence supporting a cleavage event involving trans-splicing, since SL1-type operon 3’ ends are marked by Ser-2p and CstF-64, similar to the patterns seen in SL2-type operons. Moreover, SL1-type operons required CstF-50 for processing their 3’ ends and recruiting CstF-64, similar to SL2-type operon 3’ ends. These results

are consistent with the experimental results presented by Williams et al. (1999), suggesting that 3’ end formation at SL1-type operons can occur through the canonical 3’ end formation machinery.

Para mi abuelo Guillermo, quien me introdujo la pasión por la ciencia a muy temprana edad, y que con su experiencia me indico el camino que debía tomar. Tenias razón Papapa!

ACKNOWLEDGEMENTS

I will like to thank everyone who has helped and supported me during the seven years in graduate school. Without their consistent advice and support I would have never made it through the most challenging and most exciting period of my life. I am also eternally thankful to my family who taught me that with dedication and effort anything in life is possible.

I will like to thank my advisor and mentor, Tom Blumenthal, for teaching me how to think like a scientist. He is without a doubt the best scientific mentor there is in the field. I admired his constant ability to think creatively about models explaining results and his love for science. I also would like to thank all the past and present members of the lab who were always willing to help and discuss science and beyond. I feel proud and honored to have been trained by such an exceptional scientist and being part of such a wonderful lab. Also I will like to thank all the members of my committee who were helpful in providing advice throughout my project.

CONTENTS

CHAPTERS

FIGURES

CHAPTER I

INTRODUCTION

Eukaryotic gene expression is a complex process that starts in the nucleus with transcription of the DNA into mRNA and ends in the cytoplasm with translation of these mRNAs into proteins. RNA polymerase II (RNAPII) is responsible for transcribing all protein coding genes into pre-mRNAs, which are processed into mature mRNAs while RNAPII is still engaged in transcription (co-transcriptionally). Therefore, transcription is tightly coupled to pre-mRNA processing in order to efficiently process the mRNA before it exits the nucleus for translation. However, this coupling makes it difficult to study pre-mRNA processing and transcription events independent from each other.

In this thesis I used the unique gene organization of the *C. elegans* genome to study pre-mRNA processing at sites where processing has been naturally separated from transcription events. I show how operons can be used to differentiate co-transcriptional capping sites at 5' ends of outrons from SL2 trans-splice sites at internal 5' ends. Furthermore, I use operons to differentiate 3' end formation sites from transcription termination.

Pre-mRNA processing

Pre-mRNA processing is composed of three distinct steps needed from yeast to humans for efficient gene expression. The first pre-mRNA processing event to occur is capping, then splicing and finally 3' end formation. *In vitro* these processing events can occur independent of transcription, but in the nucleus they are interconnected with each

other and with transcription allowing increased processing efficiency and accuracy (Beyer and Osheim 1988; Bauren et al., 1998; Rasmussen and Lis 1993).

Addition of a 7-methyl guanosine cap occurs early in the transcription cycle when the nascent RNA is only 20-25 nt long (Shuman 2001). In the nucleus the cap functions in protecting mRNAs from exonucleases, and promotes polyadenylation, splicing, and nuclear export (Lewis and Izaurralde 1997; Gu and Lima 2005). The cap is recognized by the nuclear cap-binding complex (CBC) that functions in mediating these nuclear effects (Visa et al., 1996; Topisirovic et al., 2011). Once the mRNA is exported into the cytoplasm the nuclear CBC is replaced by eIF4E and associated proteins, which function in recruiting the mRNA to the small ribosomal subunit for initiating translation (reviewed in Topisirovic et al., 2011).

Splicing consists of the removal of non-coding regions called introns from the pre-mRNA. EM visualization of pre-mRNA splicing using drosophila embryos showed that most introns, but not all, are removed co-transcriptionally (Beyer and Osheim 1988). *In vitro* splicing experiments on pre-made transcripts show how the spliceosome assembles on the RNA in a stepwise manner. U1 small nuclear ribonuclearprotein (snRNP) first binds the 5’ splice site (consensus GURAGU) followed by the binding of U2 and U2AF to the branch point and 3’ splice site (consensus UUUUYAG), respectively. Then the tri-snRNP U4-U6/U5 binds and U1 and U4 are discarded, resulting in the formation of the catalytically active spliceosome. This type of splicing will be referred to throughout the thesis as cis-splicing.

Another important pre-mRNA processing event is the formation of 3’ ends by cleavage and polyadenylation. This type of processing and its relationship to

transcription termination are extensively studied in this thesis. Therefore, it is the major focus of this introduction.

3’ end formation

With the exception of some histone genes, all protein-coding genes have a poly-A tail that is formed by cleavage of the pre-mRNA at the 3’ end followed by subsequent polyadenylation of the free 3’ hydroxyl group. The poly-A tail’s primary function is to protect 3’ ends of mRNAs from exonucleases present in the nucleus and cytoplasm. Besides its role in mRNA stability, 3’ end formation is also required for other RNA processing events such as terminal intron removal, mRNA export, transcription termination (see below), and translation initiation (Proudfoot et al., 2002; 2004). Defects in 3’ end formation are associated with a number of human diseases (Danckwardt et al., 2008) highlighting the importance of proper processing of mRNA 3’ ends in eukaryotic gene expression.

Assembly of the 3’ end formation complex

Surprisingly, 3’ end formation requires a large multisubunit complex for performing an apparently simple cleavage and polyadenylation reaction (Figure I-1). The need for such a large number of proteins is indicative of the high level of regulation required for accurately processing 3’ ends. The “core” of the 3’ end formation complex, determined by proteomic analysis (Shi et al., 2009), contains two highly conserved multi-protein subunits, the cleavage and polyadenylation specificity factor (CPSF) and the cleavage stimulatory factor (CstF).

Yang and Doublié 2011

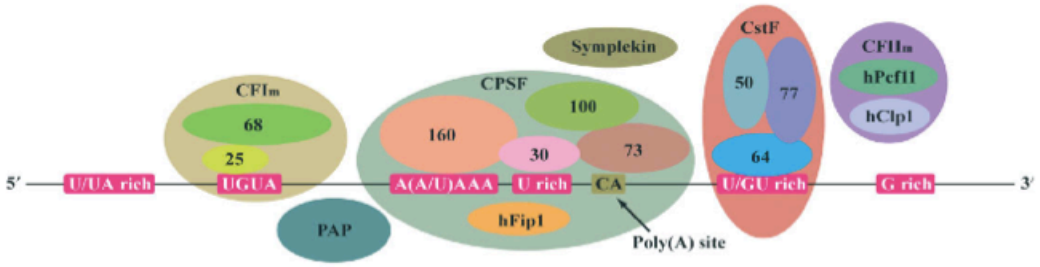


Figure I-1. Schematic representation of the mammalian 3' end formation complex with their corresponding mRNA binding sites.

Assembly of a stable complex at the 3' end of the pre-mRNA is initiated by the cooperative binding of CPSF and CstF to highly conserved cis-regulatory elements located on the RNA (Figure I-1). CPSF binds the poly A signal motif (AAUAAA) located 10-30 nt upstream of the cleavage site. In *C. elegans* this motif is present in the vast majority of 3' ends (Mangone et al., 2010). The CstF trimeric complex binds a more variable and less defined U or U/G rich region located less than 30 nt downstream of the poly A site (Figure I-1), which is conserved in *C. elegans* (Graber et al., 2007).

The CPSF complex bound at the poly-A signal bridges across the putative cleavage site contacting the CstF complex bound at the U-rich region, positioning the site of polyadenylation after a CA dinucleotide (Figure I-1). How the other 3' end formation factors are recruited to the RNA is not known, but *in vitro* experiments suggests the initial binding of CPSF-CstF complex is sufficient for the binding of additional key factors including CFI, CFII, and PAP (Venkataraman et al., 2005). Therefore, an extensive network of RNA-protein and protein-protein interactions is

needed for the proper assembly of the 3' end formation machinery at 3' ends of pre-mRNAs.

The core 3' end formation complex: CPSF and CstF

CPSF is composed of 5 subunits (30 KDa, 73 KDa, 100 KDa, 160 KDa, and Fip1) needed for both steps in the 3' end formation reaction. CPSF-160 is responsible for the binding to the poly-A signal consensus sequence (AAUAAA) with remarkable specificity, as any single mutation in the hexamer motif strongly prevents processing (Sheets et al., 1990; Keller et al., 1991). *In vivo* and *in vitro* studies have shown that CPSF-73 is the endonuclease responsible for the cleavage step in the 3' end formation reaction (Ryan et al., 2004; Mandel et al., 2006). The 100 KDa subunit is structurally similar to the 73 KDa subunit but lacks the zinc-binding motif needed for endonuclease activity (Mandel et al., 2006). The function of CPSF-100 remains unclear.

At least two other CPSF subunits are involved in RNA binding, CPSF-30 and Fip1 (Barabino et al., 1997; Kaufmann et al., 2004). The C-terminal arginine-rich domain of human Fip-1 shows preference for U-rich sequences, which are needed for stimulating poly-A polymerase (PAP) activity (Kaufmann et al., 2004). CPSF-30 also shows specificity for poly-U rich sequences through its zinc finger domain (Barabino et al., 1997). Since regions surrounding the cleavage site of pre-mRNAs tend be rich in poly-U sequences (Figure I-1), Fip1 and CPSF-30 may provide additional RNA-protein contacts stabilizing the interaction of the core 3' end formation complex with the RNA.

Across the poly-A site, the CstF trimeric complex is composed of three subunits (50 KDa, 64 KDa, and 77 KDa), which are required only for the cleavage step (Colgan

and Manley 1997; Mandel et al., 2008). Intriguingly, each subunit of the complex is represented as a dimer based on Drosophila genetics (Simonelig et al., 1996), *in vivo* interaction experiments (Takagaki and Manley 2000), and crystal structure studies (Moreno-Morcillo et al., 2011). These results provide strong evidence that the CstF complex functions in 3' end formation as a heterohexamer, but the biological function for this dimerization is unknown.

The 64 KDa subunit of the CstF complex is responsible for binding the RNA downstream of the cleavage site. This binding is mediated through the highly conserved N-terminal RNA-recognition domain (RRM). *In vitro* binding studies have shown that the RRM motif is sufficient for binding the RNA at both U-rich and U/G-rich regions, but only when these RNA motif are situated less than 30 bp downstream of the cleavage site (Takagaki and Manley 1997). CstF-64 plays a key role in the choice of cleavage site as shown in studies done during B-cell differentiation, in which increased levels of CstF-64 during activation of B-cells was sufficient to switch the IgM heavy chain mRNA expression from membrane-bound form to secreted form (Takagaki and Manley 1998).

In addition to the RRM domain, CstF-64 contains two highly conserved regions called the hinge and C-terminal domain. The hinge domain is a 100 amino acid region needed for its interaction with CstF-77 and symplekin, as well as for its nuclear localization (Qu et al., 2007). Also involved in protein binding, the C-terminal domain of CstF-64 is needed for its interaction with Pcf-11 and the transcription co-activator/repressor PC4 (Calvo and Manley 2001, Hockert et al., 2010). The RRM, hinge, and C-terminal domain of CstF-64 are essential for polyadenylation *in vivo*, highlighting the importance of protein-protein interactions in 3' end formation (Hockert et

al., 2010). Interestingly, the putative *C. elegans* ortholog of the CstF-64 protein (CPF-2) is smaller compared to other organisms (Figure I-2). This is due to the absence of a portion of the C-terminal region known to contain a central proline/glycine rich domain with unknown function (Hatton et al., 2010). In humans, embedded within this proline/glycine rich domain is 12 repeats of the pentapeptide sequence consensus, MEARA/G, which is absent in worms and yeast (Hatton et al., 2000).

CstF-77 is crucial for the assembly of the CstF complex, as it bridges the interaction of both CstF-64 and CstF-50 via its proline-rich domain, since CstF-64 and CstF-50 do not directly bind each other (Takagaki and Manley 1994; 2000). The crystal structure of CstF-77 revealed that this protein is built around 11 conserved Half a TPR (HAT) domains used for homodimerization (Legrand et al., 2007). This observation is consistent with earlier biochemical studies showing that CstF-77 and its yeast homolog RNA-14 are present as a dimer (Takagaki and Manley 2000; Noble et al., 2004).

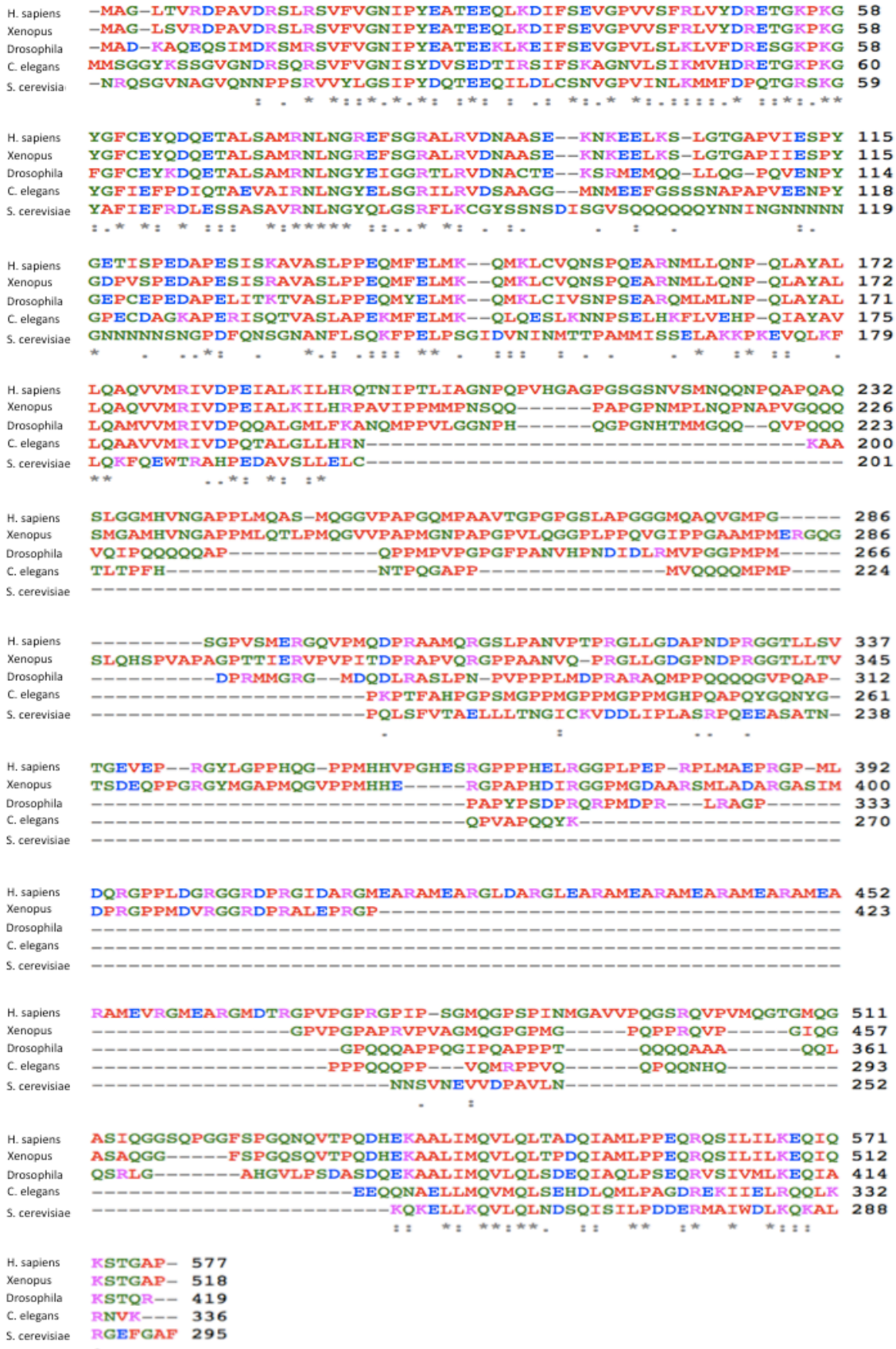

H. sapiens
Xenopus
Drosophila
C. elegans
S. cerevisiae

Figure I-2. CstF-64 protein alignment comparing humans, Xenopus, flies, *C. elegans* and yeast. Each amino acid has been colored coded according to its chain chemical properties. Red indicate small hydrophobic, blue acidic, magenta basic, green hydroxyl + sulfhydryl + amine, and gray indicates unusual amino acid. An * (asterisk) indicates positions which have a single, fully conserved residue. A : (colon) indicates conservation between groups of strongly similar properties. A . (period) indicates conservation between groups of weakly similar properties. Alignments were generated using the free-online ClustalW2 algorithm.

The 50 KDa subunit of the CstF complex dimerizes with itself via its N-terminal region. Indeed, when the N-terminal region of CstF-50 was used for crystal formation, a dimer structure was obtained suggesting that the N-terminal region is sufficient for self-association (Moreno-Morcillo et al., 2011). Moreover, *in vitro* interaction experiments showed that CstF-50 interacts with CstF-77 through its WD repeat-domains located at the C-terminal region of the protein (Takagaki and Manley 2000). Furthermore, *in vitro* binding experiments showed that the N-terminal region of CstF-50 binds equally well to a peptide composed of either phosphorylated or unphosphorylated full-length CTD of RNAPII (Fong and Bentley 2001). Importantly, this interaction functions in 3' end formation *in vivo,* since overexpression of the CstF-50 N-terminal region has a dominant negative effect on cleavage (Fong and Bentley 2001). Therefore, CstF-50's interaction with the CTD of RNAPII may aid in the recruitment of the CstF complex to the elongating RNAPII (Fong and Bentley 2001).

Biochemical purification methods and proteomics approaches have identified most, if not all, the protein components of the 3' processing multisubunit complex (Mandel et al., 2008; Shi et al., 2009). Interestingly, the stoichiometries of the different multi-protein subunits within the core 3' end formation complex are not represented in equal ratios. The CPSF-CstF multisubunit complexes are represented in a 1:2 ratio,

while the other 3’ end processing factors dynamically associate with the core complex (Shi et al., 2009).

The *C. elegans* genome encodes clear homologs of all the subunits of the 3’ end formation complexes. The cis-regulatory sequences these factors recognize are also highly conserved, indicative of stringent conservation of the mechanism of 3’ end formation. Indeed, in a suppressor screen to identify proteins that play roles at processes at the 3’ end of genes, several CPSF and CstF factors were identified (Cui et al., 2008).

RNA Polymerase II (RNAPII)

In all organisms DNA-dependent RNA polymerases have the critical role of transcribing the DNA into RNA needed for proper cellular development and homeostasis. Eukaryotes have evolved three types of nuclear RNA polymerases (I, II and III), each responsible for the transcription of different classes of RNAs. RNAPII is responsible for transcribing all protein coding genes and most non-coding RNAs, including small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), microRNAs (miRNAs), cryptic unstable transcripts (CUTs) and stable unannotated transcripts (SUTs) (Kuehner et al., 2011). In contrast to the other DNA-dependent RNA polymerases found in the nucleus, RNAPII contains a unique C-terminal domain that allows the coupling of transcription with RNA processing.

The carboxy-terminal domain (CTD) of RNAPII

RNAPII is composed of 12 subunits. The largest subunit, RPB1, contain a unique carboxy-terminal domain (CTD) that is composed of numerous heptad repeats with the consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$. The heptads repeat sequence is highly conserved among all eukaryotes (Stiller and Hall 2002). However, the number of these repeats varies among species correlating with genomic complexity. For example, the budding yeast *S. cerevisiae* contains 26 repeats, *C. elegans* contains 42 repeats and mammals have 52 repeats (Allison et al., 1988; Rosonina and Blencowe 2004). The function of the CTD has been proposed to mediate the coupling of transcription with pre-mRNA processing, since deletion of the CTD impairs capping, splicing and 3' end formation (McCracken et al., 1997a and b). Therefore, the CTD is thought to act as a "landing pad" for the recruitment of different RNA processing factors to the transcription site, thus allowing pre-mRNA processing to occur co-transcriptionally (Zorio and Bentley 2004).

CTD phosphorylation

During the transcription cycle (initiation, elongation and termination) the CTD of RNAPII is dynamically and reversibly phosphorylated, giving the CTD specificity for recruitment of different factors. *In vivo* the CTD heptad repeat is phosphorylated at five of the seven residues, but only Ser-2, Ser-5 and recently Ser-7 have been extensively studied to provide insights into their function in transcription (Bartkowiak and Greenleaf 2011). The extent and transcriptional function of the other CTD phosphorylation residues (Tyr-1 and

Thr-4) as well as other post-translational modifications of the CTD (i.e. glycosylation) remains unknown.

In a simplified model for a phospho-CTD cycle, RNAPII is recruited to promoters in an unphosphorylated state forming a pre-initiation complex (PIC). Indeed, phosphorylation of the CTD prior to RNAPII promoter binding has an inhibitory effect on transcription (Hengartner et al., 1998). Following PIC formation at the promoter, the CTD is phosphorylated at Ser-5 by the CTD kinase subunit of the TFIIH complex, CDK7/cyclinH in metazoans. As RNAPII clears the promoter Ser-5p levels start to decrease while Ser-2p levels start to increase, yielding a CTD that contains a mix of doubly phosphorylated repeats in the center of the gene (Buratowski 2009). Ser-2 is phosphorylated by the positive transcription elongation complex b (pTEFb), coinciding with RNAPII entry into productive elongation. As RNAPII approaches the 3 end of the gene, Ser-5 specific phosphatases further decrease Ser-5p, leaving the CTD phosphorylated at Ser-2 to terminate transcription (Bartkowiak et al., 2011).

In vivo and *in vitro* evidence suggests that these phosphorylations of the CTD heptad repeats facilitate the recruitment of processing factors to the RNAPII ternary complex (Perales and Bentley 2009; Phatnani and Greenleaf 2006). It has been proposed that Ser-5p is high at 5' ends of genes in order to recruit capping enzymes needed to cap the nascent RNA co-transcriptionally (McCracken et al., 1997a; Cho et al., 1997; Komarnitsky et al., 2000; Schroeder et al., 2000; Shuman 2001; Glover-Cutter et al., 2008). Inactivation of the yeast Ser-5p kinase (Kin-28) resulted in a significant decrease of capping enzymes binding to the 5' end of genes, consistent with Ser-5p being required for the recruitment of capping enzymes (Schroeder et al., 2000). At 3'

ends of genes it is believed that Ser-2p is high in order to recruit 3' end formation/termination factors needed for co-transcriptional cleavage and polyadenylation of the pre-mRNA (Licatalosi et al., 2002; Ahn et al., 2004; Glover-Cutter et al., 2008). Consistent with Ser-2p being needed for the recruitment of 3' end processing factors is the finding that inactivation of the yeast Ser-2p kinase (Ctk-1) lead to a significant decrease of 3' end processing factors but had no effect on elongation factors such as components of the TFIIS, Paf, TREX and FACT complexes (Ahn et al., 2004). Therefore, as RNAPII traverses a gene modulation of CTD phosphorylation may help in coordinating the assembly of pre-mRNA processing factors on the CTD. Co-transcriptional recruitment of processing factors allows the nascent pre-mRNA to be efficiently processed as it emerges from the RNA exit channel.

Transcription termination

Transcription termination and 3' end processing

The last step in the transcription cycle is transcription termination, where the transcribing RNAPII ceases RNA synthesis and releases both the nascent mRNA and the DNA template. Transcription termination is crucial for maintaining gene expression in the cell by preventing RNAPII molecules from colliding with each other, and for recycling RNAPII for subsequent use.

The eukaryotic cell has evolved mechanisms to ensure that RNAPII stops transcription only when the 3' end of the gene has been transcribed by coupling 3' end formation (cleavage and polyadenylation) with termination. Indeed, the same cis-regulatory signals present at 3' ends of genes required for cleavage and

polyadenylation are also the signals required for termination (Whitelaw and Proudfoot 1986; Logan et al., 1987; Connelly and Manley 1988; Zaret and Sherman 1982). In addition, several yeast factors required for cleavage and polyadenylation such as Rna-14, Rna-15, Pcf-11, and Yhh1 are also required for termination (Birse et al., 1998). Therefore, at the 3' end of genes the 3' end formation processing reaction communicates with RNAPII to terminate transcription.

In *C. elegans*, RNAPII transcription proceeds up to 1 kb beyond the poly-A site (Haenni et al., 2009), and more than 1.5 Kb in human genes (Iwamoto et al., 1986; Dye and Proudfoot 1999; Glover-Cutter et al., 2008). This is in contrast to *S. cerevisiae* in which RNAPII terminates in close proximity to the poly-A site (Zaret and Sherman 1982; Birse et al., 1998; Kim et al., 2004; Birse et al., 1997). However, the molecular mechanism that directs RNAPII termination at the end of genes is poorly understood. Two non-mutually exclusive models have been proposed to explain the mechanism connecting 3' end formation and transcription termination: the torpedo model and the allosteric model.

Termination models

The torpedo model proposes that RNAPII can transcribe through the termination signals of a gene, but upon cleavage of the nascent pre-mRNA at the poly-A site, a 5' to 3' exonuclease starts degrading the downstream RNA, which is still attached to the elongating RNAPII (Connelly and Manley 1988). Upon reaching the transcribing RNAPII, the exonuclease somehow destabilizes the transcription elongation complex (TEC) and causes RNAPII to terminate. Evidence supporting the torpedo model comes

from nuclear run-on (NRO) experiments performed in yeast using temperature-sensitive cleavage and polyadenylation mutants. In these experiments, only mutations affecting the factors involved in the cleavage reaction, and not factors involved in polyadenylation, were required for termination (Birse et al., 1997). In addition, inactivation of the yeast and mammalian nuclear 5' to 3' exonuclease (Rat-1 in yeast and Xrn-2 in mammals) caused stabilization of RNA downstream from the cleavage site, indicative of termination defects (Kim et al., 2004; West et al., 2004).

On the other hand, the allosteric model proposes that as RNAPII transcribes the poly-A signal, it undergoes a conformation change in the TEC destabilizing RNAPII association with the template DNA. This destabilization results in termination. Destabilization could be due to the binding of negative elongation factors or by the dissociation of an anti-terminator factor (Logan et al., 1987). Evidence for this model comes from the finding that Pcf-11, a subunit of the cleavage and polyadenylation complex in yeast (CF1A), was able to dismantle *in vitro* assembled TEC (Zhang et al., 2005). This result suggested that Pcf-11 functions in communicating a termination signal to the active site of RNAPII by binding both the RNA and CTD.

A vast amount of data supports the torpedo model of termination at 3' end of protein-coding genes (Kim et al., 2004; West et al., 2004; Ujvari et al., 2002; Luo et al., 2006). However, most likely both methods are used. Indeed, a combined torpedo/allosteric model has been proposed in which the degradation of the RNA downstream of the cleavage site by an exonuclease is not sufficient to cause RNAPII release. Interestingly, instead the exonuclease enhances the recruitment of PCF-11 and

RNA-15 (yeast CstF-64 ortholog) to 3’ ends in order to cause RNAPII to terminate (Luo et al., 2006; Dengl and Cramer 2009).

RNAPII pause sites and transcription termination

An additional class of termination element can act to enhance termination besides the cis-regulatory 3’ end formation signals found at 3’ ends of genes. RNAPII pausing enhances termination by slowing down the elongating RNAPII in order to give enough time for the exonuclease to reach the TEC making it terminate (Plant et al., 2005; Gromak et al., 2006). Interestingly, molecular dissection of pause site-dependent transcription termination of the mammalian RNAPII revealed that this event occurs at G-rich sequences located downstream of the poly-A site, leading to the formation of RNA:DNA hybrids (R-loops) behind the TEC (Skourti-Stathaki et al., 2011). Resolution of these hybrids by a specialized helicase (Senataxin) allows access to the 5’-3’ exonuclease following cleavage at the 3’ end and consequently causing RNAPII termination (Skourti-Stathaki et al., 2011). Therefore, the exonuclease is in kinetic competition with the TEC, so when RNAPII pauses at 3’ end of genes, the equilibrium is tilted towards RNA degradation resulting in termination.

Operons

Operons represent a type of gene organization in which a group of genes are under the control of a single promoter located at the 5’ end of the cluster. Operons are present in bacteria, as well as several eukaryotes such as nematodes, flatworms, primitive chordates, protists, and hydra (Blumenthal 2002). In *C. elegans*, around 15% of protein

coding genes are organized into operons that are up to eight genes long (Blumenthal 2012).

Transcription of these clusters produces polycistronic pre-mRNAs that are further processed into individual mature cistrons by cleavage and polyadenylation at the 3' end of the upstream gene and SL2 trans-splicing at the 5' end of the downstream gene. Most likely, polycistronic resolution into individual cistrons happens co-transcriptionally, because it is usually impossible to detect polycistronic transcripts from total RNA preparations.

The Lin-15 operon contains two genes that encode proteins involved in vulva development. The n765 allele contains a mutation within the first gene in this operon that introduces another gene 3' end causing a multivulva phenotype. Reduced function of both proteins is needed for the development of the multivulva phenotype. Cui et al. (2008) used this operon mutation as a suppressor screen to identify proteins involved in transcription termination. As expected several CPSF and CstF subunits were identified, since these protein complexes are known to be required for 3' end formation and transcription termination in other organisms (Whitelaw and Proudfoot 1986; Logan et al., 1987; Connelly and Manley 1988; Birse et al., 1998). Interestingly, factors that previously were not known to be involved in termination were also identified, three containing a CTD-interacting domain (CID) and one SR protein (SRp20). Therefore, operons provide a unique opportunity that could be used to uncover unknown protein factors involved in termination.

SL trans-splicing

Trans-splicing is a reaction similar to cis-splicing and is catalyzed by many of the same U snRNPs (U2, U4, U5, and U6). In addition to using the same snRNPs, trans-splicing uses the same RNA sequences to indicate splice sites. In order for trans-splicing to occur it depends on the presence of a 5' splice site sequence and branch point on one RNA molecule, and the polypyrimidine tract and the 3' splice site on the other RNA molecule (Blumenthal 2005). The consensus sequence of these elements varies among organisms (Schwartz et al., 2008), but within the same organism trans-splice sites and cis-splice sites tend to be the same.

About 70% of *C. elegans* mRNAs are processed by trans-splicing (Lasda and Blumenthal 2011; Allen et al., 2011), which adds a 22 nt sequence, the spliced leader (SL), to the 5' end of genes (Figure I-3). The SL is donated by a 100 nt RNA that exists as snRNP and is consumed in the trans-splicing reaction. The majority of trans-splicing utilizes a spliced leader known as SL1, which replaces the outron, an intron-like sequence located at the very 5' end of mRNA. This SL1 trans-splicing has been observed in genes organized in operons and in non-operon genes. In operon genes, it is used to trans-splice the first gene in the operon and not in downstream genes in operons except for "hybrid" operons (see below).

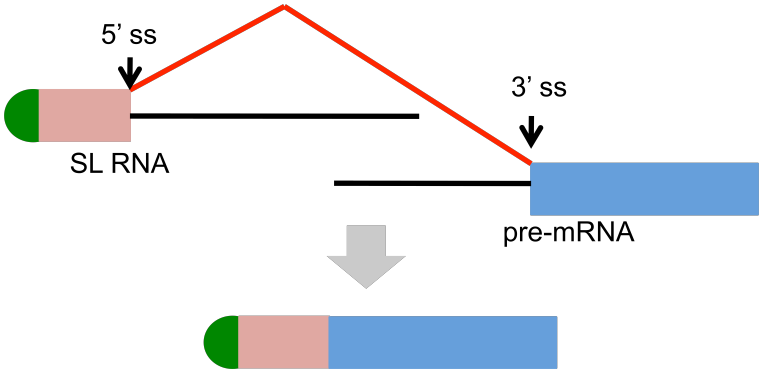


Figure I-3. Trans-splicing joins the specialized SL exon from the SL RNA to the 5’ end of some pre-mRNAs. Boxes represent exons; black lines represent introns and outrons. Green semicircle represents the cap on the SL exon; 5′ and 3′ splice sites are marked; and red lines connecting exons indicate splicing. Figure adapted from Lasda and Blumenthal 2011.

In contrast, the SL2 RNA is used exclusively to trans-splice downstream genes in polycistronic pre-mRNAs, although downstream mRNAs are sometimes trans-spliced to both SL1 and SL2. Indeed, increased use of SL1 trans-splicing to downstream operon genes has been shown to be due to the presence of an extra promoter in the intercistronic region (ICR), creating what has been termed a "hybrid" operon (Allen et al., 2011; Blumenthal 2012). However, most operons contain short ICRs (median 129 nt) and downstream genes are exclusively trans-spliced to SL2 (Allen et al., 2011).

In addition to SL2-type operons, there is a rare type of operon in which downstream genes are trans-spliced to SL1 instead of SL2. These operons are called SL1-type (Williams et al., 1999). Only 23 examples of SL1-type operons are present in the genome (Blumenthal 2012). These operons are characterized by SL1 trans-splicing to downstream genes and by containing no ICR between genes. Based on mutational analysis of a single transgenic SL1-type operon it was proposed that 3’ end formation and trans-splicing were in competition in this type of operon, so that only the upstream or downstream gene can be expressed, but not both (Williams et al., 1999). It is currently unknown how SL1-type operon 3’ ends are processed and if they are treated differently than SL2-type 3’ ends.

Polycistronic pre-mRNA processing

The exact mechanism by which polycistronic pre-mRNAs are processed is not clear, but a model has been proposed based on accumulation of a processing intermediate from a single synthetic operon (Liu et al., 2003). This intermediate, called the Ur-RNA, accumulated only when trans-splicing was inhibited, suggesting that following cleavage at the 3' end of the upstream gene, a 5' to 3' exonuclease degrades the precursor RNA stopping when it reaches a factor bound at the Ur-element (Liu et al., 2003). Consistent with the torpedo model for termination, the Ur-element functions as a "road blockage" for the passage of the exonuclease, consequently preventing it from reaching the elongating RNAPII.

The Ur-element is a U-rich sequence located in the ICR of operons approximately 50 nt upstream of the trans-splice site (Figure I-4), which has been shown to be necessary and sufficient for SL2 trans-splicing of the downstream gene (Huang et al., 2001). It is currently unclear what proteins or RNAs bind the Ur-element that will allow SL2 specific trans-splicing. Interestingly, based on mutational analysis and bioinformatics the Ur-element was further defined as a stem-loop followed by a UAYYUU sequence motif, which was predicted to hybridize with the 5' splice site on the SL2 RNA (Lasda et al., 2011). By a mechanism that is not fully understood, the SL2 could be recruited to the ICR of polycistronic pre-mRNAs by the Ur element, which also provides the SL2 specificity for trans-splicing. In theory the bound SL2 could act as a blockage for preventing the exonuclease from reaching the TEC and causing it to terminate.

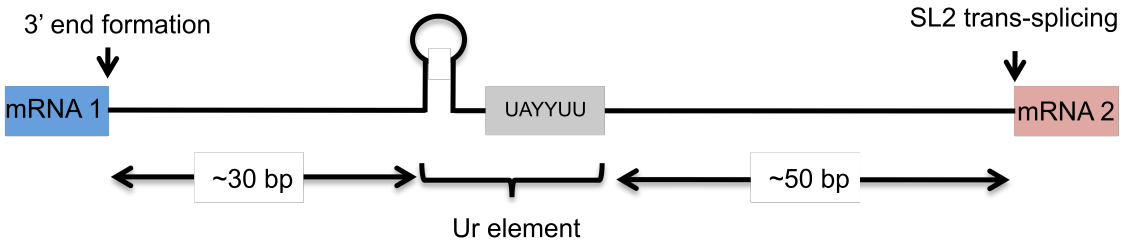


Figure I-4. The Ur element is located approximately 50 bp upstream of the SL2 trans-splice site. mRNA 1 (blue) represents the transcript from the upstream gene in the operon and mRNA 2(orange) represents the transcript from the downstream gene in the operon. The black line represents the ICR between genes in operons. Figure adapted from Lasda and Blumenthal 2011.

C. elegans operons provide an excellent model for studying termination in a unique context, since 3' end formation must be uncoupled from termination in operon genes. The poly-A signal at the ends of upstream genes in operons must be prevented from signaling transcription termination in order for proper transcription of the entire polycistronic pre-mRNA. All cleavage and polyadenylation factors as well the RNA sequences they recognize have been conserved in *C. elegans*. However, how transcription termination is prevented at upstream genes in operons is unclear. I believe that by understanding how RNAPII is prevented from terminating at internal genes in operons, I might gain insights into how RNAPII terminates at 3' ends of genes in general.

In this thesis

In this work I take advantage of *C. elegans* unique gene organization to provide insights into pre-mRNA processing and transcriptional events, which are naturally separated in operons. Moreover, I investigate *in vivo* the mechanisms for 3' end formation at 3' ends of a rare type of operon.

In Chapter III, I use ChIP experiments to test if Ser-5 and Ser-2 phosphorylation of the RNAPII CTD marks pre-mRNA processing sites or transcriptional events. I show that Ser-5p is associated with promoter locations, but not with locations specifying 5' ends of mRNAs distant from promoters. I also show that Ser-2p is associated with all 3' ends, even those at large distances from transcription termination sites.

In Chapter IV, I use ChIP-seq experiments to find if CstF-64 functions in 3' end formation or transcription termination. I show that CstF-64 is associated with all 3' ends, even those in which transcription termination does not occur following 3' end formation. Then I demonstrate that CstF-64 colocalizes with Ser-2p and paused RNAPII at each 3' end in operons.

In Chapter V, I use ChIP experiments and RT-PCR to test if cleavage at SL1-type operon 3' ends occurs by trans-splicing. I show that this type of 3' end is marked by Ser-2p and bound by CstF-64, similar to SL2-type 3' ends. Also, I demonstrate that CstF-50 is needed for processing SL1-type operon 3' ends and for CstF-64 recruitment to this type of 3' ends.

CHAPTER II

MATERIALS AND METHODS

Worm culture and growing conditions

Mixed stage and synchronized Bristol (N2) worms were maintained and grown as described by Sulston and Brenner (1974). Worms for ChIP experiments were synchronized and grown in liquid culture until the young adult stage with few eggs. For other assays, worms were grown in NGM plates. The worms were then washed three times in water and the bacteria were cleared by sucrose flotation.

Formaldehyde in vivo cross-linking

Worms were frozen in liquid nitrogen and grounded to a powder with a mortar and pestle. The resulting worm powder was transferred to cross-linking buffer (1 mM PMSF, 1 mM EDTA, 1 mM EGTA, 1% formaldehyde, PBS) for 10 min at room temperature. The reaction was quenched for 5 min at room temperature by addition of glycine to a final concentration of 125 mM, and the mixture was sedimented at 4000 g. Pellets were washed three times with FA buffer + 0.1% SDS (50 mM HEPES pH 7.5, 1 mM EDTA, 1% Triton, 0.1% deoxycholic acid, 150 mM NaCl, 0.1% SDS) containing one protease inhibitor cocktail tablet (Roche). Each wash was sedimented at 4000 g. Then pellets were divided into 500 ul aliquots and stored at -80 ºC.

Chromatin Immunoprecipitation (ChIP)

Each aliquot was resuspended in 1.5 ml of FA buffer + 0.3% SDS containing protease (Roche complete cocktail tablets 11697498001) and phosphatase inhibitors (GBiosciences 786-450). The samples were sonicated with a Virsonic digital 600 sonicator using a microtip (20 pulses of 11 s each at 30% amplitude with bursts of 0.9 s on and 0.5 s off), to generate DNA fragments of approximately 500 bp, as determined experimentally. Samples were sedimented at 13000 g for 15 min at 4 ºC, and the supernatant transferred to a new tube, which was then diluted to 4.5 ml with FA buffer containing protein and phosphatase inhibitors. The extract was divided into four 1 ml samples.

Immunoprecipitation and elution were performed according to Lee et al. (2006) with some modifications: For each 1 ml extract, 100 ul of protein A Dynabeads (Invitrogen), conjugated with antibody, were added. After a 4 ºC overnight incubation, the beads were washed five times with RIPA buffer (50 mM HEPES pH 7.5, 1 mM EDTA, 1% NP-40, 0.7% deoxycholic acid, 0.5 M LiCl) and once with TE + 50 mM NaCl. The beads containing the protein-DNA complex were transfer using cold TE to a clean eppendorf tube before elution. After removal of any TE buffer, 210 ul of elution buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS) was added and incubated at 65 ºC for 30 min, with intense mixing briefly every 5 min. The elution was separated from the beads by applying the magnet. 10 ul of proteinase K (20 mg/ml) was added to the eluted fraction and incubated at 55 ºC for 2-3 hrs. Alternatively, in recent versions of the protocol the RNA was degraded from the IP DNA to improve the efficiency of the qPCR reaction. In these case 30ug of affinity purified RNAse A (Ambion) was added to each

tube and incubated at 37 ºC for ~2hrs. Then 10 ul of proteinase K (20 mg/ml) was added to each eluted and incubates at 55 ºC for 2hrs. The tubes in either version were transferred to 65 ºC overnight to reverse the crosslinks.

DNA isolation and purification

For ChIP-qPCR the bound DNA was purified using a Qiagen PCR purification kit eluted twice with 50 ul of water. For ChIP-on-chip the DNA was purified using the Affymetrix cDNA cleanup kit eluted twice with 20ul of elution buffer. For ChIP-seq the DNA was purified using a Qiagen column and eluted twice with 30 ul of water.

Real-time PCR.

A fraction of the DNA was used as a template in real-time PCR reactions. The primers used were designed with Roche LightCycler Probe design software 2.0 and their sequences are shown in Supplementary Table 1. PCR products were between 75 and 150 bp. PCR reactions were performed in a total volume of 10 ul containing 1X SYBR Green Mix (Applied Biosynthesis), 1/200 fraction of the ChIP-enriched DNA, and 100 nM primers in a 384-well plate using an Eppendorf epMotion 5070 robot. Standard curves were generated using sonicated genomic DNA samples run concurrently with ChIP samples. All standards and samples were run in triplicate. Plates were read in an Applied Biosynthesis 7900HT Real-time PCR machine (Absolute Quantification Method). Enrichment of ChIP DNA was calculated by using the standard curve method and the numbers were corrected by removing outlier points from samples that had a

greater than 17% coefficient of variation. Input DNA values were used to normalize results following subtraction of the control without antibody.

