Dissertation zur Erlangung des Doktorgrades

der Fakultät für Chemie und Pharmazie

der Ludwig-Maximilians-Universität München

Structural and functional studies on the Negative Elongation Factor

David Pöllmann

aus

Weiden in der Oberpfalz, Deutschland

2015

Erklärung

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Patrick Cramer betreut.

Eidesstattliche Versicherung

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München den 23.8.2015

_____

David Pöllmann

Dissertation eingereicht am: 10.09.2015

1. Gutachter: Prof. Dr. Patrick Cramer

2. Gutachter: PD Dr. Dietmar Martin

Mündliche Prüfung am: 30.09.2015

Dictum

per aspera ad astra

Acknowledgements

First of all I want to thank Lina for her support, particularly in difficult times, and my parents and family for always backing and allowing me to become the person I am.

All my thanks to Patrick. For the opportunity to work in an inspiring and exciting place, for giving me the chance to grow and learn many valuable lessons and for being a great and patient supervisor.

I´m grateful to all people in the Cramer lab: To Claudia and Stefan for their support and patience when I started working in the lab and after. To all my colleagues for their help and advices, the great atmosphere and the fun we had together. Special thanks to Dirk for his advice on data processing and model building.

I´m obliged to all my collaborators for their help, effort and reliability. First and most of all to Seychelle who was more than helpful to me. To Tom Zimniak, Franz Herzog and Henning Urlaub. To the people from the Max Planck Institut of Biochemistry crystallization and core facility, in particular Karina Valer and Sabine Pleyer.

I´m grateful to my examiners Dietmar Martin, Klaus Förstemann, Roland Beckmann, Mario Halic and Karl-Peter Hopfner and the membes of my thesis committee Dirk Eick and Katja Lammens for their support, time and effort.

Finally, I deeply want to thank my friends Alwin, Sebastian and Stefan. I don´t have to say why.

Summary

Promoter-proximal pausing is a wide-spread phenomenon in metazoans. RNA polymerase II is stably paused after transcribing 20-60 nucleotides of a gene and awaits controlled release. First considered a rare phenomenon, this mechanism was recently recognized as a key regulatory step in controlling transcription for the majority of genes. The four-subunit negative elongation factor (NELF) is essential for establishing promoter-proximal pausing. Despite years of study, structural information is known only for the NELF-E RRM domain and is lacking for the rest of the complex. A high-resolution structure of NELF would greatly contribute to understanding the role of NELF in promoter-proximal pausing and provide an important basis for further research.

In this work, a highly conserved NELF subcomplex consisting of NELF-A (6-188) and NELF-C (183-590) was identified, crystallized and its structure solved to 2.8 Å resolution. Both subunits interact extensively. NELF-C adopts a horse-shoe shaped conformation including a CTD-interacting domain (CID)-like domain at its C-terminus. The NELF-A C-terminal region (111-182) stretches across NELF-C. The NELF-A N-terminal region (6-110) forms a highly conserved domain with structural similarity to the HIV integrase-binding domain in human PC4 and SFRS1-interacting protein.

Analysis of the NELF-AC surface revealed that the exterior of the complex contains several large, positively charged patches. Fluorescence anisotropy experiments demonstrated that NELF-AC specifically interacts with single stranded nucleic acids in a strongly sequence dependent manner. Mutation of the positive patches confirmed the importance of surface residues in nucleic acid binding. Three previously identified *in vivo* NELF-AC phosphorylation sites are located close to the surface regions involved in nucleic acid binding and phosphomimetic mutations of these sites effectively decrease affinity of NELF-AC for nucleic acids. NELF activity is tightly regulated by the kinase positive transcription elongation factor (P-TEFb). Treatment of NELF-AC with P-TEFb revealed two novel phosphorylation sites that have a strong, inhibitory effect on nucleic acid binding by NELF-AC. Finally, crosslinking of a four-subunit NELF complex coupled with mass-spectrometry elucidated the architecture of the complete NELF complex and confirmed that the nucleic acid binding surfaces identified in the NELF-AC crystal structure are accessible in solution.

These results suggest a possible model for NELF action at the molecular level. NELF-AC contributes to pausing by binding to nascent RNA in a sequence dependent manner to additionally stabilize the pausing complex and further enhance promoter-proximal pausing.

Publication

Part of this work has been published:

Contents

I   Introduction

1   Transcription of genes in eukaryotes

1.1 DNA-dependent RNA polymerases

To adapt to a changing environment and maintain basic metabolic processes, a cell needs to transcribe the information contained in its DNA into various types of RNA continuously and dynamically. In most organisms, production of RNA is performed by DNA-dependent RNA polymerases, with the exception of some viruses that use RNA-dependent RNA polymerases to synthesize RNA (te Velthuis, 2014). DNA-dependent RNA polymerases consist of one or more subunits and are tightly regulated by transcription factors (Cramer, 2002a; Thomas and Chiang, 2006). Different than in viruses and organelles, complex RNA polymerases composed of several different subunits evolved in all living organisms (Cramer, 2002b). Bacteria and Archaea contain only one multi-subunit RNA polymerase consisting of five and twelve subunits, respectively (Werner and Grohmann, 2011).

In eukaryotes, five multi-subunit RNA polymerases (Pol) I–V produce different kinds of RNA (Haag and Pikaard, 2011; Werner, 2007). Whereas Pol I transcribes only one gene, the 45S ribosomal RNA (rRNA), Pol II synthesizes all pre-messenger RNAs (mRNA), micro RNAs (miRNA) and most small nuclear RNAs (snRNA) and Pol III produces many untranslated RNAs like translator RNAs (tRNA) and 5S rRNA (Roeder and Rutter, 1970; Thomas and Chiang, 2006). Pol IV and V exist only in plants and are involved in the biogenesis of siRNA and heterochromatin formation (Haag and Pikaard, 2011).

Pol I, II and III are composed of 14, 12 and 17 subunits, respectively. Ten subunits are structurally homologous among all three Pols and represent the conserved core of the enzyme (Table 1) (Vannini and Cramer, 2012). Pol I and III contain additional subunits not present in the Pol II core enzyme. Homologs of these unique Pol I and Pol III subunits are transiently associated with Pol II during transcription and are known as initiation specific transcription factors (TFs) (Table 1) (Vannini and Cramer, 2012). Permanent association of homologous Pol II transcription factor homologues to Pol I and III likely results from the greater transcriptional specialization of Pol I and Pol III. The mechanism of transcription initiation between Pol I-III is similar in that all are recruited by auxiliary factors to the correct transcription start site (TSS) (Table 1) (Vannini and Cramer, 2012). A plethora of additional proteins regulate Pol II activity in a cell and gene specific manner (Sikorski and

Buratowski, 2009; Thomas and Chiang, 2006).

Table 1: RNA polymerase subunits and initiation factor homologues in yeast. Adapted from (Vannini and Cramer, 2012).