Affymetrix tiling arrays

I used the GeneChip *C. elegans* Tiling 1.0R array from Affymetrix. Each array is composed of 3.2 million probes spanning the whole non-repetitive worm genome. The probes are tiled at an average resolution of 25 base pairs, measured from the center of the adjacent probe. The preparation of the IP DNA for hybridization was done according to the manufacturer's protocol. Briefly, PCR was performed on the purified IP DNA for dUTP incorporation; cycle number is dependent on the quality of the antibody used and maintained of the enrichment post-amplification was determined experimentally by qPCR. Next, the amplified IP DNA was enzymatically fragmented (uracil DNA glycosylase and APE-1) into smaller pieces (<100bp) and labeled with biotin at their 3' ends with terminal deoxynucleotidyl transferase. Finally, the fragmented and labeled DNA was hybridized to the array.

Library preparation

Preparation of the DNA library was performed according to the manufacturer's protocol (Illumina). Approximately 10ng of enriched IP DNA was used for the library preparation. The overhangs resulting from the ChIP experiment were converted into blunt ends using the end repair mix. Single "A" nucleotides were added to 3' ends of the blunt fragments to prevent them from self-ligating. Different sequence adapters were ligated to the 5' and 3' ends, follow by gel extraction of ligation products between 250-300 bp. PCR of

20 cycles was used to enrich DNA fragments containing the adapters, allowing the DNA to bind the flow cell. The library was first validated by fluorescence Qubit (Invitrogen) and Agilent Technologies 2100 Bioanalyzer, and then sequenced using the Illumina Hiseq 2000 platform, multiplexing two samples in a single lane.

Galaxy: reads manipulations

Galaxy is an open, web-based platform for accessible, reproducible, and transparent computational biomedical research. I used Galaxy for processed and mapped all sequencing reads. A total of 171,614,331 raw reads were obtained from a single lane, N2 anti-CstF-64 having 86,067,449 reads and CstF-50 RNAi anti-CstF-64 having 85,546,882 reads. Bases from reads were trimmed from either end that had a quality score less than 31 (scale 0-40), then the entire read length was filtered by a quality score no less than 20 and allowed 0 bases outside this range. Around 60% and 59% of the N2 and RNAi reads, respectively, met the above conditions. Adapters and barcodes were cleaved from the ends of each read, and mapped to the *C. elegans* genome (ce10) using Bowtie (default settings). In the N2 sample 98% of the reads mapped to the genome, while in the RNAi'ed sample 95% of the reads mapped the genome.

Peak calling algorithm

The model-based analysis of ChIP-seq (MACS 1.0.1) algorithm was used from the galaxy web browser to find statistically significant peaks. Bam files containing aligned reads were used as input files for MACS, which outputs a wiggle file at 1 bp resolution (used for python script) and a Bed file. Based on the parameters used on galaxy, MACS slides 500 bp windows across the genome to find regions containing reads 2-fold enriched relative to a background level. For each candidate peak a background level is calculated by counting the number of reads in a 5 kb, or 10 kb window centered from the peak location. Using a *p-value* cutoff of 1e-5, the candidate peaks are selected and reported as fold enrichment over the local background level.

Python script

Will Kruesi from the B. Meyer lab at the University of California at Berkley wrote the script on python called “average profiles”. It is used to create metagenes around a feature (poly A site), or across a gene by scaling them to the same size. The wiggle file created from MACS was used as input files. The script calculates average fold enrichment by adding the level of CstF-64 at each base pair and divides it by the number of genes. If a gene 3’ end is not present in the input wiggle file then nothing will be added to the total, but it will still be counted towards the average.

CstF-50 knocked-down

Strain containing a balanced deletion allele for CstF-50 (tm 4163) was obtained from the Caenorhabditis Genetic Center (CGC). I used RNAi by feeding to significantly knocked-down CstF-50 levels in this balance deletion strain. Bacteria expressing dsRNA corresponding to part of the CstF-50 locus were used to feed synchronized starved L1s either on plates or liquid culture. For the RNAi done in plates, the bacteria was induced by adding log-phase grown bacteria to NGM plates containing ampicillin and IPTG and left at room temperature for at least 2 hrs before adding the starved L1s. For the RNAi done in liquid, the bacteria was previously induced before being fed to the starved L1s. Briefly, a single colony of the CstF-50 RNAi bacteria was used to inoculate a 3 ml overnight LB broth culture. Then, 100 ul of this overnight was used to inoculate 1 L of LB broth and grown at 37 ºC until log-phase (~ 3 hrs). Next, IPTG was added to a final concentration of 1 mM and induced at 37 ºC overnight. Finally, the bacteria were sedimented at 4000 g and fed to the starved L1s. In both plates and liquid the worms were allowed to grown on the RNAi bacteria for 3-4 days at 20 ºC.

Antibodies. α-Ser-5p and α-Ser-2p antibodies were from Bethyl Laboratories (A300-655A and A300-654A, respectively). In some experiments, where indicated, peptides antibodies against Ser-5p and Ser-2p were used, which were a gift from David Bentley. The "total RNAPII" antibody is either a rabbit polyclonal antibody raised against the recombinant mouse CTD (52 repeats) protein, which was also a gift from David Bentley, or the 8wg16 antibody that is commercially available from Millipore (05-952). α-CstF-64 antibody is a rabbit polyclonal antibody raised against the recombinant full-length *C.*

elegans protein. Rabbit #1 is used on all ChIP-qPCR and array experiments, and rabbit #2 used on ChIP analyzed by next generation sequencing. α-H3K9ac is a commercially available rabbit polyclonal antibody from Abcam (4441).

CHAPTER III

RNA POLYMERASE II C-TERMINAL DOMAIN PHOSPHORYLATION PATTERNS IN *CAENORHABDITIS ELEGANS* OPERONS, POLYCISTRONIC GENE CLUSTERS WITH ONLY ONE PROMOTER BUT MULTIPLE 3' ENDS

Introduction

Pre-mRNAs of protein coding genes must be processed into mature mRNAs for translation. This transcription is carried out by RNA Polymerase II (RNAPII) in association with a wide range of nuclear proteins that serve at different stages in the transcription cycle. Shortly after the nascent RNA emerges from RNAPII, its 5' end is co-transcriptionaly capped (Coppola et al., 1983; Rasmussen and Lis 1993; Moteki and Price 2002; Chiu et al., 2002). At the other end of the gene, the pre-mRNA is co-transcriptionaly cleaved by the 3' end formation machinery composed of the multisubunit proteins, CPSF and CstF, as well as several additional proteins. However, transcription does not terminate until the polymerase has continued synthesizing RNA for an additional kilobase or more (Iwamoto et al., 1986; Dye and Proudfoot 1999; Glover-Cutter et al., 2008; Haenni et al., 2008). The 3' end formation machinery, and perhaps pre-mRNA cleavage itself, plays a key role in the termination event. One popular idea is that cleavage exposes a free 5' phosphate end on the downstream RNA, thereby allowing access to the 5' to 3' exonuclease, XRN2 (Dye and Proudfoot 2001; Kim et al., 2004; Teixeira et al., 2004; West et al., 2004).

The CTD of RNAPII is a unique and flexible tail-like domain needed to accommodate the large number of proteins required for these and other co-

transcriptional events. The CTD is composed of numerous heptad repeats with the consensus sequence, $Y_1S_2P_3T_4S_5P_6S_7$, a sequence conserved among all eukaryotes (Stiller and Hall 2002). Deletion of the CTD in mammalian cells inhibits co-transcriptional capping, splicing, 3' end cleavage and polyadenylation, suggesting that the CTD functions in coupling transcription with pre-mRNA processing (Allison et al., 1988; Rosonina and Blencowe 2004; McCracken et al., 1997a and b).

The CTD is dynamically and reversibly modified during transcription (Komarnitsky et al., 2000; Perales and Bentley 2009), predominantly by phosphorylations at heptad repeat positions serine 5 (Ser-5p) and serine 2 (Ser-2p) (Dahmus 1996; Zhang and Corden 1991). *In vivo* and *in vitro* evidence suggests that these phosphorylations of the RNAPII CTD heptad repeats facilitate recruitment of processing factors to the transcription complex. Ser-5p, phosphorylated primarily by the cyclin dependent kinase 7 (cdk7) is required for binding capping enzymes to RNAPII at the 5’ ends of genes (Schroeder et al., 2000; Cho et al., 1997; McCracken et al., 1997a; Ho et al., 1998; Komarnitsky et al., 2000). On the other hand, phosphorylation of Ser-2 by positive transcription elongation factor b (pTEFb) is required for binding 3’ end formation/termination factors to RNAPII at the 3’ end of genes (Komarnitsky et al., 2000; E.J. Cho et al., 2001). In addition, ChIP experiments using antibodies specific for these phospho-epitopes in mammals and yeast have shown that Ser-5p is present at higher levels at the 5’ ends of genes while Ser-2p levels peak closer to the 3’ ends (Gomes et al., 2006; Kim et al., 2009; Komarnitsky et al., 2000; Rosonina et al., 2006). These observations have led to the proposal that these phosphorylated serine residues guide

the co-transcriptional processing of the pre-mRNA at different stages of the transcription cycle (Perales and Bentley 2009; Phatnani and Greenleaf 2006).

In the nematode *C. elegans*, many genes are organized into operons (Blumenthal 2005). These operons can contain from 2 to 8 genes, and each gene's mRNA is independently cleaved and polyadenylated at the 3' end of the upstream gene and trans-spliced by SL2 at the 5’ end of the downstream gene. Transcription termination has been shown to be prevented from occurring at these internal 3’ ends in order for the downstream gene to be expressed (Haenni et al., 2009). Moreover, we know that trans-splicing at internal 3’ ends in operons are co-transcriptional, since it is normally difficult or impossible to detect the polycistronic precursors. Thus, downstream operon transcripts are provided with a cap co-transcriptionaly by trans-splicing rather than by direct processing by the capping enzymes. As a result, they are predicted to be processed without any need for co-transcriptional binding of capping enzymes to the CTD, and so Ser-5p at these 5' ends should be unnecessary, assuming the SL2 snRNP does not bind Ser-5p. However, all operon 3’ ends are formed co-transcriptionaly by the normal mechanism, which would presumably require Ser-2p at the 3' end of each gene in the cluster. This provides a unique opportunity to test whether Ser-5p is indeed associated only with sites requiring co-transcriptional association of capping enzymes, or whether it occurs also at 5' ends of genes expressed downstream in the cluster at sites distant from the promoter. Furthermore, it provides an opportunity to test whether Ser-2p is associated with all RNA 3' ends in the cluster or only those accompanied by transcription termination at the 3' end of the entire cluster.

In this chapter I present high resolution mapping of Ser-5p and Ser-2p by ChIP-qPCR experiments in three different *C. elegans* operons. In all cases, Ser-5p is associated with promoter locations, but not with locations specifying 5' ends of mRNAs distant from promoters, whereas Ser-2p is associated with all 3' ends, even those at large distances from transcription termination sites. These data provide strong support for the idea that these phosphorylation events mark RNA processing sites rather than gene ends, and also provide a novel demonstration that genes in *C. elegans* operons are co-transcribed as predicted.

Results

Ser-5 phosphorylation in a non-operon gene

ChIP experiments done in mammals and yeast have shown that RNAPII found at 5’ ends of genes contains high levels of Ser-5p, and that this phosphorylation decreases as RNAPII moves across the gene (Phatnani and Greenleaf 2006; Buratowski 2009). However, it is unknown if in *C. elegans* RNAPII is phosphorylated at Ser-5 in a similar manner to other model organisms.

In order to determine whether the CTD of RNAPII is phosphorylated in a pattern similar to yeast and mammalian RNAPII, I performed ChIP qPCR experiments using an antibody specific to Ser-5p in a non-operon *C. elegans* gene, r10e4.2. This gene was chosen because of its isolated genomic location and its relatively high transcript level (Thierry-Mieg D. and Thierry-Mieg J. 2006). By quantifying the immunoprecipitated DNA using multiple primer sets spanning the entire transcription unit, I was able to map Ser-5 phosphorylations across the gene.

As shown in Figure III-1, the pattern of Ser-5p is high near the promoter and stays high throughout the body of the gene, although its level decreases in the latter half of the gene. Moreover, plotting the data from Figure III-1 as a ratio of a ChIP with an antibody against total RNAPII CTD, a clear 5' bias is observed (Figure III-2). These results indicate that the r10e4.10 behaves much like yeast and mammalian genes with respect to Ser-5 phosphorylation (Chapman et al., 2007; Morris et al., 2005; Komarnitsky et al., 2000; Gomes et al., 2006).

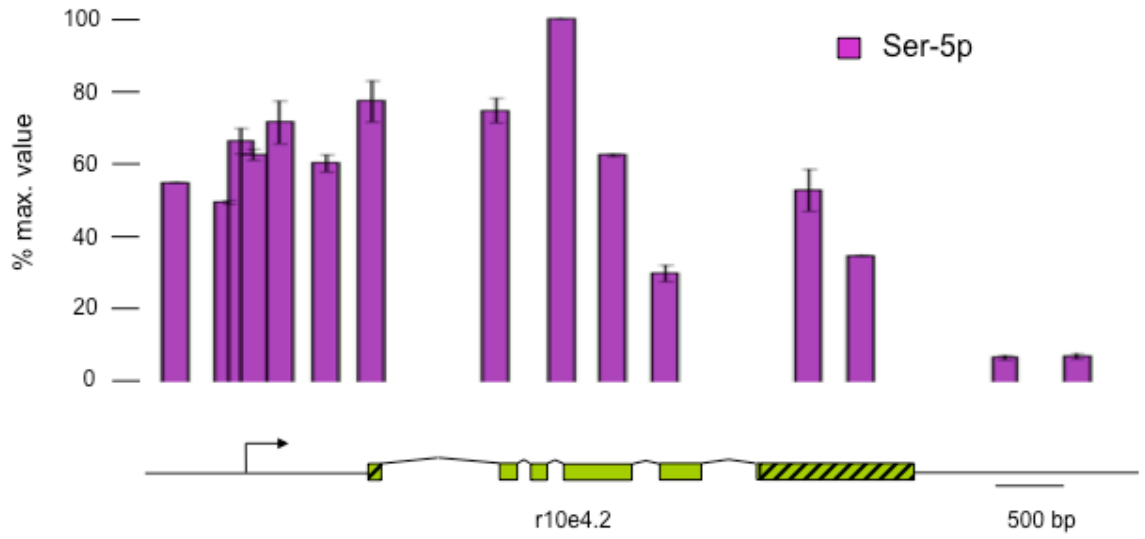


Figure III-1: Ser-5 phosphorylation across a *C. elegans* gene. ChIP qPCR experiment using an antibody against Ser-5p (Bethyl) in the r10e4.2 gene, each normalized to the highest value. The gene is depicted by filled boxes representing exons, with untranslated regions hatched, and introns shown as angled lines. Flanking intergenic regions are shown as lines. The arrow denotes the location of the promoter. The results from each primer set are positioned immediately above the corresponding genomic location. Error bars represent standard error of the mean of two independent immunoprecipitation experiments.

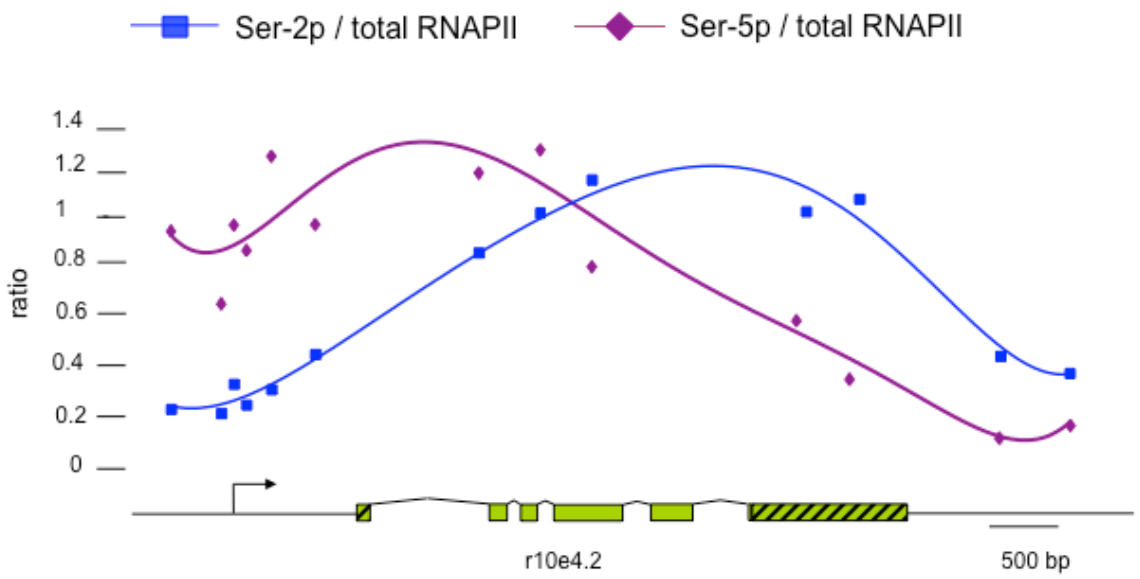


Figure III-2: Ser-5 and Ser-2 phosphorylation normalized to total RNAPII across r10e4.2. ChIP qPCR signals of Ser-5p and Ser-2p as a ratio of total RNAPII. Normalized ChIP signals of Ser-5p and Ser-2p were divided by the normalized total RNAPII at each primer set. The graph shows best fit lines. See Figure III-1 caption for details.

Ser-2 phosphorylation in a non-operon gene

In contrast to Ser-5p, in yeast and mammals the phosphorylation of Ser-2 is low at 5' ends of genes and increases as RNAPII moves across the gene, with maximal phosphorylation at the 3' ends of genes (Phatnani and Greenleaf 2006; Buratowski 2009). In *C. elegans* it is unknown if Ser-2 phosphorylation resembles the pattern seen in other model organisms.

In order to determine if the CTD of RNAPII is also phosphorylated at Ser-2 in a pattern similar to yeast and mammals, I performed ChIP-qPCR experiments with an antibody specific to Ser-2p on r10e4.2. As shown in Figure III-3, phosphorylation of Ser-2 is relatively low near the promoter and gradually increases towards the 3' end of the gene. In contrast to Ser-5p, when the Ser-2p data from Figure III-3 is plotted as a ratio of a ChIP with an antibody against total RNAPII CTD, the bias is towards the 3' end of the gene (Figure III-2). These results are consistent with the phosphorylation pattern for

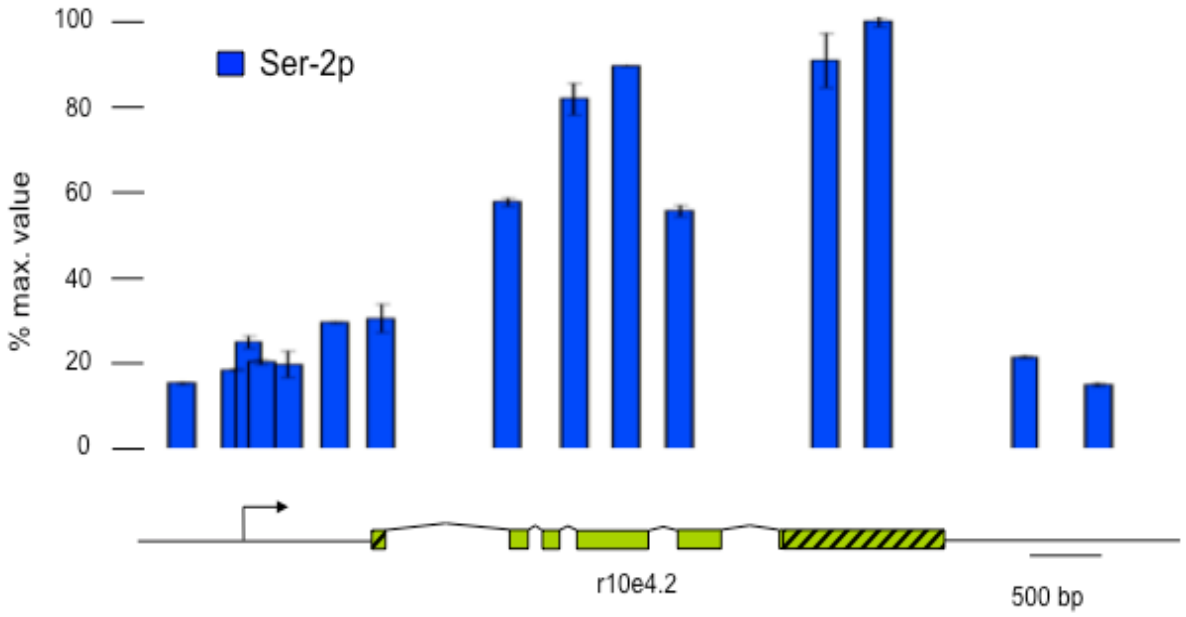


Figure III-3: Ser-2 phosphorylation across a *C. elegans* non-operon gene. ChIP qPCR experiment using an antibody against Ser-2p in r10e4.2. See Figure III-1 caption for details.

Ser-2 shown in other organisms. In sum, my experiments demonstrate that *C. elegans* non-operon genes are similar to yeast and mammalian genes with respect to Ser-2p and Ser-5p (Chapman et al., 2007; Morris et al., 2005; Komarnitsky et al., 2000; Gomes et al., 2006).

Ser-5 phosphorylation is associated with 5' ends, but only those close to promoters

In general, Ser-5 phosphorylation marks the 5' ends of genes where the site of mRNA capping and promoters occur. Therefore, Ser-5p could be correlated with either sites of mRNA 5' end formation or promoters. Operons provide a unique opportunity to test whether Ser-5 CTD phosphorylation is specific to promoter regions or whether it marks all sites of mRNA 5' end formation, whether they occur by transcription initiation or by trans-splicing.

I used ChIP-qPCR experiments with an antibody specific to Ser-5p in three *C. elegans* operons, in order to reveal if phosphorylation correlates with promoters or mRNA 5' end formation. In all cases I found that Ser-5p was associated with promoters and not with 5' ends far from promoters.

In a two-gene operon (CEOPX012), Ser-5p is highest at the promoter and drops unevenly throughout the body of the operon (Figure III-4). The reason for the uneven drop of Ser-5p within the body of the operon is unclear. One likely possibility is that the levels of Ser-5p drops to close to background levels downstream of the promoter, making difficult the quantification of Ser-5 phosphorylation by qPCR.

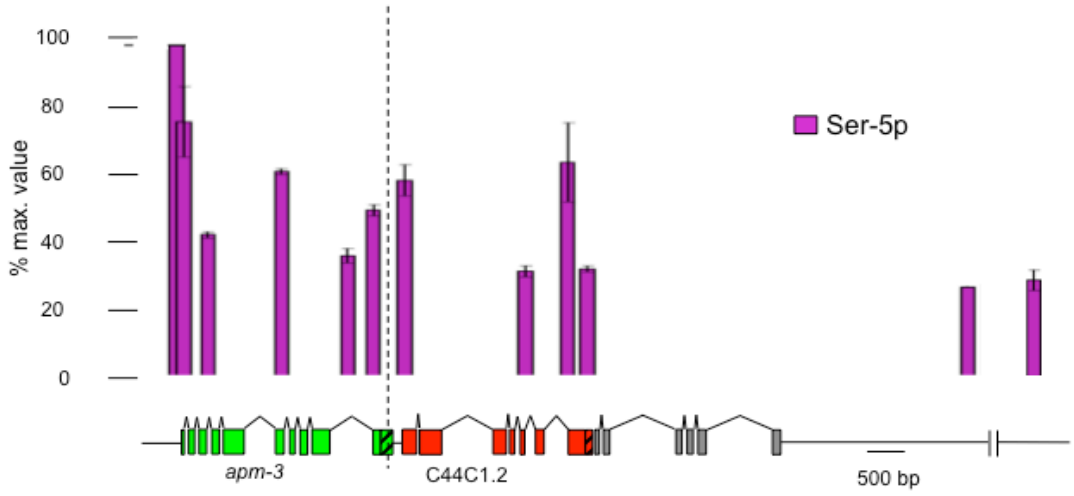


Figure III-4: In a two-gene operon (CEOP X012) Ser-5p is high at the promoter but not at an internal 5' end. ChIP qPCR signals of Ser-5p in a two-gene operon, each normalized to the highest value. The dashed line separates genes. A gene in the opposite orientation, just 3' of C44C1.2, is expressed at a very low level, and primers that query this region were not tested. The most 3' primer pair is a negative control located in the center of a region lacking annotated genes; it is not adjacent to this operon. See Figure III-1 caption for details.

In a four-gene operon (CEOP3156), Ser-5p is high near the promoter and decreases throughout the body of the operon (Figure III-5). Importantly, Ser-5p RNAPII does not peak at most of the internal 5' ends (trans-splice sites), which are not close to promoters, except for the second gene in the operon (see below). In addition, I have confirmed this result by using a different Ser-5 phosphorylation antibody from the Bentley lab (Figure III-6). Although the patterns are very similar showing the highest Ser-5p at the promoter of the operon, they are not identical (see below).

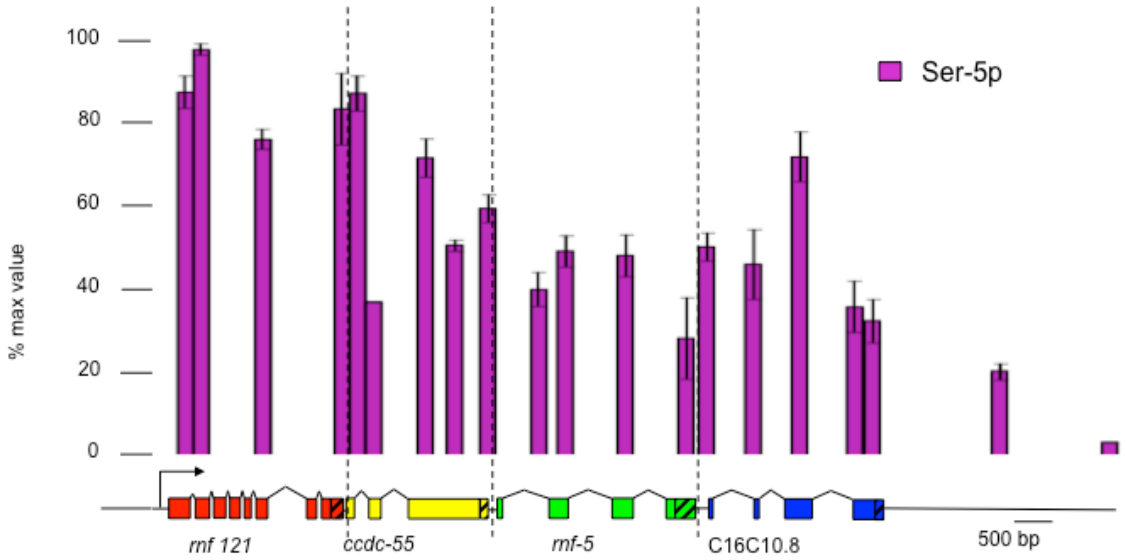


Figure III-5: In a four-gene operon (CEOP 3156) Ser-5p is high at the promoter but not at internal 5' ends. ChIP qPCR signals of Ser-5p in a four-gene operon, each normalized to the highest value. Error bars represent standard error of the mean from three independent immunoprecipitation experiments. The dashed lines separate genes. See Figure III-1 caption for details.

As shown in Figure III-6, the Bentley lab antibody (yellow) shows a second prominent peak at the 5’ end of the internal *ccdc-55* gene, which is barely visible when using the Bethyl antibody (magenta). In addition, the last gene in the operon seems to have a small peak too only present with the Bentley antibody. The reason for this is unknown but may be due to a PCR artifact. For example, at the 5’ end of the ccdc-55 gene the overlapping primer right next to it does not show such an elevated signal, which may be an indication of a problem quantifying the IP DNA. Similarly, for the primer set showing the peak at the 5’ end of the c16c10.8 gene, the error bar is big indicating high variability in quantifying the IP DNA. Alternatively, there could be some small level of transcription initiation at these internal sites causing Ser-5p to be elevated (see below), which can be better distinguished with the Bentley Ser-5p antibody. Allen et al., (2011) showed that increased used of SL1 trans-splicing to the downstream mRNA correlated with the presence of an internal promoter between genes in operons. Indeed, RNA-seq experiments showed that the sites where there is a possible Ser-5p peak, rnf-121/ccdc-55 and rnf-5/c16c10.8, SL1 trans-splicing occurs in 6% and 3% of the reads, respectively (Allen et al., 2011). However, the 5’ end of rnf-5 within this operon that has no Ser-5p present (Figure III-6) shows similar low levels of SL1 trans-splicing (2%), suggesting that it is unlikely that RNAPII initiates at these internal sites.

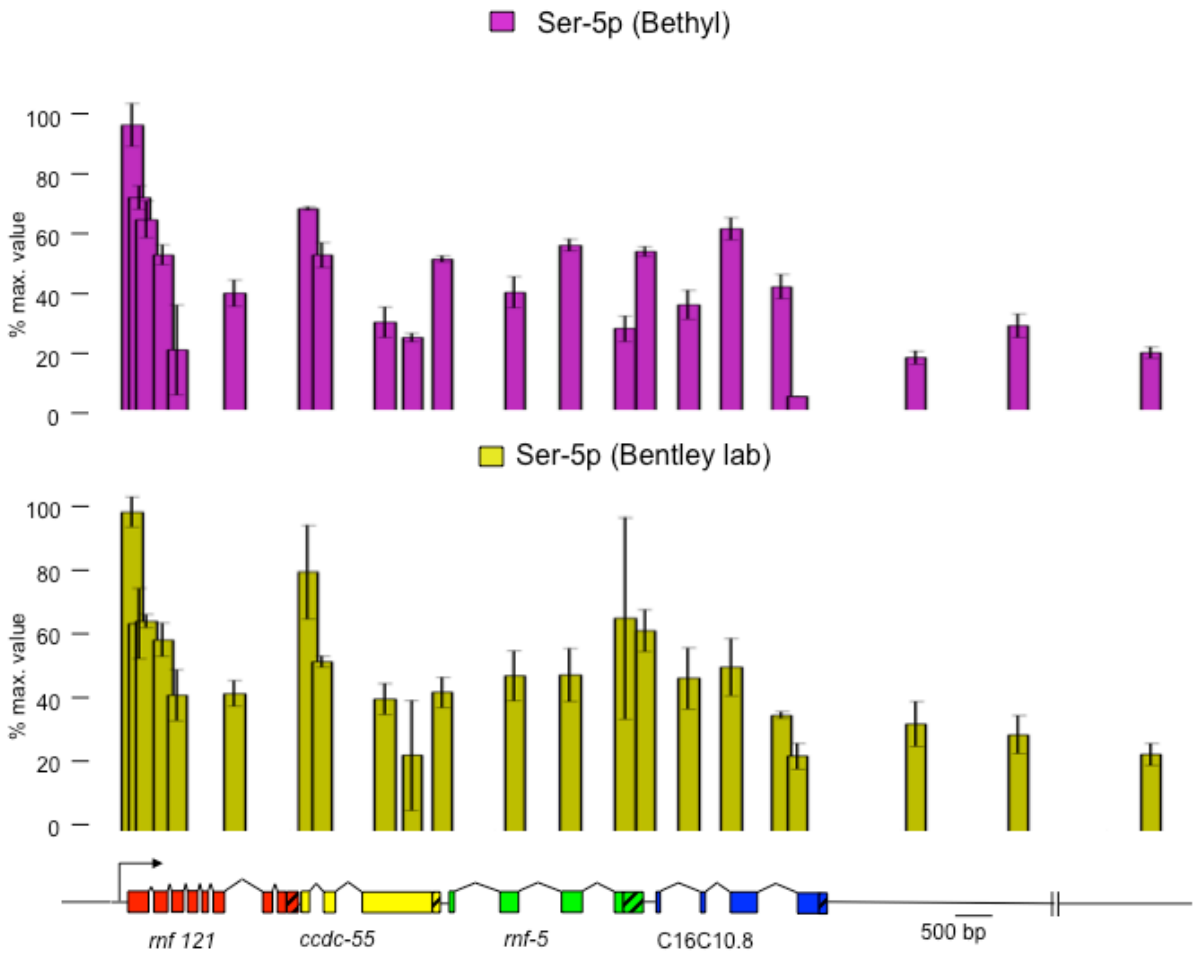


Figure III-6: Similar patterns using two different Ser-5p antibodies in a four-gene operon. Magenta ChIP-qPCR (top) was done with the commercially available Ser-5p antibody. The yellow ChIP-qPCR (bottom) was done with a Ser-5p antibody from the Bentley lab. Error bars represent standard error of the mean from three independent immunoprecipitation experiments. See Figure III-1 caption for details.

In an eight-gene operon (CEOP1484) Ser-5p is high near the promoter of the operon and then drops throughout the first five genes. However, there is a second Ser-5p peak at an internal site near the 5' end of the sixth gene (Figure III-7, asterisk). Interestingly, this internal Ser-5p peak occurs at a previously uncharacterized internal promoter. Three lines of evidence support the existence of this promoter: First, H3K9ac is enriched at this internal location based on a ChIP-on-chip experiment (Figure III-8). Promoter accessibility or “openness” has been correlated with the presence of H3K9ac at 5’ ends of genes, needed for efficient and accurate gene activation (Roh et al., 2005; Agalioti et al., 2002). Second, there is a peak of paused RNAPII in starved worms that are released upon feeding, which have been shown to be characteristic of promoters (Baugh et al., 2009). Third, the r05d11.7 trans-splice site shows a high level of SL1 trans-splicing (Allen et al., 2011). Generally SL1 trans-splicing occurs at the first gene downstream of a transcription start site, but not at downstream genes (Hillier et al., 2009). Thus, Ser-5p peaks at promoter locations, even a promoter situated within an operon, but it does not peak at most positions where 5’ ends are formed by trans-splicing.

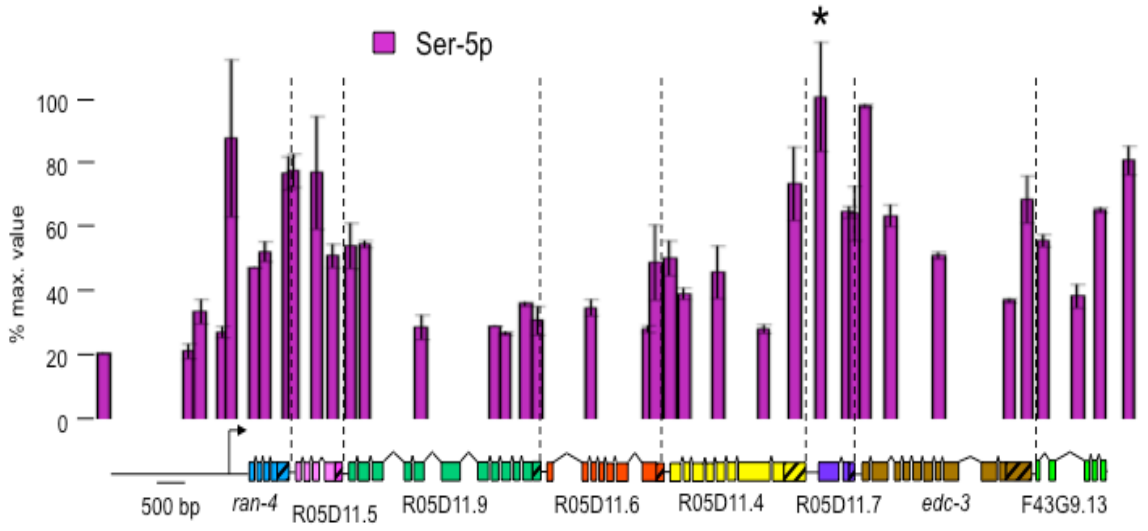


Figure III-7: In an eight-gene operon (CEOP 1484) Ser-5p is high at the promoter and at a previously uncharacterized internal promoter (asterisk). ChIP qPCR signals of Ser-5p were normalized to the highest value. Error bars represent standard error of the mean from three independent immunoprecipitation experiments. The dashed lines separate genes. See Figure III-1 caption for details.

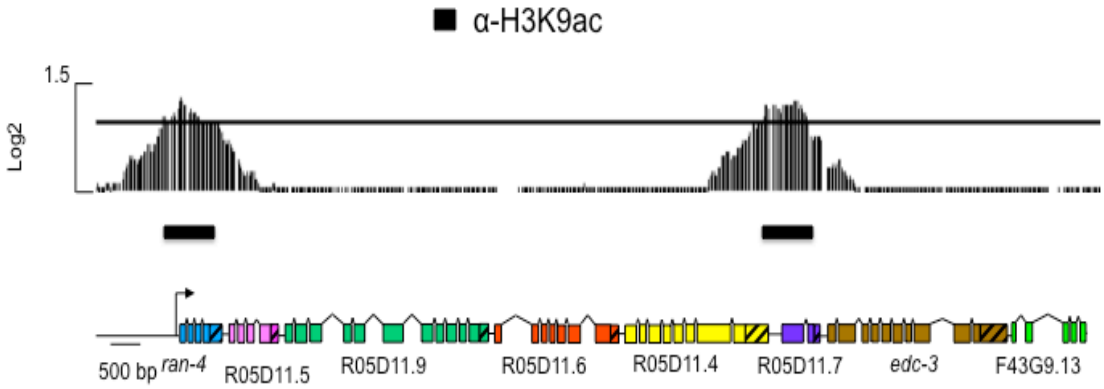


Figure III-8: H3K9ac in the eight-gene operon confirms the presence of an internal promoter. H3K9ac ChIP-on-chip is shown in black, marking the promoters. Each vertical line represents data from an individual probe, and the height of the bar is proportional to the amount of hybridization. The horizontal black line under the graph shows the region of the peak with a p-value of 0.05.

In addition, I noted that Ser-5p does appear to peak at a few 3' ends, especially those near the 3' end of this operon (Figure III-7), a result for which I lack an explanation. One possibility is that they represent RNAPII transcription in the anti-sense direction, so that the 3' end of the gene becomes the start of transcription. An alternative possibility is that the peaks of Ser-5p at the 3' end of the three last genes in the operon represent RNAPII pause sites. However, a ChIP experiment with total RNAPII antibody on this operon revealed that there are no significant pause sites at these 3' ends (see Figure III-12). Therefore, the peaks of Ser-5p present at the 3' ends of the three genes in this operon do not appear to result from RNAPII pausing at these locations, and so probably represent real Ser-5p peaks.

Ser-2 phosphorylation is associated with 3' end formation sites

Ser-2 phosphorylation has been shown to be associated with 3' ends of genes where 3' end formation and transcription termination occur. However, this phosphorylation could either be correlated with the site of mRNA 3' end processing or transcription termination. Operons are able to separate these two events, since at internal 3' ends in operons 3' end formation occurs without transcription termination. Therefore, operons provide a unique opportunity to test if Ser-2p is associated with sites of 3' end formation or with transcription termination.

I used ChIP-qPCR experiments with an antibody specific to Ser-2p in three *C. elegans* operons to determine if Ser-2p peaks at sites where termination does not occur. In all cases I found that Ser-2p was present at each 3' end, suggesting that Ser-2p is marking 3' end formation sites, rather than transcription termination.

In a two-gene operon (CEOPX012) Ser-2p is low at the 5' end of the operon, peaks at the 3' end of the first gene, then drops within the second gene, and peaks a second time at the 3' end of the operon (Figure III-9). Note that here and in subsequent figures the peaks at 3' ends of operon genes often extend to the 5' end of the next gene. We presume this is due to the fact that the genes are only 100 bp apart, and ChIP fragment size averages ~500 bp. Thus, the experiments lack the resolution to enable us to determine whether Ser-2p decreases immediately following the 3' end formation site.

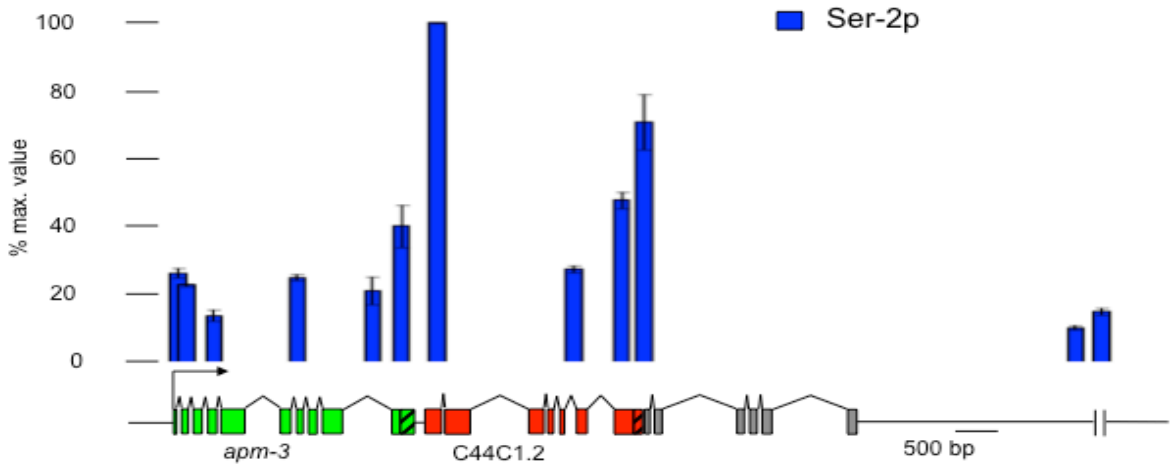


Figure III-9: In a two-gene operon (CEOP X012) Ser-2p is high at both 3' ends. ChIP qPCR signals of Ser-2p in a two-gene operon, each normalized to the highest value. The dashed lines separate genes. See Figure III-4 caption for details.

In a four-gene operon (CEOP 3156) Ser-2p peaks at all four 3' ends in the operon, dropping to a lower level following each 3' end (Figure III-10A). Do these peaks represent pause sites where total RNAPII would be expected to increase as well, or do they represent increases in the fraction of RNAPII that is Ser-2 phosphorylated? To distinguish between these two possibilities I determined the amount of total RNAPII with

each primer set and then divided the ChIP signal of Ser-2p by the ChIP signal for total RNAPII. This ratio clearly delineates the four peaks of Ser-2p representing the 3' end of each gene in the operon (Figure III-10B), consistent with these peaks representing a specific enrichment of Ser-2 phosphorylation. Another interesting question is whether RNAPII pauses at internal 3' end formation sites. Although the data in this chapter does not resolve this question, chapter IV shows RNAPII pausing at all 3' ends in operons correlating with Ser-2p and CstF-64 recruitment.

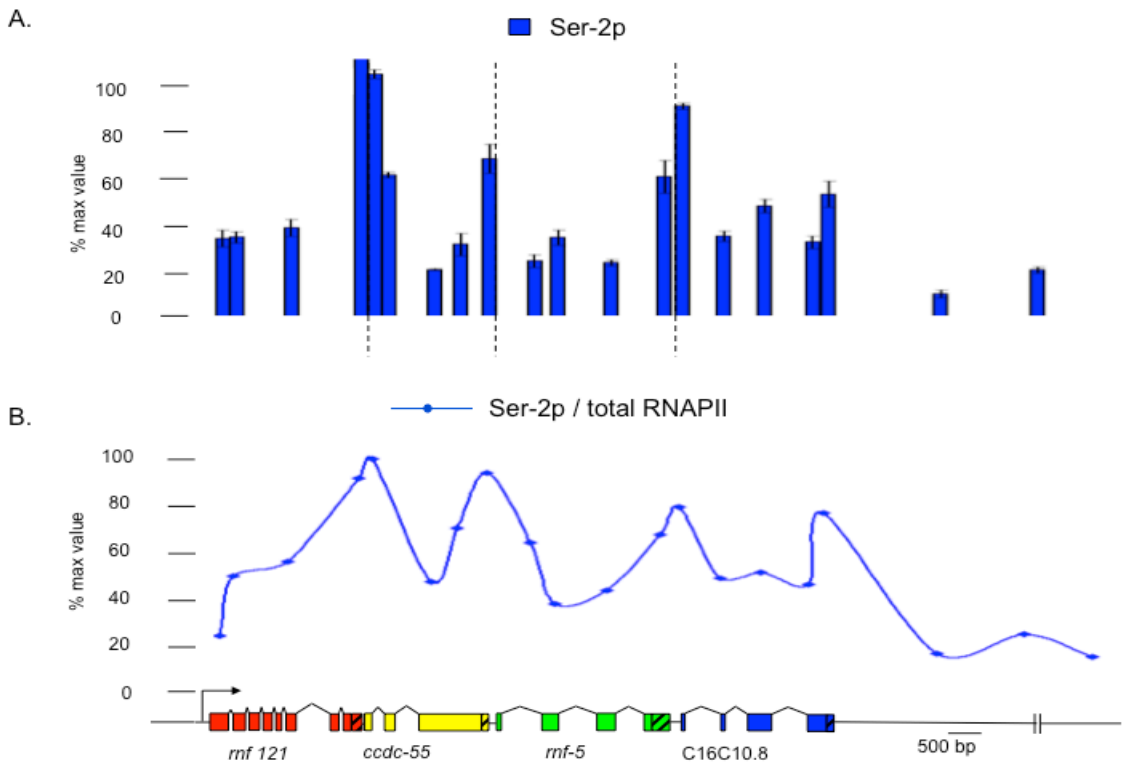


Figure III-10: In a four-gene operon (CEOP 3156) Ser-2p is high at each 3' end. A. ChIP qPCR signals of Ser-5p in a four-gene operon, each normalized to the highest value. Error bars represent standard error of the mean from three independent immunoprecipitation experiments. The dashed lines separate genes. See Figure 1 caption for details. B. ChIP qPCR signals of Ser-2p as a ratio of total RNAPII in a four-gene operon. Normalization was done as in Figure III-3. The line connects adjacent points.

In an eight-gene operon (CEOP1484) Ser-2p peaks near each gene 3’ end, falling to a lower level within each gene, and rising again at each gene 3' end (Figure III-11). Importantly, the Ser-2p pattern seen in the eight-gene operon is consistent with the previous operon examples shown.

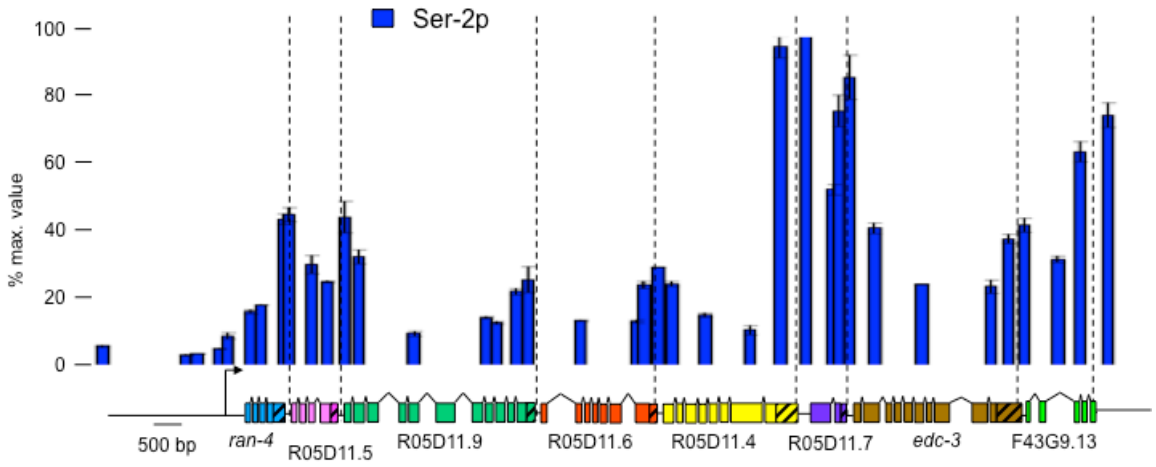


Figure III-11: In an eight-gene operon (CEOP 1484) Ser-2p is high at each 3’ end. ChIP qPCR signals of Ser-2p were normalized to the highest value. Error bars represent standard error of the mean from three independent immunoprecipitation experiments. The dashed lines separate genes. See Figure III-1 caption for details.

In summary, these data provide strong support for the idea that each gene is treated as a separate entity with respect to Ser-2 phosphorylation, supporting the idea that Ser-2 phosphorylation is associated with RNA 3' end formation. However, one could argue that a small population of RNAPII molecules does terminate at internal sites, allowing the possibility that Ser-2p could be associated with sites of transcription termination. To investigate this possibility, we performed ChIP qPCR experiments using an antibody to total RNAPII in the eight-gene operon (Figure III-12). Clearly, total RNAPII levels do not drop from one gene to the next in the operon, providing evidence that premature transcription termination is not occurring at significant levels at most 3'

end formation sites. However, the increase in total RNAPII levels following the internal promoter is small, suggesting this promoter may compensate for some transcription termination following r05d11.4.

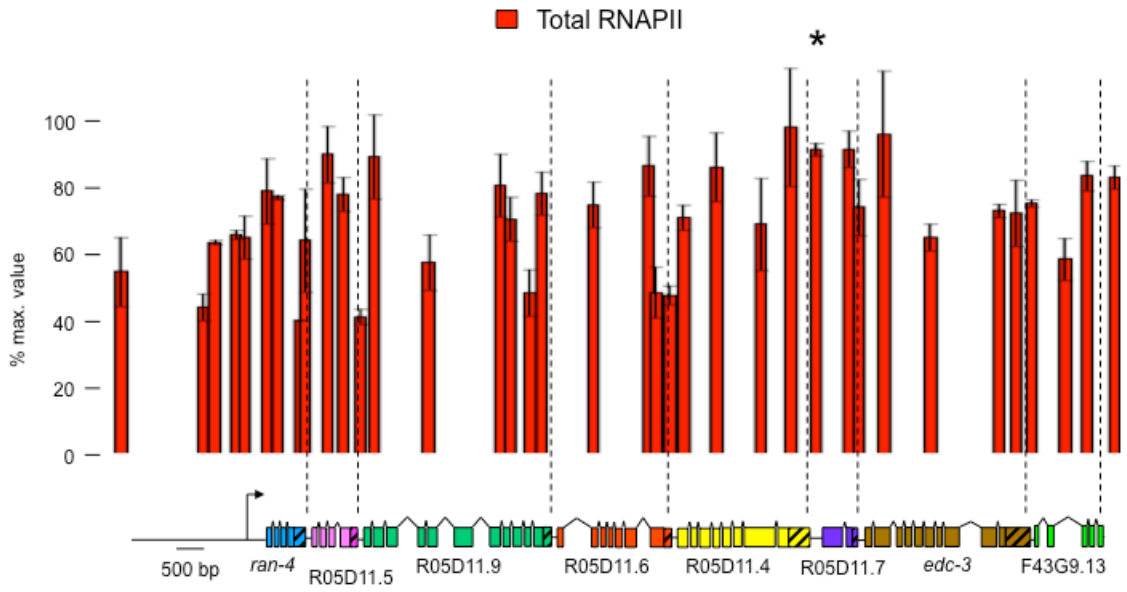


Figure III-12: In an eight-gene operon (CEOP 1484) total RNAPII levels are constant. ChIP qPCR experiment using an antibody against total RNAPII normalized to the highest value. The location of a previously unknown internal promoter is marked by an asterisk. Error bars represent standard error of the mean of two independent immunoprecipitation experiments. The dashed lines separate genes. See Figure III-1 caption for details.

Discussion

C. elegans non-operon genes are treated similarly to genes of other model organisms with respect to RNAPII CTD phosphorylation

The phosphorylation of the RNAPII CTD is dynamic and highly complex. Each heptad repeat can get phosphorylated at five out of the seven residues creating a CTD code, which is further complicated by proline isomerization and glycosylation. Out of all these post-translational modifications, significant attention has been paid to the phosphorylation of serine 2 and 5 (reviewed in Bartkowiak et al., 2011). Several ChIP experiments done in yeast and mammals have shown that Ser-5p is high at 5' ends of genes, while Ser-2p is high at 3' ends. Here I tested if the pattern of Ser-5p and Ser-2p in a *C. elegans* non-operon gene is similar to the pattern reported for other organisms.

I showed that in a non-operon gene Ser-5p is high near the promoter and relatively low at the 3' end of the gene. In contrast, Ser-2p is low near the promoter but high at the 3' end of the gene. The bias of Ser-5p and Ser-2p towards the 5' ends and 3' ends of non-operon genes, respectively, becomes more apparent when the data is plotted as a ratio of a ChIP with an antibody against total RNAPII. Therefore, *C. elegans* genes are treated similarly to those of other model organisms with respect to Ser-5 and Ser-2 phosphorylation (Bartkowiak et al., 2011).

RNAPII CTD phosphorylation correlates with RNA processing events

The patterns of phosphorylation of Ser-2p and Ser-5p on RNAPII CTD correlate with RNA processing events that occur near gene ends: Ser-5 is phosphorylated near the 5' ends of genes, whereas Ser-2 is phosphorylated near 3' ends. Because some pre-mRNA processing enzymes can bind to Ser-5p or Ser-2p, it has been postulated that these processing events are facilitated by binding of processing proteins to the CTD phosphorylated at these sites (Bartkowiak et al., 2011; Perales and Bentley 2009; Phatnani and Greenleaf 2006). However, this idea requires additional support and in more experimental systems; it remains possible these phosphorylation events are associated with sites of pre-mRNA processing, but do not direct them. Capping enzymes do bind preferentially to RNAPII containing Ser-5p, which peaks close to the 5' ends of genes, the site of pre-mRNA capping. However, this site also occurs close to the promoter. Since in general all 5' ends of genes occur at the promoter, Ser-5p could be correlated with either promoters or sites of mRNA 5' end formation. Similarly, Ser-2p peaks near 3' ends of genes, but mRNA 3' end formation and transcription termination are in general inextricably linked (Rosonina et al., 2006). Thus, it is possible that Ser-2p plays a role in transcription termination as well as mRNA 3' end formation. The events simply occur too close together to allow them to be distinguished based solely on chromatin immunoprecipitation in yeast or mammalian experimental systems. In this section I used *C. elegans* operons to answer these questions.

Ser-5p in operon genes is associated with promoters

I showed that Ser-5p marks the location where pre-mRNAs are co-transcriptionaly capped near the transcription start site. However, Ser-5p does not mark downstream positions, where capped 5' ends are formed by trans-splicing (Figures III-4 – III-6). The only site within any of the three operons I studied with a large Ser-5p peak was at a site in the eight-gene operon where there is a promoter (Figure III-7, asterisk). The existence of an internal promoter was not entirely surprising since internal promoters in *C. elegans* operons have been noted previously (Huang et al., 2007; Whittle et al., 2008). Thus our results support the idea that Ser-5 phosphorylation serves to facilitate co-transcriptional capping, presumably by binding capping enzymes.

How are internal 5’ ends in operons co-transcriptionally processed by SL2, if they contain low levels of Ser-5p? One likely possibility is that the SL2 trans-splicing machinery is not recruited by Ser-5p, but instead by the 3’ end formation machinery needed to process the upstream 3’ end. Indeed, 3’ end formation occurring at the 3’ end of the upstream gene has been shown to be mechanistically coupled to SL2 trans-splicing occurring at the 5’ end of the downstream gene (Kuersten et al., 1997; Liu et al., 2001). Moreover, a complex consisting of CstF-64 and the SL2 RNA was purified from *C. elegans* extracts (Evans et al., 2001). Therefore, SL2 trans-splicing might occur as a result of 3’ end formation occurring at the 3’ end of the upstream gene. Alternatively, the SL2 trans-splicing machinery could be co-transcriptionally recruited by other CTD post-translational modifications.

Ser-2p in operon genes is associated with 3’ end formation sites

Similarly, my data support the conclusion that Ser-2p serves to facilitate RNA 3' end cleavage, presumably by binding 3' end formation proteins to RNAPII, since I found a peak of Ser-2p at all 3' end formation sites I examined, many of which occur at large distances from sites of transcription termination (Figure III-9 – III-11). Indeed, it was sometimes the case that the lowest peak of Ser-2p occurred at the 3' end of the entire gene cluster. Of course it is possible that some RNAPII is terminating at each internal poly-A site, but this possibility is made unlikely by the fact that total RNAPII levels remained flat throughout the eight-gene operon (Figure III-12). If transcription were terminating at internal sites, I would have expected to see a gradual reduction of RNAPII levels from the 5' to the 3' end of the operon. These results argue strongly that Ser-2p facilitates RNA 3' end formation as previously proposed, although they do not eliminate the possibility that Ser-2p could facilitate transcription termination directly as well. If that is the case, my results indicate that Ser-2 phosphorylation is not sufficient to direct transcription termination, since the internal Ser-2p peaks occur at sites quite distant from transcription termination.

Dynamic phosphorylation and dephosphorylation cycle across operons

Interestingly, it is clear that with respect to Ser-5p the entire operon is treated as a single gene, while with respect to Ser-2p it is treated as a cluster of individual genes. The mechanism by which the phosphorylation and dephosphorylation enzymes can make this distinction will be an interesting subject for further study. For example, one of the phosphatases for Ser-5p, Rtr1, is required for the transition from Ser-5p to Ser-2p

enrichment in yeast genes (Mosley et al., 2009). Therefore it will be of interest to determine whether the worm Rtr1 ortholog dephosphorylates Ser-5p in operon genes without causing an immediate increase in Ser-2p. In addition, the observation that Ser-2p peaks at each 3' end and then drops to lower levels in the body of the next gene before peaking again at the next 3' end, argues that the heptad Ser-2 residues are actively dephosphorylated following 3' end formation and then re-phosphorylated near the 3' end of the next gene. Clearly, it would be interesting to examine the patterns of Ser-5 and Ser-2 phosphatases as well as their kinases across these operons once antibodies capable of recognizing the *C. elegans* versions of these proteins become available. Furthermore, I presume that the capping enzymes and 3' end formation proteins are interacting with Ser-5p and Ser-2p CTD, respectively, as RNAPII traverses the operons. As shown in chapter IV, CstF-64 binding to operons and non-operons genes matches the location of Ser-2p, suggesting that this phosphorylation is indeed needed for recruitment of processing enzymes.

Ser-5p can be used for showing co-transcription of gene clusters

The evidence for the existence of *C. elegans* operons is overwhelming, but circumstantial. It has rested on the very strong association of SL2-accepting trans-splice sites with downstream positions in unusually tightly linked genes. Virtually all SL2 trans-splicing occurs at such positions, and typically only ~100 bp separates the site of 3' cleavage of the upstream gene and the SL2 trans-splice site of the downstream gene (Blumenthal et al., 2002; Allen et al., 2011). This observation, however, does not demonstrate co-transcription of gene clusters, although this has been demonstrated in a

few cases (Tanaka et al., 1997; Spieth et al., 1993). The data presented here provides a different kind of evidence for co-transcription. The fact that Ser-5p does not peak at these SL2 trans-splice sites suggests strongly that transcription is not initiating there, implying that these are not sites of co-transcriptional capping. We conclude that, at least for the three operons investigated here, the gene clusters are in fact co-transcribed. Only one of these operons also has an internal promoter.

Why don't internal 3' end formation sites in operons result in transcription termination?

I show here that the RNAPII CTD is phosphorylated on Ser-2 residues near internal 3' end formation sites just like terminal sites are, so something else must differentiate internal from terminal sites. Internal sites could have a sequence that prevents termination (Graber et al., 2007), or they might lack a sequence needed for termination. It has recently been shown that cis-splicing of the first intron of downstream genes in operons may play a role in preventing transcription termination from occurring at these internal sites (Haenni et al., 2009). Alternatively, the trans-splicing event itself, which generally occurs quite close to the 3' end cleavage event, could prevent transcription termination. If the torpedo model for transcription termination (Connelly and Manley 1988; Kim et al., 2004; West et al., 2004) is at least partly correct, then the cap provided by trans-splicing would be expected to prevent 5' to 3' degradation of downstream transcripts in operons, thereby preventing termination until after the final cleavage event.

Conclusions:

In sum, ChIP analysis of RNAPII CTD phosphorylation on several *C. elegans* operons has provided strong support for the idea that these modifications facilitate co-transcriptional processing in *C. elegans*, just as they do in mammals. The Ser-5p data also provide a novel demonstration that the genes in *C. elegans* operons are in fact parts of a single transcription unit. Some *C. elegans* operons also have internal promoters, and these can be revealed by Ser-5p peaks.

CHAPTER IV

ASSOCIATION OF CstF-64 WITH 3' ENDS IN OPERONS

Introduction

Evolution has ingeniously connected 3' end formation with transcription termination to ensure only RNAPII molecules that transcribe the end of the gene are capable of terminating. This connection was first hinted at when it was reported that the same RNA sequences needed for 3' end formation, were also needed for termination (Logan et al., 1987; Zaret and Sherman 1982; Whitelaw and Proudfoot 1986; Connelly and Manley 1988). The interdependency of 3' end formation and termination was further reinforced when it was found that some cleavage/polyadenylation factors were required for termination (Birse et al., 1998; Dye and Proudfoot 1999; Yonaha and Proudfoot 2000; Proudfoot et al., 2002). Therefore, is not clear what factors are exclusively needed for 3' end formation or transcription termination. In this chapter, I use operons to investigate CstF-64 binding at 3' ends of genes where 3' end formation has been naturally uncoupled from transcription termination.

RNA processing is composed of three distinct events that need to be completed in the nucleus, prior to mRNA export into the cytoplasm for translation. Among these is 3' end formation, which consists of an endonucleolytic cleavage at the 3' end of a pre-mRNA followed by the subsequent polyadenylation of the free 3' end. Concurrent with 3' end formation at the 3' end of genes, transcription termination must also take place, but termination tends to occur far downstream of the poly-A site (Logan et al., 1987; Bauren et al., 1998; Dye and Proudfoot 1999; Haenni et al., 2009; Glover-Cutter et al., 2008).

Transcription termination is defined as the cessation of RNA synthesis and the release of RNAPII from the DNA. Termination is needed to prevent read-through transcription of non-expressed regions and to effectively recycle RNAPII for further use.

Two models exist to explain the linkage between 3' end formation and transcription termination (Chapter I). A considerable amount of experimental evidence exists supporting both of these models for termination. However, the exact molecular mechanism that directs RNAPII termination is poorly understood. Experiments done by Glover-Cutter et al., 2008 have shown that RNAPII pauses 1-2 kb downstream of the poly-A site, with high levels of Ser-2p prior to termination. 3' end processing factors such as CstF-64 are also preferentially bound at the RNAPII termination site (Glover-Cutter et al., 2008). Moreover, this pausing downstream of the 3' end of genes appears to enhance transcription termination (Gromak et al., 2006). Therefore, 3' end formation factors such as CstF-64 might actually play a role in transcription termination since they are localized at the termination site, far downstream of the 3' end formation site.

A multisubunit complex that recognizes highly conserved cis-regulatory sequences within the RNA orchestrates the processing of 3' ends of pre-mRNAs. According to proteomics analysis, the cleavage and polyadenylation specificity factor (CPSF) and the cleavage stimulatory factor (CstF) represent the "core" of the 3' end formation machinery (Shi et al., 2009). Indeed, the formation of a stable complex at the 3' end of pre-mRNAs is initiated by the cooperative binding of CPSF-CstF with the cis-regulatory sequences on the RNA. A subunit of CPSF binds the poly-A signal (AAUAAA) located upstream of the cleavage site. This same subunit also bridges across the cleavage site to contact the CstF trimeric complex. This complex binds a

highly variable G/U rich region located 30-40 nt downstream of the cleavage site through its 64 KDa subunit. Therefore, CstF-64 is recruited at the 3' end of genes where 3' end formation and transcription termination occur close together.

Operons can be used to study the binding of CstF-64 in a unique context in which 3' end formation is uncoupled from termination. Operons are gene clusters containing two to eight genes, controlled by a single promoter located at the 5' end of the operon. Upon transcription, a single polycistronic pre-mRNA is made, which is further processed into mature mRNAs representing individual cistrons by 3' end formation occurring at the 3' end of the upstream gene and SL2 trans-splicing occurring at the 5' end of the downstream gene. Therefore, operons separate 3' end formation from transcription termination by allowing RNAPII to transcribe several functional poly-A signals without causing transcription to terminate. The mechanism by which termination is prevented at internal sites in operons is unknown. However, the most likely mechanism is that SL2 trans-splicing provides a cap to the downstream mRNA, resulting in the blockage of the exonuclease needed to terminate transcription.

Here I report an investigation on the mechanism for preventing transcription termination at internal genes in operons. I show by ChIP-seq that CstF-64 is likely directly involved in 3' end formation, since it is present at internal and terminal 3' ends in operons. Importantly, the CstF-64 binding seen at all 3' ends colocalizes with elevated Ser-2p (Chapter III) and with paused RNAPII, suggesting that RNAPII at internal genes in operons might be capable of terminating, but does not do so. My results are consistent with the proposed anti-termination model by Liu et al., 2003, in which SL2 trans-splicing prevents RNAPII from terminating following cleavage at the 3' end of the

upstream gene, by providing a cap and thus preventing torpedoing of RNAPII by the exonuclease.

Results

The CstF-64 worm homolog is enriched around the poly-A site of non-operon genes

ChIP can be used to map RNA-binding proteins relative to the DNA as an indirect way for revealing the site of association with the pre-mRNA during RNA processing. Identifying sites of association for RNA processing factors using ChIP can reveal important sites of action for these proteins during transcription, because many of these factors associate with the transcription elongation complex (TEC). They are therefore, in close proximity with the DNA (Swinburne et al., 2006). Based on ChIP experiments done in mammals, CstF-64 binding is biased towards 3' ends of genes, consistent with its known function in 3' end processing (Swinburne et al., 2006; Glover-Cutter et al., 2008). The *C. elegans* proteome contains a clear ortholog of CstF-64, based on protein sequence similarity. This ortholog is much smaller than CstF-64 found in other model organisms, due its lack of part of the C-terminal region. However, whether this CstF-64 ortholog also functions in 3' end formation has not yet been tested.

In order to determine if the recruitment of this protein correlates with the site of 3' end formation, I performed ChIP-seq experiments with an antibody against the worm protein. I found that it is significantly enriched at 3' ends of non-operon genes, similar to the reported binding for CstF-64 in other model organisms (Swinburne et al., 2006; Glover-Cutter et al., 2008). Figures IV-1 – IV-3 show representative examples of CstF-

64 association with 3' ends in three non-operon genes. The heights of the peaks are proportional to the number of aligned reads mapping to the region. The horizontal red line under each graph represents statistically significant CstF-64 binding (*p-value* of 1e-5) detected by the MACS algorithm. This algorithm predicts true binding sites (p-value 1e-5) based on fold enrichment over a statistically established background (see methods). In addition, for each example the position of the promoter is marked by the presence of H3K9ac based on ChIP-on-chip analysis performed on the same samples. The H3K9ac modification has been shown to mark active promoters in other organisms (Roh et al., 2005; Agalioti et al., 2002). CstF-64 is enriched at 3' ends of non-operon genes coinciding with the site of 3' end formation and with the CstF-64 pattern seen in other organisms (Swinburne et al., 2006; Glover-Cutter et al., 2008; Kim et al., 2004).

In Figure IV-1, the *ama-1* locus is shown with a single peak of H3K9ac marking the presumed location of the promoter. Importantly, CstF-64 is strongly enriched at the 3' end of the gene (Figure IV-1). Upon closer examination of the CstF-64 binding at the 3' end of the *ama-1* gene (zoom-in Figure IV-1), several overlapping peaks of CstF-64 are revealed with no clear boundaries between the peaks. One likely explanation is that the multiple peaks present at the 3' end of the *ama-1* gene correspond to several 3' end formation sites, which further validates CstF-64 role in 3' end formation. As many as 43% of all *C. elegans* genes contain 2 or more alternative 3' end formation sites (Mangone et al., 2010; Bartel et al., 2010) including the *ama-1* locus, which has four different alternative poly-A sites (asterisks in Figure IV-1), all located within 400 bp of the 3' end of the *ama-1* gene according to the UTRome dataset (Mangone et al., 2010).

However, assigning specific CstF-64 peaks to individual poly-A sites is not possible, presumably due to the limitation of ChIP resolution.

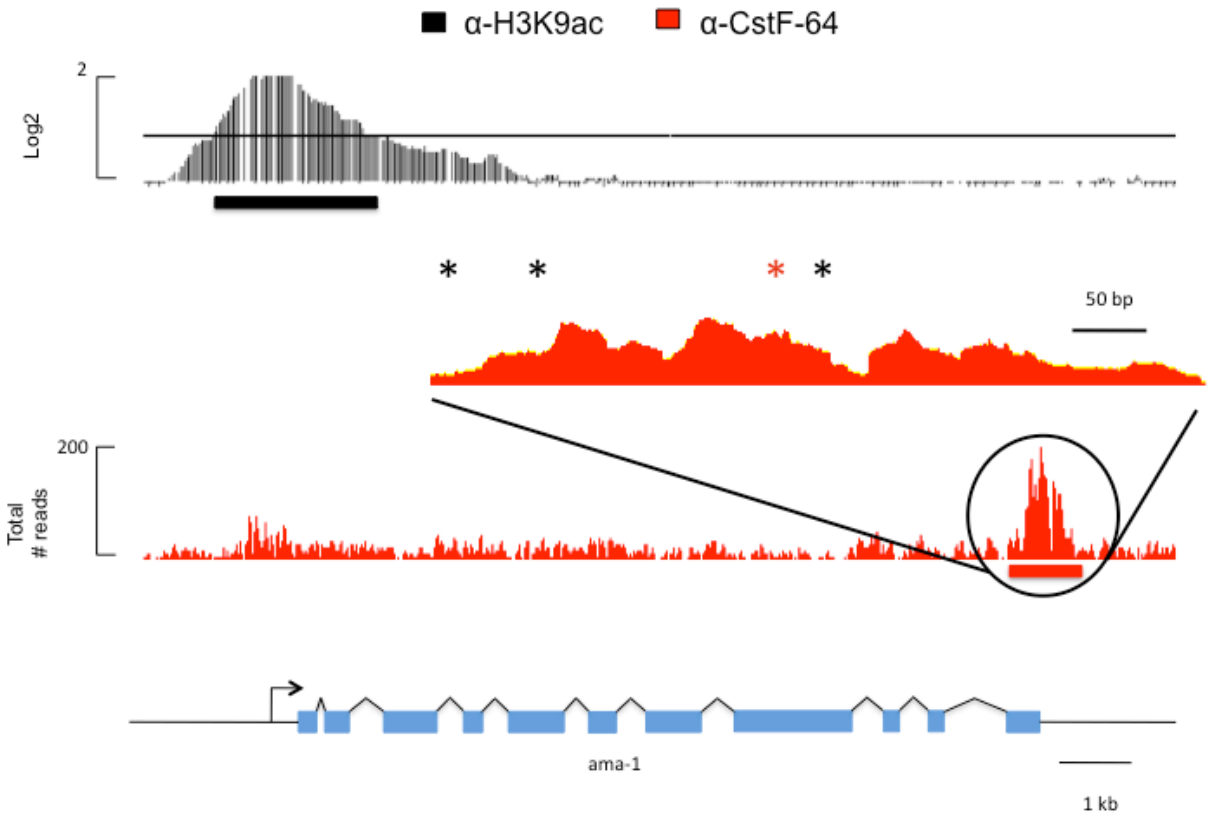


Figure IV-1: CstF-64 is enriched at the 3' end of the *ama-1* gene. H3K9ac ChIP-on-chip is shown in black, marking the promoter. Each vertical line represents an individual probe and the height of the bar is proportional to the amount of total hybridization. The horizontal black line under graph shows the region of the peak with a p-value of 0.05. CstF-64 ChIP-seq enrichment (red) is shown for the same gene. The height of the peak is proportional to the number of aligned reads matching the region. The horizontal red line under the CstF-64 peak represents statistically significant binding (p-value of 1e-5) detected by the MACS algorithm. Zoom-in of the CstF-64 peak is shown and the location of the four annotated poly A sites (UTRome) are marked by asterisks above the graph. The red asterisk indicates the major isoform based on experimental evidence across different developmental stages and supported by more than one detection method. The gene analyzed is depicted by filled boxes representing exons and introns shown as angled lines. The arrow denotes the location of the promoter.

In addition, CstF-64 levels are slightly elevated at the 5' end of the *ama-1* gene (Figure IV-1). However, there is an adjacent 3' end located less than 500 bp from the *ama-1* trans-splice site, which could be responsible for the 5' end peak at the *ama-1* gene. This genomic arrangement in *C. elegans* is very common, sometimes making it difficult to correctly assign CstF-64 peaks to specific genes. Alternatively, this could be a true 5' peak of CstF-64 as has been seen in other organisms (Glover-Cutter et al., 2008; Swinburne et al., 2006) (see below).

Figure IV-2 displays a second example of CstF-64 binding, this one at the 3' end of the *hsp-3* gene. According to the UTRome, *hsp-3* contains multiple poly-A sites that are positioned within a 250 bp window at the 3' end of the gene, again not allowing individual binding sites to be resolved (asterisks in Figure IV-2). Importantly, in this gene the CstF-64 peak is localized slightly downstream of the 3' end formation sites, matching the location of the previously analyzed CstF-64 binding regions (Figure IV-2) (Weiss et al., 1991; MacDonald et al., 1994; Takagaki and Manley 1997). In addition, a smaller CstF-64 peak is present 400 bp downstream of the primary 3' end peak. This peak may be due to the 3' end of *c45b2.8*, which is arranged convergently.

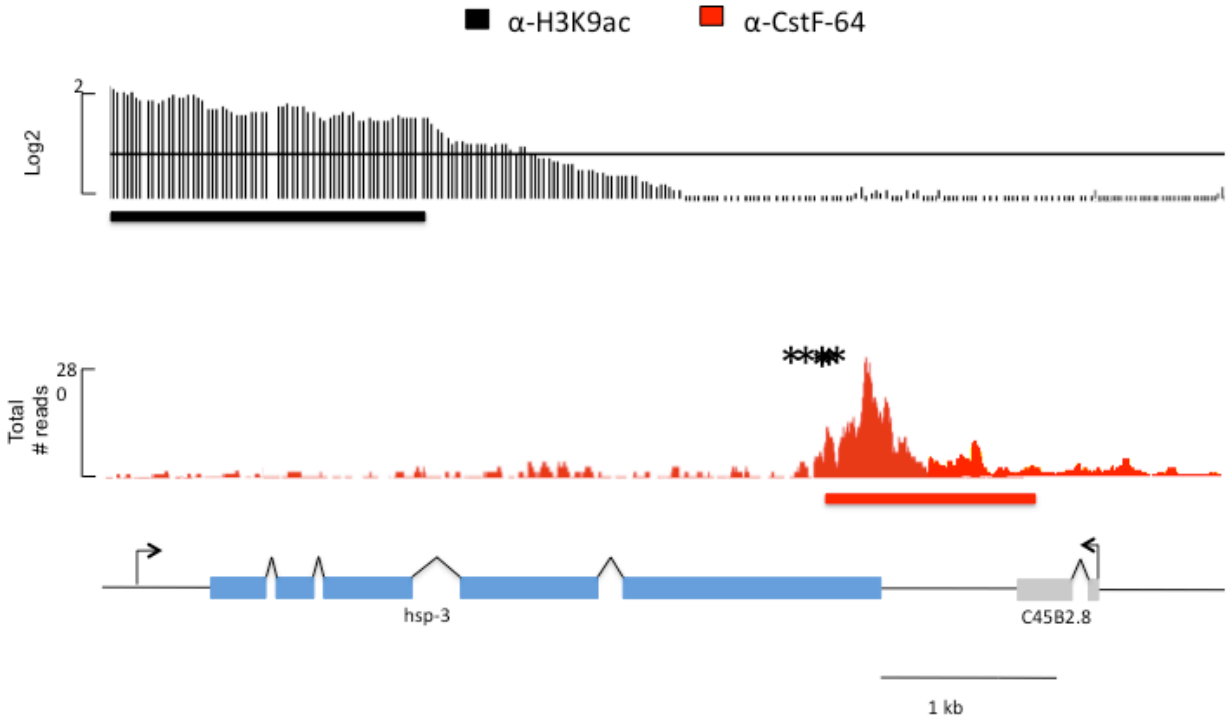


Figure IV-2: CstF-64 is enriched at the 3' end of the hsp-3 gene. H3K9ac ChIP-on-chip is shown in black marking the promoter and CstF-64 ChIP-seq is shown in red. See Figure 1 caption for details. There is a gene on the opposite strand, c45b2.8.

In contrast to the *ama-1* locus, the promoter region of *hsp-3* does not contain a CstF-64 peak. Considering *hsp-3* does not have an adjacent upstream 3' end, this result suggests that CstF-64 may only associate with 3' ends of genes. However, some genes may contain CstF-64 present far up-stream of the 3' end (see below). Several 3' end formation factors, including CstF-64, have been shown by ChIP to be present at the 5' end of many mammalian and yeast genes (Murthy and Manley 1995; Glover-Cutter et al., 2008; Wang et al., 2010; Medler et al., 2010), but its role at the promoter, if any, remains unclear.

In another non-operon example, f34d10.4 was chosen because of its genomic isolation from genes that could interfere with the accurate identification of CstF-64 binding sites associated with *f34d10.4*. The position of a single promoter located at the 5' end of the gene is shown in black by the presence of H3K9ac modification. Similar to

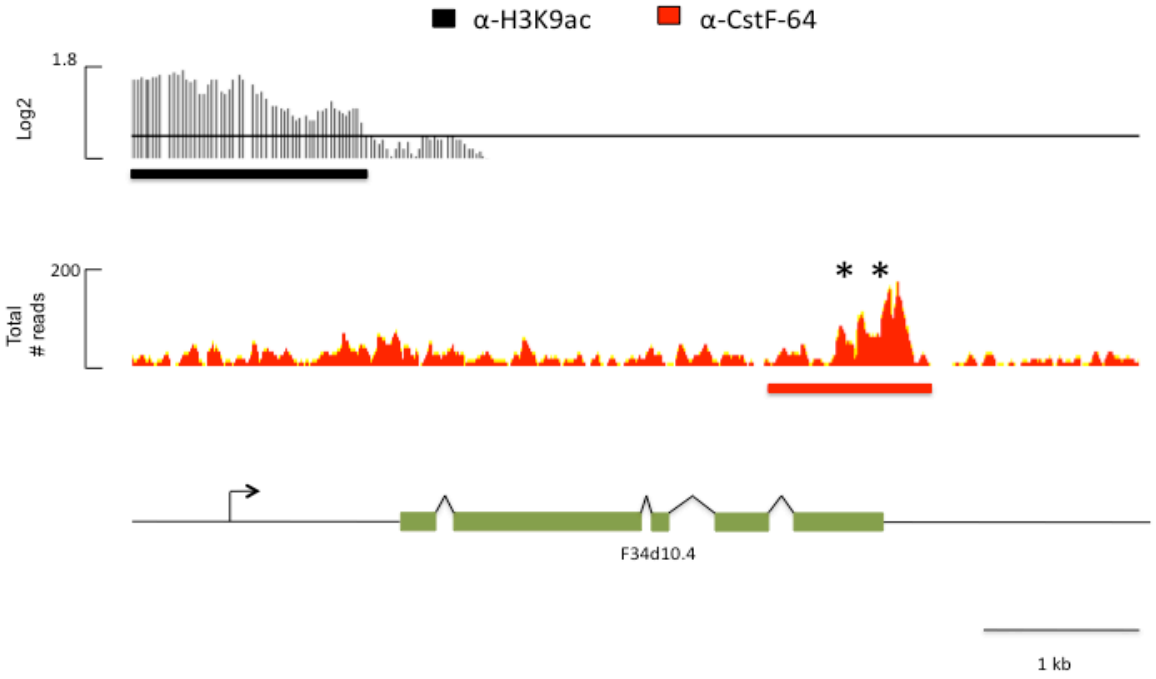


Figure IV-3: CstF-64 is enriched at the 3' end of the f34d10.4 gene. H3K9ac ChIP-on-chip is shown in black marking the promoter and CstF-64 ChIP-seq is shown in red. See Figure 1 caption for details.

the previous non-operon examples, CstF-64 is only convincingly associated with the 3' end of the gene, suggesting a role in 3' end formation (Figure IV-3). This gene's 3' end contains two alternative poly-A sites separated by 150 bp, which is a distance too close for the two peaks to be resolved, given the resolution of the ChIP. In addition, the summit of the CstF-64 peak is slightly downstream of the two alternative poly-A sites, again consistent with the location of CstF-64 binding in other organisms.

The three non-operon genes shown above (Figures IV-1 – IV-3) are representative examples of the CstF-64 profile seen in this type of gene. In all three cases the binding of CstF-64 was specifically restricted to the 3' ends of genes, coinciding with the site of 3' end processing. Importantly, mammalian ChIP experiments show a similar distribution of CstF-64 and other 3' end processing factors (Glover-Cutter et al., 2008), providing additional evidence that CstF-64 is the mammalian CstF-64 ortholog. Moreover, the summit of the CstF-64 peak occurs at a short distance (<50 bp) from the alternative poly-A site, matching the predicted CstF-64 binding site on the RNA. Therefore, these three examples are consistent with a correlation of CstF-64 binding and 3' end formation. However, these specific examples cannot indicate if CstF-64 associates at all 3' ends, similar to mammalian CstF-64.

Genome-wide analysis of CstF-64 binding in non-operon genes

In order to reveal the CstF-64 pattern across all non-operon genes in the genome, the levels of CstF-64 were averaged at single nucleotide resolution along the 3' ends of genes by dividing each gene into a metagene (Figure IV-4). The metagenes were created by taking all non-operon genes in the genome greater than 2 kb (n=8400) and scaling the region starting at +500 from the promoter to -500 from the 3' end.

As shown in Figure IV-4, CstF-64 seems to be associated at both ends of non-operon genes, with the peak at the 3' end being much higher than the one at the promoter. This result recapitulates the CstF-64 enrichment bias towards the 3' ends of the three non-operon examples shown above (Figures IV-1 – IV-3). This result is also consistent with the genome-wide association pattern of CsfF-64 in yeast and mammals,

showing a bias towards the 3' ends of genes (Swinburne et al., 2006; Glover-Cutter et al., 2008).

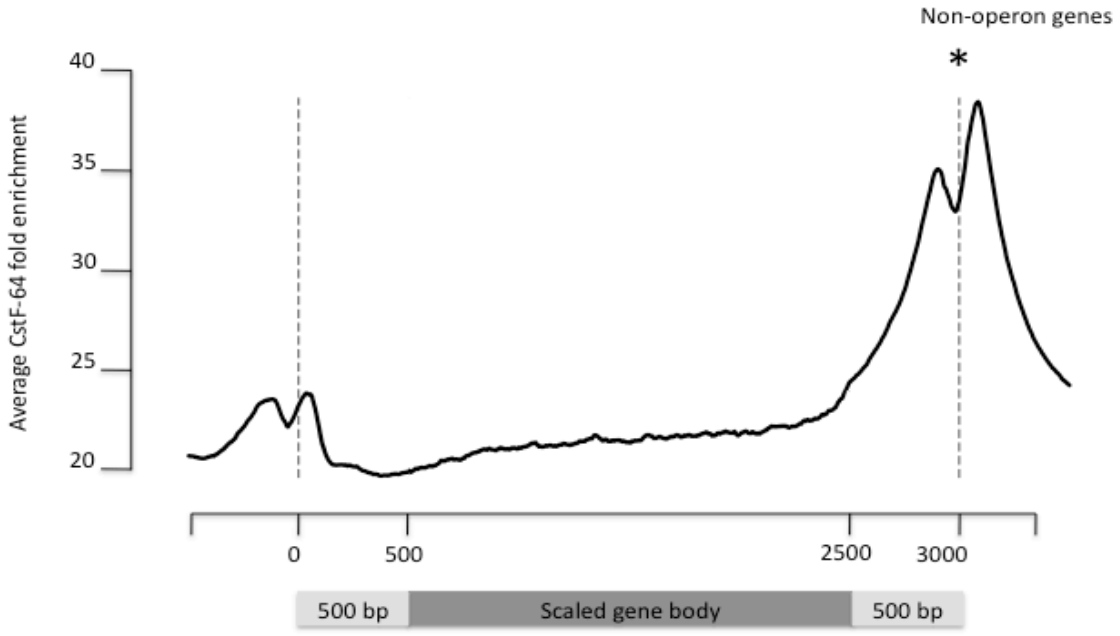


Figure IV-4. CstF-64 is significantly enriched in non-operon genes around the poly-A site: CstF-64 ChIP-seq experiment plotted along the metagene shown at the bottom of the slide, with the 0 and 3000 mark representing the transcription start site and poly-A site respectively (asterisks), which are marked by the doted lines. The metagene was made by taking all genes in the genome greater than 2 kb (n=8236) and scaling the region starting at +500 from the promoter to -500 from the 3' end (dark gray box). The y-axis represents fold enrichment for each nucleotide over background.