Pol I	Pol II	Pol III
Polymerase Core		
A190	Rpb1	C160
A135	Rpb2	C128
AC40	Rpb3	AC40
Rpb5	Rpb5	Rpb5
Rpb6	Rpb6	Rpb6
Rpb8	Rpb8	Rpb8
A12.2 N-ribbon	Rpb9	C11 N-ribbon
Rbp10	Rpb10	Rpb10
AC19	Rpb11	AC19
Rbp12	Rpb12	Rpb12
Polymerase Stalk		
A14	Rpb4	C17
A43	Rpb7	C25
Pol II transcription factors and homologues in Pol I and Pol III		
A49 N-terminal domain	Tfg1 (TFIIFα)	C37
A34.5	Tfg2 (TFIIFβ)	C53
	Tfa1 (TFIIEα)	C82
A49 C-terminal domain	Tfa2 (TFIIEβ)	C34
		C31

1.2 Eukaryotic RNA polymerase II

Eukaryotic Pol II is a highly conserved protein complex with a mass greater than 500 kDa and consists of 12 subunits in *S.cerevisiae* (Figure 1) (Armache et al., 2005; Cramer et al., 2001). The yeast Pol II crystal structure (Cramer et al., 2001), initiation complex (Cheung et al., 2011; Liu et al., 2011; Sainsbury et al., 2015; Sainsbury et al., 2013) and elongation complex (Figure 1) (Gnatt et al., 2001; Kettenberger et al., 2004) as well as other functional complexes (Cheung and Cramer, 2011; Kostrewa et al., 2009; Plaschka et al., 2015; Sydow et al., 2009) have been characterized. Mammalian Pol II has not been crystallized yet, but medium-resultion structures of human Pol II complexes obtained by

cryo-electron microscopy studies are available (Bernecky et al., 2011; He et al., 2013; Kassube et al., 2013).
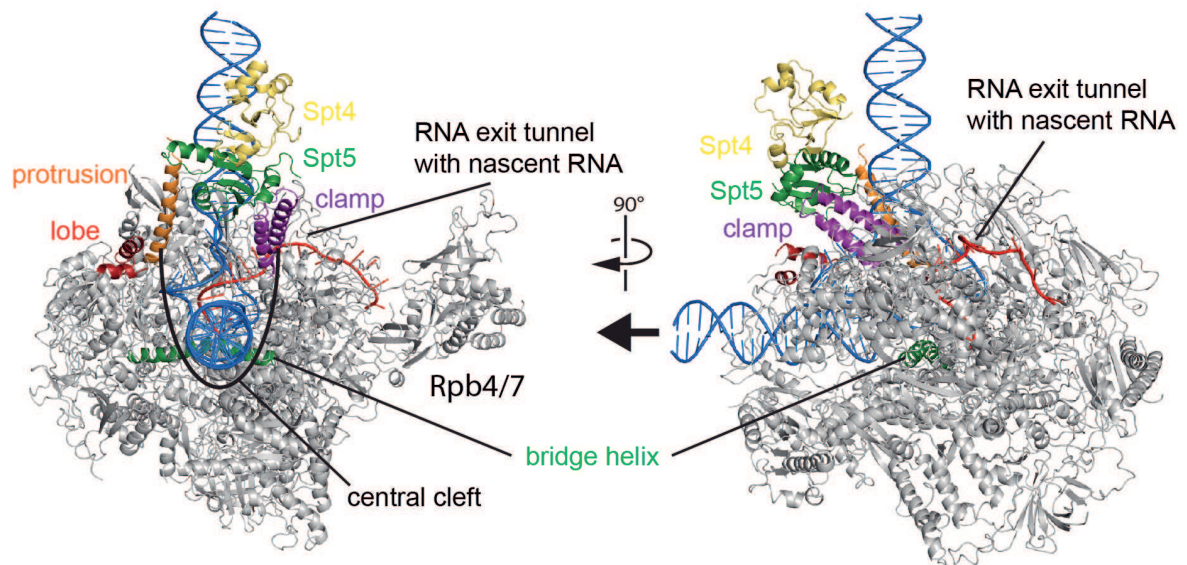
Figure 1: Model of eukaryotic Pol II-DSIF elongation complex.
Adapted from (Martinez-Rucobo et al., 2011). Both perspectives are related by a 90° turn around the vertical axis. Parts of Pol II domains interacting with Spt4/5 (clamp coiled coil) or in close proximity (protrusion and lobe) are colored. Rpb4/7 and the bridge helix are labeled for orientation. The arrow indicates the direction of movement of transcribing Pol II. Pol II, Spt4 and Spt5 are coloured in grey, yellow and green, respectively. DNA and RNA are blue and red, respectively.

Ten subunits invariantly constitute the core enzyme. Subunits Rbp 4 and Rpb 7 form the peripheral 'stalk' subcomplex (Figure 1) that is required for transcription initiation and can dissociate from the core enzyme in yeast (Edwards et al., 1991).

Pol II comprises a 'central cleft' that encompasses the DNA template and harbors the highly conserved 'bridge helix' – the active site – at its base (Figure 1) (Cramer et al., 2001; Weinzierl, 2011). The cleft is defined by the two largest subunits Rbp1 and Rpb2 and confined by three distinctive domains called 'clamp', 'lobe' and 'protrusion'. Newly synthesized RNA exits the cleft through a RNA exit-tunnel located between the active site and the clamp and resurfaces near the clamp (Figure 1) (Andrecka et al., 2008; Kettenberger et al., 2004).

The clamp (Figure 1) is a highly conserved structural feature of Pol II. The mobile clamp adopts an 'open' conformation in the ten-subunit complex (Cramer et al., 2001) and switches to a 'closed' conformation after Pol II binds DNA encircling the template (Gnatt et al., 2001). The closed conformation is stabilized by Rpb4/7 (Armache et al., 2003). The

inner side of the clamp interacts with DNA (Gnatt et al., 2001), the outer side with interchanging accessory factors during the transcription cycle (Grohmann et al., 2011) like TFIIE (Chen et al., 2007), Spt4/5 (Klein et al., 2011; Martinez-Rucobo et al., 2011) or the human hepatitis virus delta antigen (HDAg) (Yamaguchi et al., 2007).

The C-terminal repeat domain (CTD) of Rbp1 is unique to Pol II. The CTD serves as a binding platform for RNA modifying enzymes positioned close to the RNA exit tunnel and coordinates RNA synthesis and co-transcriptional processing (Martinez-Rucobo et al., 2015; Munoz et al., 2010; Perales and Bentley, 2009; Proudfoot et al., 2002). The CTD is also important to recruit transcription regulation factors (Napolitano et al., 2014; Proudfoot et al., 2002). The CTD consists of a repetitive heptapeptide sequence (consensus sequence $Y_1S_2P_3T_4S_5P_6S_7$) that is variably phosphorylated during the course of the transcription cycle (Heidemann et al., 2013). The dynamic phosphorylation pattern is thought to be the molecular basis for the recruitment of different factors at various stages of the transcription process (Buratowski, 2009).

2 The transcription cycle of RNA polymerase II

Transcription of eukaryotic genes by Pol II is divided into three highly regulated steps:

(I) initiation, (II) elongation and (III) termination (Figure 2) (Hahn and Young, 2011; Nechaev and Adelman, 2011). The elongation step can be subdivided in early elongation IIa and productive elongation IIb, respectively.
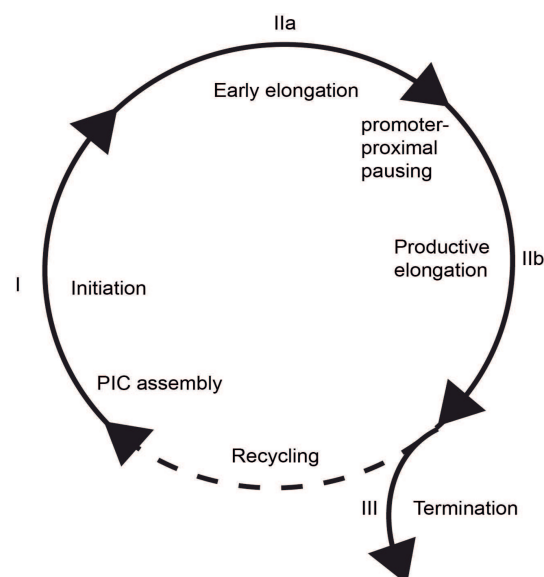
Figure 2: The transcription cycle of eukaryotic RNA polymerase II.
Adapted from: (Nechaev and Adelman, 2011).