In order to take a closer look at the CstF-64 binding at 3' ends of non-operon genes, all annotated non-operon genes in the genome (n=17500) containing significant levels of CstF-64 were aligned by their poly-A sites. The levels of CstF-64 were averaged at a single nucleotide resolution 1000 bp upstream and 1000 bp downstream of the cleavage site (Figure IV-5).

As shown in Figure IV-5, the CstF-64 average profile at 3' ends of non-operon genes reveals two CstF-64 peaks flanking the poly-A site, both of which are approximately 90-95 bp from the cleavage site. The downstream peak is more prominent than the upstream peak, reflecting higher levels of CstF-64 and coinciding with the known CstF-64 binding site located downstream of the cleavage site. Interestingly, several mammalian 3' end formation factors show a similar double peak pattern at 3' ends of genes, but the mammalian peaks are positioned further downstream of the poly-A site (Glover-Cutter et al., 2008). The reason for this double CstF-64 peak at the 3'ends of genes is not known.

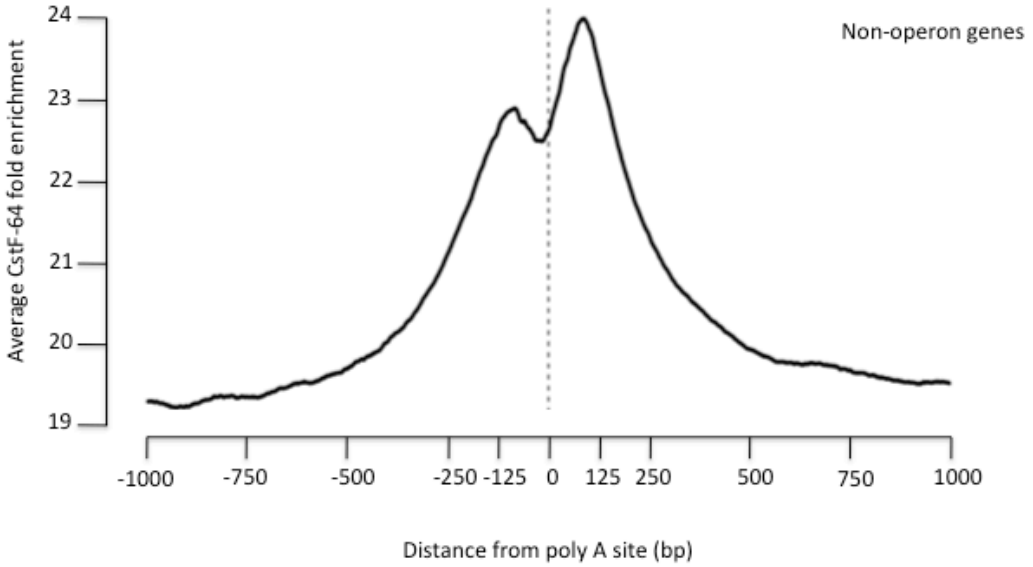


Figure IV-5. CstF-64 at the 3' end of non-operon genes consists of a double peak flanking the poly-A site: All non-operon genes (n=17063) containing significant levels of CstF-64 in ChIP-seq are plotted. The levels of CstF-64 were averaged at a single nucleotide resolution 1000 bp upstream and 1000 bp downstream of the poly A site (dashed line). The X-axis indicates distance from the poly A site which is set at zero. The y-axis shows the fold enrichment for each nucleotide over background. MACS algorithm was used to calculate CstF-64 enrichment over background (local) and a python script (average profiles) was used to manipulate data for plotting.

The majority of genes have CstF-64 binding downstream of the poly-A site, explaining the higher CstF-64 peak, but there could be genes that need extra CstF-64 for efficient processing of their 3' ends. Indeed, additional U-rich cis-regulatory elements in the RNA have been identified upstream of the poly-A signal (AAUAAA) in viruses, yeast, plants and humans (Carswell and Alwine 1989; Brown et al., 1991; Valsamakis et al., 1991; Moreira et al., 1995; Arhin et al., 2002; Natalizio et al., 2002; Hall-Pogar et al., 2005). These extra U-rich sequences have been suggested to function as an additional anchor for the 3' end formation machinery at 3' ends containing weak RNA cis-regulatory elements (Millevoi and Vagner 2010). Since CstF-64 binds to U-rich sequences, these could represent extra CstF-64 binding sites. Therefore, it may be the case that containing extra CstF-64 binding sites upstream of the AAUAAA may compensate for 3' ends containing weak poly-A signals. If true, this would predict that removing genes with weak poly-A signals might eliminate the upstream peak. This calculation has not been done, but I am planning on doing it in the future.

Using computational analysis to identify over-represented cis-acting sequences at the 3' ends of worm genes, Graber et al. (2007) revealed a bimodal representation of putative CstF-64 binding sites flanking the poly-A site. It is possible that this upstream

CstF-64 peak may be involved in 3' end formation in combination with the downstream CstF-64 peak.

Interestingly, the position of the CstF-64 peak downstream of the poly-A site does not match the highly variable CstF-64 binding site (U/G rich region), located less than 20 bp downstream of the cleavage site (Graber et al., 2007). The reason for this may be that in my assay CstF-64 is associated with RNAPII, and this peak represents paused RNAPII at +95, on average. Since ChIP is capable of pulling-down CstF-64 associated with either RNAPII or with the RNA, it is possible that for some reason I am only visualizing CstF-64 bound to RNAPII. In contrast, at 3' ends of mammalian genes the double peak of CstF-64 and other 3' end formation factors occurs further downstream of the poly-A site (500 bp–1000 bp). The presence of peaks closer to the site of cleavage in worms is consistent with the more compact nature of the worm genome. It is reasonable to suggest that the CstF-64 peak located downstream of the poly-A site functions in 3' end formation in the majority of non-operon genes (see Figure IV-5).

CstF-64 is enriched at some 5' ends in non-operon genes

As shown in the metagene analysis of non-operon genes (see Figure IV-4), CstF-64 is present at the 5' end of some non-operon genes. This wasn't surprising considering CstF-64 and other 3' end processing factors are also present at promoters in other organisms (Murthy and Manley 1995; Glover-Cutter et al., 2008; Wang et al., 2010). However, due to the compact genome of *C. elegans*, genes are arranged in close proximity to each other making it impossible to accurately assign a CstF-64 peak to an

individual 3’ end. Therefore, a likely explanation for the CstF-64 binding at the 5’ end of genes is that this peak represents CstF-64 binding to a 3’ end of another gene located in close proximity to the 5’ end of the gene analyzed.

In order to find out if CstF-64 binds promoters, I looked at CstF-64 binding in divergent gene arrangements that are located close to each other. This type of gene arrangement in which the two genes are located on opposite strands sharing a single promoter region is ideal for answering this question, since the possibility for a 3’ end between the genes is impossible. I found that some divergent genes contain CstF-64 at the promoter. However, this does not seem to be the case for all genes, since I found some examples in which CstF-64 is not present at the promoter.

As shown in Figure IV-6, CstF-64 is present betweens *c08h9.2* and *f26c11.1*. Importantly, the CstF-64 peak present at the 5’ end of these two genes matches the H3K9ac peak, indicative of the presence of an active promoter. In addition, the CstF-64 peak present at the 5’ end is smaller than the CstF-64 peak at the 3’ end of the c08h9.2 gene, similar to the metagenes analysis shown in Figure IV-4.

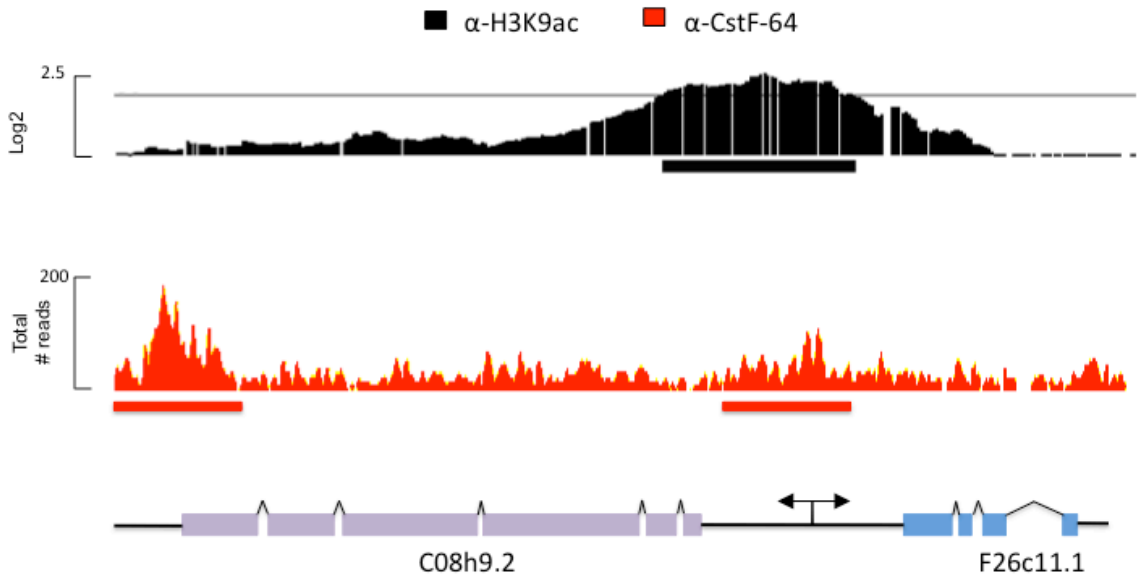


Figure IV-6. CstF-64 is present at the promoter of the divergent pair c08h9.2/f26c11.1: H3K9ac ChIP-on-chip is shown in black, marking the promoter. CstF-64 ChIP-seq enrichment (red) is shown for the same gene. The arrow denotes the location of the promoter. See Figure IV-1 caption for details.

In a second example, CstF-64 is also present between *mei-2* and f57b10.4, which also colocalizes with H3K9ac that marks the location of the promoter. As shown in Figure IV-7, the peak of CstF-64 is significantly smaller at the promoter than at the 3' end of the *mei-2* gene, similar to the previous example. This may be due to the fact that in my ChIP experiments I am only capturing CstF-64 bound to RNAPII, so the peaks of CstF-64 represent RNAPII pause sites (see below and discussion).

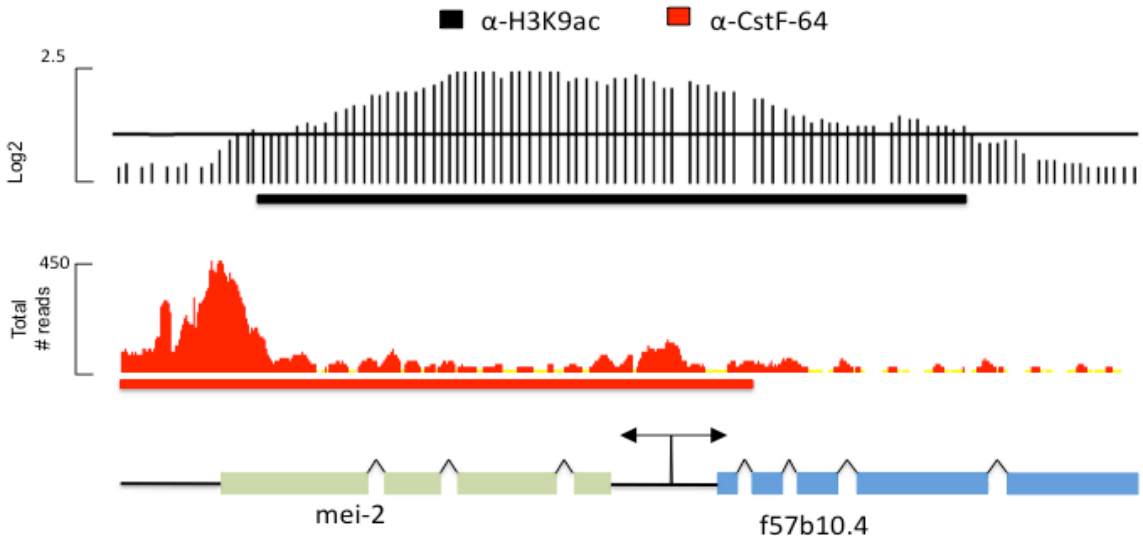


Figure IV-7. CstF-64 is present at the promoter of the divergent pair mei-2/f57b10.4: H3K9ac ChIP-on-chip is shown in black, marking the promoter. CstF-64 ChIP-seq enrichment (red) is shown for the same gene. The arrow denotes the location of the promoter. See Figure IV-1 caption for details.

In a third example, CstF-64 is also present at the 5' end of this divergent gene pair (Figure IV-8). Similar to the previous examples, the CstF-64 peak at the 5' end matches the location of the promoter as indicated by the presence of the H3K9ac mark. In the *lpd-7* gene, CstF-64 starts low at the promoter, then it drops to undetectable levels along the body of the gene and finally it peaks at the 3' end of the gene. In contrast, for the *rpl-9* gene, CstF-64 is low at the promoter and is present throughout the body of the gene, until the 3' end were CstF-64 is significantly enriched. The difference of CstF-64 binding within the body of these genes may be due to differences in expression levels, so that CstF-64 association with RNAPII is revealed only when genes are highly expressed. Indeed, *rpl-9* is expressed more than 7-fold higher than *lpd-7*, consistent with seeing CstF-64 binding in the body of *rpl-9* and not *lpd-7* (Thierry-Mieg D. and Thierry-Mieg J. 2006). Importantly, both peaks at the 3' end of this divergent gene pair are bigger than the peak present at the promoter, consistent with the previous examples and the metagenes analysis (Figure IV-6 – IV-8 and IV-4).

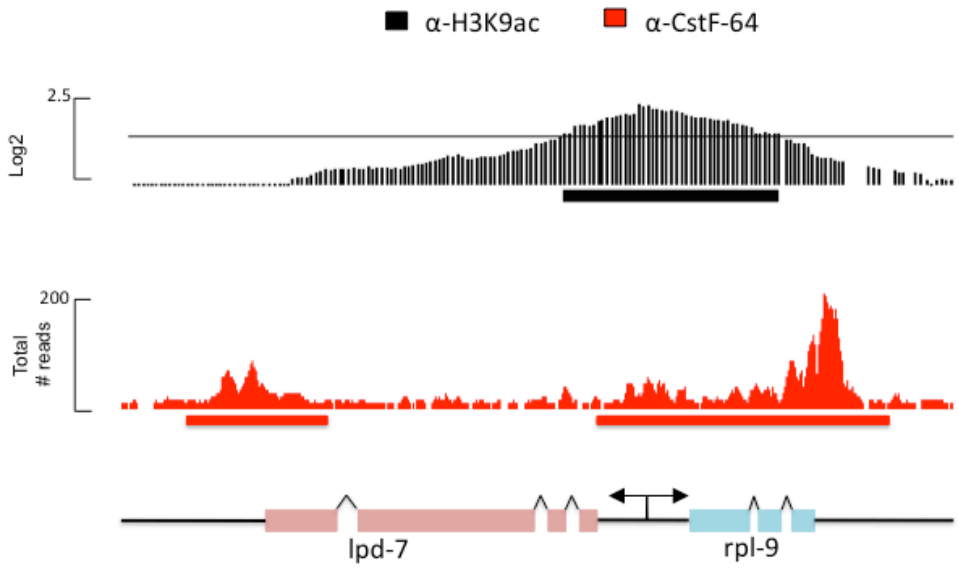


Figure IV-8. CstF-64 is present at the promoter of the divergent pair lpd-7/rpl-9: H3K9ac ChIP-on-chip is shown in black, marking the promoter. CstF-64 ChIP-seq enrichment (red) is shown for the same gene. The arrow denotes the location of the promoter. See Figure IV-1 caption for details.

However, I was also able to find many examples of divergent gene arrangements in which CstF-64 is not present at the promoter (Figure IV-9 – IV-11). Importantly in all cases analyzed CstF-64 was present at the 3' ends of the genes, consistent with CstF-64 recruitment late in the transcription cycle. It is currently unknown, why some genes have CstF-64 at both ends of the gene, while other genes have CstF-64 only at the 3' end.

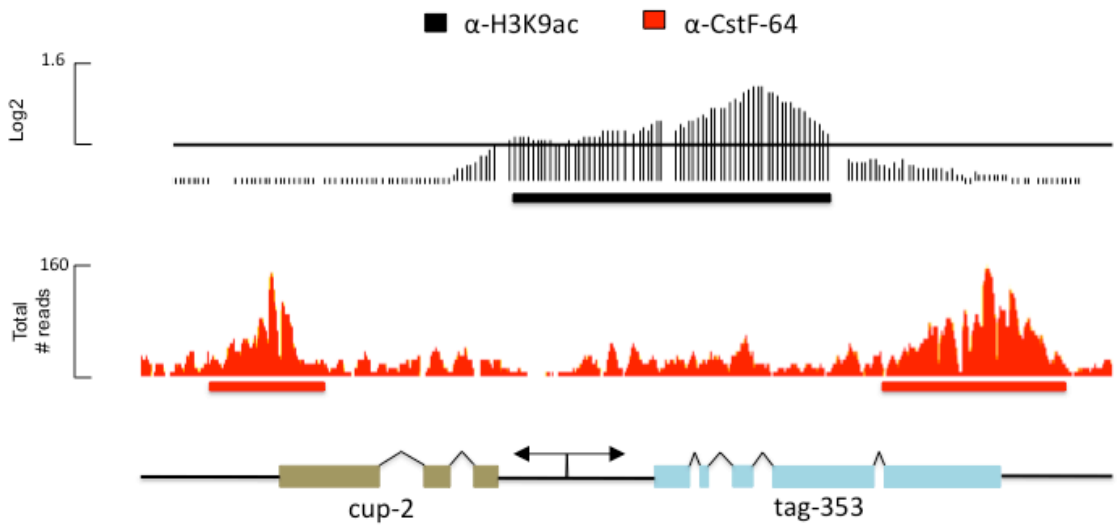


Figure IV-9. CstF-64 is not present at the promoter of the divergent pair cup-1/tag-353: H3K9ac ChIP-on-chip is shown in black, marking the promoter. CstF-64 ChIP-seq enrichment (red) is shown for the same gene. The arrow denotes the location of the promoter. See Figure IV-1 caption for details.

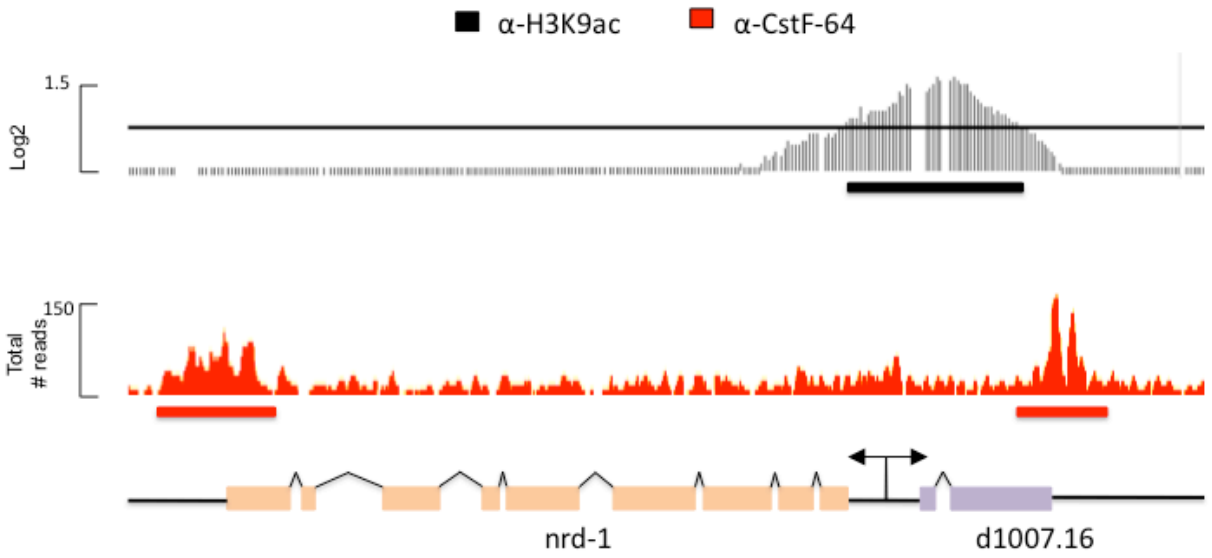


Figure IV-10. CstF-64 is not present at the promoter of the divergent pair nrd-1/d1007.16: H3K9ac ChIP-on-chip is shown in black, marking the promoter. CstF-64 ChIP-seq enrichment (red) is shown for the same gene. The arrow denotes the location of the promoter. See Figure IV-1 caption for details.

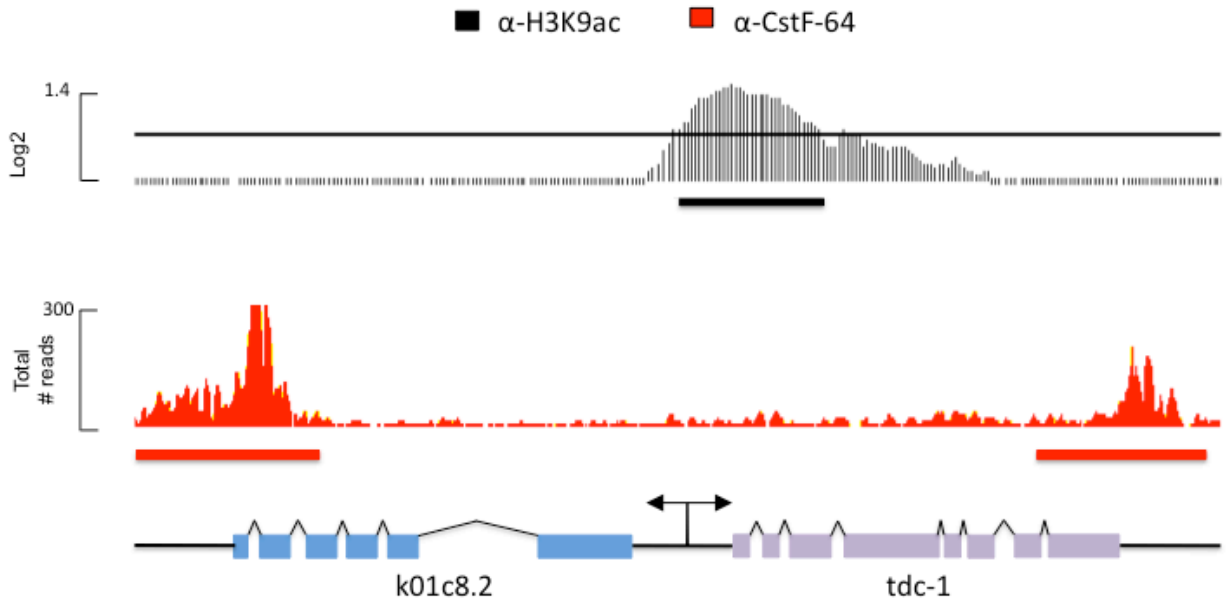


Figure IV-11. CstF-64 is not present at the promoter of the divergent pair k01c8.2/tdc-1: H3K9ac ChIP-on-chip is shown in black, marking the promoter. CstF-64 ChIP-seq enrichment (red) is shown for the same gene. The arrow denotes the location of the promoter. See Figure IV-1 caption for details.

CstF-64 is enriched at 3' ends of genes where 3' end formation occurs but termination does not

3' end formation is needed for correct transcription termination. Several mRNA cleavage and polyadenylation factors required for 3' end formation are also required for transcription termination (Birse et al., 1998; Dye and Proudfoot 1999; Yonaha and Proudfoot 2000; Proudfoot et al., 2002). Further, both 3'end formation and termination depend on the same RNA sequences (Whitelaw and Proudfoot 1986; Connelly and Manley 1988). However, genes arranged in operons contain 3' end formation sites in which termination does not occur. How is RNAPII able to transcribe functional poly-A sites without triggering RNAPII termination?

As shown in Chapter III, RNAPII CTD is phosphorylated on Ser-2 residues at both internal and terminal 3' ends in operons. In addition, Graber et al. (2007) showed that the same canonical 3' end formation sequences were present at both internal and terminal 3' ends. These results suggest that internal 3' ends contain the necessary RNAPII modifications and RNA cis-regulatory sequences needed for efficient 3' end formation. Both of these observations predict that the 3' end formation machinery is present at, and acts at, these internal sites. In order to ask if canonical RNA processing factors are recruited to the 3' ends that don't result in termination, I assayed for the presence of CstF-64 by ChIP-seq at the 3' ends of genes within operons. If the results show that CstF-64 is only associated with terminal sites, it would suggest that transcription termination is prevented at internal 3' ends by inhibiting CstF-64 recruitment. In contrast, if the results show that CstF-64 is associated with all 3' ends in operons, it would suggest that transcription termination might be prevented by a

mechanism other than modulating CstF-64 binding. As shown in Figures IV-12 – IV-14, CstF-64 is indeed enriched at 3' ends in operons to an extent similar to that of non-operon genes (Figure IV-1 – IV-5). An example is shown in Figure IV-12 with the four peaks of CstF-64 corresponding to the four 3' ends in this four-gene cluster. In this operon the single H3K9ac peak at the 5' end of the cluster (CEOP 3244) confirms that transcription begins only upstream of the first gene, indicative of co-expression of genes. Interestingly, the position of the CstF-64 peak with respect to the poly-A site seemed to differ between internal and terminal 3' ends. At internal sites the CstF-64 peak is located downstream of the poly-A site, but at this terminal 3' end the CstF-64 peak is more spread around the poly-A site (Figure IV-12). According to the UTRome, each gene in this operon contains a single poly-A site (asterisks). Thus the spread of CstF-64 at the terminal 3' end (*c26e6.3*) is likely not due to multiple alternative poly-A sites.

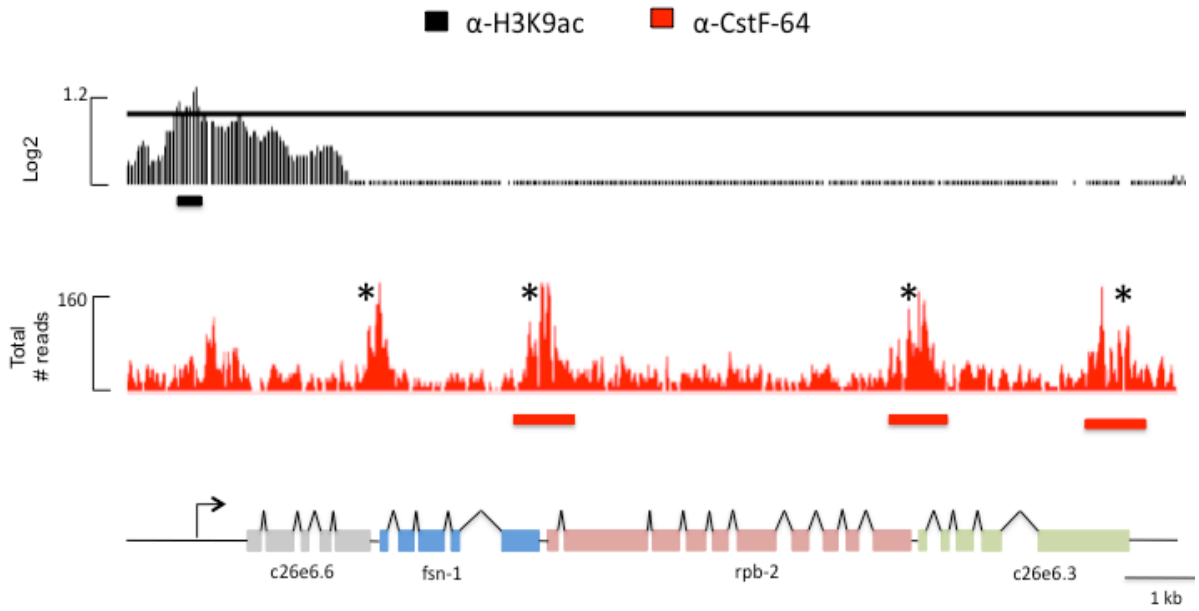


Figure IV-12. CstF-64 is enriched at all 3' ends in a four-gene operon (CEOP3244). H3K9ac ChIP-on-chip is shown in black, marking the position of the promoter and CstF-64 ChIP-seq is shown in red. See Figure IV-1 caption for details.

In another operon example, again CstF-64 is associated with each 3' end of a two-gene operon (CEOP 4649) (Figure IV-13). A single H3K9ac peak located at the 5' end of the cluster is indicative of co-expression of these two genes. The CstF-64 peak at the internal 3' end of *y73b6bl.33* is much more pronounced than the terminal 3' end of *smg-3*, similar to the previous example (Figure IV-12). The reason for this is likely due to the gradual decrease of expression seen across operons based on ESTs and publicly available transcriptome data. Terminal genes tend to be expressed at a lower level than internal genes in operons, which is consistent with the internal *y73b6bl.33* gene having more CstF-64 at the 3' end (Figure IV-13). Also, at the internal 3' end (*y73b6bl.33*) the peak of CstF-64 is located a short distance downstream of the annotated poly-A site (asterisk). In contrast, at the terminal 3' end (*smg-3*) the peak of CstF-64 is located closer to the poly-A site, but without the CstF-64 signal spreading around the poly-A site seen on the previous example (compare terminal genes between Figure IV-12 and Figure IV-13).

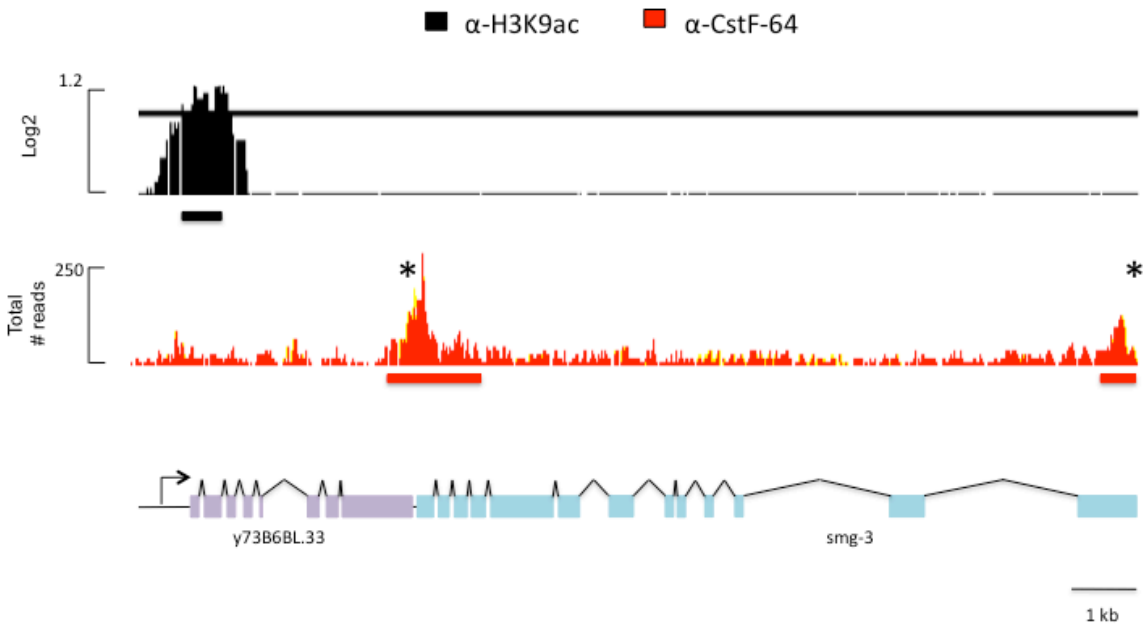


Figure IV-13. CstF-64 is enriched at all 3' ends in a two-gene operon (CEOP 4649). H3K9ac ChIP-on-chip is shown in black marking the position of the promoter and CstF-64 ChIP-seq is shown in red. See Figure IV-1 caption for details.

A third example is a two-gene operon (CEOP 1576), again with a single H3K9ac peak (Figure IV-14). Similar to the previous examples (Figures IV-12 and IV-13), CstF-64 is present at internal and terminal 3' ends in this two-gene operon. In this example, the internal 3' end (*eif-3.c*) again contains higher levels of CstF-64 compared to the lower level and more spread out signal present at the 3' end of the terminal gene in the operon (*t23d8.3*).

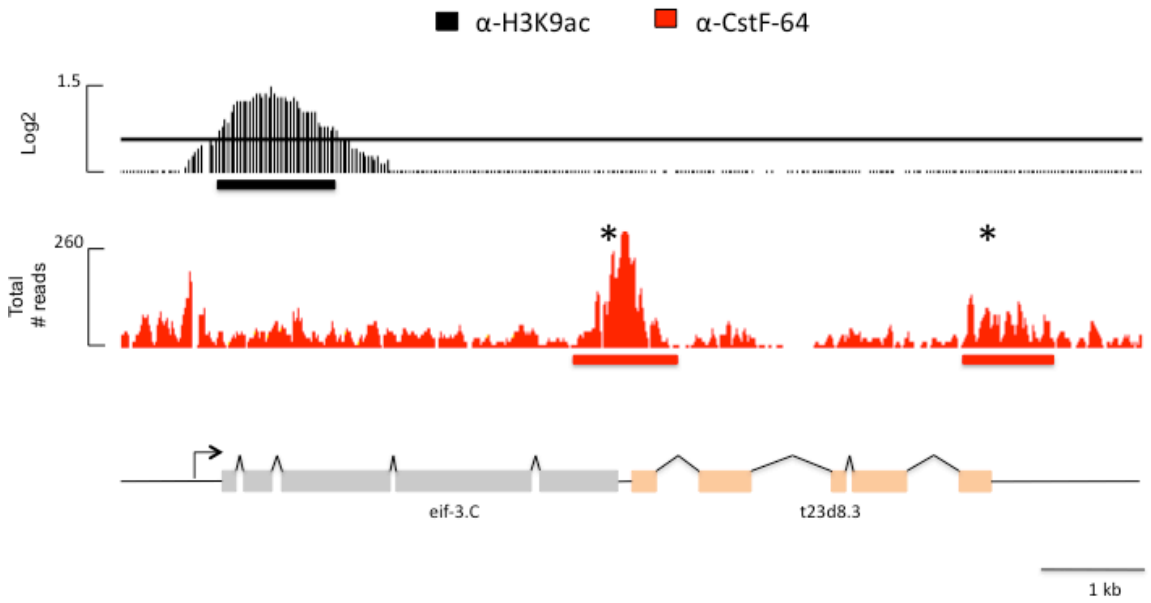


Figure IV-14. CstF-64 is enriched at all 3' ends in a two gene operon (CEOP 1576). H3K9ac ChIP-on-chip is shown in black marking the position of the promoter and CstF-64 ChIP-seq is shown in red. See Figure IV-1 caption for details.

The three representative operon examples shown above (Figures IV-12 – IV-14) reveal that CstF-64 is associated with all the 3'ends in these operons, demonstrating that internal and terminal 3' ends are both bound by CstF-64. This result indicates that

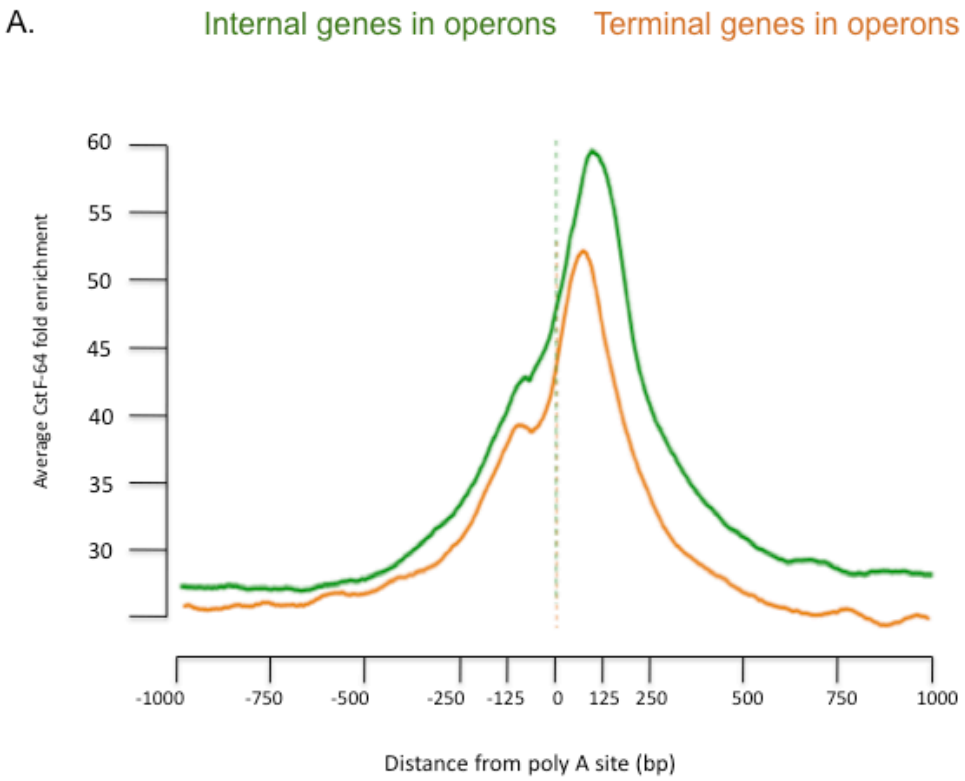
CstF-64 associates with 3'ends where termination is prevented, as well as sites where termination occurs, showing that the presence of CstF-64 is not sufficient for termination. However, in all three examples the CstF-64 binding pattern present at internal 3' ends differs from the pattern present at terminal 3' ends (Figures IV-12 – IV-14). At internal 3' ends the levels of CstF-64 tend to be high, with the summit of the CstF-64 peak located slightly downstream of the poly-A site. In contrast, at terminal 3' ends the levels of CstF-64 are relatively low and are spread around the poly-A site. Interestingly, the pattern of CstF-64 binding in non-operon genes resembles the binding pattern shown for terminal 3' ends (see below and Figures IV-1 – IV-3). Therefore, these different patterns could be relevant to how 3' end formation occurring near the site of termination is different from 3'end formation at sites lacking termination. However, these differences are based on a limited set of genes.

Genome-wide analysis of CstF-64 binding in operons

In order to determine if the binding pattern of CstF-64 in the above examples applies to all operons in the genome, I aligned the ChIP-seq results of all internal (*n*=1880) and terminal 3' ends *(n*=1220) by their poly-A site. All of these genes in operons containing significant levels of CstF-64 were averaged within a 2000 bp window centered at the poly-A site, and plotted on the same graph (Figure IV-15A). I found that the levels of CstF-64 at internal 3' ends are indeed higher than the levels present at terminal 3' ends, consistent with the above examples (Figure IV-12 – IV-15A).

In addition, 3' ends of terminal genes in operons contain a small CstF-64 peak upstream of the poly-A site, similar to the levels of CstF-64 present at internal genes in

operons (Figure IV-15A). The function of this small CstF-64 peak is unknown, but it seems to play a general role at all 3' ends in the genome, since it is also present at non-operon gene 3' ends. Intriguingly, the levels of CstF-64 upstream of the poly-A site in non-operon gene 3' ends are higher than the levels seen for operon gene 3' ends, both internal and terminal (Figure IV-15A and B). Operon genes (internal and terminal) are more similar to each other regarding CstF-64 levels upstream of the cleavage site, while non-operon genes are quite different from terminal operon genes (Figures IV-15A and B). These results are surprising since 3' end formation at 3' ends of terminal genes in operons and non-operon genes are both accompanied by transcription termination.



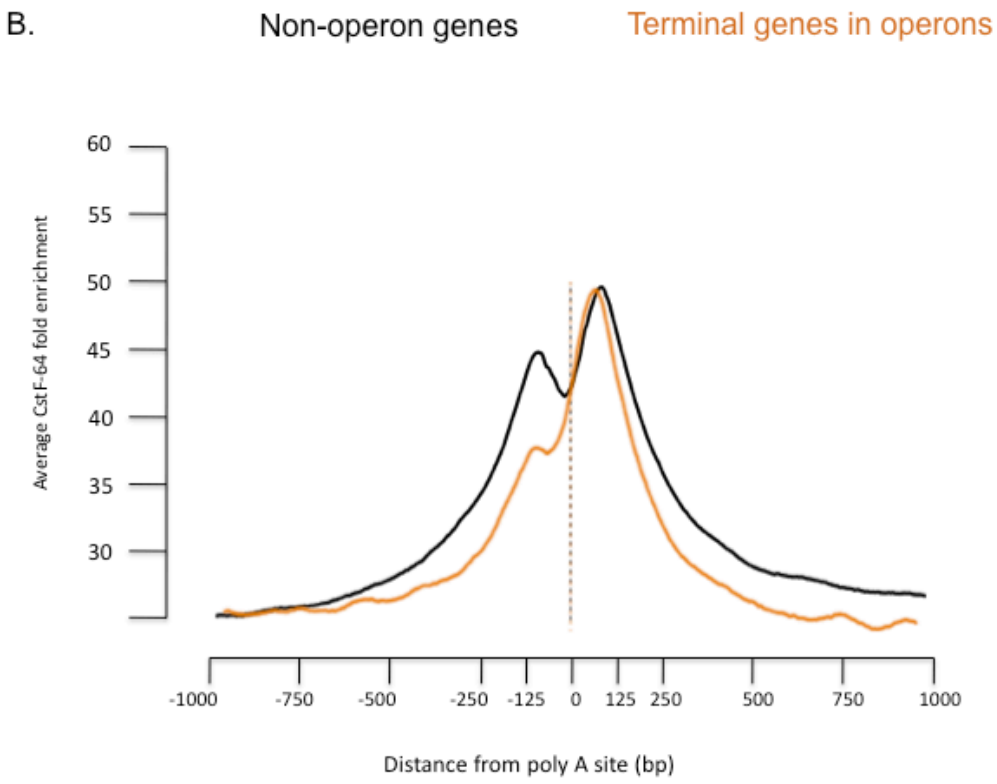


Figure IV-15. A. CstF-64 enrichment at 3' ends is similar for internal and terminal genes in operons: All internal (n=1880) and all terminal genes in operons (n=1220) containing significant levels of CstF-64 (ChIP-seq) are aligned by their poly A site (3' cleavage). The X-axis indicates distance from the poly A site which is set at zero. The y-axis shows the fold enrichment for each nucleotide over background. The dashed line indicates the position of the poly A site. B. CstF-64 present upstream of the poly-A site is similar at 3' ends of non-operon genes and terminal genes in operons. All non-operon genes (n=17063) and all terminal genes in operons (n=1220) containing significant levels of CstF-64 (ChIP-seq) are aligned by their poly A site (3' cleavage).