Prior to initiation the pre-initiation complex (PIC) consisting of general transcription initiation factors (Table 2) and Pol II assembles at the promoter DNA and creates a short region of single stranded (ss) DNA – the transcription bubble – where transcription can take place (He et al., 2013; Muhlbacher et al., 2014; Sainsbury et al., 2015). A minimum of five general transcription factors (GTFs) are necessary to recruit Pol II to the promoter and start transcription (Table 2) (Sikorski, Buratowski 2009, Thomas and Chiang 2006). Aided by the non-essential factor TFIIA (Thomas and Chiang, 2006) the TATA-binding protein (TBP) subunit of the multi-protein complex TFIID recognizes the AT-rich promoter TATA-box sequence and upon binding bends the DNA by 90° (Kim et al., 1993; Tsai and Sigler, 2000). TBP is also important for activating TATA-less genes and initiation of Pol I and Pol III (Sadowski et al., 1993; White and Jackson, 1992). TFIIB stabilizes the TFIID-DNA complex and recruits a Pol II – TFIIF complex to the promoter site (Bushnell et al., 2004; Kostrewa et al., 2009; Orphanides et al., 1996; Sainsbury et al., 2013). Consecutive binding of TFIIE and TFIIH completes PIC assembly (Forget et al., 2004; Watanabe et al., 2003).

Transcription initiation starts with unwinding DNA and promoter melting stimulated by TFIIE and TFIIH resulting in a transcription bubble with a 15 nt single stranded DNA (Holstege et al., 1996; Kim et al., 2000). When single stranded DNA is available in the active site of the open promoter complex, incorporation of the first nucleotides can begin.

Once the newly synthesized RNA has reached a length of 10 nt transcribing Pol II escapes from the promoter region. The initially unstable transcription process gradually stabilizes and after 25 nt the early elongation phase begins (Jonkers and Lis, 2015; Margeat et al., 2006; Shandilya and Roberts, 2012). Phosphorylation of the Pol II CTD at serine 5 is a critical step in the transition from initiation to elongation (Buratowski, 2009).

After promoter escape of Pol II a part of the PIC – TFIIA, -D, -E and -H – remains bound to the promoter to enhance recruitment of another Pol II molecule and facilitate reinitiation of transcription (Yudkovsky et al., 2000).

For many metazoan genes, Pol II pauses after transcribing 20-60 nt and remains stably bound to DNA and RNA (Kwak and Lis, 2013). Additional factors are required before elongation is resumed (Chiba et al., 2010). This process is called promoter-proximal pausing and will be discussed in more detail later. Once promoter-proximally paused Pol II is released productive elongation ensues.

During elongation, nucleosomes are a major obstacle to transcribing Pol II that is overcome by extensive histone modification and displacement mediated by numerous chromatin-modifying enzymes (Kulaeva et al., 2013; Saunders et al., 2006). Chromatin of

activated genes exhibits a distinct pattern of post-translational histone modifications (Li et al., 2007; Shilatifard, 2006). Methylations H3K27me3 and H3K9me3 as well as other modifications that lead to silencing are removed during activation and replaced by acetylation of histones H3 and H4, methylations H3K36me3 and H3K79me3, phosphorylation H3S10P and other modifications. Furthermore mRNA splicing at intron-exon junctions constitutes a transcription rate-decreasing process (Jonkers and Lis, 2015; Kwak et al., 2013).

Termination is predominantly mediated through the poly(A)-dependent pathway (Kuehner et al., 2011). A highly conserved poly(A) signal sequence positioned upstream (AAUAAA) and a G/U-rich sequence positioned immediately downstream of the genes 3' end decelerate and finally pause Pol II. Joint action of several factors including cleavage stimulation factor (CstF), cleavage and polyadenylation specificity factor (CPSF) and polyadenylate-polymerase (poly(A)-Pol) lead to cleavage of the nascent transcript and 3' polyadenylation of the transcribed mRNA by addition of 100-200 adenosines (Kuehner et al., 2011; Nag et al., 2007).

Table 2: General transcription factors in human and their functions. Adapted from (Thomas and Chiang, 2006).

Factor	Subunits	Function
TFIIA [a]	3	Stabilization of TATA-TBP complex
TFIIB	1	TSS selection, recruits pol II/TFIID, stabilization of TATA-TBP complex
TFIID	15	Core promoter-binding, coactivator, protein kinase, histone acetyltransferase, TBP is a subunit
TFIIE	2	Recruits TFIIH, formation of an initiation-compentent pol II, promoter clearance
TFIIF	2	Pol II binding to promoter, recruits TFIIE and -H, TSS selection and promoter escape
TFIIH	10	Helicase activity, phosphorylate CTD

[a] non-essential for transcription initiation

3 Promoter-proximal pausing

3.1 Prevalence and relevance

Gene expression of many metazoan genes is regulated during early elongation by promoter-proximal pausing (ppp) (Kwak and Lis, 2013). Elongating Pol II stably pauses at a position 20-60 bp downstream of the TSS and is released only upon phosphorylation of Pol II and other factors by the kinase complex positive transcription elongation factor b (P-TEFb) (Yamaguchi, Shibata, Handa, 2012, Adelman, Lis 2012). It was long thought that gene expression was primarily regulated during transcription initiation. Recently promoter-proximal pausing was recognized as a key event in the regulation of many genes during transcription elongation (Jonkers and Lis, 2015; Li and Gilmour, 2011; Yamaguchi et al., 2013).

Promoter-proximal pausing was first discovered by *in vivo* analysis of the *Drosophila melanogaster* heat shock genes *hsp70* and *hsp26* that exhibited Pol II accumulation downstream of the TSS prior to induction (Giardina et al., 1992; Gilmour and Lis, 1986; Rasmussen and Lis, 1993; Rougvie and Lis, 1988). The accumulated Pol II is transcriptionally engaged as demonstrated by permanganate footprinting and is able to resume transcription (Core et al., 2008; Rougvie and Lis, 1988). The cleavage factor TFIIS is prevented from cleaving the RNA transcript (Cheung and Cramer, 2011; Palangat et al., 2005). Therefore the Pol II is in fact stably paused instead and not terminated. Promoter-proximally paused Pol II has also been found and studied in more detail at immediate early genes like *junB* (Aida et al., 2006), *c-myc* (Krumm et al., 1992; Schneider et al., 1999) and *c-fos* (Fivaz et al., 2000; Plet et al., 1995). Subsequent studies showed promoter-proximally paused Pol II to be a common phenomenon that occurs at the majority of genes in the fruit fly D*rosophila melanogaster* (Guenther et al., 2007; Muse et al., 2007) and in human (Core et al., 2008; Gilmour, 2009; Guenther et al., 2007). Similar phenomena have been described in yeast (Venters and Pugh, 2009) and the nematode *C.elegans* (Baugh et al., 2009). However, no distinct proof for promoter-proximal pausing has been found in these organisms and promoter-proximal pausing is considered a process specific to higher metazoa. A comparable process of transient polymerase pausing following transcription initiation has been described in bacteria (Greive and von Hippel, 2005; Larson et al., 2014; Vvedenskaya et al., 2014).

Promoter-proximal pausing is important for controlling signal-responsive pathways (Aida et al., 2006; Krumm et al., 1992; Plet et al., 1995), developmental processes in multicellular organisms (Amleh et al., 2009; Keegan et al., 2002; Zeitlinger et al., 2007),

cell differentiation and reprogramming (Guenther et al., 2007; Williams et al., 2015), and expression of genes of the human immunodeficiency virus (HIV) (Natarajan et al., 2013; Zhang et al., 2007).

3.2 Involved factors and regulation

Three protein complexes are involved in regulating promoter-proximal pausing. The DRB sensitivity inducing factor (DSIF) (Wada et al., 1998; Yamaguchi et al., 1999b), P-TEFb (Cheng and Price, 2007; Chiba et al., 2010; Price, 2000) and the negative elongation factor (NELF) (Pagano et al., 2014; Yamaguchi et al., 1999a). All three factors have also been detected *in vivo* simultaneously on the gene loci of *hsp70* (Andrulis et al., 2000; Lis et al., 2000; Wu et al., 2003) or *junB* (Aida et al., 2006) containing promoter-proximally paused Pol II.

3.2.1 DSIF

DSIF is a heterodimer composed of the human Spt4 (13,2 kDa) and Spt5 (121 kDa) homologs. Spt5 (NusG in bacteria) is the only transcription factor conserved in all three domains of life (Werner, 2012). The Nus-G N-terminal (NGN) domain of Spt5 together with Spt4 constitutes the conserved core of the complex in archaea and eukaryotes (Belogurov et al., 2007; Guo et al., 2008; Martinez-Rucobo et al., 2011; Wenzel et al., 2010; Zhou et al., 2009a). The Spt4/5 core binds to the highly conserved Pol II clamp coiled coil motif (Hirtreiter et al., 2010; Martinez-Rucobo et al., 2011) and to lobe and protrusion on the opposite side of the cleft closing the Pol II cleft (Figure 1) (Klein et al., 2011; Martinez-Rucobo et al., 2011). This contacts prevent DNA release from transcribing Pol II thus increasing processivity. Spt5 further contacts the non-template DNA and stabilizes the transcription bubble (Artsimovitch and Landick, 2002). In eukaryotes Spt5 possesses a 814 amino acid long C-terminal tail (273-1087) including five Kyrpides-Ouzounis-Woese (KOW) domains and an unstructured C-terminal region (CTR) similar to the Pol II CTD (Kyrpides et al., 1996; Yamaguchi et al., 1999b). The KOW-domains and the CTR serve as binding platform for RNA-processing factors and contribute to integrating RNA-synthesis and -processing (Mayer et al., 2012; Werner, 2012).

Unphosphorylated DSIF suppresses transcription whereas P-TEFb-dependent phosphorylation of the CTR causes a functional reversion (Yamada et al., 2006). CTR phosphorylation is a critical step in recruitment of elongation factors and progression into productive elongation and is preserved in all eukaryotes (Chen et al., 2009; Liu et al.,

2009; Qiu et al., 2006; Squazzo et al., 2002; Zhou et al., 2009b).

3.2.2 NELF

NELF is a multi subunit complex that represses transcription elongation (Wu et al., 2003; Yamaguchi et al., 2002) and is associated with chromatin (Wu et al., 2005). NELF consists of four subunits NELF-A, -B, -C or its splicing variant -D, and -E. NELF is present in many metazoans such as human, zebrafish or fruit fly (Narita et al., 2003) but is also observed in the single celled organism *Dictyostelium discoideum* (Chang et al., 2012). NELF has not been found in yeast, the nematode *C. elegans* or plants (Narita et al., 2003). NELF is highly conserved with the exception of a few regions of NELF-A, NELF-C and NELF-E (Figure 3A). The four NELF subunits interact in a linear manner and the peripheral subunits NELF-A and NELF-E have been functionally characterized (Narita et al., 2003). The N-terminus of NELF-A binds to NELF-C (125-188) and contains a region that is known to associate with Pol II (189-248) (Narita et al., 2003). The NELF-A C-terminal region is conserved but is presently uncharacterized. The NELF-A•Pol II binding region exhibits a weak sequence similarity with HDAg and possibly interacts with Pol II in a similar way like HDAg by binding the Pol II clamp (Figure 3B) (Yamaguchi et al., 2001; Yamaguchi et al., 2007). NELF-E contains a structurally characterized RNA recognition motif (RRM) that binds RNA in a sequence-dependent manner (Figure 3A) (Pagano et al., 2014; Rao et al., 2006; Rao et al., 2008). Except for the RRM of NELF-E no structural information for NELF is available.

NELF is essential for the embryogenesis of higher metazoa (Amleh et al., 2009; Wang et al., 2010; Williams et al., 2015) and is required for expression of genes of the human immunodeficiency virus (Natarajan et al., 2013; Zhang et al., 2007). NELF has not only been associated with HIV but was also implicated in the etiology of other viral infections (Palermo et al., 2011; Toth et al., 2012), genetic diseases like the Wolf-Hirschhausen syndrome which is characterized by multiple malformations (Kerzendorfer et al., 2012; Wright et al., 1999) or multiple types of chancer (Iida et al., 2012; McChesney et al., 2006; Sun et al., 2008; Ye et al., 2001).

A
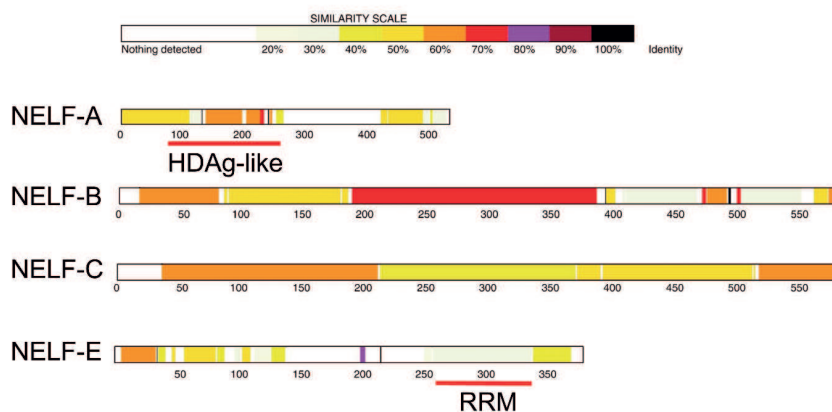
B
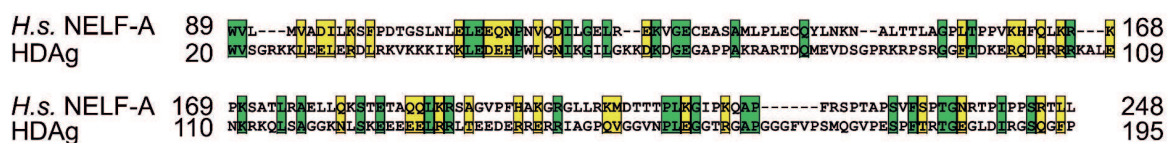
Figure 3: Conservation of NELF subunits

(A) Conservation of human NELF subunits relative to *Drosophila* NELF. NELF-A region with sequence similarity to HDAg (Figure 3B) and NELF-E RRM are indicated. Adapted from (Wu et al., 2005).

(B) Sequence alignment between human (*H.s.*) NELF-A HDAg-like region and HDAg (P0C6L3). Identical and conserved residues are colored in green and yellow, respectively. The alignment is based on (Yamaguchi et al., 2001) and was generated with CluwtalW (Larkin et al., 2007).