Interestingly, the level of CstF-64 present downstream of the cleavage site is similar between genes in which termination occurs following 3' end formation (non-operon and terminal genes). As shown in Figure IV-15B, the location and height of the CstF-64 peak downstream of the cleavage site in non-operon genes and terminal genes in operons occurs in the same location and are about the same height. However, in

internal genes in operons this peak is further downstream and the height of the peak is considerably higher than terminal and non-operon genes (Figure IV-15A and B).

Currently it is unknown if these differences in CstF-64 binding are associated with transcription termination. Since the CstF-64 peak present downstream of the poly-A site is clearly present in all three types of 3' ends, it is unlikely the downstream peak is related to transcription termination. However, the CstF-64 peak present upstream of the poly-A site varies significantly between 3' ends, being much diminished at internal 3' ends in operons (Figure IV-15A and B). This suggests the upstream CstF-64 peak could have a role in termination, since this peak is significantly reduced at internal 3' ends where termination does not occur. The small peak of CstF-64 present upstream of the poly-A site at internal 3' ends (Figure IV-15A) may be due to internal termination sites between genes. It is not known whether in some cases termination may occur at internal 3' ends.

Another issue is why terminal genes in operons do not look like non-operon genes with respect to CstF-64 upstream of the poly-A site. Since termination is thought to occur following 3' end formation with both types of 3' ends, they might have been expected to show similar patterns. The answer to this is unknown, but may be because non-operon genes need extra CstF-64 binding for processing their 3' ends.

In summary, CstF-64 is associated with all 3' ends in operons indicating that its recruitment to internal 3' ends is insufficient for transcription termination (Figures IV-12 – IV-14). Furthermore, as shown in Figure IV-15A, terminal operon 3' ends contain a double peak flanking the poly-A site, with the major peak present 70 bp – 100 bp downstream of the poly-A site, in the vicinity of the predicted CstF-64 binding site

(Graber et al., 2007). The CstF-64 binding downstream of the poly-A site may be an indication of its role in 3' end formation. However, whether both CstF-64 peaks are part of a single functional unit involved in 3' end formation or if they represent different functions performed by CstF-64 at 3' ends of genes is unclear. Interestingly, at 3' ends in which termination is prevented, the levels of CstF-64 present upstream of the poly-A site are low, in contrast to the higher levels present at 3' end sites in which termination occurs (Figures IV-4 and IV-15A and B).

CstF-64 colocalizes with Ser2p and paused RNAPII at 3' ends of genes

In mammals, RNAPII competent for termination is in a paused conformation, colocalizing with maximal Ser-2 phosphorylation and bound by high levels of 3' end formation factors (Glover-Cutter et al., 2008). In a similar situation, internal 3' ends in operons are both maximally phosphorylated at Ser-2 (Chapter III) and contain high levels of CstF-64 (Figures IV-12 – IV-15). However do Ser-2p and CstF-64 binding colocalize at 3' ends where termination is prevented? Furthermore, do these events colocalize with RNAPII pausing? This could answer whether RNAPII pausing at 3' ends is a property of 3' end formation or transcription termination.

I chose three operon examples for comparing the previous ChIP signal of Ser-2p (Chapter III) and CstF-64 (Chapter IV) to test for colocalization (Figures IV-16 – IV-18). Previous ChIP-qPCR experiments using a Ser-2p antibody (blue) were aligned with the ChIP-qPCR and ChIP-seq data for CstF-64 (red). For the ChIP-qPCR, each bar represents an individual primer used to quantify the IP DNA as a percent of the maximal value. Each bar is positioned above its corresponding genomic location (Figure IV-16 –

IV-18). The prediction is that Ser-2p and CstF-64 might coincide since the CstF-64 partner, CstF-50 (cpf-1 in worms), has been shown to bind the phosphorylated CTD of RNAPII (Fong and Bentley 2001). Indeed, Ser-2p at both internal and terminal 3' ends of operons colocalizes with maximal recruitment of CstF-64 (Figure IV-16 – IV-18).

In a four-gene operon (CEOP 3184), CstF-64 binding colocalizes with Ser-2p at all 3' ends in the operon consistent with the recruitment of the CstF complex by Ser-2p. At the first 3' end of the operon (*snfc-5*) CstF-64 and Ser-2p are both significantly increased. Due to the lack of primers around the 3' end of the second gene in the operon (*rnp-4*), no conclusion can be made for this site other than that CstF-64 is bound (Figure IV-16). However, the 3' end of the third gene in the operon (*prdx-3*) lacks a CstF-64 peak, which is consistent with the absence of Ser-2p at this same 3' end and further supports the colocalization of these two events. Could it be that some internal 3' ends in operons may not be processed by the same factors? In addition, the terminal 3' end (*r07e5.1*) in this operon also shows a clear colocalization of Ser-2p and CstF-64, similar to internal 3' ends (Figure IV-16).

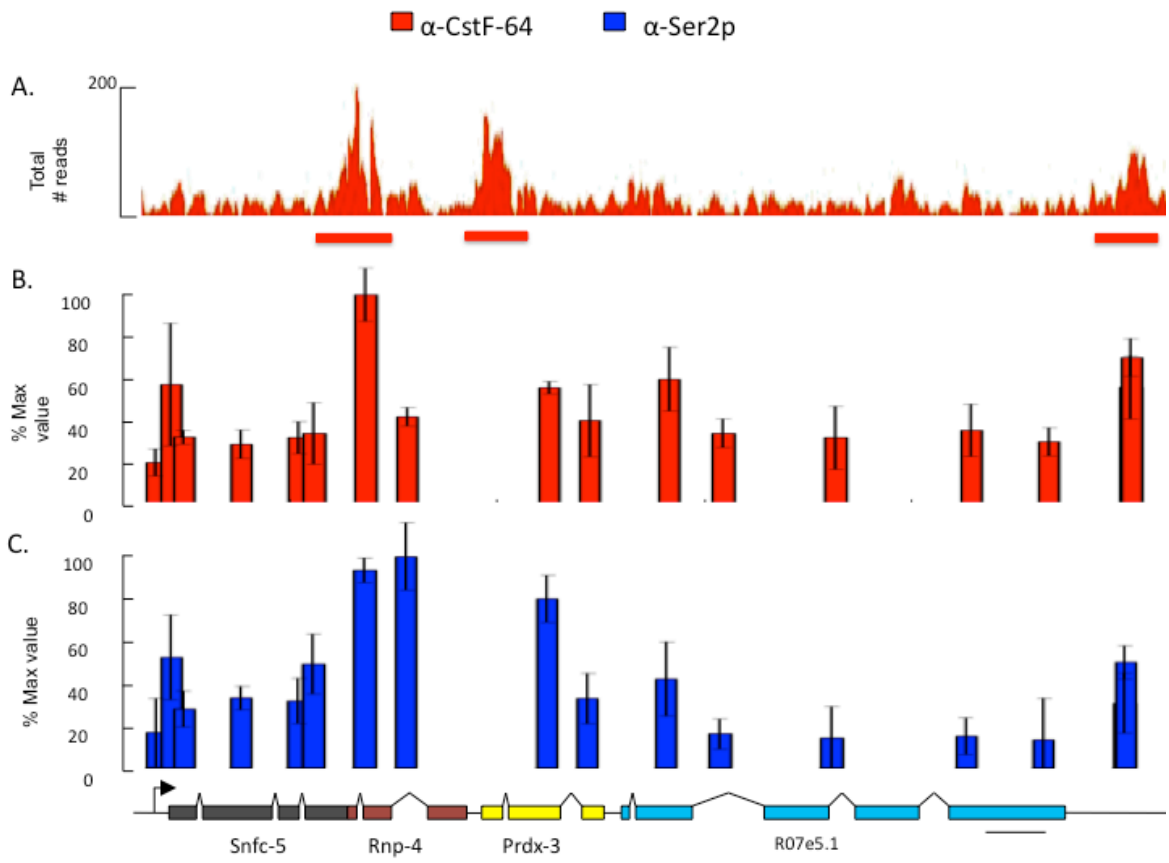


Figure IV-16. CstF-64 peaks match regions with high Ser2p in a four gene operon (CEOP 3184). A. CstF-64 ChIP-seq experiment. The height of the peak is proportional to the number of reads and the horizontal red line under the graph represents statistically significant binding (p-value of 1e-5) detected by the MACS algorithm. B. CstF-64 ChIP-qPCR experiment. Each bar represents an individual PCR amplicon used to quantified the immunoprecipitated DNA. The results from each primer set are positioned immediately above the corresponding genomic location and normalized to the highest value. Error bars represent percent error of three PCR reactions from a single immunoprecipitated experiment. C. Ser2p ChIP-qPCR experiment. Primer sets used for quantification and error bars are the same as in B.

In a second operon example (Figure IV-17), Ser-2p and CstF-64 also colocalize at internal and terminal 3' ends in this three-gene operon (CEOP 3412), consistent with the example shown in Figure IV-16. In this case, no significant Ser-2p or CstF-64 peak is detected at the 3' end of the first gene in the operon (*k04g7.11*), similar to the *prdx-3*

gene 3' end shown in the previous example (Figure IV-16). In the second internal 3' end (*rnp-7*), qPCR results suggest that Ser-2p and CstF-64 peak colocalize at this internal location (Figure IV-17B and C). However, the presence of a CstF-64 peak at this location is not significant in the ChIP-seq experiment (Figure IV-17A). The cause for this is unknown but could be due to low expression levels associated with this operon. Finally, at the 3' end of the operon both Ser-2p and CstF-64 colocalize similar to the internal 3' end and the previous example (Figure IV-16).

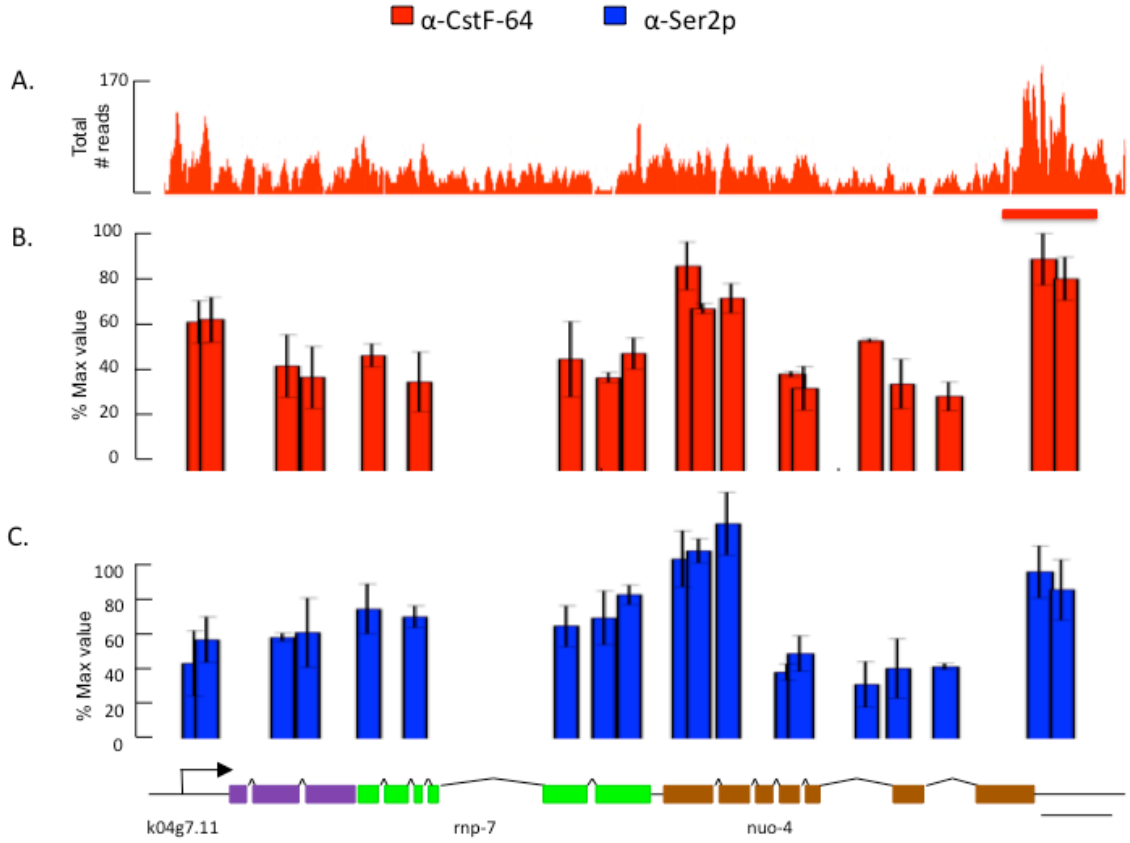


Figure IV-17. CstF-64 peaks match regions with high Ser2p in a three gene operon (CEOP 3412). A. CstF-64 ChIP-seq experiment. B. CstF-64 ChIP-qPCR experiment. C. Ser2p ChIP-qPCR experiment. See Figure IV-10 caption for details.

In another example (Figure IV-18), a three-gene operon also shows colocalization of Ser-2p with CstF-64 at an internal 3' end. In this case, the first gene in the operon (*f21d5.7*) contains high levels of CstF-64 at the 3' end, which coincides with elevated levels of Ser-2p (Figure IV-18). At the 3' end of the second gene in the operon no Ser2p or CstF-64 binding is present consistent with the idea that Ser-2p may be needed for CstF-64 recruitment. The terminal gene in this operon (*f21d5.8*) lacks the primer sets needed to show a colocalization of Ser2p with CstF-64. Indeed, the ChIP-seq experiment suggests the location of the 3' end peak is located downstream of the position of the PCR primers sets used (Figure IV-18). Interestingly, CstF-64 is present at the 5' end of this cluster (Figure IV-18A and B). This may be due to an adjacent 3' end located around 500 bp upstream of the promoter. At this location high Ser-2p levels coincide with elevated CstF-64 binding, suggesting the intriguing possibility that at 3' ends of genes, CstF-64 is associated with RNAPII phosphorylated at Ser-2.

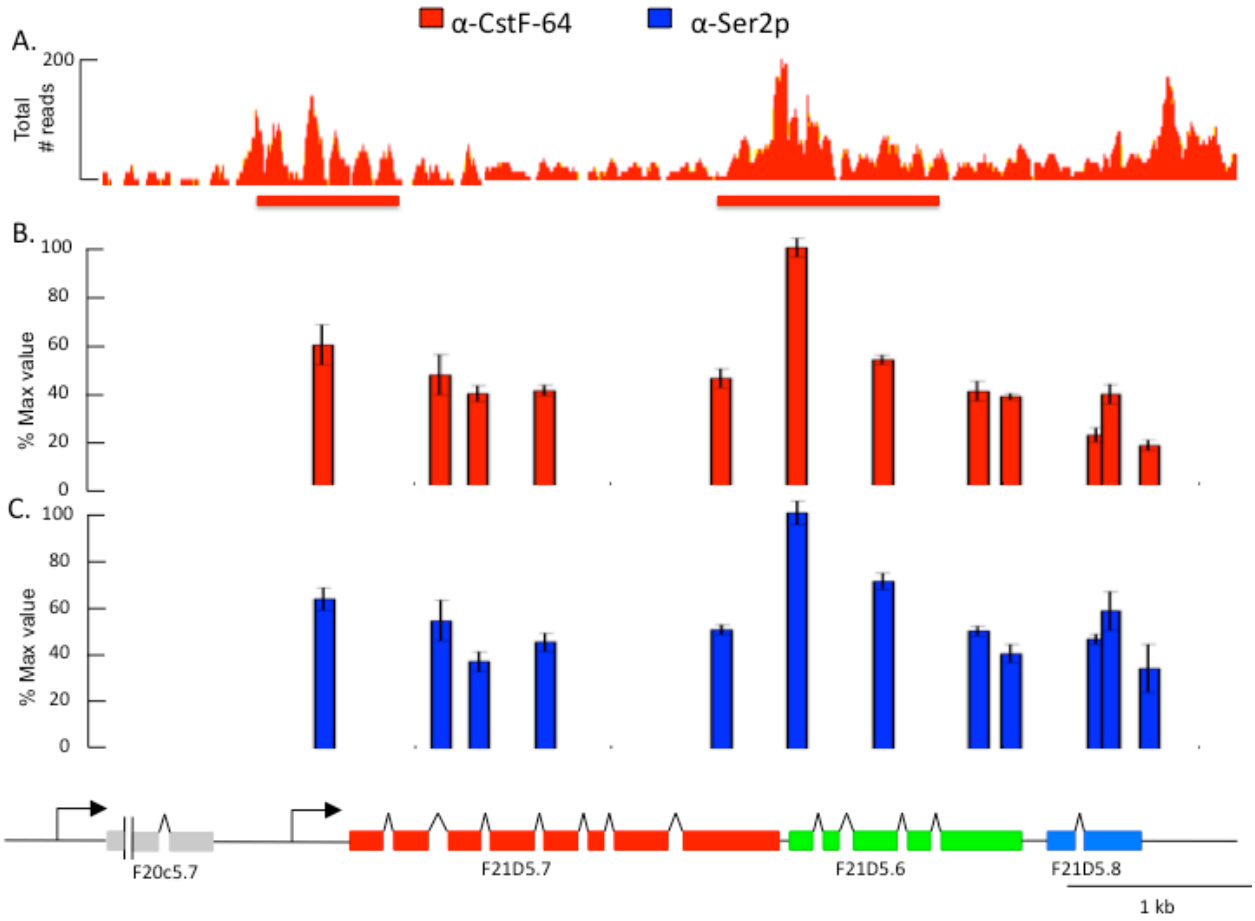


Figure IV-18. CstF-64 peaks match regions with high Ser2p in a three gene operon (CEOP 4304). A. CstF-64 ChIP-seq experiment. B. CstF-64 ChIP-qPCR experiment. C. Ser2p ChIP-qPCR experiment. See Figure IV-10 caption for details.

Based on the three examples shown above (Figure IV-16 – IV-18), Ser-2p colocalizes with CstF-64 at internal and terminal 3' ends in operons, indicating that Ser-2p may indeed be needed for CstF-64 recruitment to the transcription site. This is similar to mammals, in which RNAPII marked by Ser-2p coincides with maximal CstF-64 recruitment at 3' ends of genes. Importantly, termination-competent RNAPII complexes are paused at 3' ends of genes, colocalizing with elevated Ser-2p and CstF-64 (Glover-Cutter et al., 2008).

The internal peaks of CstF-64 and Ser-2p are RNAPII pause sites

Next, I wanted to know if RNAPII pauses at 3' ends of internal genes in operons. RNAPII pausing is associated with enhancing termination by slowing down the elongating RNAPII, so the exonuclease is capable of reaching RNAPII and causing it to fall off the DNA. Therefore, I thought that RNAPII might not pause at internal 3' ends in order to prevent the exonuclease from reaching the elongating RNAPII, thus allowing expression of the downstream genes. Here I compare unpublished GRO-seq data from Will Kruesi at the Meyer Lab (University of California at Berkeley) which measures engaged RNAPII at operon genes, with my data on CstF-64 localization (Figure IV-19 and IV-20). Global run-on sequencing (GRO-seq) is a technique used to measure the position, amount and orientation of engaged RNAPII at a genome-wide scale. In all cases analyzed, RNAPII was found to pause at internal and terminal 3' ends in operons, providing evidence that RNAPII pausing is not sufficient for termination. Moreover, the maximal CstF-64 recruitment present at each 3' end in the operons coincides with RNAPII pausing at these locations (Figure IV-19 and IV-20).

As shown in Figure IV-19, there is a peak of RNAPII at each 3' end of the three-gene operon (CEOP 3232), presumably a reflection of RNAPII pausing. This suggests that pausing is not sufficient for termination, since no termination occurs at the internal sites. The location of the only promoter is marked with H3K9ac at the 5' end of the cluster, providing evidence for co-expression. Importantly, at each 3' end in the operon the presumed pause site colocalizes with CstF-64 maximal recruitment, indicating that 3' end formation occurs in the context of RNAPII pausing. In addition, the gradual decrease in CstF-64 levels seen across this operon 3' ends (Figure IV-19) and other examples shown above (Figures IV-12 – IV-14) matches the amount of RNAPII, suggesting that CstF-64 may actually be bound to RNAPII at 3' ends in operons.

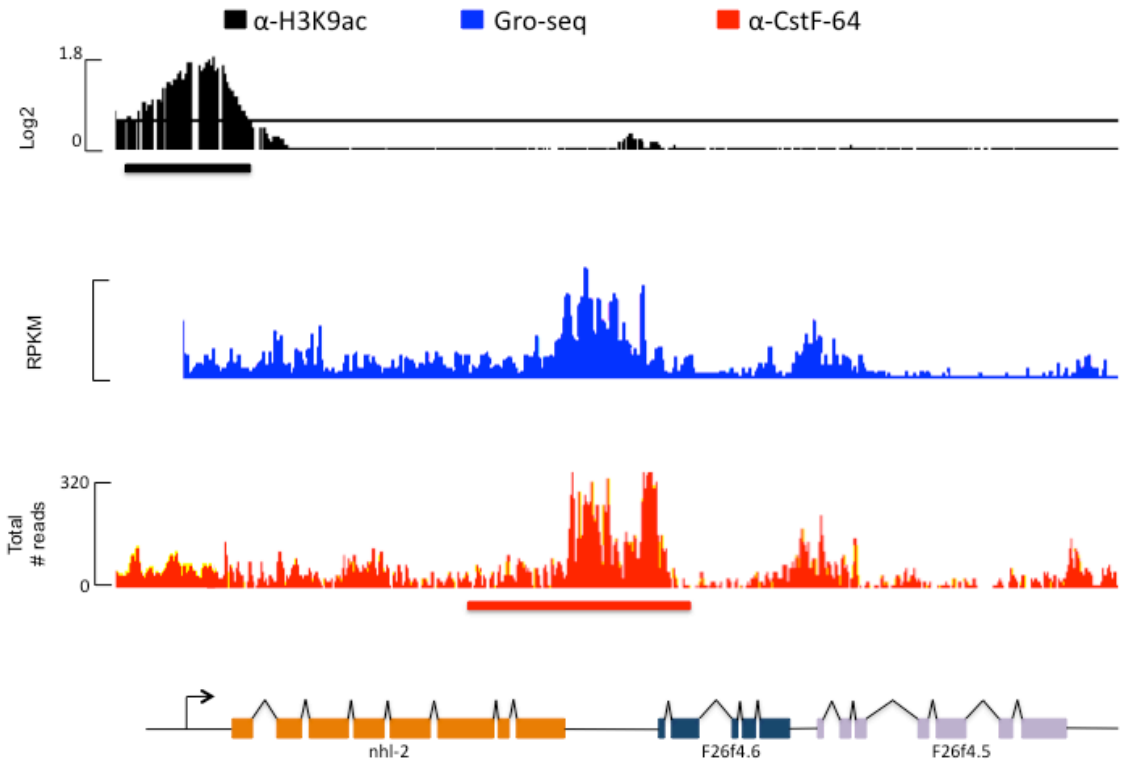


Figure IV-19. Comparison of GRO-seq with CstF-64 ChIP-seq in a three gene operon (CEOP 3232). H3K9ac ChIP-on-chip is shown in black marking the promoter, GRO-seq is in blue and CstF-64 ChIP-seq in red. The GRO-seq experiment was done by Will Kruesi at the University of California at Berkley (Meyer Lab). The height of the peak represents the level of engaged RNAPII (RPKM). In the H3K9ac ChIP-on-chip each vertical line represents an individual probe and the height of the bar is proportional to the amount of total hybridization. The horizontal black lines under H3K9ac graph represent region with a p-value of 0.05 and. In the CstF-64 ChIP-seq graph the height of the peak is proportional to the number of aligned reads matching the region. The horizontal red line under the graph represents statistically significant binding (p-value of 1e-5) detected by the MACS algorithm.

In the second operon example (Figure IV-20), RNAPII pauses at both internal and terminal 3' ends in this two-gene operon (CEOP 3384), similar to the example shown above (Figure IV-19). As seen in Figure IV-20, there are H3K9ac peaks at each end of the cluster. The one at the 5' end is at a promoter presumably controlling the co-expression of these two genes, while the 3' end H3K9ac peak is at a promoter driving transcription of a downstream gene (Figure IV-20). The RNAPII pause site between the two operon genes is located a small distance downstream of the maximal CstF-64 peak. In contrast, the terminal pause site is colocalized with the CstF-64 peak. There is a small peak of RNAPII at the promoter of this operon that does not appear to show a CstF-64 peak, suggesting that in some genes CstF-64 may only associate with RNAPII at 3' end pause sites.

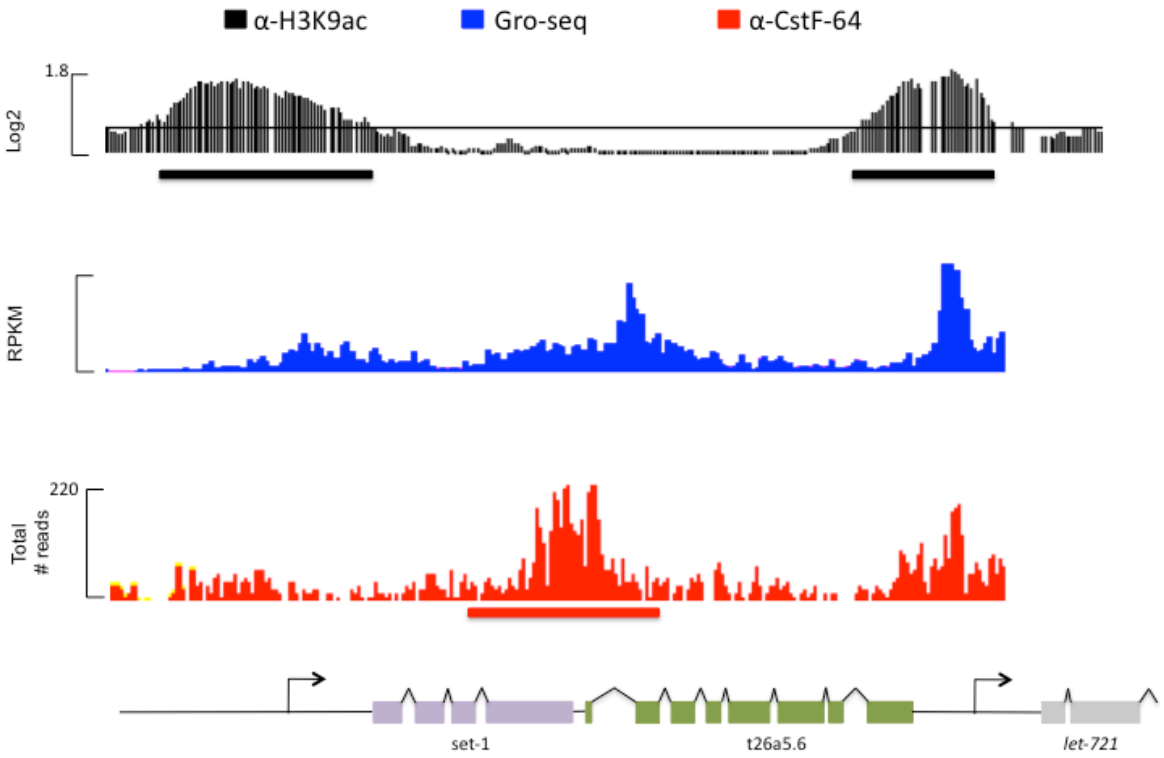


Figure IV-20. Comparison of GRO-seq with CstF-64 ChIP-seq in a two-gene operon (CEOP 3384). H3K9ac ChIP-on-chip is shown in black marking the promoter, GRO-seq is in blue and CstF-64 ChIP-seq in red. See Figure IV-13 caption for details.

RNAPII was also found to pause at 3' ends of non-operon genes and to colocalize with maximal CstF-64 binding there, in a similar manner to both internal and terminal 3' ends in operons (Figure IV-21 – IV-22). In the *ama-1* gene, engaged RNAPII accumulates at the 3' end of the gene (Figure IV-21). The small level present at the 5' end in this example may be due to the close proximity of an upstream gene 3' end. Similar to the ama-1 gene, RNAPII pauses at the 3' end of the cul-1 gene, colocalizing with CstF-64 binding (Figure IV-22), as in both 3' ends found in operons (Figure IV-19 – IV-20).

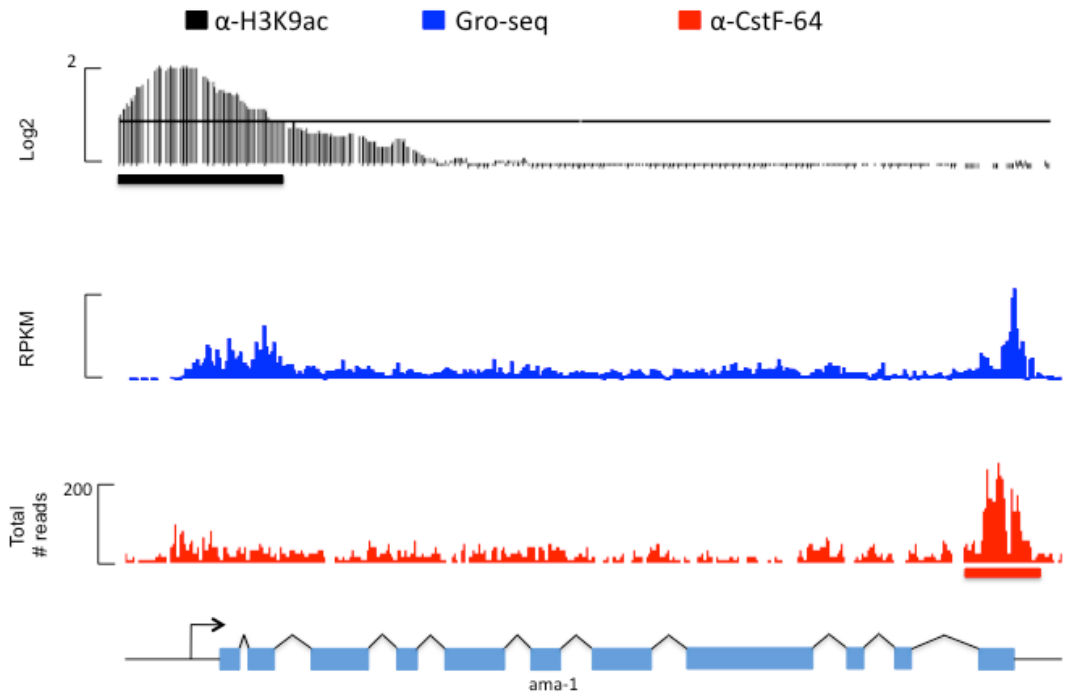


Figure IV-21. Comparison of GRO-seq with CstF-64 ChIP-seq in ama-1 non-operon gene. H3K9ac ChIP-on-chip is shown in black marking the promoter, GRO-seq is in blue and CstF-64 ChIP-seq in red. See Figure IV-13 caption for details.

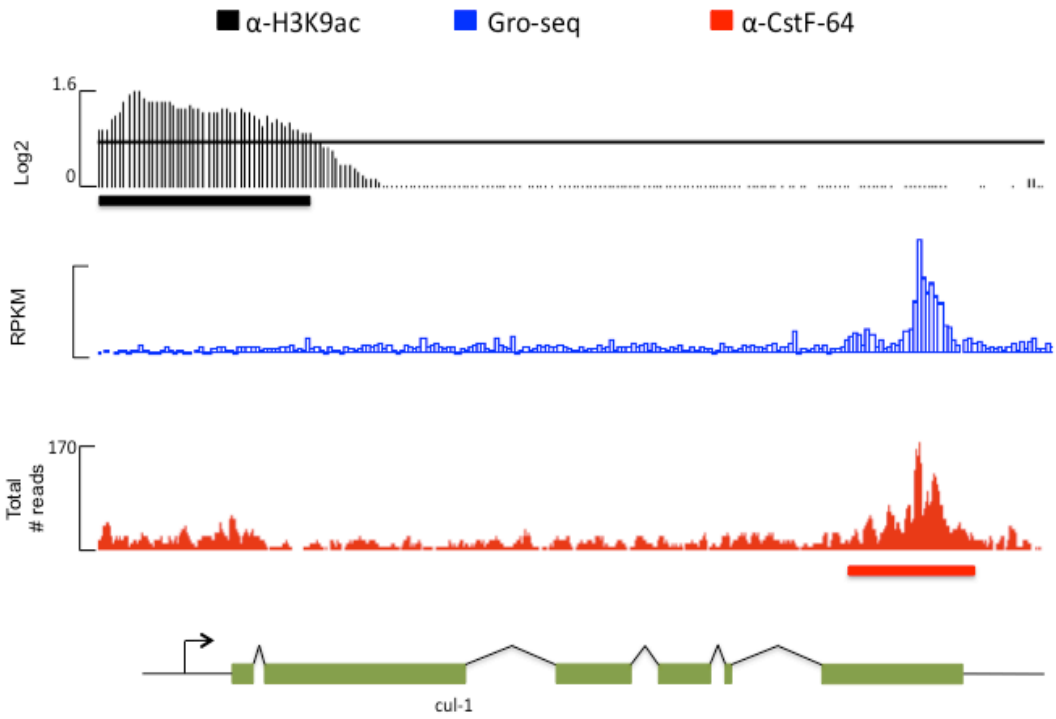


Figure IV-22. Comparison of GRO-seq with CstF-64 ChIP-seq in cul-1 non-operon gene. H3K9ac ChIP-on-chip is shown in black marking the promoter, GRO-seq is in blue and CstF-64 ChIP-seq in red. See Figure IV-13 caption for details.

Finally, I wanted to compare the levels of engaged RNAPII with CstF-64 binding along all genes in the genome (Figure IV-23), in order to provide genome-wide evidence for their colocalization. These metagene profiles were created by taking all genes in the genome greater than 2 kb (n=8400) and scaling the region starting at +500 from the promoter to -500 from the 3' end. Only for the GRO-seq results the amount of engaged RNAPII (RPKM) along metagenes was subdivided into expression levels, which are denoted by the different colored lines (Figure IV-17). Consistent with the examples above (Figure IV-19 – IV-22), CstF-64 colocalizes with paused RNAPII at both ends of genes with a significant bias towards 3' ends (Figure IV-23). Furthermore, the CstF-64 double peak present at all types of 3' ends (Figure IV-4, IV-5, and IV-15) is also shown

by paused RNAPII found at 3' ends of genes (Figure IV-23). This suggests that CstF-64 is bound to the RNAPII paused complex at both sides of the poly-A site.

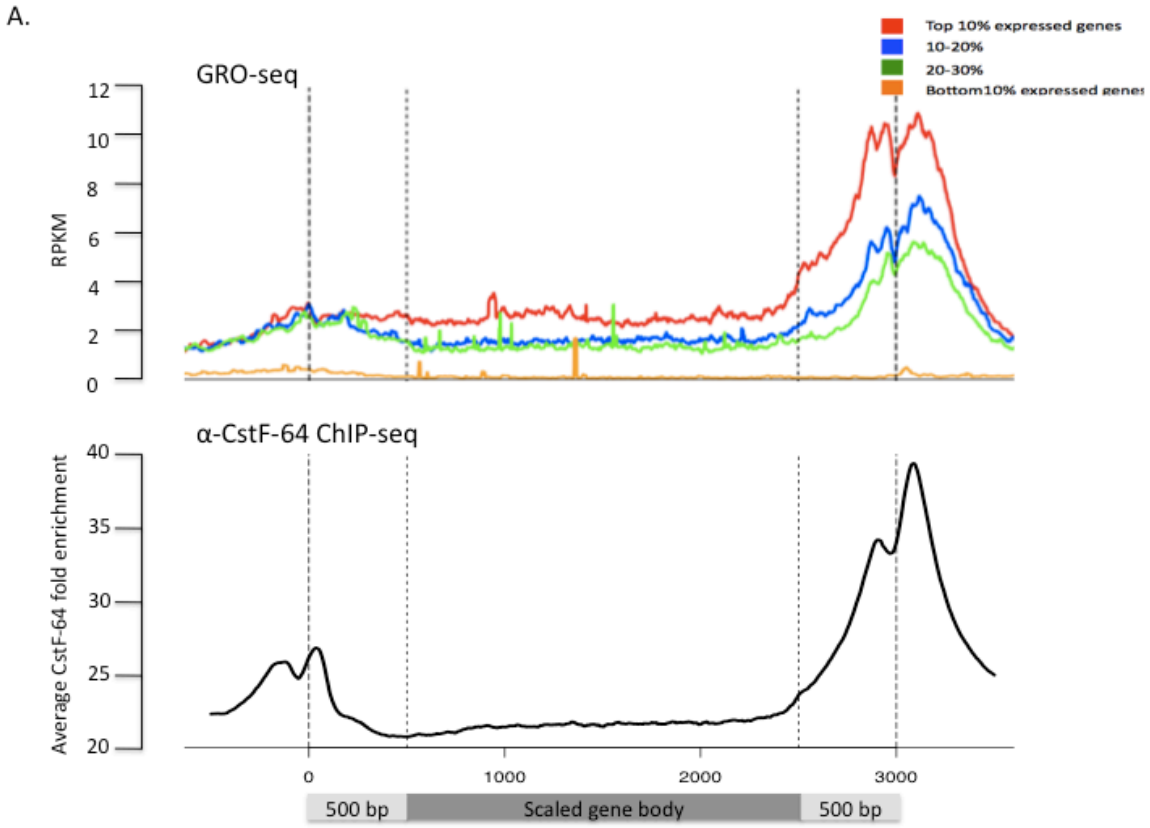


Figure IV-23. The Cstf-64 peak at 3' ends of genes is correlated with engaged RNAPII: A. GRO-seq (top) and CstF-64 ChIP-seq (bottom) experiments are plotted along the same metagenes shown at the bottom of the slide, with the 0 and 3000 mark representing the transcription start site and poly A site respectively. Metagenes are made by taking all genes in the genome greater than 2 kb and scaling the region starting at +500 from the promoter to -500 from the 3' end (dark gray box). The y-axis indicates the levels of engaged RNAPII (RPKM) and the different line colors denote genes with varying expression levels. For the CstF-64 ChIP-seq experiment the y-axis shows the fold enrichment for each nucleotide over background.

In summary, in this section I have shown that elevated Ser-2p at 3' ends of internal and terminal operon genes colocalizes with maximal CstF-64 binding indicative of CstF-64 recruitment by Ser-2p. Moreover, RNAPII pauses at all 3' end in operons

eliminating the possibility that pausing is prevented at internal 3' ends, in order to prevent termination. I conclude that in both internal and terminal 3' ends, co-transcriptional 3' end formation occurs in the context of a paused RNAPII that is phosphorylated at Ser-2 and contains maximal levels of CstF-64. Therefore, at internal sites it seems that RNAPII contains all the necessary modification and protein factors needed for termination, suggesting that a yet unidentified mechanism is necessary to prevent termination.

Discussion

The protein encoded by *cpf-2* is CstF-64

The *C. elegans* genome contains a clear CstF-64 ortholog, CPF-2, identified solely based on amino acid sequence similarity. However, the worm CstF-64 encoded by *cpf-2* is significantly smaller than CstF-64 found in other organisms (Figure I-2) and has not been shown to function in 3' end formation. ChIP experiments in mammals indicate that CstF-64 is specifically recruited to 3' ends of genes coinciding with the site of 3' end formation (Swinburne et al., 2006; Glover-Cutter et al., 2008). In mammals CstF-64 is maximally recruited 0.5 – 1.5 kilobases downstream of the poly-A site (Glover-Cutter et al., 2008). In contrast, yeast CstF-64 binding occurs in close proximity to the poly-A site (Kim et al., 2004). One aim of this work was to determine if the somewhat disparate *C. elegans* CstF-64 is also a good candidate for functioning in 3' end formation.

ChIP-seq experiments demonstrated that CstF-64 is significantly enriched at the 3' ends of genes, coinciding with the site of 3' end formation. In addition, worm CstF-64 binding at 3' ends of non-operon genes occurs in close proximity to the poly-A site,

similar to the binding reported for budding yeast (Kim et al., 2004). These results suggest that the worm CstF-64 is likely to be involved in 3' end formation as its ortholog is in other organisms.

CstF-64 binding flanks the poly-A site in non-operon genes

Next, I showed that the CstF-64 binding present at non-operon genes consists of a double peak flanking the poly-A site. Whether both sites of CstF-64 binding function in 3' end formation is unknown. Interestingly, the peak occurring downstream of the poly-A site (~100 bp) does not match the region of the predicted worm CstF-64 binding site, located less than 20 nt downstream of the cleavage site (Graber et al., 2007). One possibility is that the ChIP experiment is only capturing the CstF-64 bound to the RNAPII complex, but is unable to capture CstF-64 bound to the RNA. Further support for this idea is the finding that the CstF-64 binding seen at 3' ends of genes colocalizes with paused RNAPII that is maximally phosphorylated at Ser-2 (see below). Why RNAPII with CstF-64 is pausing on both sides of the poly-A site is an interesting question that needs further study.

An alternative possibility for explaining why CstF-64 binding at 3' ends of genes does not match the known CstF-64 binding site on the RNA might be due to the fact that the pre-mRNA is tethered to RNAPII. In other words, CstF-64 association with the RNA will not match the genomic location assayed by ChIP, since the CstF-64-pre-mRNA complex is tethered through RNAPII. Thus, the peaks of CstF-64 seen by ChIP on the DNA are changed relative to the true binding site for CstF-64 on the RNA. If this is the case, then the CstF-64 peak located downstream of the poly-A site could represent the

CPSF-CstF complex needed for 3' end formation. Moreover, the CstF-64 peak present upstream of the poly-A site could then represent extra CstF-64 needed on some genes with weak 3' end formation signals. This way the genes with poor matches to the consensus will increase its affinity for the binding of CstF-64, required for the assembly of the 3' end formation complex. Upstream of the poly-A site, the CstF-64 could be used for converting a weak 3' end formation signal into a stronger site by providing extra support for the 3' end formation complex (see below).

CstF-64 may have a role in transcription initiation of some genes

I demonstrated that CstF-64 is also enriched at the 5' end of some genes, although to a much lower level than the binding seen at 3' ends (Figure IV-4 and IV-6 – IV-8). In mammals and yeast many 3' end formation factors have been found at promoters (Murthy and Manley 1995; Glover-Cutter et al., 2008; Wang et al., 2010), suggesting CstF-64 may have a role in transcription initiation. The function of 3' end formation factors at promoters is unclear, but one popular idea is that this association is due to gene loops. Gene loops are physical interactions between promoters and DNA termination regions, which result in cross-linking of the 5' to 3' end in ChIP and 3C (chromosome conformation capture) experiments (Ansari and Hampsey 2005; Singh et al., 2009; O'Sullivan et al., 2004; Singh and Hampsey 2007; Tan-Wong et al., 2009; Perkins et al., 2008). Alternatively the presence of 3' end formation factors at promoters could be explained by the fact that certain cleavage/polyadenylation factors interact with general transcription factors (GTF) found at promoters (Dantonel et al., 1997; Murthy and Manley 1995; Calvo and Manley 2003). For example, CPSF is recruited to the pre-

initiation complex (PIC) by TFIID, and after transcription starts, CPSF dissociates from TFIID and binds the elongating RNAPII (Dantonel et al., 1997). Therefore, it is unknown if 3' end formation factors at promoters have a role in initiation or whether they could be just a consequence of gene loops.

Interestingly, some genes do not have CstF-64 bound to promoters (Figure IV-9 – IV-11), suggesting that in some genes CstF-64 might have a function in transcription initiation and in others not. Alternatively, the reason why some genes show CstF-64 at promoters may have something to do with expression levels. Since it is likely that my ChIP experiments only show CstF-64 association with RNAPII, thus highly expressed genes will only be capable of revealing CstF-64 association with RNAPII. Further analysis is require, to find out if highly expressed genes correlate with CstF-64 presence at the promoter.

CstF-64 binding is not sufficient to cause transcription termination

Operons provide a unique opportunity to test if the CstF-64 binding present at 3' ends of genes is sufficient to cause transcription termination. While internal genes in operons undergo 3' end formation, transcription termination must be prevented for the downstream genes to be expressed. In chapter III, I showed that Ser-2p of RNAPII CTD marks 3' ends of internal genes in operons. Furthermore, Graber et al., 2007 showed that both internal and terminal 3' ends contain all the required 3' end formation signals needed for the CPSF-CstF core complex to bind. Therefore, neither Ser-2p nor the 3' end formation signals present on the RNA are sufficient for causing termination. In this section I ask if CstF-64 recruitment occurs at internal 3' ends in operons.

I discovered that most internal 3' ends in operons are bound by CstF-64, demonstrating that CstF-64 recruitment is not sufficient for termination. Importantly, the pattern of CstF-64 binding present at internal genes in operons is similar to the CstF-64 pattern seen at terminal genes in operons. At both internal and terminal 3' ends in operons the location of the major CstF-64 peak relative to the poly-A site is less than 100 bp, which is in close proximity to the known CstF-64 binding site on the RNA (Figure IV-15A). These results suggest that internal 3' ends are processed similarly to 3' ends in which termination occurs, both involving CstF-64 binding.

CstF-64 binding differences between internal and terminal 3' ends

Interestingly, two main differences were found between CstF-64 binding at internal and terminal 3' ends. The first obvious difference is the height of the CstF-64 peak. This peak is distinctly higher at internal genes versus terminal genes in operons. The most likely explanation is that RNAPII pauses longer at internal 3' ends for unknown reasons (see below). Alternatively, more CstF-64 at internal 3' ends could also be explained by being expressed at a higher level. Co-expression of genes from a single promoter might predict that all genes within the same operon should be expressed equally. However, according to ESTs and transcriptome data this is not the case, since internal genes are expressed at a higher level than terminal genes in operons. Why operon genes are not expressed equally is unknown, but could be due either to differential RNA stability or to some level of internal RNAPII termination.

The second difference in CstF-64 binding between 3' ends of internal and terminal genes in operons is on the amount of CstF-64 present upstream of the poly-A site (Figure IV-15A). This CstF-64 peak is distinguishable at terminal 3' ends, but is barley detected at internal 3' ends. Intriguingly, CstF-64 binding at 3' ends of non-operon genes also contains a prominent CstF-64 peak upstream of the poly-A site, similar to the one present at terminal 3' ends (Figures IV-15A and B). Therefore, an interesting possibility is that the CstF-64 peak present upstream of the poly-A site has something to do with termination, since it is only prominent at 3' ends where termination occurs. The small peak of CstF-64 present at internal 3' ends may be due to low levels of internal termination between genes in operons.

Why do 3' ends where termination occurs contain CstF-64 binding upstream of the poly-A site?

One possibility to explain the binding of CstF-64 upstream of the poly-A site at 3' ends associated with termination could be due to increased 3' end formation efficiency, thus making sure termination occurs. For example, extra U-rich sequences at the 3' ends of genes containing weak 3' end formation signals has been shown to provide extra binding sites for the 3' end formation complex, thus increasing 3' end formation efficiency (Millevoi and Vagner 2010). In addition, Graber et al. (2007) showed that a putative CstF-64 binding site was located on both sides of the poly-A site in internal and terminal genes in operons. Moreover, previous studies have shown that poly-A signals that deviate from the consensus sequence result in less efficient processing (Sheets et al., 1990). Therefore, the presence of CstF-64 upstream of the poly-A site could reflect the necessity of extra CstF-64 binding sites to efficiently process 3' ends with weak 3' end formation signals.

One prediction from this hypothesis is that the poly-A signal at 3' ends of internal genes in operons may contain better matches to the consensus AAUAAA when compared to terminal 3' ends, since internal genes in operons contain low levels of CstF-64 upstream of the poly-A site. However, when Graber et al., 2007 compared poly-A signals between internal and terminal genes in operons, they found the opposite to be true. Terminal poly-A signals preferentially match the consensus poly-A signal, whereas internal poly-A signals match more weakly. This result argues against the prediction that the smaller CstF-64 peak upstream of the poly-A site at internal genes in operons is indicative of strong 3' end formation signals. However, the possibility remains that the

CstF-64 peak present upstream of the poly-A site may function to provide additional binding sites for 3' end formation factors since the efficiency of 3' end formation signals is not solely dependent on the poly-A signal (Sheets et al., 1990).

Alternatively, the binding of CstF-64 upstream of the poly-A site may be a common characteristic of all genes, thus having nothing to do with transcription termination. Evans et al. (2001) showed the existence of a CstF-64 complex with the SL2 RNA, which is the specialized spliced leader used in trans-splicing downstream genes in operons. Moreover, Erica Lasda in the lab has shown that an oligonucleotide composed of the intercistronic region sequence needed for SL2 trans-splicing, called the U-rich element, is capable of pulling-down CstF-64 from embryonic extract (unpublished results). Therefore, an interesting possibility is that internal genes in operons do not require extra CstF-64 binding because their 3' end arrangement is strengthened by whether CstF-64 is attracted by SL2 trans-splicing or by the Ur-element.

Termination-competent RNAPII at 3' ends of internal genes in operons

ChIP experiments done in mammals have shown that RNAPII continues transcribing downstream of the poly-A site for another 0.5 kb – 1.5 kb (Glover-Cutter et al., 2008). Prior to termination, RNAPII is in a paused conformation, colocalizing with maximal Ser-2 phosphorylation and bound by high levels of CstF-64 (Glover-Cutter et al., 2008). In chapter III, I demonstrated that internal 3' ends in operons are maximally phosphorylated at Ser-2. Furthermore, I have also shown that internal 3' ends contain high levels of CstF-64. Therefore, I tested if CstF-64 colocalizes with Ser-2p and

paused RNAPII at 3' ends of internal genes in operons, in order to provide insights into how RNAPII is prevented from terminating at these internal sites.

First I showed that CstF-64 colocalizes with Ser-2p at all 3' ends in operons, suggesting a functional overlap. Indeed, the 50 KDa subunit of the CstF complex binds the phosphorylated form of RNAPII CTD (Fong and Bentley 2001), so CstF-64 present at 3' end of genes may be present as a complex bound to Ser-2p.

Finally, I showed that RNAPII pauses at each 3' end of genes in operons, colocalizing with CstF-64 and Ser-2p, similar to the characteristics associated with termination-competent RNAPII shown in mammals (Glover-Cutter et al., 2008). Therefore, it is likely that in my ChIP experiments I am only seeing CstF-64 bound to RNAPII, since CstF-64 and RNAPII colocalize. In contrast to mammals, the worm peak of paused RNAPII containing Ser-2p and CstF-64 occurs much closer to the poly-A site, which may be due to a more compact genome. Overall, these results provide evidence that RNAPII is pausing at all 3' ends in operons, suggesting that termination at internal 3' ends in operons must be prevented in a way other than modulating RNAPII pausing. Furthermore, these results provide evidence that the paused RNAPII associated with gene 3' ends is not sufficient for transcription termination, since paused RNAPII is present at 3' ends where termination does not occur.

Previous work from the lab has shown that 3' end formation and SL2 trans-splicing are mechanistically connected. Evidence from *in vivo* experiments using a synthetic operon showed that mutating the poly-A signal of the upstream gene reduced SL2 trans-splicing downstream (Kuersten et al., 1997). However, when the trans-splice site was mutated this had no effect on 3' end formation occurring upstream, even

though no downstream RNA accumulated (Kuersten et al., 1997). Moreover, a complex containing CstF-64 and SL2 snRNP has been purified from worm crude extract, consistent with the coupling of these two events at internal positions in operons (Liu et al., 2001). Based on these data it was proposed that termination might be prevented at internal genes in operons by SL2 trans-splicing, since SL2 addition adds a cap, already present on the SL2. Therefore, following cleavage at the 3' end of the upstream RNA, the downstream RNA is uncapped and contains a free 5' phosphate, which is substrate for 5'-3' exonucleases. According to the torpedo model, termination occurs when the exonuclease catches up to the elongating RNAPII and somehow causes it to fall off the DNA. However, when the downstream RNA is capped by SL2 trans-splicing then this blocks the passage of the exonuclease from reaching the elongating RNAPII allowing it to transcribe to the end of the next gene. The data presented in this chapter provide further support for this idea, since RNAPII ready for termination contain the same termination competent characteristics as RNAPII at internal genes in operons where termination does not occur.

CHAPTER V

IN VIVO ANALYSIS OF 3' END FORMATION AT SL1-TYPE OPERONS

Introduction

The *C. elegans* genome contains a unique gene arrangement known as operons. These multi-gene clusters represent an elegant response to the evolutionary pressure for a compact genome, by allowing a single promoter to control the expression of up to eight genes. In this chapter, I investigate the mechanism of 3' end formation in a rare type of operon with no intercistonic DNA between genes.

Operons are transcribed as polycistronic clusters that are further processed into individual cistrons by 3' end formation at the 3' end of the upstream mRNA, and trans-splicing at the 5' end of the downstream mRNA. In *C. elegans*, there are two types of SL trans-splicing that are used in different genomic arrangements. SL2 trans-splicing is exclusively used for downstream genes in operons, while the majority of SL1 trans-splicing is used to replace the outron at the 5' end of genes, very close to promoters. Typically genes in operons are separated by an intercistonic region (ICR) of ~100 bp, that strongly correlates with SL2 trans-splicing at these locations (Allen et al., 2011). These types of operon are called SL2-type and represent the vast majority of operons in *C. elegans*.

A rare type of operon called SL1-type does not follow these general rules. Interestingly and in contrast to SL2-type operons, downstream genes are trans-spliced to SL1 and not SL2. Further, genes in SL1-type operons lack an ICR, and the upstream gene contains an extremely long polypyrimidine tract (Williams et al., 1999). Very few

examples (23) of SL1-type operons have been identified in the genome, as defined by the above characteristics (Blumenthal 2012).

The mechanism for 3' end formation at SL1-type operon 3' ends is unknown. In contrast to SL2-type operons, the Blumenthal lab hypothesized that polycistronic pre-mRNA arising from these operons only needs a single RNA processing event (SL1 trans-splicing) for cistron maturation. In SL1-type operons, both the 3' end formation and trans-splicing machinery share a common processing site. Therefore, in order for both genes to be expressed simultaneously, SL1 trans-splicing must occur first to prevent 3' end formation from destroying the trans-splice site. In support of this model, hundreds of ESTs (wormbase) and RNA-seq reads (modencode) show cleavage and subsequent polyadenylation occur at the trans-splice site dinucleotide (AG) of the downstream gene. Interestingly, a genome-wide study of *C. elegans* polyadenylation sites revealed that the AG dinucleotide is one of the least common cleavage sites by the canonical 3' end formation machinery for poly-A addition. Furthermore, 3' ends of upstream genes in SL1-type operons contain a perfect match to the consensus AAUAAA (Blumenthal 2012), but these operons lack a CstF binding site (Graber et al., 2007). These observations are consistent with cleavage at the 3' end of the upstream gene occurring in a CstF independent manner (SL1 trans-splicing), followed by CPSF-dependent polyadenylation of the free 3' end, created by SL1 trans-splicing of the downstream gene.

However, in contrast to the transcriptome data, limited mutational analysis of a transgenic SL1-type operon demonstrated that mutation of the trans-splice site did not prevent 3' end formation of the upstream gene (Williams et al., 1999). This result raised

the possibility that in these operons 3' end formation is in competition with SL1 trans-splicing, so that both genes are not expressed from the same pre-mRNA. Although this hypothesis suggests that cleavage may not be occurring by SL1 trans-splicing, the possibility exists that both CPSF-CstF and SL1 trans-splicing dependent 3' end formation can occur at these 3' ends. Therefore, a novel mechanism of 3' end formation by SL1 trans-splicing may exist in *C. elegans*.

Here I report an investigation into how 3' ends are formed in SL1-type operons, hopefully resolving the apparent discrepancy between the transcriptome data and the experimental results. I have used ChIP and RT-PCR experiments to test whether 3' ends in SL1-type operons are bound by the same factors as 3' ends of SL2-type operons. I show that Ser-2p and CstF-64 are present at both SL1-type operon and SL2-type operon 3' ends. Moreover, when CstF-50 was knocked-down, the levels of CstF-64 at both types of 3' ends were significantly decreased, and consequently cleavage was prevented. My results provide support for the processing of SL1-type operon 3' ends by the same 3' end formation mechanism as 3' ends found at SL2-type operons.

Results

SL1-type operon 3' ends are marked by Ser-2p

In chapter III, I showed that Ser-2p of RNAPII CTD is associated with 3' end formation sites in SL2-type operons, suggesting this phosphorylation is indeed required for the recruitment of 3' end formation factors. *In vivo* and *in vitro* experiments have shown that reducing the levels of Ser-2p at gene 3' ends affects the binding of 3' end processing factors and alters the site of cleavage/polyadenylation (Ahn et al., 2004). Therefore, if

cleavage at 3’ ends of SL1-type operons is performed by SL1 trans-splicing and not by CstF associated proteins, then Ser-2p at this type of 3’ end might not occur.

To test if the 3’ ends in SL1-type operons are marked by Ser-2p, I performed ChIP-qPCR experiments using antibodies against Ser-2p in three different SL1-type operons (Figure V-1 – V-3). Each bar represents an individual primer amplicon used to quantify the IP DNA, positioned above its corresponding genomic location. I found that Ser-2p is present at all three SL1-type operon 3’ ends, similar to the binding seen for SL2-type and terminal 3’ ends.

The first SL1-type operon analyzed is a two-gene cluster (CEOP 3666), which consists of *mev-1* internal 3’ end (SL1-type) and the *ced-9* terminal 3’ end (Figure V-1). The levels of Ser-2p are low at the promoter and increase gradually, peaking downstream of the *mev-1* 3’ end (SL1-type). Then the levels of Ser-2p drop and are maintained at a low level throughout the body of *ced-9*, without an obvious peak at its 3’ end. The reason there is no clear peak at the 3’ end of *ced-9* is unknown, but could be due to the lack of primer sets covering that region. However, it is clear from this example that the levels of Ser-2p are high at the *mev-1* internal 3’ end (SL1-type), suggesting a need for the recruitment of 3’ end formation factors.

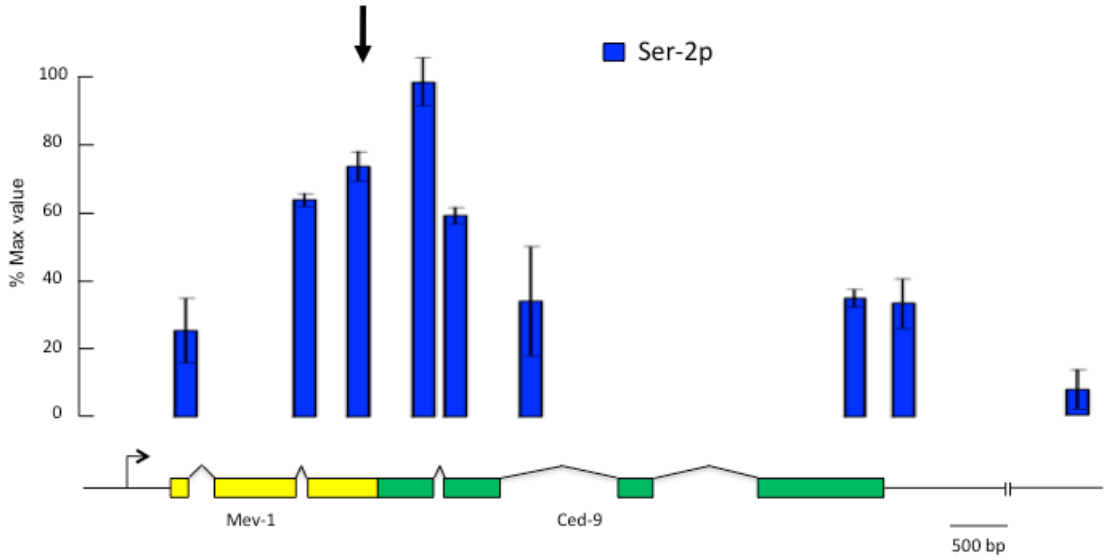


Figure V-1. Ser2p of RNAPII CTD marks the SL1-type operon 3' end in a two-gene operon (CEOP 3666): ChIP-qPCR experiments using anti-Ser2p antibody. The genes are depicted by filled boxes representing exons and introns shown as angled lines. The results from each primer set are positioned immediately above the corresponding genomic location and normalized to the highest value. Error bars represent percent error of three PCR reactions from a single immunoprecipitation experiment. The most 3' primer pair is a negative control located in the center of a region lacking annotated genes; it is not adjacent to this operon. The vertical arrow denotes the location of the 3' end in SL1 type operons.

The second operon analyzed is a four-gene cluster (CEOP 3184) that consists of the *snfc-5* internal operon 3' end (SL1-type) followed by two SL2-type internal operon 3' ends (*rnp-4* and *prdx-3*), and the *r07e5.1* terminal 3' end (Figure V-2). Ser-2p levels within this operon peak at the *snfc-5* 3' end (SL1-type) as well as at the other canonical 3' ends located within the operon. Again for unknown reasons the peak of Ser-2p appears more pronounced in the SL1-type operon 3' end than in the other 3' ends. The peak at the 3' end of *snfc-5* may represent a RNAPII pause site, suggesting a differential level of pausing at 3' ends in this operon.

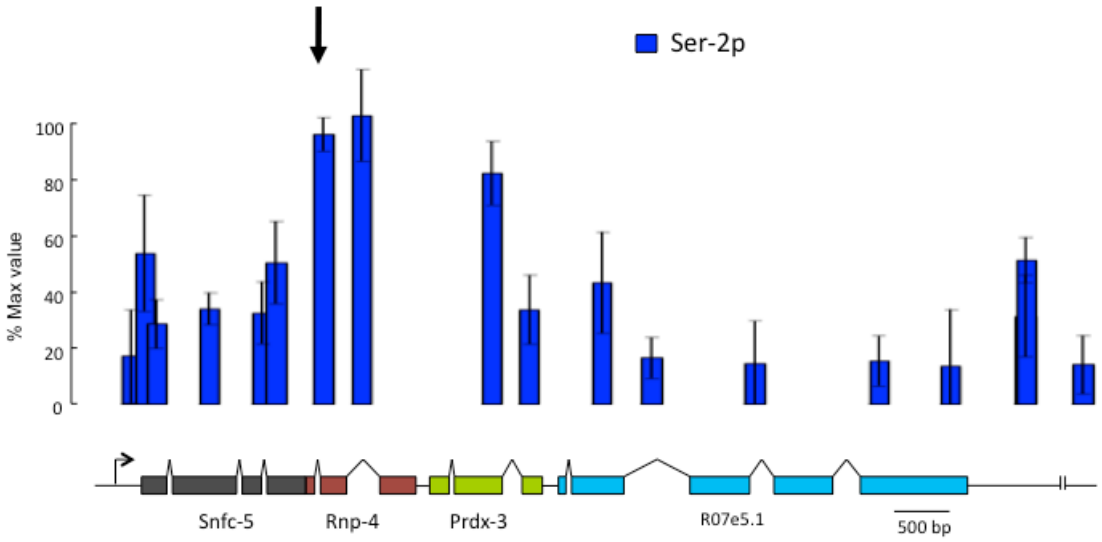


Figure V-2. Ser2p of RNAPII CTD marks the SL1-type operon 3' end in a four-gene operon (CEOP 3184): ChIP-qPCR experiments using anti-Ser2p antibody. See Figure V-1 caption for details.

The last SL1-type operon analyzed is a three-gene operon (CEOP 3412) composed of *k04g7.11* internal 3' end (SL1-type), *rnp-7* internal 3' end (SL2-type), and the *nuo-4* terminal 3' end (Figure V-3). The levels of Ser-2p seemed to be elevated throughout the first two genes in the operon with a slight increase over the *k04g7.11* 3' end (SL1-type). As seen in the above examples, Ser-2p is clearly peaking at the canonical 3' ends of *rnp-7* and *nuo-4*. In contrast, there is no prominent peak of Ser-2p at the *k04g7.11* 3' end (SL1-type), maybe due to lack of primer sets in that region. In any case, the levels of Ser-2p around this 3' end is clearly above background levels confirming the presence of Ser-2p at this location. This result is consistent with the operons analyzed above, providing *in vivo* evidence that Ser-2p peaks at SL1-type operon 3' ends. Peaks of Ser-2p occur at SL1-type and SL2-type internal 3' ends

indistinguishably.

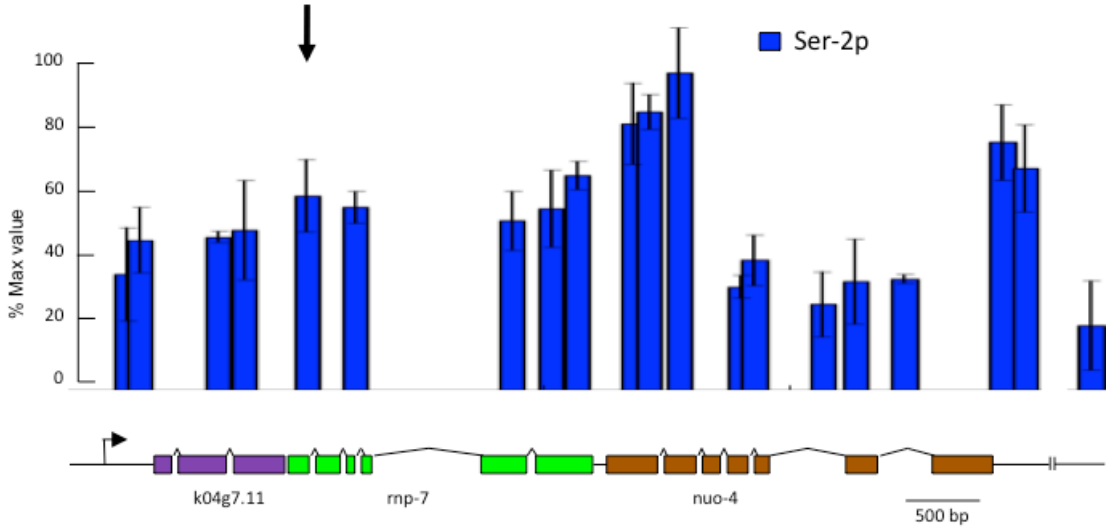


Figure V-3. Ser2p of RNAPII CTD marks the SL1-type operon 3' end in a three-gene operon (CEOP 3412): ChIP-qPCR experiments using anti-Ser2p antibody. See Figure V-1 caption for details.

SL1-type operon 3' ends are bound by CstF-64

Next, I asked if the Ser-2p mark seen at SL1-type operon 3' ends is associated with the binding of CstF-64. Genome-wide bioinformatic analysis of SL1-type operon 3' ends reveals a lack of the known CstF-64 binding sites downstream of the cleavage site (Graber et al., 2007). Therefore, the Ser-2p mark at SL1-type operon 3' ends may not be associated with CstF-64 binding. Absence of CstF-64 would also be consistent with cleavage at 3' ends of SL1-type operons occurring in a CstF-independent manner.

To address this issue, I took a genome-wide approach to determine all CstF-64 binding sites using ChIP experiments analyzed by array, sequencing and qPCR methods (Figure V-4 – V-6). ChIP-on-chip experiments were done by hybridizing the CstF-64 immunoprecipitated DNA into an Affymetrix tiling array, consisting of 3.2 million

probes spanning the whole non-repetitive worm genome. This technique gave me inconsistent results and poor resolution over putative CstF-64 binding sites, making difficult the interpretation of the results. In order to overcome these major drawbacks associated with hybridization based methods, a fraction of the CstF-64 immunoprecipitated DNA was analyzed by next generation sequencing (ChIP-seq) using the Illumina Hiseq2000 platform. For each example shown in Figure V-4 – V-6 the binding of CstF-64 was further confirmed by ChIP-qPCR experiments and the positions of transcriptionally active promoters are shown by the presence of H3K9ac. As exemplified in Figure V-4 – V-6, I found that CstF-64 is present at SL1-type operon 3' ends, colocalizing with Ser-2p.

In the *mev-1* operon, CstF-64 was found at the *mev-1* internal 3' end (SL1-type) and at the *ced-9* terminal 3' end using all three quantification methods (Figure V-4). Importantly, this pattern matches very closely the Ser-2p pattern seen for this operon (Figure V-1), both showing a much more pronounced peak at the internal 3' end (SL1-type).

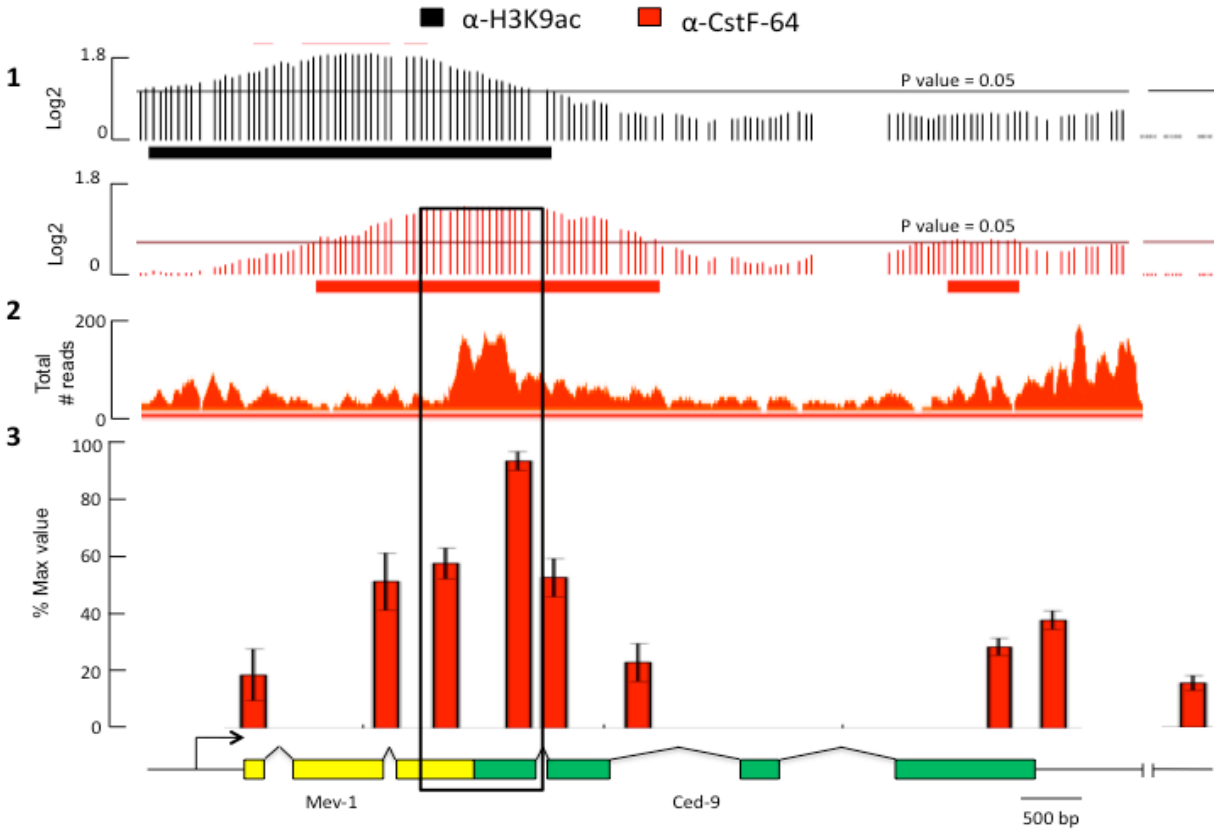


Figure V-4. SL1-type operon 3' ends are bound by CstF-64 in a two-gene operon (CEOP 3666): ChIP experiments using antibodies against CstF-64 (red) and H3K9ac (black). The genes are depicted by filled boxes representing exons and introns shown as angled lines. Flanking intergenic regions are shown as lines. The arrow denotes the location of the promoter. Black boxes denotes the location of 3' ends in SL1 type operons. 1. ChIP-chip experiments using Affymetrix tiling arrays. Each vertical line is an individual probe and the height of the bar is proportional to the amount of total hybridization. Horizontal lines under each graph shows the regions with a p-value of 0.05. 2. ChIP-seq experiment using the Hiseq 2000 platform from Illumina. The height of the peak is an indication of the number of reads in that region. 3. ChIP-qPCR experiment. The results from each primer set are positioned immediately above the corresponding genomic location and normalized to the highest value. Error bars represent percent error of three PCR reactions from a single immunoprecipitation experiment.

In the second operon analyzed, CstF-64 binding was also detected at the SL1-type 3' end (*snfc-5)* using all three methods (Figure V-5). However, all 3' ends in the operon differed in the levels of CstF-64 bound. The highest amount was present at the

3’ ends of *snfc-5* (SL1-type) and *rnp-4* (SL2-type), suggesting that higher levels of CstF-64 are not specific to SL1-type operons 3’ ends. In addition, there was an extra peak not associated with any apparent 3’ end at the 4th exon of *r07e5.1*. I was unable to find any evidence supporting the possibility of a premature 3’ end that could account for the peak of CstF-64 seen in the middle of this gene. Alternatively, the peak of CstF-64 present at the 4th exon of *r07e5.1* could represent RNAPII pausing at this location (see Chapter IV Discussion).

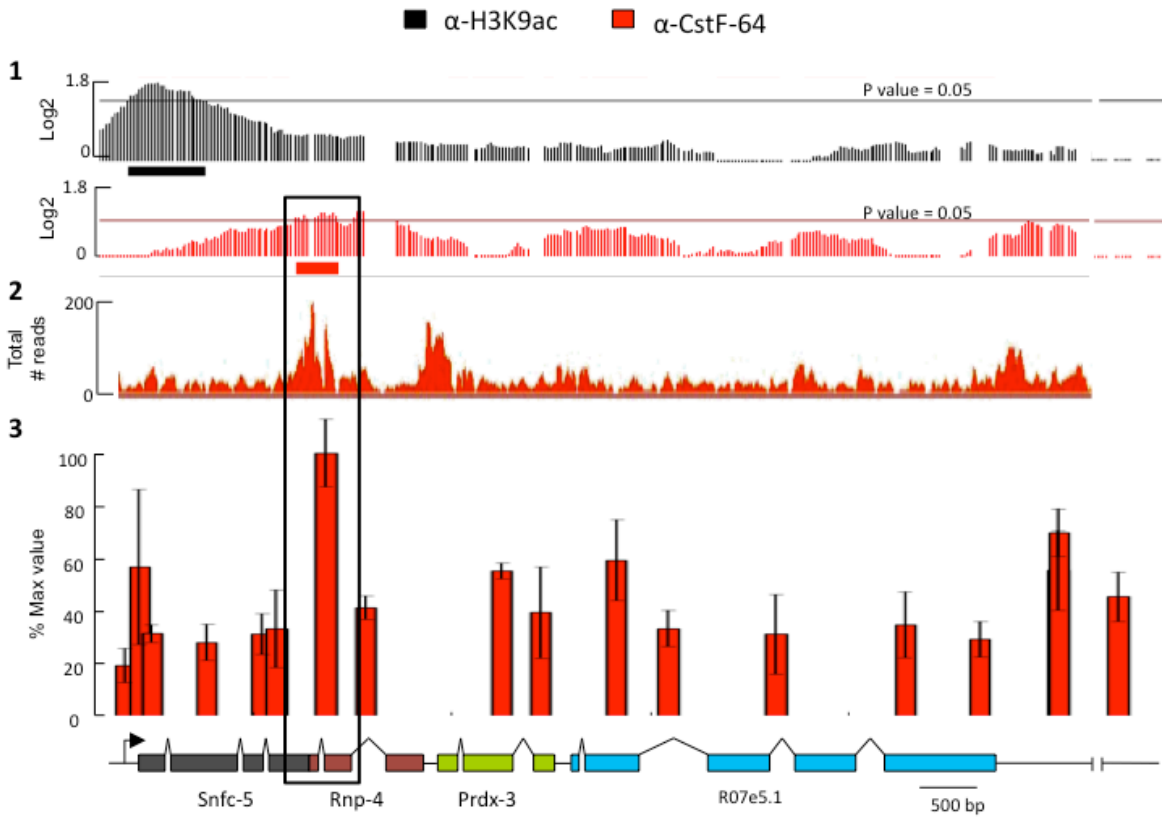


Figure V-5: SL1-type operon 3’ ends are bound by CstF-64 in a four-gene operon (CEOP 3184): ChIP experiments using antibodies against CstF-64 (red) and H3K9ac (black). See Figure V-4 caption for details.

The last SL1-type operon example is the *k04g7.11* three-gene operon (Figure V-6). In this case there is a discrepancy between methods on CstF-64 binding at the SL1-type operon 3' end (*k04g7.11*). As shown in Figure V-6, by sequencing and qPCR methods there was no clear peak evident at this 3' end, in contrast to the array-based method, which showed a statistically significant CstF-64 peak at this same location. The reason for this is unknown but most likely reflects non-specific hybridization of IP DNA fragments to this genomic location. This discrepancy was not found on the other 3' ends within the operon. Importantly, at the SL1-type operon 3' end there was a low level of Ser-2p that seemed to be spread around the entire gene body (*k04g7.11*), consistent with the low pattern seen for CstF-64 binding at this same 3' end. This suggests that RNAPII is phosphorylated at Ser-2 throughout the body of the *k04g7.11* gene and not localized to a discrete peak at the 3' end, providing an explanation for the low level of CstF-64 binding. The reason for this is unknown, but it could be a reflection of CstF-64 traveling with the elongating RNAPII, in which case the distinct peaks represent RNAPII pause sites (Glover-Cutter et al., 2008). Why RNAPII pauses at some but not all 3' ends in operons is an interesting question that was discussed in Chapter IV. In any case, this operon demonstrates that the pattern of CstF-64 matches very well the pattern seen for Ser-2p. Overall, these experiments together with the Ser-2p data for this operon (see Figure V-3) are consistent with a phosphorylation-dependent recruitment of CstF-64 at SL1-type operon 3' ends.

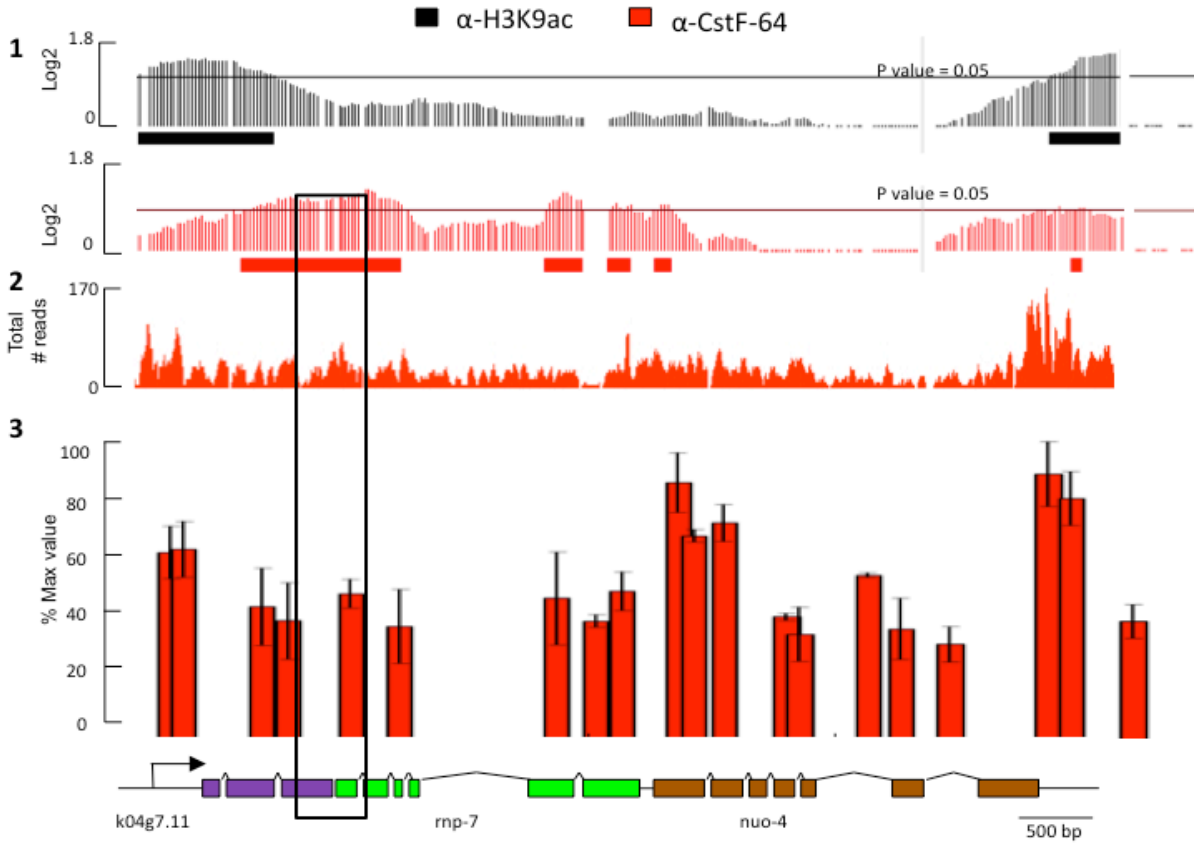


Figure V-6: SL1-type operon 3' ends are bound by CstF-64 in a three-gene operon (CEOP 3412): ChIP experiments using antibodies against CstF-64 (red) and H3K9ac (black). See Figure V-4 caption for details.

Genome-wide analysis of CstF-64 binding at 3' ends of SL1-type operons

In order to further investigate the binding of CstF-64 in SL1-type operons, 3' ends corresponding to all SL1-type operons in the genome containing significant levels of CstF-64 were aligned by their poly-A signal (Figure V-7). The levels of CstF-64 were averaged at a single nucleotide resolution 1000 bp upstream and 1000 bp downstream of the cleavage site. As a control I used 3' ends from downstream genes of the same SL1-type operons, which are either internal SL2-type or terminal 3' ends. As shown in

Figure V-7, the CstF-64 peaks look similar, consistent with the examples above (Figures V-1 – V-6) and with cleavage occurring through the canonical CPSF-CstF mechanism.