3.2.3 P-TEFb

P-TEFb is a heterodimeric cyclin-dependent kinase composed of cyclin-dependent kinase 9 (CDK9) and cyclins T1, T2, or K (Peterlin and Price, 2006). P-TEFb was initially identified as the primary target of a drug, 5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole (DRB), which prevents Pol II from producing full-length transcripts (Marshall et al., 1996; Marshall and Price, 1992, 1995; Zhu et al., 1997). It was later shown that P-TEFb counteracts promoter-proximal pausing and enables productive elongation by phosphorylating the Ser-2 sites of the Pol II CTD (Cheng and Price, 2007), the Spt4/5 CTR (Yamada et al., 2006) and NELF (Fujinaga et al., 2004). P-TEFb also contributes to establishing open chromatin by phosphorylating histones (O'Brien et al., 2010). Conversely, inhibition of P-TEFb activity almost completely blocks Pol II transcription elongation (Henriques et al., 2013; Jonkers et al., 2014). Crystal structures of P-TEFb alone and in complex with drugs and small peptides are available to provide a mechanistic basis for its role in transcription (Baumli et al., 2012; Baumli et al., 2008; Schulze-Gahmen et al., 2014; Schulze-Gahmen et al., 2013; Tahirov, 2010). P-TEFb is essential for

the replication of the human immunodeficieny virus (He et al., 2010; Sobhian et al., 2010).

3.3   Mechanism and regulation of promoter-proximal pausing

Association of DSIF and NELF with elongating Pol II is sufficient to induce promoter-proximal pausing (Missra and Gilmour, 2010; Narita et al., 2003; Renner et al., 2001; Yamaguchi et al., 1999b). NELF requires a preformed Pol II-DSIF complex for stable binding. The position of Pol II pausing on the gene is a function of the rate of transcription elongation and NELF binding efficiency (Li et al., 2013), which is influenced by competition with the transcription factor TFIIF (Renner et al., 2001). NELF and DSIF bind to the elongation complex once nascent RNA is longer than 18 nt and emerges from Pol II surface (Andrecka et al., 2008; Missra and Gilmour, 2010). Despite a strong association of the NELF-E RRM with RNA (Pagano et al., 2014; Rao et al., 2008), it was reported that DSIF, but not NELF contacts nascent RNA (Missra and Gilmour, 2010). Several models how promoter-proximal pausing is established have been suggested (Kwak and Lis, 2013).

The kinetic model (Figure 4A) posits that the sequence dependent rate of Pol II transcription competes with the efficiency of pausing factor recruitment to thereby affect extent and location of pausing (Bai et al., 2004; Li et al., 2013; Nechaev et al., 2010). So far no common cis-element at human pausing sites has been discovered, albeit promoters with paused Pol II share some characteristics: the respective promoters are CpG rich, lack a TATA box and contain Pol II CTD Ser-5 but not Ser-2 phosphorylations (Core et al., 2008; Kininis et al., 2009; Mavrich et al., 2008). Furthermore, the +1 nucleosome is shifted downstream at promoters containing paused Pol II and histones carry a unique H3K4 and H3K27 methylation (Bernstein et al., 2006; Schones et al., 2008).

The nucleosome barrier model (Figure 4B) assumes the first nucleosome after the TSS to prevent elongating Pol II from proceeding further into the gene. Indeed nucleosomes contribute to promoter-proximal pausing (Gilchrist et al., 2010; Gilchrist et al., 2008; Jimeno-Gonzalez et al., 2015) but are not necessary to pause Pol II. For example the highly paused *Drosophila melanogaster hsp70* gene contains a 5' nucleosome free region (Fuda et al., 2009; Gilchrist et al., 2010; Yamaguchi et al., 1999a).

The interaction model (Figure 4C) relies on sequence specific interaction between pausing factors and DNA/RNA as observed in bacteria (Wang et al., 1997). Indeed, the NELF-E RRM binds to RNA in a sequence dependent manner (Pagano et al., 2014) and thereby may contribute to promoter-proximal pausing (Yamaguchi et al., 2002).

In some cases estrogen-dependent recruitment of NELF or sequence-dependent recruitment of DSIF to specific sites is pivotal for setting up promoter-proximal pausing (Aiyar et al., 2004; Amir-Zilberstein et al., 2007).
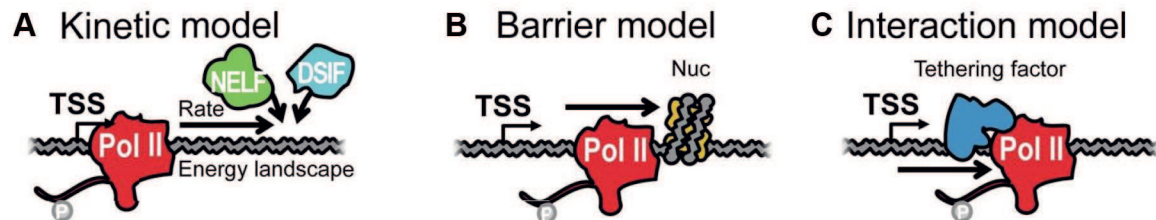
Figure 4: Models of possible mechanisms of promoter-proximal pausing.
Adapted from (Kwak and Lis, 2013). Possible mechanisms that impede Pol II elongation and induce pausing.

Phosphorylation of NELF, DSIF and Pol II CTD by P-TEFb is critical for pause release (Figure 5) and promoter-proximal pausing is mainly regulated by spatiotemporal regulation of P-TEFb catalytic activity (Chiba et al., 2010; Peterlin and Price, 2006; Yamaguchi et al., 2013).

P-TEFb is recruited to specific genomic target sites by sequence specific factors such as the heat shock factor (hsf) (Lis et al., 2000) or the bromodomain-containing protein Brd4 (Hargreaves et al., 2009; Jang et al., 2005). Brd4 binds to acetylated histones and represents a prevalent mechanism how P-TEFb is recruited to active genes (Hargreaves et al., 2009; Yang et al., 2005). P-TEFb is also recruited by NF-κB (Barboric et al., 2001; Luecke and Yamamoto, 2005), the viral transactivator Tat (Price, 2000) and others (Gargano et al., 2007; Oven et al., 2007).

The equilibrium between the active and the inactive state of P-TEFb is regulated by association with the small nuclear ribonucleic protein (snRNP) 7SK snRNP consisting of 7SK snRNA and the proteins HEXIM, LARP7 and MEPCE (Figure 5) (Chen et al., 2008; Jeronimo et al., 2007; Markert et al., 2008; Peterlin and Price, 2006; Yik et al., 2003). P-TEFb release from this inhibitory complex is controlled by various factors in response to external stimuli (Chen et al., 2004; Li et al., 2005), such as the direct interaction with transcriptional coactivators like Brd4 or the HIV-1 Tat protein (Barboric et al., 2007; Krueger et al., 2010; Sedore et al., 2007; Tahirov et al., 2010; Yang et al., 2005).
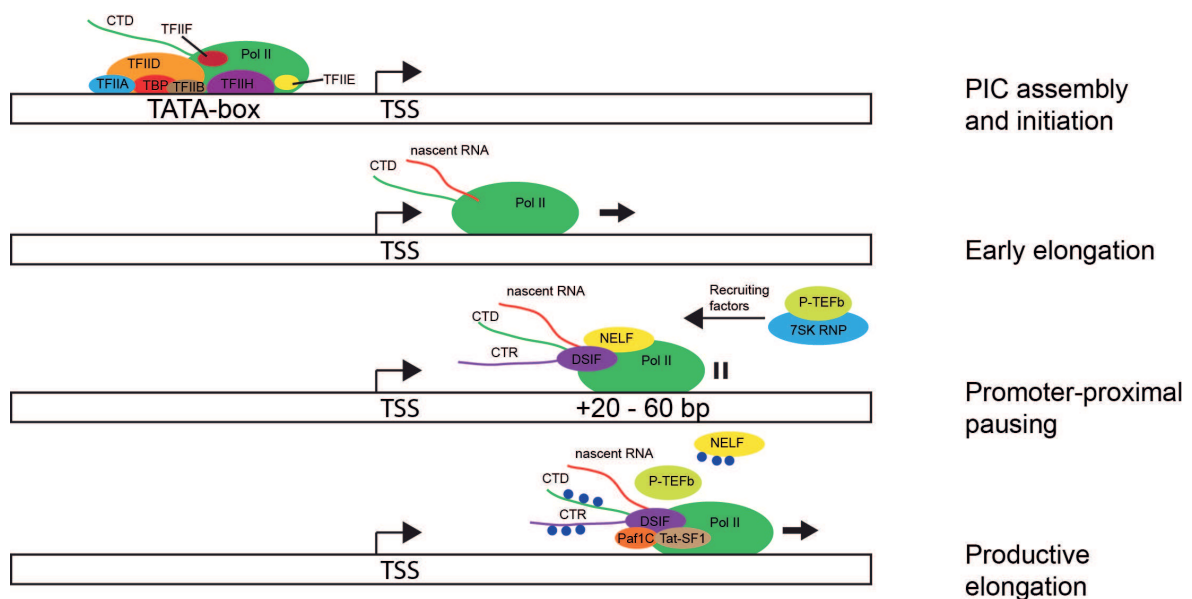
Figure 5: Overview of the early transcription phase and the factors involved in initiation and promoter-proximal pausing.
Blue circles indicate P-TEFb-mediated phosphorylation that are requisite for pause-release. Before Pol-II commences stable elongation, phosphorylated NELF dissociates and elongation factors Paf1C and Tat-SF1 associate with Pol II.