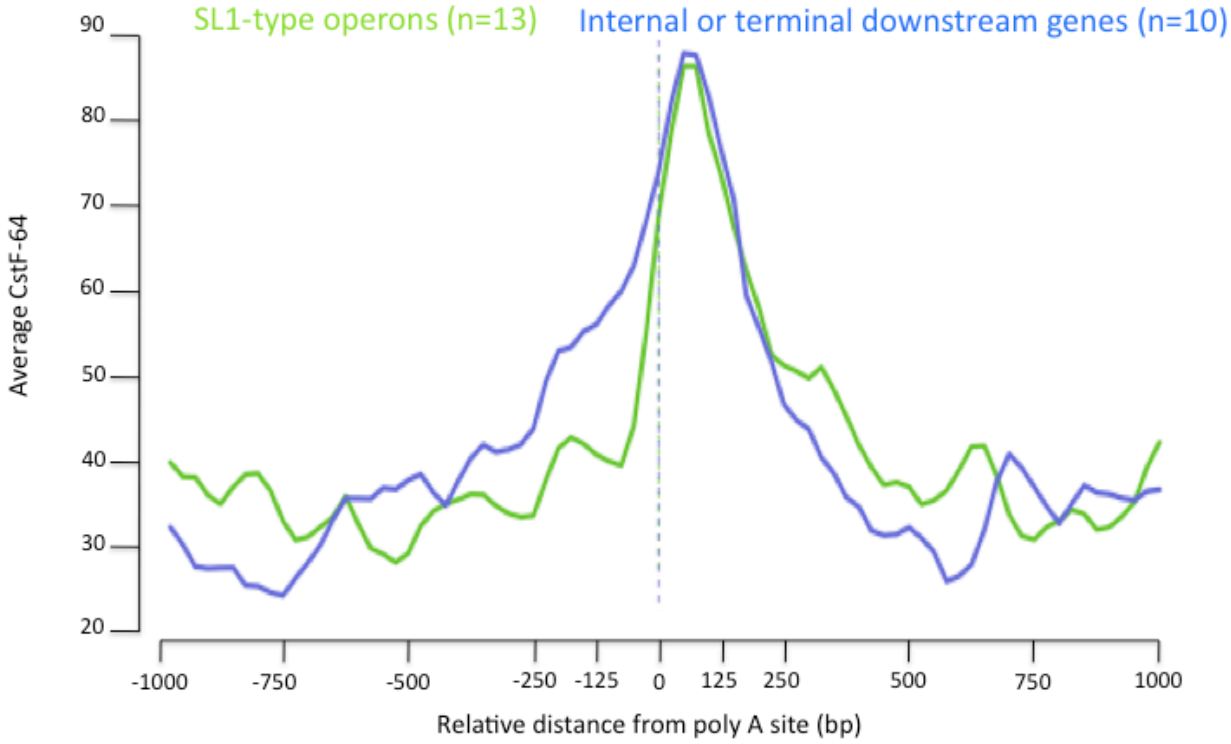


Figure V-7: CstF-64 association with SL1-type operons 3' ends is similar to internal or terminal downstream genes 3' ends: All SL1-type operons genes (n=13) containing significant levels of CstF-64 are plotted (green line). Internal or terminal downstream genes of the SL1-type operons are also plotted (blue line) serving as a control. The levels of CstF-64 were averaged at a single nucleotide resolution 1000 bp upstream and 1000 bp downstream of the poly-A site (dashed line). The x-axis indicated distance from the poly-A site which is set at zero. The y-axis shows the fold enrichment for each nucleotide over background. MACS algorithm was used to calculate CstF0-64 enrichment over background and a python script was used to manipulate data for plotting.

Surprisingly, CstF-64 binding in SL1-type operons seems to be present only downstream of the poly-A site. In contrast, CstF-64 is present at both sides of the poly-A site in the control genes (Figure V-7). This result is not predicted by the bioinformatics analysis of known CstF-64 binding sites in SL1-type operons (Graber et al., 2007). This

study showed that SL1-type operons lack a CstF-64 binding site on the RNA downstream of the cleavage site, predicting that CstF-64 should not be present at these locations (Graber et al., 2007). Instead these operons contain long poly-Y tracts located upstream of the poly-A site that could potentially function as CstF-64 binding sites (Williams et al., 1999; Graber et al., 2007). However, CstF-64 was present only downstream of the poly-A site in SL1-type operons 3' ends as shown in Figure V-7. Therefore, this data provides evidence that the CstF-64 detected by ChIP represents CstF-64 bound to RNAPII, but may not reveal CstF-64 bound to the RNA, since CstF-64 can't bind the RNA downstream of the poly-A site in SL1-type operons 3' ends.

In summary, I have used ChIP experiments in three SL1-type operons to show that their 3' ends are marked by Ser-2p and are bound by CstF-64. In addition, CstF-64 binding at SL1-type operon 3' ends is similar to the binding present at the control 3' ends. These results are consistent with the co-transcriptional recruitment of 3' end processing factors by CTD phosphorylation. These data demonstrate that CstF-64 is present at 3' ends of SL1-type operons and is thus available to perform a role in their cleavage.

CstF-50 is needed for the localization of CstF-64 at SL1-type operon 3' ends

The Bentley lab demonstrated that the CstF subunits are independently recruited to the transcription site; CstF-64 is recruited at 5' ends while CstF-77 is recruited later towards the 3' ends of genes (Glover-cutter et al., 2008). This suggests that the trimeric CstF complex may assemble on the CTD during transcription, and that each independent subunit may have roles other than 3' end formation when not associated with each

other. For example, CstF-64 recruitment at 5’ ends of genes may be due to an unknown function in transcription initiation. Therefore the possibility exists that CstF-64 is at 3’ ends of SL1-type operons because of its association with the 5’ end of the downstream gene rather than for 3’ end formation of the upstream gene. Is the association of CstF-64 with 3’ ends of SL1-type operons present as the CstF complex, which is required for 3’ end formation?

In order to discover whether the CstF-64 is present as a CstF complex at 3’ ends of SL1-type operons, I RNAi’ed the CstF-50 subunit and used ChIP-seq to measure binding of CstF-64. I chose to perform RNAi on the CstF-50 subunit because a balanced deletion allele of CstF-50 was available from the Caenorhabditis Genetics Center (CGC). Therefore, I could perform RNAi on the balanced strain allowing me to reduce the levels of CstF-50 significantly. RT-PCR experiments in three biological replicates were used to confirm CstF-50 knocked-down. As shown in Figure V-8A, in all cases the levels of CstF-50 mRNA are significantly reduced compared to the control. Unfortunately, the lack of a CstF-50 antibody makes it impossible to test how much CstF-50 protein is left in the cell after RNAi.

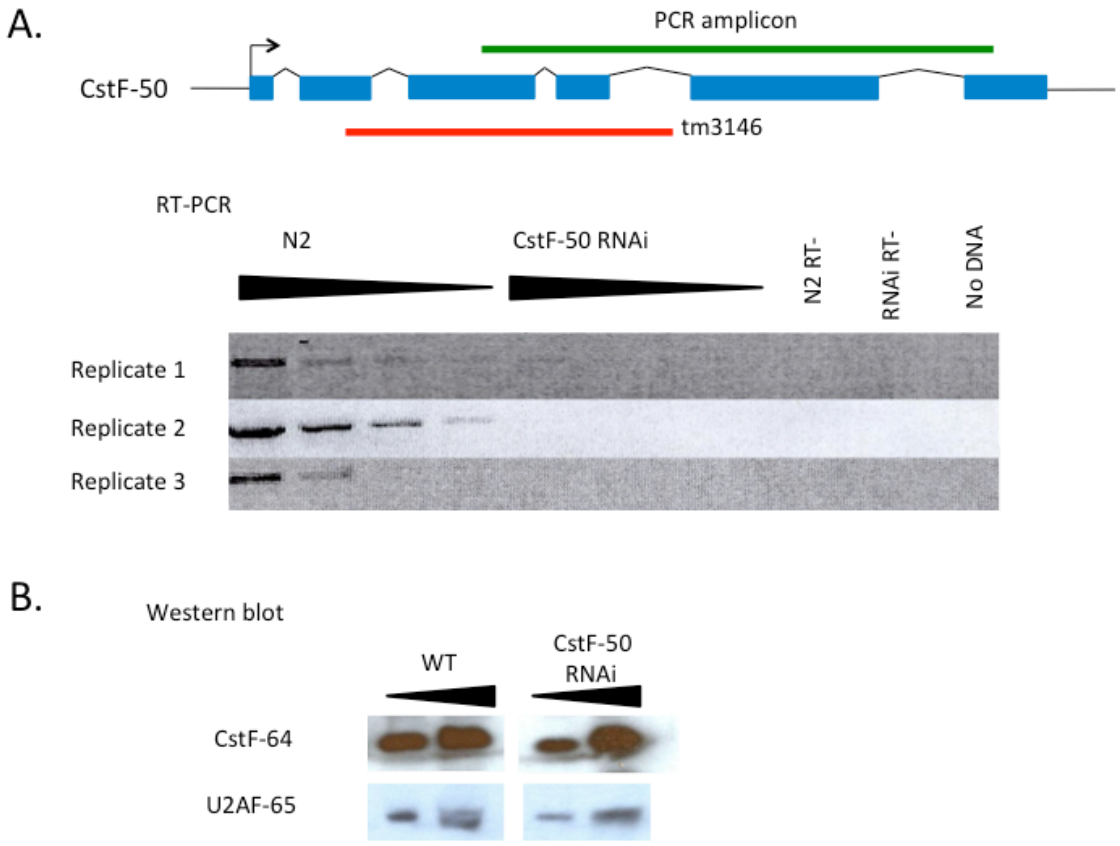


Figure V-8: RNAi of CstF-50 in a balanced deletion allele (tm 3146): A. RT-PCR experiments using oligo dT primed cDNA in three biological replicates. The CstF-50 locus is shown above the gels with the promoter represented by a bend arrow. The genes are depicted by filled boxes representing exons and introns shown as angled lines. The green line on top of the gene model represents the PCR amplicon used for detecting the CstF-50 mRNA. The red line on the bottom of the gene model shows the deleted region of CstF-50 in the tm 3146 allele. The horizontal triangles on the top of the gels represent 3-fold serial dilutions of the cDNA samples prior to PCR. Each replicate contains a minus RT control (no reverse transcriptase) and no DNA control shown on the right of the gel. B. Western blot analysis of WT and RNAi samples probed with anti-CstF-64 and anti-U2AF-65 antibodies. The horizontal triangles on top of the gels indicate different amounts of protein extract loaded in each lane.

The co-depletion of CstF factors by reduction of a single subunit of the complex has been shown to occur in flies (Sullivan et al., 2009), and is sometimes the case for other protein complexes. As shown in Figure V-8B, CstF-50 knockdown did not have an effect on the level of CstF-64. The reason for this is unclear but may be due to the fact

that in worms CstF-64 is not always part of the trimeric complex, consistent with unpublished results from our lab (see Discussion). Alternatively, the reason CstF-64 does not show a reduction by Western blot after CstF-50 knockdown could be that there is an autoregulatory mechanism that upregulates transcription for keeping CstF-64 levels constant. In any case, if the levels of CstF-64 at SL1-type operon 3' ends are unchanged by RNAi of CstF-50, this will suggest that CstF-64 at these 3' ends is not present as part of the CstF complex. As shown in Figures V-9 – V-13, five SL1-type operons are used as examples and in all cases CstF-64 levels were reduced when CstF-50 was RNAi'ed. Note that the first two operons are the same examples used throughout this chapter.

The first SL1-type operon example is the *mev-1/ced-9* operon (Figure V-9). In this two-gene operon the CstF-64 present at the 3' end of the *mev-1* gene (SL1-type) is significantly decreased upon CstF-50 knockdown. A similar reduction in CstF-64 binding is seen at the terminal 3' end (*ced-9*) when CstF-50 has been RNAi'ed. The reason CstF-64 is not completely eliminated when CstF-50 has been knocked-down is unclear and is further discussed below. One likely possibility is that the CstF-50 that was not eliminated upon RNAi is sufficient for CstF-64 recruitment. Importantly, CstF-64 binding is reduced at both types of 3' end upon CstF-50 knockdown, suggesting a similar 3' end processing mechanism.

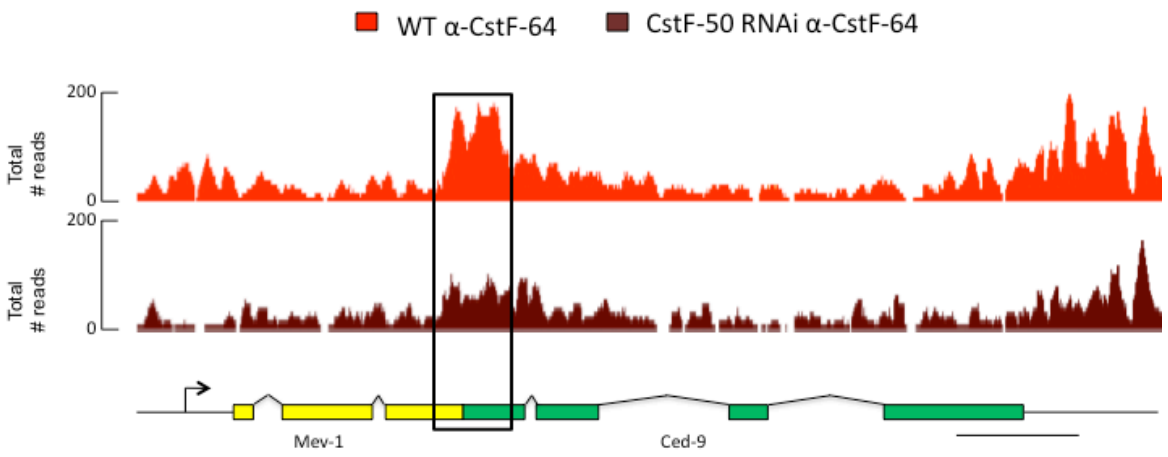


Figure V-9. CstF-64 binding to 3' ends in a two-gene operon (CEOP 3666) depends on CstF-50: ChIP-seq experiment with an antibody against CstF-64 in WT (red) and CstF-50 RNAi samples (dark red). Black boxes denotes the location of 3' ends in SL1 type operons. The height of the peak is an indication of the number of reads in a region.

The second example is the *snfc-5* four-gene operon (Figure V-10). In this case the reduction of CstF-50 nearly eliminated CstF-64 binding at the SL1-type operon 3' end (*snfc-5*). Importantly, CstF-64 binding was also severely reduced at the SL2-type 3' end (*rnp-4*) and terminal 3' end (*r07e5.1*), consistent with the previous example.

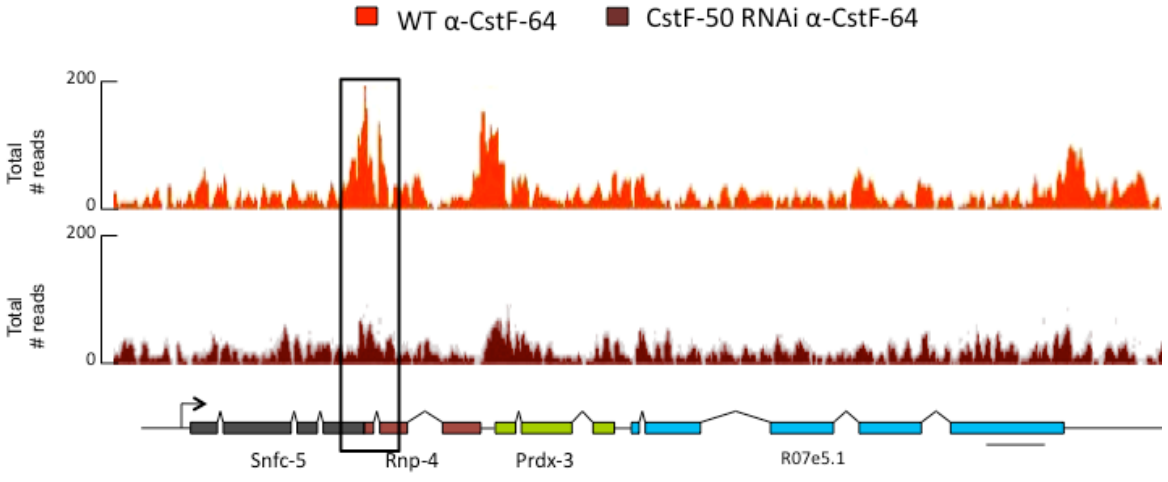


Figure V-10. CstF-64 binding to 3' ends in a four-gene operon (CEOP 3184) depends on CstF-50: ChIP-seq experiment with an antibody against CstF-64 in WT (red) and CstF-50 RNAi samples (dark red). See Figure V-11 caption for details.

The third example is a three-gene operon containing an SL1-type 3' end (*gna-1*), an SL2-type 3' end (b0024.11) and a terminal 3' end (b0024.10) (Figure V-11). CstF-64 binding at the SL1-type operon 3' end is completely eliminated when CstF-50 has been knocked-down, consistent with the formation of the CstF complex at this 3' end. Importantly, CstF-64 levels were significantly reduced at the SL2-type 3' end (b0024.11) and at the terminal 3' end (b024.10), suggesting that 3' end formation at this type of 3' end occurs in a similar manner to SL1-type operon 3' ends. However, in the SL2/terminal 3' ends CstF-64 levels are not eliminated. The reason for the differential result of CstF-64 binding upon CstF-50 knockdown between SL1-type 3' ends and SL2/terminal 3' ends is not known (but see below).

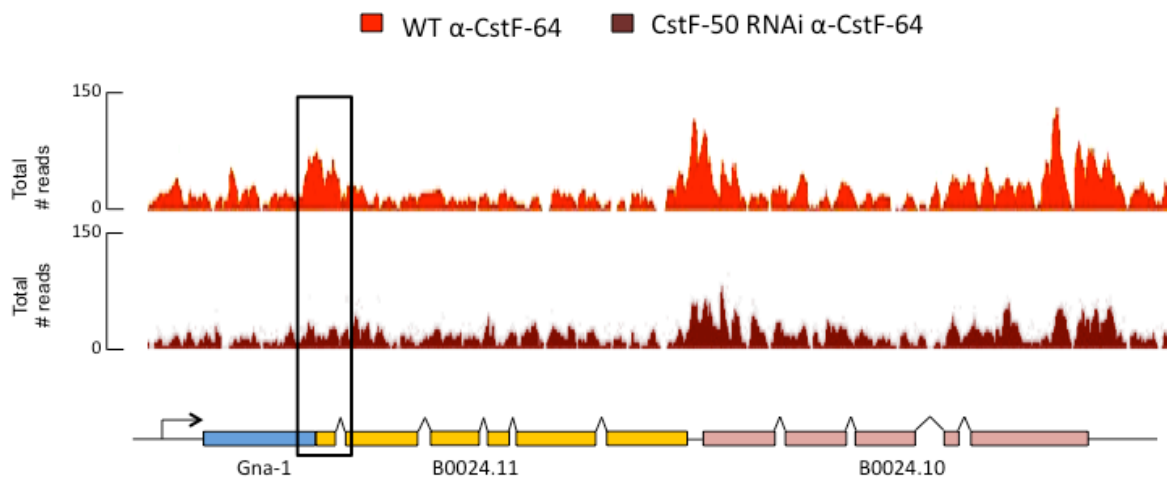


Figure V-11. CstF-64 binding to 3' ends in a three-gene operon (CEOP 5252) depends on CstF-50: ChIP-seq experiment with an antibody against CstF-64 in WT (red) and CstF-50 RNAi samples (dark red). See Figure V-11 caption for details.

In the *mes-6* three-gene operon, CstF-64 binding at the SL1-type operon 3' end (c09g4.4) was significantly decreased upon CstF-50 knockdown (Figure V-12), consistent with the previous examples. Similarly, the levels of CstF-64 are also reduced at the SL2-type 3' end (*mes-6*) and at the terminal 3' end (*cks-1*), suggesting the presence of the CstF trimeric complex in all types of 3' ends.

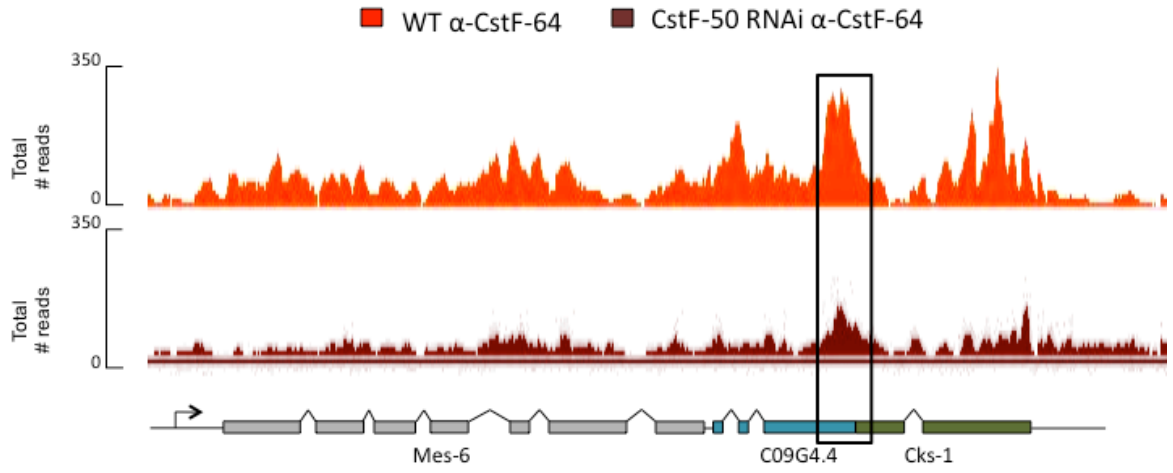


Figure V-12. CstF-64 binding to 3' ends in a three-gene operon (CEOP 4294) depends on CstF-50: ChIP-seq experiment with an antibody against CstF-64 in WT (red) and CstF-50 RNAi samples (dark red). See Figure V-11 caption for details.

The last SL1-type operon example is a two-gene operon (Figure V-13). In this example the low level of CstF-64 present at the SL1-type operon 3' end (*cdd-2*) is eliminated upon CstF-50 knockdown. At the terminal 3' end (*pam-1*), the level of CstF-64 after CstF-50 knockdown is reduced but not eliminated. This example is consistent

with all previous examples showing that CstF-64 recruitment to all types of 3' end is dependent on the levels of CstF-50.

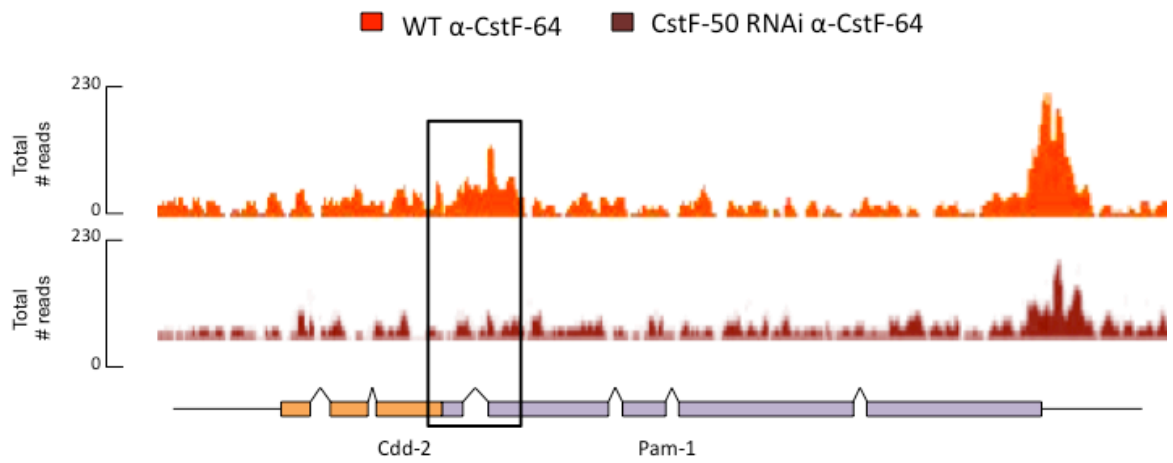


Figure V-13. CstF-64 binding to 3' ends in a three-gene operon (CEOP 4228) depends on CstF-50: ChIP-seq experiment with an antibody against CstF-64 in WT (red) and CstF-50 RNAi samples (dark red). See Figure V-11 caption for details.

In summary, I found that the levels of CstF-64 in all SL1-type operon 3' ends show a large reduction when CstF-50 has been knocked-down, suggesting formation of the CstF complex at these 3' ends (Figure V-9 – V-13). Importantly, this same reduction is seen at canonical 3' ends (SL2-type and terminal), consistent with the presence of CstF-64 as part of the CstF complex needed for 3' end formation at these 3' ends.

Some 3' ends are more drastically reduced than others, and this effect does not seem to be associated with a particular type of 3' end. For example, the levels of CstF-64 at the *mev-1* and *c09g4.4* internal 3' ends (SL1-type) are reduced but not gone upon CstF-50 knockdown. The same is true for three SL2-type operon 3' ends (b0024.11, *rnp-4*, and *rnp-7*) and four terminal 3' ends (i.e. *ced-9*, *nuo-4*, b0024.10, and *pam-1*). On the other hand, CstF-64 levels are completely eliminated on some SL1-type operon 3'

ends as well as canonical 3' ends. This is the case for the *snfc-5*, k04g7.11, *gna-1*, and *cdd-2* internal 3' ends (SL1-type), and for canonical internal 3' ends (SL2-type) like *mes-6* and terminal 3' ends like r07e5.1, and *cks-1*. The reason for the difference between 3' ends in the degree of reduction in CstF-64 upon CstF-50 knockdown is not completely clear. A likely possibility is that gene 3' ends need different levels of CstF-50 for the recruitment of CstF-64. In other words, since RNAi is unable to eliminate all the CstF-50 protein from the cell, the remaining CstF-50 may be sufficient for the CstF-64 recruitment seen in certain 3' ends but not for others. Alternatively, some 3' ends may be able to bind CstF-64 independently of CstF-50.

I conclude that CstF-64 recruitment at SL1-type operon 3' ends depends on the levels of CstF-50, similar to the dependency seen at canonical 3' ends within the same operons. This result is consistent with the findings of Fong and Bentley (2001), showing that CstF-50 binds the CTD of RNAPII and that this interaction is needed for the co-transcriptional formation of the CstF complex. Importantly, this result provides evidence that the trimeric CstF complex is assembled at this rare type of 3' end, most likely for functioning in 3' end formation.

CstF-50 is needed for cleavage at SL1-type operon 3' ends

Chromatin immunoprecipitation is a powerful technique that can be used for *in vivo* localization of RNA binding proteins bound to the transcription site (Swinburne et al., 2006). However, interpreting ChIP results can be taken only so far because the binding of processing factors does not indicate function. Therefore, I took a reverse genetic approach in order to test if the cleavage event in SL1-type operon 3' ends is processed

in a CstF-dependent manner. I expected that if CstF is playing a functional role at these 3' ends, then knockdown of an essential component of this complex should interfere with 3' end formation of SL1-type operons.

I used RNAi to knockdown the 50 KDa subunit of the CstF complex (Figure V-8). Then, I used RT-PCR experiments to measure cleavage and mRNA levels at four SL1-type operon 3' ends. If SL1 trans-splicing performs the cleavage at SL1-type operon 3' ends, then 3' end formation at these sites should be unaffected by CstF-50 knockdown. However, if CstF plays an important role, the knockdown might cause unprocessed RNA to accumulate if SL1 trans-splicing is slow. I found that at all 3' ends the reduction of CstF-50 caused accumulation of unprocessed RNA.

As controls, I tested by RT-PCR if the worm homolog of CstF-50 is needed for 3' end formation at canonical 3' ends (Figure V-14). Cleavage was tested using random primed cDNA with a PCR amplicon spanning the cleavage site as defined by ESTs. Using the same RNA samples, I measured the levels of polyadenylated mRNA in oligo-dT cDNA with primers within an exon of the upstream gene. As shown in Figure V-14, at two internal 3' ends (SL2-type) and at two terminal 3' ends the level of uncleaved RNA increased upon Cst-50 knockdown. It should be noted that the 4 genes tested for cleavage are not the same genes used to measure the levels of processed RNA, except for b0024.11. The reason for this is the unavailability of primer sets covering exons in these genes. In any case, the levels of processed RNA were reduced when CstF-50 was knocked-down, which most likely is due to an effect on cleavage. These results confirmed that the worm CstF-50 is needed for 3' end cleavage both at SL2-type and terminal 3' ends, consistent with its function in other organisms.

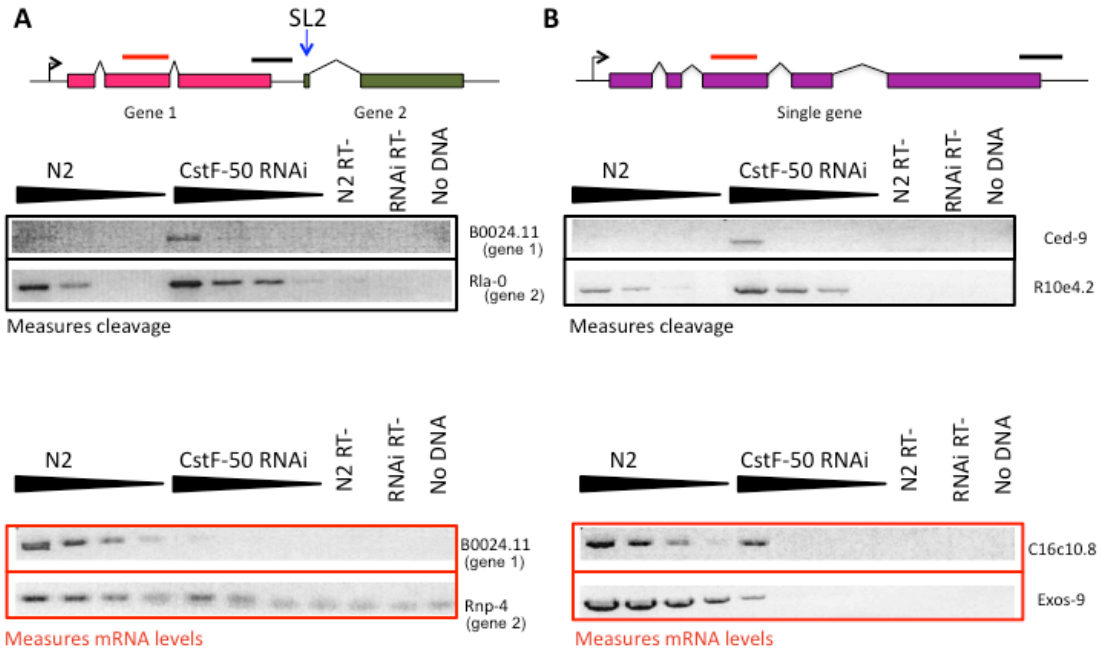


Figure V-14. CstF-50 is needed for 3' end formation in SL2-type operons and single genes: RT-PCR experiments in SL2-type operons (A) and in single genes (B) when CstF-50 levels were reduced. Gene model examples are shown at the top of the figure with the location of the SL2 trans-splice site shown with a blue arrow. Gels in black boxes represent experiments where cleavage was measured using random primed cDNA and primers spanning the cleavage site (black line). Gels in red boxes represent experiments where processed mRNA was measured using oligo dT cDNA and primers within an exon (red line). Gene names are shown to the right of the gels.

Next, I tested whether CstF-50 plays a role in cleavage in SL1-type operon 3' ends (Figure IV-15). For these experiments, I isolated the CstF-50 -/- worms from the balanced deletion strain using a worm sorter, so no RNAi was needed. I used RT-PCR experiments to measure 3' end cleavage in SL1-type operon 3' ends. I found that in all four SL1-type operon 3' ends the levels of uncleaved RNA increased when CstF-50 was deleted. For example, cleavage at the *mev-1* internal 3' end (SL1-type) was at least 3 fold decreased compared to the WT control. In addition, using the same RNA samples I found that the mRNA levels of upstream genes in SL1-type operons were significantly

reduced when CstF-50 was knocked out, consistent with CstF being required for 3' end formation in these genes. However, I can not rule out the possibility that CstF-50 may be needed for SL1 trans-splicing in this type of operon. Attempts to measure SL1 trans-splicing in downstream genes of SL1-type operons when the levels of CstF-50 were knocked-down gave me inconsistent results. The effect on 3' end formation that I see on SL1-type operon 3' ends is similar to the one seen at canonical 3' ends. Therefore, I conclude that SL1-type operon 3' ends need CstF-50 for 3' end formation.

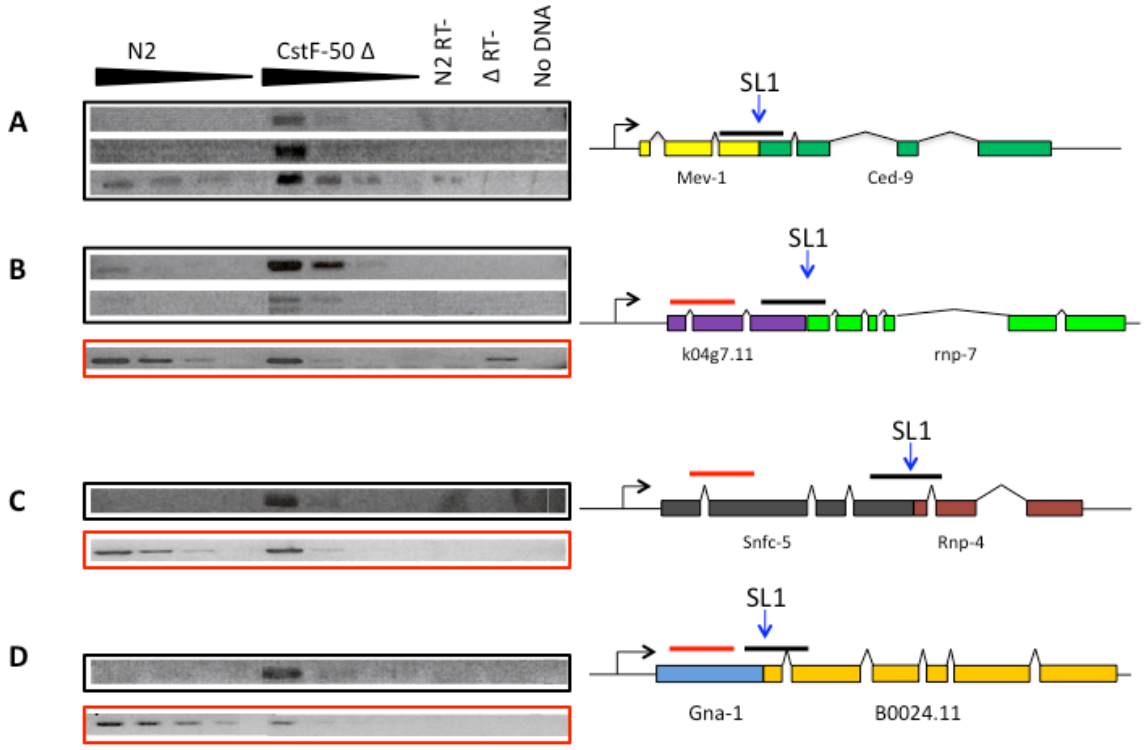


Figure V-15. CstF-50 is needed for 3' end formation in SL1-type operons: RT-PCR experiments in SL1-type operons when CstF-50 levels has been either reduced or knocked out. Operons analyzed are shown to the right of the gels with the SL1 trans-splice site shown with a blue arrow. Gels in black boxes represent experiments where cleavage was measured using random primed cDNA and primers spanning the cleavage site (black line). Gels in red boxes represent experiments where processed mRNA was measure using oligo dT cDNA and primers within an exon (red line). Multiple gels in a single box represent independent RT repeats.

Discussion

SL1-type and SL2-type operon 3' ends appear to be processed by similar mechanisms

According to the transcriptome data, cleavage and subsequent polyadenylation of SL1-type operon 3' ends occurs right at the trans-splice site. Therefore, trans-splicing instead of the canonical 3' end formation machinery may be responsible for cleavage at SL1-type operon 3' ends. If this were the case, then I would have expected to find differences between SL1 and SL2-type operon 3' end formation mechanisms. In chapter III, I showed that Ser-2p could be used to indirectly mark sites where 3' end formation occurs by the canonical 3' end formation machinery. Therefore, in order to provide insights into the cleavage mechanism at 3' ends of SL1-type operons, I tested if this type of 3' end is treated differently from SL2-type operon 3' ends regarding Ser-2 phosphorylation and CstF-64 binding.

First I showed that Ser-2p is indeed associated with SL1-type operon 3' ends. Importantly, this association is indistinguishable from SL2-type operon 3' ends, indicating that both types of 3' ends are processed by similar mechanisms. Therefore, this result suggests that Ser-2p is marking both types of 3' ends perhaps for the recruitment of 3' end processing factors.

Phosphorylation of Ser-2 is also needed for recruitment of co-transcriptional chromatin modifiers to the transcription site, such as the H3K36 methyltransferase Set-2 (Li et al., 2002). So the presence of Ser-2p *per se* is not capable of demonstrating a particular mode of 3' end formation. Furthermore, a bioinformatic analysis of SL1-type operons showed absence of the canonical CstF-64 binding site downstream of the poly-

A site (Graber et al., 2007), consistent with cleavage occurring in a CstF independent manner. Therefore, I looked directly for the presence of CstF-64 at SL1-type operon 3' ends. I found that CstF-64 is recruited to the 3' ends of SL1-type operons, similar to the binding present at SL2-type operon 3' ends. Importantly, CstF-64 and Ser-2p are colocalized at SL1-type operon 3' ends, suggesting that Ser-2p's role at this type of 3' end may be for the co-transcriptional recruitment of the CstF complex. Since the canonical CstF-64 binding site is missing downstream of the cleavage site, the CstF-64 present at SL1-type operon 3' ends may be bound through its interaction with CPSF, the upstream poly-Y tract, or RNAPII.

I conclude that SL1-type operon 3' ends behave similarly to SL2-type operon 3' ends with respect to CTD phosphorylation and recruitment of CstF-64. These observations predicted that cleavage at SL1-type operon 3' ends would occur through a similar mechanism to the cleavage event at SL2-type operon 3' ends.

The CstF-64 recruited to SL1-type operon 3' ends is part of a complex

The tri-subunit CstF complex is not known to exist in worms, although all three subunits have clear homologs in the genome. Experiments done by Peg MacMorris in the lab suggested that the CstF subunits are not always present as a complex in worms, as immunoprecipitation with either anti-CstF-64 or anti-CstF-50 antibodies precipitated little of the other subunit (unpublished results). Interestingly, a related observation was noted in mammals, in which ChIP experiments with anti-CstF-64 and anti-CstF-77 antibodies showed a differential recruitment of these proteins. CstF-64 was recruited at the promoter while CstF-77 was only recruited at the 3' end of genes (Glover-Cutter et al.,

2008). The presence of CstF-64 at SL1-type operons independent of the complex would be an indication that CstF-64 may be playing a role other than cleavage. Therefore, the possibility existed that CstF-64 could have been present at SL1-type operon 3' ends independent of the other subunits. Thus, I tested whether the CstF-64 was present at 3' ends of SL1-type operons as part of the CstF complex known to be involved in cleavage by the 3' end formation machinery.

Knockdown of one subunit within a complex often leads to destabilization of the remaining subunits in the complex. For example, knockdown of one of the subunits of the tetrameric clathrin adaptor complex causes a codepletion of the other subunit (Motley et al., 2003). Importantly, this is also the case for the CstF complex in flies, in which knockdown of any of the subunits within the complex caused a codepletion of the other subunits (Sullivan et al., 2009). Thus, I RNAi'ed the CstF-50 subunit in an attempt to destabilize the CstF complex and ask by ChIP if CstF-64 is still being recruited to SL1-type operon 3' ends.

For reasons that are still unclear, knockdown of CstF-50 did not deplete CstF-64 as measured by Western blot analysis. Interestingly, I found by ChIP that, in all five SL1-type operon 3' ends analyzed, CstF-64 binding was dependent on CstF-50 levels, consistent with a destabilization of the CstF complex associated with the RNAPII holoenzyme. Moreover, this result provides further support for the co-transcriptional recruitment of the CstF complex by the interaction of CstF-50 with the CTD of RNAPII (Fong and Bentley 2001). So when CstF-50 is not present the CstF-64 subunit can't interact with RNAPII and consequently is not recruited to the transcription site.

Importantly, I found that CstF-64 binding was also dependent on CstF-50 at SL2-type and terminal 3' ends in operons. Therefore, the CstF-64 binding seen at SL1-type operon 3' ends is indistinguishable from the binding seen at SL2-type operon 3' ends, providing evidence that cleavage at SL1-type operons 3' ends may occur through the canonical 3' end formation machinery.

The CstF-50 worm ortholog (cpf-1) is involved in 3' end formation

The worm CstF-50 ortholog has been established based on amino acid conservation, but it has never been shown that the worm CstF-50 product plays a role in 3' end formation, as it is known to do in other organisms. I used RNAi on a deletion-balanced allele (CGC) of CstF-50 to significantly reduce the levels of CstF-50 in the cell. RT-PCR experiments verified that the mRNA of CstF-50 was reduced at least 10 fold.

I used RT-PCR to measure cleavage at 3' ends of non-operon genes and SL1-type operon genes, in order to provide evidence for CstF-50 function in 3' end formation. I showed by RT-PCR that the worm CstF-50 ortholog is indeed involved in 3' end formation. Uncleaved RNA from both SL2-type operon 3' ends and non-operon 3' ends accumulated upon CstF-50 knockdown. Moreover, the amount of correctly processed RNA was reduced, consistent with lack of 3' end formation at these 3' ends. Therefore, I concluded that the worm CstF-50 ortholog plays a role in 3' end formation as in other organisms.

The role of CstF-50 in forming SL1-type operon 3' ends

Next, I used genetics to provide functional evidence that CstF is actually providing the cleavage at SL1-type operons 3' ends. Since I can successfully knockdown CstF-50 in worms, I asked if uncleaved RNA accumulated at SL1-type operon 3' ends. If cleavage is occurring through trans-splicing at SL1-type operon 3' ends, then knocking-down CstF-50 should have had no effect on cleavage.

Surprisingly, all SL1-type operon 3' ends tested needed CstF-50 for processing of their 3' ends, similarly to SL2-type operon and single genes 3' ends. I found that upon knocking-down CstF-50 the levels of uncleaved RNA accumulated. Moreover, I showed that for the same genes the amount of correctly processed mRNA decreased, indicative of failure to properly process the pre-mRNA. Indeed, in *S. cerevisiae,* knockdown of RNA-15 (CstF-64), RNA-14 (CstF-77) or PAP1 causes 3' end processing defects that leads to down-regulation of the transcript by the nuclear exosome (Schmid and Jensen 2008). Therefore, these results are consistent with CstF-50 playing a role in cleavage at SL1-type operon 3' ends. Unfortunately, SL1 trans-splicing to the downstream mRNA can't be accurately measured since its transcript level will be affected by CstF-50 knockdown.

How are SL1-type operon 3' ends cleaved?

The results presented in this chapter provide evidence that cleavage at SL1-type operons occurs through the canonical CPSF-CstF complex, similar to other 3' ends. Furthermore, the data is consistent with the model proposed by Williams et al., (1999) based on mutational analysis of a transgenic SL1-type operon. According to this model both genes in this type of operons are not expressed from the same pre-mRNA molecule, because 3' end formation is in competition with trans-splicing. If cleavage occurs first at SL1-type operon 3' ends, this will destroy the trans-splice site sequence causing the downstream precursor to be uncapped and consequently degraded. However, if SL1 trans-splicing occurs first then this will presumably leave the upstream RNA branched which could result in its degradation. So only the downstream gene would be expressed. Thus the Ser-2p and CstF-64 binding seen at SL1-type operon 3' ends may be used by the 3' end formation machinery for cleaving the pre-mRNA, but perhaps only when the competition is pushed towards 3' end formation.

Competition at SL1-type operon 3' ends between trans-splicing and 3' end formation could be a way to regulate gene expression at the level of RNA processing. The choice between which gene is expressed may be determined by the relative amount of SL1 snRNP and 3' end formation factors present in the cell. If more SL1 snRNP molecules are present around the transcription site, the system would be pushed towards trans-splicing and the downstream gene would predominate. In contrast, if an excess of 3' end formation factors exist then the system would be pushed towards 3' end formation and the upstream gene will be predominant.

The experiments presented in this section suggest that the canonical 3' end formation machinery cleaves SL1-type operons. However, these experiments do not rule out the possibility that sometimes SL1 trans-splicing could be providing the cleavage at this rare type of gene arrangement, followed by CPSF-CstF dependent polyadenylation of the free 3' end. It is important to note that my experiments were done on whole worms, which contain many different cell-types. Therefore, these experiments are based on a mixed population of cells in which SL1-type operons could be potentially processed differently in different cell types and/or at different developmental stages or even on different pre-mRNAs in the same cell. However, more experiments are required to determine if SL1-type operons are on some occasions cleaved by trans-splicing.

Why does 3' end formation at SL1-type operon 3' ends occur at the same dinucleotide (AG) used by trans-splicing?

The transcriptome data showed that polyadenylation at SL1-type operon 3' ends occurs after the AG dinucleotide. Since cleavage after an AG dinucleotide is not commonly used by the canonical 3' end formation machinery (Chen et al., 1995; Mandel et al., 2008), it was suggested that trans-splicing might be providing the cleavage at this type of 3' end. However, the data presented in this chapter challenges this possibility since SL1-type operon 3' ends are treated similarly to SL2-type operon 3' ends with respect to Ser-2p, recruitment of CstF-64 and the need for CstF-50 in processing their 3' ends. Therefore, why are SL1-type operon 3' ends processed at the same site as trans-splicing?

A reasonable possibility is that 3' end formation at the AG dinucleotide of SL1-type operon 3' ends represents only the cleavage products from trans-splicing. The data presented in this chapter does not eliminate the possibility that in some cases SL1 trans-splicing is providing the cleavage at SL1-type operon 3' ends. So it might be the case that 3' end formation at SL1-type operon 3' ends can occur either through the canonical 3' end formation machinery or through the trans-splicing machinery, which is also consistent with the experimental results of Williams et al. (1999). Indeed, not all the reads in the transcriptome data are polyadenylated at the AG dinucleotide, suggesting that sometimes cleavage occurs at others locations. Hence, the RNA polyadenylated at the AG dinucleotide might represent a cleavage product from trans-splicing, whereas the RNA polyadenylated at nearby locations might represent a cleavage product from the canonical 3' end formation machinery. Further experiments are needed in order to test this idea.

An alternative model could be that the canonical 3' end formation machinery might be capable of cleaving after an AG dinucleotide at SL1-type operon 3' ends. Could this be related to the fact that SL1-type operons lack a CstF-64 binding site downstream of the cleavage site (Graber et al., 2007)? Since, CstF-64 binding downstream of the cleavage site has been shown to be needed for normal cleavage at 3' ends of genes (Edwalds-Gilbert and Milcarek 1995; Takagaki et al., 1996; Chan et al., 2011), perhaps lack of CstF-64 binding site, combined with the presence of CstF-64 bound to the RNAPII CTD, results in a different mode of cleavage. The absence of CstF-64 on the RNA downstream of the cleavage site could provide some flexibility for the 3' end formation machinery to cleave the pre-mRNA at a non-canonical dinucleotide

sequence. Finally, the presence of the SL1 trans-splicing event occurring at this location could influence the site of canonical CstF-dependent cleavage.

CHAPTER VI

SUMMARY, CONCLUSIONS AND FUTURE DIRECTIONS

The discovery that 3' end formation is mechanistically linked to transcription termination (Whitelaw and Proudfoot 1986) introduces a unique problem for *C. elegans* operons. Operons are composed of many 3' end formation signals under the control of a single promoter, however transcription does not terminate until it reaches the end of the operon (Blumenthal 2005). The mechanism of how RNAPII is able to transcribe internal poly-A signals without causing transcription termination remains unknown. Therefore, understanding how 3' end formation can occur in the absence of termination will provide invaluable insights into both of these processes.

CHAPTER III: Does RNAPII CTD phosphorylation mark sites of pre-mRNA processing or transcriptional events?

In vivo and *in vitro* experiments have shown that several pre-mRNA processing factors bind the phosphorylated form of RNAPII CTD, leading to the proposal that these phosphorylations may guide the co-transcriptional processing of the pre-mRNA (Corden 1990; Greenleaf 1993; Perales and Bentley 2009; Phatnani and Greenleaf 2006). However, the possibility remains that these phosphorylations events are not needed for guiding pre-mRNA processing, but instead for events following pre-mRNA processing, such as transcription termination. Operons separate pre-mRNA processing sites from transcriptional events, thus allowing me to test in a unique context if these phosphorylation events of RNAPII CTD correlate with pre-mRNA processing or

transcriptional events. In this work I provide evidence answering the following two questions regarding CTD phosphorylation in operons:

1. Does Ser-5p associate with 5' ends processed by co-transcriptional capping or all 5' ends, even those distant from promoters?

   ChIP experiments done in yeast and mammals have shown that Ser-5p is associated with 5' ends of genes (Gomes et al 2006; Kim et al., 2009; Komarnitsky et al., 2000; Rosonina et al., 2006; Perales and Bentley 2009). Thus, Ser-5p could be either associated with co-transcriptional capping or with all 5' ends, even those distant from promoters. I found that Ser-5p was associated with sites processed by co-transcriptional capping, but not with sites distant from promoters. This result provides a novel and independent support for the conclusion that Ser-5p is needed for the co-transcriptional recruitment of capping enzymes. Furthermore, the 5' ends processed by SL2 trans-splicing does not colocalize with high levels of Ser-5p, indicating that this processing event might occur without the need of Ser-5p.
2. Does Ser-2p associate with 3' end formation or transcription termination?

   In contrast to Ser-5p, ChIP experiments done in other organisms have shown that Ser-2p is associated with 3' ends of genes (Gomes et al 2006; Kim et al., 2009; Komarnitsky et al., 2000; Rosonina et al., 2006; Perales and Bentley 2009). However, in this case Ser-2p could be either functioning in 3' end formation or transcription termination. I found that Ser-2p is associated with all 3' ends in operons, even those not associated with transcription termination. This result provides novel and independent evidence that, again, this phosphorylation may be

functioning in 3' end formation rather than transcription termination. Furthermore, this indicates that Ser-2p cannot be sufficient for causing transcription termination, since Ser-2p is present at 3' ends in which termination does not occur following 3' end formation.

Further experiments could be performed in order to show that similar phosphorylation patterns are found associated with other genes and operons in the genome. Do all genes in the genome contain patterns of Ser-5p and Ser-2p consistent with the examples shown in Chapter III?, or do they differ in their phosphorylation pattern? If they do, then it will be interesting to find out if genes that have similar phosphorylation patterns correlate with some type of gene characteristic (i.e. length, expression level, operon vs. non-operon). I plan to answered these questions by performing ChIP-seq experiments with Ser-5p and Ser-2p antibodies.

Revealing the association of Ser-7p and other CTD phosphorylations (i.e. Tyr-1 and Thr-4) and modifications (i.e. glycosylation and isomerization) within operons will provide invaluable insights in revealing the function of these post-translation modifications. For example, the ChIP signal for Ser-7p is biased towards the 5' ends of genes, similarly to Ser-5p (Kim et al., 2009; Glover-Cutter et al., 2009). Thus it will be interesting to use operons to test if these phosphorylations events colocalize with each other, or if Ser-7p marks a separate event associated with 5' ends.

Studies revealing the chromatin association of the kinases/phosphatases responsible for the phosphorylation/dephosphorylation of the bulk of Ser-5 (CDK-7/RTR-1) and Ser-2 (CDK-9/FCP-1) could provide further support for the conclusions of

found in this chapter. Importantly, these experiments should provide insights into the mechanisms of CTD phosphorylation and dephosphorylation along genes. For example, since Ser-2p peaks at each 3' end within an operon, will cdk-9 mimic this pattern or will it be recruited at the beginning of the operon and released at the terminal 3' end. Similar questions for the Ser-5p kinase (cdk-7) as well as for the RTR-1 and FCP-1 phosphatases will be interesting to pursue.

Furthermore, genetic experiments designed for knocking-down the kinases/phosphatases responsible for Ser-5p and Ser-2p could provide strong support for the specificities of the Ser-5p and Ser-2p antibodies. Also, these experiments might provide insights into whether other kinases or phosphatases are responsible for Ser-5 and Ser-2 phosphorylation/dephosphorylation.

CHAPTER IV: Does CstF-64 associate with 3' end formation or transcription termination?

In higher eukaryotes 3' end formation of the mRNA leads to transcription termination one to several kilobases downstream. For example in mammals, ChIP experiments have shown that CstF-64 binding is biased towards the 3' end of genes (Glover-Cutter et al., 2008), making it impossible to conclude for certain whether CstF-64 is functioning in 3' end formation or transcription termination. I used operons to show that one of the subunits of the CstF 3' end formation complex, CstF-64, is present at 3' ends of internal genes, thus suggesting that CstF-64 is not sufficient for transcription termination. Importantly, CstF-64 colocalizes with Ser-2p at 3' ends of genes, indicating that CstF-64 could be recruited by Ser-2p at 3' end of *C. elegans* genes.

Also in this chapter I provide *in vivo* evidence showing that RNAPII present at internal 3' ends in operons shows the characteristics associated with RNAPII at the 3' end of mammalian genes (Glover-Cutter et al., 2008). I found that at 3' ends of internal genes in operons RNAPII is paused (Figure IV-19 and IV-20), suggesting that RNAPII pausing is not sufficient for termination. Moreover, RNAPII pausing at internal 3' ends colocalizes with maximal Ser-2p and with CstF-64, similar to the pattern seen at terminal 3' ends in which termination does occur. Therefore, these results are consistent with the model proposed by Liu et al. (2003), in which RNAPII is being prevented from terminating at internal 3' ends in operons by a still unidentified factor bound to the Ur element, which is believed to block the passage of the exonuclease capable of torpedoing RNAPII off the DNA, presumably XRN-2 (West et al., 2004; Kim et al., 2004).

Further experiments need to be performed in order to find if the other CstF subunits are also associated with CstF-64 at 3' ends of genes. The Blumenthal laboratory is currently making antibodies against the 50 and 77 KDa subunits, in order to reveal the association of the other CstF subunits with genes. These experiments should also provide insights into the order of recruitment of the individual subunits needed for the assembly of the CstF complex at 3' ends of genes. Moreover, they should provide evidence for how the CstF complex is assembled at 3' ends of genes. For example, is the CstF complex recruited as a pre-assembled complex or are the CstF subunits recruited separately and assembled at the 3' end?

Revealing the association of other 3' end formation/termination factors, such as CPSF, XRN-2 and PCF-11, within operons would allow me to decipher which factors

may be involved in either 3' end formation or transcription termination. CPSF together with CstF represent the core of the 3' end formation complex needed for specifying the site of polyadenylation (Mandel et al., 2008; Shi et al., 2009; Chan et al., 2011). An antibody agaisnt CPSF-160 is currently being made by the Blumenthal laboratory. Thus, I plan to perform a ChIP experiment with CPSF-160 and compare it to the CstF-64 profile. This will provide additional support for sites in which 3' end formation is indeed occurring. Interestingly, the sites that don't show a colocalization of both proteins will indicate that whatever factor is bound might be performing a role other than 3' end formation.

The CstF-64 colocalization with the GRO-seq data from Will Kruesi (Mayer Lab at UC Berkley) provides strong support for the idea that CstF-64 bound to RNAPII pauses at 3' ends of genes, irrespective of transcription termination (Figure IV-23). Thus, it will be interesting to find out if RNAPII pauses at all other 3' ends in the genome, or only those associated with CstF-64. I have preliminary data not included in this thesis that indicates that when CstF-50 is knocked-down, CstF-64 binding at most 3' ends is reduced. Thus, it will be interesting to find out if the reduction in CstF-64 binding is due to a decrease in RNAPII pausing or if it is due to the inability of CstF-64 to be recruited in the absence of CstF-50.

CHAPTER V: In SL1-type operons are 3' ends cleaved by trans-splicing?

In this chapter, I investigated a possible novel 3' end formation mechanism in a rare type of *C. elegans* operon. SL1-type operons contain distinctive characteristics in gene organization that are not shared with SL2-type operons. Based on the transcriptome

data showing that 3' ends are formed right at the AG dinucleotide used by trans-splicing, it was proposed by this laboratory that this type of operon could be processed by a novel 3' end formation mechanism involving trans-splicing. This hypothesis has been previously tested. Nevertheless the results appear to be in conflict with the transcriptome data (Williams et al., 1999), with some results indicating 3' end formation by trans-splicing and other results not.

I took a genome-wide approach in order to resolve the discrepancy between the transcriptome data and the experimental results (Williams et al., 1999). I revealed that Ser-2p and CstF-64 are present at 3' ends of internal genes in operons, similar to the pattern seen for SL2-type operons. Importantly, I showed that CstF-50 is needed for cleavage of SL1-type operons, similar to the dependency present at SL2-type operons. My results are consistent with the experimental data (Williams et al., 1999) suggesting that 3' end formation through the canonical 3' end formation machinery is indeed occurring at this type of 3' end.

The experiments presented in this chapter do not eliminate the possibility that in some cases trans-splicing may be providing the cleavage at this type of 3' end. Further experiments need to be done to determine if following trans-splicing of the downstream RNA, the upstream RNA is capable of being polyadenylated, thus allowing both genes to be expressed from the same polycistronic transcript. I have sequenced the RNA from the CstF-50 RNAi'ed sample, data I already have, but have not yet analyzed. This data should show if polyadenylation of the upstream genes in SL1-type operons is occurring when CstF-50 has been knocked-down. Since CstF is only involved in cleavage and not in the polyadenylation step of 3' end formation, then by knocking down CstF-50,

cleavage of the RNA by the 3’ end formation machinery should not occur. This will allow me to detect polyadenylation of processed RNA by trans-splicing. Similarly, an in vitro 3’ end formation assay could be performed in which an artificial SL1-type operon is incubated with worm crude extract depleted of CstF-50. Here again, the polyadenylation of the upstream RNA following cleavage by trans-splicing could be measured, since cleavage can’t occur through the canonical 3’ end formation machinery.

Since polyadenylation occurs at the AG in SL1-type operon 3’ ends, it is currently unclear how the 3’ end formation machinery is able to cleave after an AG dinucleotide. Perhaps some 3’ ends are formed by trans-splicing and some by the conventional 3’ end formation machinery. Re-analyzing the RNA-seq data looking for nearby 3’ end processing sites could provide insights, if the 3’ end processing machinery cuts at other sites and trans-splicing cuts after the AG. If the majority of reads are found to be at the AG dinucleotide, then it is still possible that the 3’ end formation machinery is now capable of cleaving after an AG dinucleotide when it is associated with the SL1 snRNP.

REFERNCES

Agalioti T, Chen G, Thanos D. Deciphering the transcriptional histone acetylation code for a human gene. Cell. 2002 Nov 1;111(3):381-92.

Ahn SH, Kim M, Buratowski S. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3' end processing. Mol Cell. 2004 Jan 16;13(1):67-76.

Allen MA, Hillier LW, Waterston RH, Blumenthal T. A global analysis of C. elegans trans-splicing. Genome Res. 2011 Feb;21(2):255-64.

Allison LA, Wong JK, Fitzpatrick VD, Moyle M, Ingles CJ. 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. 1988 Jan;8(1):321-9.

Ansari A, Hampsey M. A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping. Genes Dev. 2005 Dec 15;19(24):2969-78. Epub 2005 Nov 30.

Arhin GK, Boots M, Bagga PS, Milcarek C, Wilusz J. Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals. Nucleic Acids Res. 2002 Apr 15;30(8):1842-50.

Barabino SM, Hubner W, Jenny A, Minvielle-Sebastia L, Keller W. The 30-kD subunit of mammalian cleavage and polyadenylation specificity factor and its yeast homolog are RNA-binding zinc finger proteins. Genes Dev. 1997 Jul 1;11(13):1703-16.

Bartkowiak B, Greenleaf AL. Phosphorylation of RNAPII: To P-TEFb or not to P-TEFb? Transcription. 2011 May;2(3):115-9.

Baugh LR, Demodena J, Sternberg PW. RNA pol II accumulates at promoters of growth genes during developmental arrest. Science. 2009 Apr 3;324(5923):92-4.

Bauren G, Belikov S, Wieslander L. Transcriptional termination in the balbiani ring 1 gene is closely coupled to 3'-end formation and excision of the 3'-terminal intron. Genes Dev. 1998 Sep 1;12(17):2759-69.

Beyer AL, Osheim YN. Splice site selection, rate of splicing, and alternative splicing on nascent transcripts. Genes Dev. 1988 Jun;2(6):754-65.

Birse CE, Minvielle-Sebastia L, Lee BA, Keller W, Proudfoot NJ. Coupling termination of transcription to messenger RNA maturation in yeast. Science. 1998 Apr 10;280(5361):298-301.

Blumenthal T. Trans-splicing and operons. WormBook. 2005:1-9.

Blumenthal T, Evans D, Link CD, Guffanti A, Lawson D, Thierry-Mieg J, et al. A global analysis of caenorhabditis elegans operons. Nature. 2002 Jun 20;417(6891):851-4.

Blumenthal T, Davis P. Operons and non-operon gene clusters in the c. elegans genome. Wormbook 2012. In preparation.

Brown PH, Tiley LS, Cullen BR. Effect of RNA secondary structure on polyadenylation site selection. Genes Dev. 1991 Jul;5(7):1277-84.

Buratowski S. Progression through the RNA polymerase II CTD cycle. Mol Cell. 2009 Nov 25;36(4):541-6.

Calvo O, Manley JL. Strange bedfellows: Polyadenylation factors at the promoter. Genes Dev. 2003 Jun 1;17(11):1321-7.

Calvo O, Manley JL. Evolutionarily conserved interaction between CstF-64 and PC4 links transcription, polyadenylation, and termination. Mol Cell. 2001 May;7(5):1013-23.

Carswell S, Alwine JC. Efficiency of utilization of the simian virus 40 late polyadenylation site: Effects of upstream sequences. Mol Cell Biol. 1989 Oct;9(10):4248-58.

Chan S, Choi EA, Shi Y. Pre-mRNA 3'-end processing complex assembly and function. Wiley Interdiscip Rev RNA. 2011 May;2(3):321-35.

Chapman RD, Heidemann M, Albert TK, Mailhammer R, Flatley A, Meisterernst M, et al. Transcribing RNA polymerase II is phosphorylated at CTD residue serine-7. Science. 2007 Dec 14;318(5857):1780-2.

Chen F, MacDonald CC, Wilusz J. Cleavage site determinants in the mammalian polyadenylation signal. Nucleic Acids Res. 1995 Jul 25;23(14):2614-20.

Chiu YL, Ho CK, Saha N, Schwer B, Shuman S, Rana TM. Tat stimulates cotranscriptional capping of HIV mRNA. Mol Cell. 2002 Sep;10(3):585-97.

Cho EJ, Takagi T, Moore CR, Buratowski S. mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. Genes Dev. 1997 Dec 15;11(24):3319-26.

Cho EJ, Kobor MS, Kim M, Greenblatt J, Buratowski S. Opposing effects of Ctk1 kinase and Fcp1 phosphatase at Ser 2 of the RNA polymerase II C-terminal domain. Genes Dev. 2001 Dec 15;15(24):3319-29.

Colgan DF, Manley JL. Mechanism and regulation of mRNA polyadenylation. Genes Dev. 1997 Nov 1;11(21):2755-66.

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

Coppola JA, Field AS, Luse DS. Promoter-proximal pausing by RNA polymerase II in vitro: Transcripts shorter than 20 nucleotides are not capped. Proc Natl Acad Sci U S A. 1983 Mar;80(5):1251-5.

Corden JL. Tails of RNA polymerase II. Trends Biochem Sci. 1990 Oct;15(10):383-7.

Dahmus ME. Reversible phosphorylation of the C-terminal domain of RNA polymerase II. J Biol Chem. 1996 Aug 9;271(32):19009-12.

Danckwardt S, Hentze MW, Kulozik AE. 3' end mRNA processing: Molecular mechanisms and implications for health and disease. EMBO J. 2008 Feb 6;27(3):482-98.

Dantonel JC, Murthy KG, Manley JL, Tora L. Transcription factor TFIID recruits factor CPSF for formation of 3' end of mRNA. Nature. 1997 Sep 25;389(6649):399-402.

Dye MJ, Proudfoot NJ. Multiple transcript cleavage precedes polymerase release in termination by RNA polymerase II. Cell. 2001 Jun 1;105(5):669-81.

Dye MJ, Proudfoot NJ. Terminal exon definition occurs cotranscriptionally and promotes termination of RNA polymerase II. Mol Cell. 1999 Mar;3(3):371-8.

Edwalds-Gilbert G, Milcarek C. Regulation of poly(A) site use during mouse B-cell development involves a change in the binding of a general polyadenylation factor in a B-cell stage-specific manner. Mol Cell Biol. 1995 Nov;15(11):6420-9.

Evans D, Perez I, MacMorris M, Leake D, Wilusz CJ, Blumenthal T. A complex containing CstF-64 and the SL2 snRNP connects mRNA 3' end formation and trans-splicing in C. elegans operons. Genes Dev. 2001 Oct 1;15(19):2562-71.

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

Glover-Cutter K, Kim S, Espinosa J, Bentley DL. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. Nat Struct Mol Biol. 2008 Jan;15(1):71-8.

Gomes NP, Bjerke G, Llorente B, Szostek SA, Emerson BM, Espinosa JM. Gene-specific requirement for P-TEFb activity and RNA polymerase II phosphorylation within the p53 transcriptional program. Genes Dev. 2006 Mar 1;20(5):601-12.

Graber JH, Salisbury J, Hutchins LN, Blumenthal T. C. elegans sequences that control trans-splicing and operon pre-mRNA processing. RNA. 2007 Sep;13(9):1409-26.

Greenleaf AL. A positive addition to a negative tail's tale. Proc Natl Acad Sci U S A. 1993 Dec 1;90(23):10896-7.

Gromak N, West S, Proudfoot NJ. Pause sites promote transcriptional termination of mammalian RNA polymerase II. Mol Cell Biol. 2006 May;26(10):3986-96.

Gu M, Lima CD. Processing the message: Structural insights into capping and decapping mRNA. Curr Opin Struct Biol. 2005 Feb;15(1):99-106.

Haenni S, Sharpe HE, Gravato Nobre M, Zechner K, Browne C, Hodgkin J, et al. Regulation of transcription termination in the nematode caenorhabditis elegans. Nucleic Acids Res. 2009 Nov;37(20):6723-36.

Hall-Pogar T, Zhang H, Tian B, Lutz CS. Alternative polyadenylation of cyclooxygenase-2. Nucleic Acids Res. 2005 May 4;33(8):2565-79.

Hatton LS, Eloranta JJ, Figueiredo LM, Takagaki Y, Manley JL, O'Hare K. The drosophila homologue of the 64 kDa subunit of cleavage stimulation factor interacts with the 77 kDa subunit encoded by the suppressor of forked gene. Nucleic Acids Res. 2000 Jan 15;28(2):520-6.

Hengartner CJ, Myer VE, Liao SM, Wilson CJ, Koh SS, Young RA. Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases. Mol Cell. 1998 Jul;2(1):43-53.

Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. Genome Res. 2009 Apr;19(4):657-66.

Ho CK, Sriskanda V, McCracken S, Bentley D, Schwer B, and Shuman S. The Guanylyltransferase Domain of Mammalian mRNA Capping Enzyme Binds to the Phosphorylated Carboxyl-Terminal Domain of RNA Polymerase II. J. Biol. Chem. 1998273: 9577-9585.

Hockert JA, Yeh HJ, MacDonald CC. The hinge domain of the cleavage stimulation factor protein CstF-64 is essential for CstF-77 interaction, nuclear localization, and polyadenylation. J Biol Chem. 2010 Jan 1;285(1):695-704.

Huang P, Pleasance ED, Maydan JS, Hunt-Newbury R, O'Neil NJ, Mah A, et al. Identification and analysis of internal promoters in caenorhabditis elegans operons. Genome Res. 2007 Oct;17(10):1478-85.

Huang T, Kuersten S, Deshpande AM, Spieth J, MacMorris M, Blumenthal T. Intercistronic region required for polycistronic pre-mRNA processing in caenorhabditis elegans. Mol Cell Biol. 2001 Feb;21(4):1111-20.

Iwamoto S, Eggerding F, Falck-Pederson E, Darnell JE Jr. Transcription unit mapping in adenovirus: regions of termination. J Virol. 1986 Jul;59(1):112-9.

Jan CH, Friedman RC, Ruby JG, Bartel DP. Formation, regulation and evolution of caenorhabditis elegans 3'UTRs. Nature. 2011 Jan 6;469(7328):97-101.

Kaufmann I, Martin G, Friedlein A, Langen H, Keller W. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. EMBO J. 2004 Feb 11;23(3):616-26.

Keller W, Bienroth S, Lang KM, Christofori G. Cleavage and polyadenylation factor CPF specifically interacts with the pre-mRNA 3' processing signal AAUAAA. EMBO J. 1991 Dec;10(13):4241-9.

Kim M, Ahn SH, Krogan NJ, Greenblatt JF, Buratowski S. Transitions in RNA polymerase II elongation complexes at the 3' ends of genes. EMBO J. 2004 Jan 28;23(2):354-64.

Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, Greenblatt JF, et al. The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. Nature. 2004 Nov 25;432(7016):517-22.

Kim M, Suh H, Cho EJ, Buratowski S. Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2, 5, and 7. J Biol Chem. 2009 Sep 25;284(39):26421-6.

Komarnitsky P, Cho EJ, Buratowski S. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. Genes Dev. 2000 Oct 1;14(19):2452-60.

Kuehner JN, Pearson EL, Moore C. Unravelling the means to an end: RNA polymerase II transcription termination. Nat Rev Mol Cell Biol. 2011 May;12(5):283-94.

Kuersten S, Lea K, MacMorris M, Spieth J, Blumenthal T. Relationship between 3' end formation and SL2-specific trans-splicing in polycistronic caenorhabditis elegans pre-mRNA processing. RNA. 1997 Mar;3(3):269-78.

Lasda EL, Blumenthal T. Trans-splicing. Wiley Interdiscip Rev RNA. 2011 May;2(3):417-34.

Lee TI, Johnstone SE, Young RA. Chromatin immunoprecipitation and microarray-based analysis of protein location. Nat Protoc. 2006;1(2):729-48.

Legrand P, Pinaud N, Minvielle-Sebastia L, Fribourg S. The structure of the CstF-77 homodimer provides insights into CstF assembly. Nucleic Acids Res. 2007;35(13):4515-22.

Lewis JD, Izaurralde E. The role of the cap structure in RNA processing and nuclear export. Eur J Biochem. 1997 Jul 15;247(2):461-9.

Li J, Moazed D, Gygi SP. Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. J Biol Chem. 2002 Dec 20;277(51):49383-8.

Licatalosi DD, Geiger G, Minet M, Schroeder S, Cilli K, McNeil JB, Bentley DL. Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. Mol Cell. 2002 May;9(5):1101-11.

Liu Y, Huang T, MacMorris M, Blumenthal T. Interplay between AAUAAA and the trans-splice site in processing of a caenorhabditis elegans operon pre-mRNA. RNA. 2001 Feb;7(2):176-81.

Liu Y, Kuersten S, Huang T, Larsen A, MacMorris M, Blumenthal T. An uncapped RNA suggests a model for caenorhabditis elegans polycistronic pre-mRNA processing. RNA. 2003 Jun;9(6):677-87.

Logan J, Falck-Pedersen E, Darnell JE Jr, Shenk T. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. Proc Natl Acad Sci U S A. 1987 Dec;84(23):8306-10.

Luo W, Johnson AW, Bentley DL. The role of Rat1 in coupling mRNA 3'-end processing to transcription termination: Implications for a unified allosteric-torpedo model. Genes Dev. 2006 Apr 15;20(8):954-65.

MacDonald CC, Wilusz J, Shenk T. The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. Mol Cell Biol. 1994 Oct;14(10):6647-54.

Mandel CR, Bai Y, Tong L. Protein factors in pre-mRNA 3'-end processing. Cell Mol Life Sci. 2008 Apr;65(7-8):1099-122.

Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, et al. Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. Nature. 2006 Dec 14;444(7121):953-6.

Mangone M, Manoharan AP, Thierry-Mieg D, Thierry-Mieg J, Han T, Mackowiak SD, et al. The landscape of C. elegans 3'UTRs. Science. 2010 Jul 23;329(5990):432-5.

McCracken S, Fong N, Rosonina E, Yankulov K, Brothers G, Siderovski D, et al. 5'-capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. Genes Dev. 1997 Dec 15;11(24):3306-18.

McCracken S, Fong N, Yankulov K, Ballantyne S, Pan G, Greenblatt J, et al. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. Nature. 1997 Jan 23;385(6614):357-61.

Millevoi S, Vagner S. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. Nucleic Acids Res. 2010 May;38(9):2757-74.

Moreira A, Wollerton M, Monks J, Proudfoot NJ. Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. EMBO J. 1995 Aug 1;14(15):3809-19.

Moreno-Morcillo M, Minvielle-Sebastia L, Mackereth C, Fribourg S. Hexameric architecture of CstF supported by CstF-50 homodimerization domain structure. RNA. 2011 Mar;17(3):412-8.

Morris DP, Michelotti GA, Schwinn DA. Evidence that phosphorylation of the RNA polymerase II carboxyl-terminal repeats is similar in yeast and humans. J Biol Chem. 2005 Sep 9;280(36):31368-77.

Mosley AL, Pattenden SG, Carey M, Venkatesh S, Gilmore JM, Florens L, et al. Rtr1 is a CTD phosphatase that regulates RNA polymerase II during the transition from serine 5 to serine 2 phosphorylation. Mol Cell. 2009 Apr 24;34(2):168-78.

Moteki S, Price D. Functional coupling of capping and transcription of mRNA. Mol Cell. 2002 Sep;10(3):599-609.

Motley A, Bright NA, Seaman MN, Robinson MS. Clathrin-mediated endocytosis in AP-2-depleted cells. J Cell Biol. 2003 Sep 1;162(5):909-18.

Murthy KG, Manley JL. The 160-kD subunit of human cleavage-polyadenylation specificity factor coordinates pre-mRNA 3'-end formation. Genes Dev. 1995 Nov 1;9(21):2672-83.

Natalizio BJ, Muniz LC, Arhin GK, Wilusz J, Lutz CS. Upstream elements present in the 3'-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals. J Biol Chem. 2002 Nov 8;277(45):42733-40.

Noble CG, Walker PA, Calder LJ, Taylor IA. Rna14-Rna15 assembly mediates the RNA-binding capability of saccharomyces cerevisiae cleavage factor IA. Nucleic Acids Res. 2004 Jun 23;32(11):3364-75.

O'Sullivan JM, Tan-Wong SM, Morillon A, Lee B, Coles J, Mellor J, Proudfoot NJ. Gene loops juxtapose promoters and terminators in yeast. Nat Genet. 2004 Sep;36(9):1014-8. Epub 2004 Aug 15.

Perales R, Bentley D. "Cotranscriptionality": The transcription elongation complex as a nexus for nuclear transactions. Mol Cell. 2009 Oct 23;36(2):178-91.

Perkins KJ, Lusic M, Mitar I, Giacca M, Proudfoot NJ. Transcription-dependent gene looping of the HIV-1 provirus is dictated by recognition of pre-mRNA processing signals. Mol Cell. 2008 Jan 18;29(1):56-68.

Phatnani HP, Greenleaf AL. Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev. 2006 Nov 1;20(21):2922-36.

Plant KE, Dye MJ, Lafaille C, Proudfoot NJ. Strong polyadenylation and weak pausing combine to cause efficient termination of transcription in the human ggamma-globin gene. Mol Cell Biol. 2005 Apr;25(8):3276-85.

Proudfoot N. New perspectives on connecting messenger RNA 3' end formation to transcription. Curr Opin Cell Biol. 2004 Jun;16(3):272-8.

Proudfoot NJ. Genetic dangers in poly(A) signals. EMBO Rep. 2001 Oct;2(10):891-2.

Proudfoot NJ, Furger A, Dye MJ. Integrating mRNA processing with transcription. Cell. 2002 Feb 22;108(4):501-12.

Qu X, Perez-Canadillas JM, Agrawal S, De Baecke J, Cheng H, Varani G, et al. The C-terminal domains of vertebrate CstF-64 and its yeast orthologue Rna15 form a new structure critical for mRNA 3'-end processing. J Biol Chem. 2007 Jan 19;282(3):2101-15.

Rasmussen EB, Lis JT. In vivo transcriptional pausing and cap formation on three drosophila heat shock genes. Proc Natl Acad Sci U S A. 1993 Sep 1;90(17):7923-7.

Roh TY, Cuddapah S, Zhao K. Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. Genes Dev. 2005 Mar 1;19(5):542-52.

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

Rosonina E, Kaneko S, Manley JL. Terminating the transcript: Breaking up is hard to do. Genes Dev. 2006 May 1;20(9):1050-6.

Ryan K, Calvo O, Manley JL. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. RNA. 2004 Apr;10(4):565-73.

Schmid M, Jensen TH. The exosome: A multipurpose RNA-decay machine. Trends Biochem Sci. 2008 Oct;33(10):501-10.

Schroeder SC, Schwer B, Shuman S, Bentley D. Dynamic association of capping enzymes with transcribing RNA polymerase II. Genes Dev. 2000 Oct 1;14(19):2435-40.

Schwartz SH, Silva J, Burstein D, Pupko T, Eyras E, Ast G. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. 2008 Jan;18(1):88-103.

Sheets MD, Ogg SC, Wickens MP. Point mutations in AAUAAA and the poly (A) addition site: Effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. Nucleic Acids Res. 1990 Oct 11;18(19):5799-805.

Shi Y, Chan S, Martinez-Santibanez G. An up-close look at the pre-mRNA 3'-end processing complex. RNA Biol. 2009 Nov-Dec;6(5):522-5.

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

Simonelig M, Elliott K, Mitchelson A, O'Hare K. Interallelic complementation at the suppressor of forked locus of drosophila reveals complementation between suppressor of forked proteins mutated in different regions. Genetics. 1996 Apr;142(4):1225-35.

Singh BN, Hampsey M. A transcription-independent role for TFIIB in gene looping. Mol Cell. 2007 Sep 7;27(5):806-16.

Singh BN, Ansari A, Hampsey M. Detection of gene loops by 3C in yeast. Methods. 2009 Aug;48(4):361-7. Epub 2009 Mar 6.

Skourti-Stathaki K, Proudfoot NJ, Gromak N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Mol Cell. 2011 Jun 24;42(6):794-805.

Spieth J, Brooke G, Kuersten S, Lea K, Blumenthal T. Operons in C. elegans: Polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions. Cell. 1993 May 7;73(3):521-32.

Stiller JW, Hall BD. Evolution of the RNA polymerase II C-terminal domain. Proc Natl Acad Sci U S A. 2002 Apr 30;99(9):6091-6.

Sullivan KD, Steiniger M, Marzluff WF. A core complex of CPSF73, CPSF100, and symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. Mol Cell. 2009 May 15;34(3):322-32.

Sulston JE, Brenner S. The DNA of caenorhabditis elegans. Genetics. 1974 May;77(1):95-104.

Swinburne IA, Meyer CA, Liu XS, Silver PA, Brodsky AS. Genomic localization of RNA binding proteins reveals links between pre-mRNA processing and transcription. Genome Res. 2006 Jul;16(7):912-21.

Takagaki Y, Manley JL. Complex protein interactions within the human polyadenylation machinery identify a novel component. Mol Cell Biol. 2000 Mar;20(5):1515-25.

Takagaki Y, Manley JL. Levels of polyadenylation factor CstF-64 control IgM heavy chain mRNA accumulation and other events associated with B cell differentiation. Mol Cell. 1998 Dec;2(6):761-71.

Takagaki Y, Manley JL. RNA recognition by the human polyadenylation factor CstF. Mol Cell Biol. 1997 Jul;17(7):3907-14.

Takagaki Y, Manley JL. A polyadenylation factor subunit is the human homologue of the drosophila suppressor of forked protein. Nature. 1994 Dec 1;372(6505):471-4.

Takagaki Y, Seipelt RL, Peterson ML, Manley JL. The polyadenylation factor CstF-64 regulates alternative processing of IgM heavy chain pre-mRNA during B cell differentiation. Cell. 1996 Nov 29;87(5):941-52.

Tanaka Y, Ohta A, Terashima K, Sakamoto H. Polycistronic expression and RNA-binding specificity of the C. elegans homologue of the spliceosome-associated protein SAP49. J Biochem. 1997 Apr;121(4):739-45.

Tan-Wong SM, Wijayatilake HD, Proudfoot NJ. Gene loops function to maintain transcriptional memory through interaction with the nuclear pore complex. Genes Dev. 2009 Nov 15;23(22):2610-24.

Teixeira A, Tahiri-Alaoui A, West S, Thomas B, Ramadass A, Martianov I, et al. Autocatalytic RNA cleavage in the human beta-globin pre-mRNA promotes transcription termination. Nature. 2004 Nov 25;432(7016):526-30.

Thierry-Mieg D, Thierry-Mieg J. AceView: A comprehensive cDNA-supported gene and transcripts annotation. Genome Biol. 2006;7 Suppl 1:S12.1-14.

Topisirovic I, Svitkin YV, Sonenberg N, Shatkin AJ. Cap and cap-binding proteins in the control of gene expression. Wiley Interdiscip Rev RNA. 2011 Mar-Apr;2(2):277-98.

Ujvari A, Pal M, Luse DS. RNA polymerase II transcription complexes may become arrested if the nascent RNA is shortened to less than 50 nucleotides. J Biol Chem. 2002 Sep 6;277(36):32527-37.

Valsamakis A, Zeichner S, Carswell S, Alwine JC. The human immunodeficiency virus type 1 polyadenylylation signal: A 3' long terminal repeat element upstream of the AAUAAA necessary for efficient polyadenylylation. Proc Natl Acad Sci U S A. 1991 Mar 15;88(6):2108-12.

Venkataraman K, Brown KM, Gilmartin GM. Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. Genes Dev. 2005 Jun 1;19(11):1315-27.

Visa N, Izaurralde E, Ferreira J, Daneholt B, Mattaj IW. A nuclear cap-binding complex binds balbiani ring pre-mRNA cotranscriptionally and accompanies the ribonucleoprotein particle during nuclear export. J Cell Biol. 1996 Apr;133(1):5-14.

Wang Y, Fairley JA, Roberts SG. Phosphorylation of TFIIB links transcription initiation and termination. Curr Biol. 2010 Mar 23;20(6):548-53.

Weiss EA, Gilmartin GM, Nevins JR. Poly(A) site efficiency reflects the stability of complex formation involving the downstream element. EMBO J. 1991 Jan;10(1):215-9.

West S, Gromak N, Proudfoot NJ. Human 5' --> 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. Nature. 2004 Nov 25;432(7016):522-5.

Whitelaw E, Proudfoot N. Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene. EMBO J. 1986 Nov;5(11):2915-22.

Whittle CM, McClinic KN, Ercan S, Zhang X, Green RD, Kelly WG, et al. The genomic distribution and function of histone variant HTZ-1 during C. elegans embryogenesis. PLoS Genet. 2008 Sep 12;4(9):e1000187.

Williams C, Xu L, Blumenthal T. SL1 trans splicing and 3'-end formation in a novel class of caenorhabditis elegans operon. Mol Cell Biol. 1999 Jan;19(1):376-83.

Yang Q, Doublie S. Structural biology of poly(A) site definition. Wiley Interdiscip Rev RNA. 2011 Sep-Oct;2(5):732-47.

Yonaha M, Proudfoot NJ. Transcriptional termination and coupled polyadenylation in vitro. EMBO J. 2000 Jul 17;19(14):3770-7.

Zaret KS, Sherman F. DNA sequence required for efficient transcription termination in yeast. Cell. 1982 Mar;28(3):563-73.

Zhang J, Corden JL. Identification of phosphorylation sites in the repetitive carboxyl-terminal domain of the mouse RNA polymerase II largest subunit. J Biol Chem. 1991 Feb 5;266(4):2290-6.

Zhang Z, Fu J, Gilmour DS. CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11. Genes Dev. 2005 Jul 1;19(13):1572-80.

Zorio DA, Bentley DL. The link between mRNA processing and transcription: Communication works both ways. Exp Cell Res. 2004 May 15;296(