3.4 Physiological significance of promoter-proximal pausing

Several biological functions of promoter-proximal pausing have been discussed (Adelman and Lis, 2012; Chiba et al., 2010).

In addition to the highly regulated process of transcription initiation, promoter-proximal pausing provides another potential for the regulation and synchronous activation of genetic activity (Boettiger and Levine, 2009). Promoter-proximal pausing enables the cell to perform transcription more flexible and elaborate.

Promoter-proximal pausing has first been observed at immediate-early genes which are capable of rapid induction upon specific signals (Aida et al., 2006; Krumm et al., 1992; Rougvie and Lis, 1990). PIC-assembly at the promoter site is a relatively slow and complex multi-step process requiring a plethora of general and specific factors (Thomas and Chiang, 2006). Conversely, once P-TEFb has been activated release of paused Pol II is a fast process. It is hence speculated that promoter-proximal pausing circumvents the time-consuming assembly of the transcription machinery and facilitates a dynamic regulation of genes that need to be activated quickly like heat shock factors (Andrulis et al., 2000; Wu et al., 2003).

To prevent nascent pre-mRNA from fast degradation by exonucleases (Hsu and Stevens,

1993) the 5' end is modified by the cap-binding complex (CBC) with a 5' guanylyl cap (Gonatopoulos-Pournatzis and Cowling, 2014; Hocine et al., 2010; Shuman, 2001). Capping is coupled to transcription and occurs during the early elongation phase when RNA reaches a length of 20-30 nt (Martinez-Rucobo et al., 2015; Perales and Bentley, 2009; Yue et al., 1997). The CBC recruits elongation factors and during translation the cap is important for recognition of the mRNA by ribosomes (Gonatopoulos-Pournatzis and Cowling, 2014; Tarun and Sachs, 1996). The CBC was shown to stimulate the early elongation phase (Kim et al., 2004), conversely NELF, DSIF and P-TEFb were observed to interact with and regulate the CBC (Mandal et al., 2004; Narita et al., 2007; St Amour et al., 2012; Wen and Shatkin, 1999). This links two important events in early elongation, promoter-proximal pausing and capping. Thus promoter-proximal pausing could also function as a checkpoint to ensure only correctly capped mRNA is fully transcribed.

The DNA sequence in promoter regions can affect their association with nucleosomes (Gilchrist et al., 2010; Gilchrist et al., 2008; Iyer and Struhl, 1995; Kaplan et al., 2009; Valouev et al., 2011). Nucleosome formation in the promoter region is often a feature of highly regulated genes whereas promoters of housekeeping genes are free of chromatin (Gilchrist et al., 2010). In promoter regions occupied by nucleosomes, promoter-proximal pausing contributes to creating a nucleosome free region (NFR) keeping the promoter permissive for binding of the transcription machinery and regulatory factors (Gilchrist et al., 2010). Indeed promoter-proximal pausing causes a downstream shift of the +1 nucleosome (Schones et al., 2008). In contrast depletion of NELF results in a loss of the NFR and a decrease of transcription rate at many genes (Gilchrist et al., 2008).

4 Significance of NELF for the treatment of HIV infections

After entering a cell the HI virus reversely transcribes its RNA-genome and inserts the generated DNA into the nuclear genome of the host cell. For viral replication, the DNA again needs to be transcribed into RNA (Klimas et al., 2008). Transcription of HIV genes starts at the long terminal repeat (LTR) region where the cellular transcription machinery and co-activators are recruited (Rohr et al., 2003). Similar to many cellular genes, elongation of Pol II is paused during the early elongation phase (Ott et al., 2011). The highly conserved sequence of the transactivation response element (TAR) is located at the 5' end of the nascent viral transcript and forms a double-stranded stem loop. NELF-E binds to the stem loop with high affinity and represses elongation (Pagano et al., 2014; Zhang et al., 2007). Consequently the transcription of viral genes remains inefficient and

the HIV protein Tat is required to release this block. Tat-mediated recruitment of P-TEFb to the pausing site results in resumption of productive elongation (Karn and Stoltzfus, 2012; Ott et al., 2011; Zhou and Yik, 2006). Tat binds the TAR stem-loop displacing NELF and P-TEFb phosphorylates the pausing complex as described above (Ott et al., 2011). The Tat protein itself is encoded by the transcribed RNA, which triggers a positive feedback loop. If Tat is mutated or Pol II pausing release is prevented by any other means, transcription of HIV RNA continues to be unproductive, preventing HIV-1 replication and causing latent infection. A latent viral infection poses a major problem to therapeutic eradication of a HIV infection due to the absence of viral activity that can be targeted by drugs (Siliciano and Greene, 2011). For that reason promoter-proximal pausing is also of medical relevance.

5 Aims and scope

Although NELF is central to promoter-proximal pausing and involved in the regulation of various important developmental and physiological processes in multicellular organisms, no structural information is known except for the NELF-E RRM. NELF-E can bind to RNA; however, it is unclear what molecular mechanism NELF uses to induce promoter-proximal pausing. The origin and evolution of NELF is also unclear. In contrast to DSIF, which is central to transcription in all living cells, NELF is only present in some eukaryotes. Sequence alignments have identified homologs in metazoans and a subset of single celled eukaryotes but not in yeast or plants.

High-resolution structures of conserved regions of the NELF complex or of the entire complex would provide new insights into its function, overall architecture, and role in promoter-proximal pausing. Discovery of structural homologs could support the search for the evolutionary origin of NELF.

The primary goal of this thesis was to crystallize the NELF complex and solve its stucture in molecular detail. To this end the NELF complex had to be expressed, purified in high quality and characterized *in vitro*. Flexible regions had to be identified and truncated in order to obtain stable and compact constructs likely to crystallize. An optimized NELF-AC complex could finally be crystallized and the structure was solved with 2.8 Å resolution.

Once a high-resolution structure was obtained it had to be analyzed biochemically to asses its function in promoter-proximal pausing. Bioinformatic analysis of the structure model revealed large positive patches across the surface. Using protein-ligand binding assays the interaction partner was identified to be single stranded nucleic acids. To identify surface residues involved in protein-nucleic acid interaction and its regulation a

series of mutant proteins had to be generated and characterized. From this studies, a nucleic acid binding face on NELF-AC previously unknown was discovered.

The third goal of this thesis was to describe the architecture of the holo-NELF complex. For this purpose a purification protocol for the complete four-subunit NELF complex had to be established. Testing different combinations of subunits in co-expression enabled purification of holo-NELF in good yield and quality. The architecture was resolved by crosslinking the complex and identification of the crosslinks by mass-spectrometry. Holo-NELF was observed to form a linear complex as reported previously (Narita et al., 2003) with both nucleic acid binding sites solvent accessible and located at opposite ends of the complex.

Based on the information obtained during this thesis it was possible to propose a new model how NELF contributes to establish promoter-proximal pausing and postulate that NELF acts partially by binding nascent RNA via its NELF-A subunit close to the RNA-exit tunnel.

In this study the first unique high-resolution structure of a NELF subcomplex could be solved and a novel function be assigned. Thus all proteins involved in promoter-proximal pausing are now structurally characterized to a substantial extent, providing the basis for investigation of its mechanism on a molecular level.

II Materials and Methods

1 Materials

1.1 Bacterial strains

Table 3: Bacterial strains used in this study

Strain	genotype	Resistance	Source
E. coli XL-1 Blue	rec1A; endA1, gyrA96; thi-1; hsdR17; supE44; elA1; lac[F'proAB lacIqZDM 15 tn10(Tetr)]	Tetracycline	Stratagene
E. coli BL21 (DE3)RIL	B; F-; ompT; hsdS(rB- mB-); dcm+; Tetr; gal λ(DE3); endA; Hte [argU, ileY, leuW, Camr]	Chloramphenicol	Stratagene

1.2 Plasmids, synthetic genes and oligonucleotides

1.2.1 Plasmids

Table 4: Vectors used for this study

ID	vector	Insert	tag [a]	Comment
DP 1	pET28a	NELF-A	N-His	[b]
DP 2	pOPIN-M	NELF-B	N-His-MBP	[b]
DP 3	pET21b	NELF-C		[b]
DP 4	pOPIN-F	NELF-E	N-His	[c]
DP 5	pET28a	NELF-A6-188	N-His	
DP 6	pET21b	NELF-C36-590		
DP 7	pET21b	NELF-C183-590		
DP 8	pET28a	NELF-A - IRES - NELF-C	N-His (at NELF-A)	
DP 9	pET28a	NELF-A6-188 - IRES - NELF-C36-590	N-His (at NELF-A)	
DP 10	pET28a	NELF-A mut R65Q, R66Q (6-188)	N-His	[d]
DP 11	pET28a	NELF-A mut K146M, K161M, K168M, R175Q (6-188)	N-His	[d]
DP 12	pET28a	NELF-A mut R65Q, R66Q, K146M, K161M, K168M, R175Q (6-188)	N-His	[d]
DP 13	pET21b	NELF-C mut R291Q, R315Q (183-590)		[d]
DP 14	pET21b	NELF-C mut K371M, K372M, K374M (183-590)		[d]
DP 15	pET21b	NELF-C mut K384M, K388M (183-590)		[d]
DP 16	pET21b	NELF-C mut R419Q, R506Q (183-590)		[d]

17

ID	vector	Insert	tag [a]	Comment
DP 17	pET21b	NELF-C mut R291Q, R315Q, K384M, K388M, R419Q, R506Q (183-590)		d
DP 18	pET21b	NELF-C mut K371M, K372M, K374M, K384M, K388M, R419Q, R506Q (183-590)		d
DP 19	pET21b	NELF-C mut K371M, K372M, K374M, R291Q, R315Q, K384M, K388M, R419Q, R506Q (183-590)		d
DP 20	pET28a	NELF-A mut T157D (6-188)	N-His	d
DP 21	pET28a	NELF-A mut T173D (6-188)	N-His	d
DP 22	pET21b	NELF-C mut T285D (183-590)		d
DP 23	pET21b	NELF-C mut Y289E (183-590)		d
DP 24	pET21b	NELF-C mut T318D (183-590)		
DP 25	pET21b	NELF-C 36-183		d
DP 26	pANY	NELF-A fl		e
DP 27	pANY	NELF-B fl		e
DP 28	pANY	NELF-C fl		e
DP 29	pET21b			f
DP 30	pET28a			f
DP 31	pOPIN-F			
DP 32	pOPIN-M			
DP 33	pET28a	NELF-A6-182	N-His	g
DP 34	pET28a	NELF-A6-193	N-His	g
DP 35	pET28a	NELF-A6-202	N-His	g
DP 36	pET28a	NELF-A9-188	N-His	g
DP 37	pET28a	NELF-A20-188	N-His	g
DP 38	pET28a	NELF-A29-188	N-His	g
DP 39	pET28a	NELF-A36-188	N-His	g
DP 40	pET21b	NELF-C36-559		g
DP 41	pET21b	NELF-C36-568		g
DP 42	pET21b	NELF-C36-573		g
DP 43	pET21b	NELF-C36-585		g
DP 44	pET21b	NELF-C30-590		g
DP 45	pET21b	NELF-C52-590		g
DP 46	pET21b	NELF-C55-590		g
DP 47	pET21b	NELF-C57-590		g
DP 48	pET28a	NELF-A mut E111A, E112A, Q113A (6-188)		g
DP 49	pET21b	NELF-C mut E138A, E139A, E141A (36-590)		g
DP 50	pET21b	NELF-C mut Q270A, E271A, K272A (36-590)		g

ID	vector	Insert	tag [a]	Comment
DP 51	pET21b	NELF-C mut E138A, E139A, E141A, K371A, K372A, K374A (36-590)		[g]
DP 52	pET21b	NELF-C mut Q270A, E271A, K272A, K371A, K372A, K374A (36-590)		[g]
DP 53	pET21b	NELF-C mut E138A, E139A, E141A, Q270A, E271A, K272A, K371A, K372A, K374A (36-590)		[g]

[a] N-"tag" indicates N-terminally attached tag

[b] synthetic gene, codon optimized for *E. coli*

[c] natural sequence from *H. sapiens* cDNA

[d] Protein used for fluorescence anisotropy experiments

[e] Template for molecular cloning

[f] Vector backbone for molecular cloning

[g] Protein used for crystallization experiments

1.2.2 Primers

Table 5: Primers used in this study

ID	Name	Sequence 5' to 3'
DP54	NELFA_fl_fwd	CGCGCGGCAGCCATATGCCTGGTCAACGTCG
DP55	NELFA_fl_rev	GGTGGTGGTGCTCGAGTTATTTCAAGACACATTCGTCATTGG
DP56	shNELF_B_Fwd	AAGTTCTGTTTCAGGGCCCGATGTTTGCCGGACTGCAGG
DP57	shNELF_B_Rev	ATGGTCTAGAAAGCTTTATTACAGAGGGGCAGGGGC
DP58	NELFC_fl_fwd	GGAGATATACATATGGCAGGTGCTGTTCC
DP59	NELFC_fl_rev	GCTCGAATTCGGATCCTCAGTTTACCATAATGAAGTTGCTTTTACAGTGAG
DP60	NELF_E_Fwd	AAGTTCTGTTTCAGGGCCCGATGTTGGTGATACCCCCCGGACT
DP61	NELF_E_Rev	ATGGTCTAGAAAGCTCTAGAAGCCATCCACAAGGTTTTCCTTGTAG
DP62	NELFA_R6_fwd	CGCGCGGCAGCCATATGGAATCTGATACCGGTCTGTGGCTG
DP63	NELF_A_Q188_rev	GGTGGTGGTGCTCGAGTTATTGTTGGGCAGTCTCGGTTG
DP64	NELFC_E36_fwd	GGAGATATACATATGGAAGGCGAAGATGATGCCGAG
DP65	NELFC_G183_fwd	AGGAGATATACATATGGGATATCAAGG CGAGATCACCTCTG

ID	Name	Sequence 5' to 3'
DP66	NELFA_fus_rev	TATATCTCCTTCTTAAAGTTAAACAAAATTATTTCAAGACACATTCGTCATTGG
DP67	NELFC_fus_fwd	TTTGTTTAACTTTAAGAAGGAGATATACCATGGCAGGTGCTGTTCC
DP68	NELFC_E36_fus_fwd	TTTGTTTAACTTTAAGAAGGAGATATACCATGGAAGGCGAAGATGATGCCGAGGTC
DP69	NELFA_Q188_fus_rev	GGTATATCTCCTTCTTAAAGTTAAACAAATTATTGTTGGGCAGTCTCGGTTG
DP70	NELFA_R65R66mut_fwd	TGGGTACACTGCATCTGCCTCAACAGACTGTGGATGAAATGAAAGG
DP71	NELFA_R65R66mut_rev	CCTTTCATTTCATCCACAGTCTGTTGAGGCAGATGCAGTGTACCCA
DP72	NELFA_K146M_fwd	TGGAATGTCAGTATCTGAACATGAACGCCCTGACTACACTGGC
DP73	NELFA_K146M_rev	GCCAGTGTAGTCAGGGCGTTCATGTTCAGATACTGACATTCCA
DP74	NELFA_K161M_fwd	GTCCTCTGACTCCACCTGTTATGCACTTCCAACTGAAACGTAA
DP75	NELFA_K161M_rev	TTACGTTTCAGTTGGAAGTGCATAACAGGTGGAGTCAGAGGAC
DP76	NELFA_K168M_fwd	AACACTTCCAACTGAAACGTATGCCGAAATCAGCGACACTGCG
DP77	NELFA_K168M_rev	CGCAGTGTCGCTGATTTCGGCATACGTTTCAGTTGGAAGTGTT
DP78	NELFA_R175Q_fwd	AACCGAAATCAGCGACACTGCAAGCCGAGCTGCTGCAAAAATC
DP79	NELFA_R175Q_rev	GATTTTTGCAGCAGCTCGGCTTGCAGTGTCGCTGATTTCGGTT
DP80	NELFC_R291Q_fwd	GTACAGCCGCTTCTTATCCTCAAGCCTGTCAGGCCCTGGGAGC
DP81	NELFC_R291Q_rev	GCTCCCAGGGCCTGACAGGCTTGAGGATAAGAAGCGGCTGTAC
DP82	NELFC_K315M_fwd	CCGATATTACCGTACTGTTTATGATGTTCACCAGCATGGACCC
DP83	NELFC_K315M_rev	GGGTCCATGCTGGTGAACATCATAAACAGTACGGTAATATCGG
DP84	KKNKmut_fwd	CTAGTGTTGTGGAAACGTGGATGATGAACATGCGTGTGTCTATTAACAAAGA

ID	Name	Sequence 5' to 3'
DP85	KKNKmut_rev	TCTTTGTTAATAGACACACGCATGTTCATCATCCACGTTTCCACAACACTAG
DP86	KSTSKmut_fwd	CTATTAACAAAGACGAACTGATGTCGACCAGCATGGCAGTGGAGACTGTCCACAA
DP87	KSTSKmut_rev	TTGTGGACAGTCTCCACTGCCATGCTGGTCGACATCAGTTCGTCTTTGTTAATAG
DP88	NELFC_R419Q_fwd	GTACACTGTATCAGTGTATTCAGTTCCCGGTTGTGGCAATGGG
DP89	NELFC_R419Q_rev	CCCATTGCCACAACCGGGAACTGAATACACTGATACAGTGTAC
DP90	NELFC_R506Q_fwd	GTATGGTACACCTGCTGTCTCAGGGTTATGTTCTGCCGGTTGT
DP91	NELFC_R506Q_ref	ACAACCGGCAGAACATAACCCTGAGACAGCAGGTGTACCATAC
DP92	NELFA_T157D_fwd	CTACACTGGCAGGTCCTCTGGATCCACCTGTTAAACACTTCCA
DP93	NELFA_T157D_rev	TGGAAGTGTTTAACAGGTGGATCCAGAGGACCTGCCAGTGTAG
DP94	T173Dmut_fwd	AACGTAAACCGAAATCAGCGGACCTGCGCGCCGAGCTGCTGCA
DP95	T173Dmut_rev	TGCAGCAGCTCGGCGCGCAGGTCCGCTGATTTCGGTTTACGTT
DP96	NELFC_T285D_fwd	AGATCACACTGGCTCTGGGTGACGCCGCTTCTTATCCTCGTGC
DP97	NELFC_T285D_rev	GCACGAGGATAAGAAGCGGCGTCACCCAGAGCCAGTGTGATCT
DP98	NELFC_Y289E_fwd	CTCTGGGTACAGCCGCTTCTGAGCCTCGTGCCTGTCAGGCCCT
DP99	NELFC_Y289E_rev	AGGGCCTGACAGGCACGAGGCTCAGAAGCGGCTGTACCCAGAG
DP100	NELFC_T285D_ Y289E_fwd	AGATCACACTGGCTCTGGGTGACGCCGCTTCTGAGCCTCGTGCCTGTCAGGCCCT
DP101	NELFC_T285D_ Y289E_rev	AGGGCCTGACAGGCACGAGGCTCAGAAGCGGCGTCACCCAGAGCCAGTGTGATCT
DP102	T318Dmut_fwd	CCGTACTGTTTAAAATGTTCGACAGCATGGACCCACCACCTGT
DP103	T318Dmut_rev	ACAGGTGGTGGGTCCATGCTGTCGAACATTTTAAACAGTACGG

ID	Name	Sequence 5' to 3'
DP104	T7_fwd	TAATACGACTCACTATAGGG
DP105	T7_rev	CTAGTTATTGCTCAGCGG
DP106	NELFA_S9_fwd	CGCGCGGCAGCCATATGTCTCCGAAAATGGCCTCAATGC
DP107	NELFA_T20_fwd	CGCGCGGCAGCCATATGACCGGTCTGTGGCTGCACA
DP108	NELFA_G29_fwd	CGCGCGGCAGCCATATGGGTGCTACCGATGAACTGTGGG
DP109	NELFA_A36_fwd	CGCGCGGCAGCCATATGGCTCCTCCGTCAATTGCTTCTCTG
DP110	NELFA_S182_rev	GCTCGAATTCGGATCCTCATGATTTCGGTTTACGTTTCAGTTGGAAG
DP111	NELFA_S193_rev	GCTCGAATTCGGATCCTCATGATTTTTGCAGCAGCTCGGC
DP112	NELFA_R202_rev	GCTCGAATTCGGATCCTCAGCGTTTCAGTTGTTGGGCAGTC
DP113	NELFC_N559_rev	GCTCGAATTCGGATCCTCAGTTTTCCAGAATCGGCAGAAACAGC
DP114	NELFC_T568_rev	GCTCGAATTCGGATCCTCAAGTTTTAATGGTGCCGGCAATGC
DP115	NELFC_D573_rev	GCTCGAATTCGGATCCTCAATCGTGCTCGCCTTCAGTTTTAATGG
DP116	NELFC_N585_rev	GCTCGAATTCGGATCCTCAGTTGCTTTTACAGTGAGCAATGAATTCGG
DP117	NELFC_Q30_fwd	GGAGATATACATATGCAAGAGGACGATTCCGGTGAAG
DP118	NELFC_S52_fwd	GGAGATATACATATGTCTACCCGTGACTATATCATGGAGCC
DP119	NELFC_D55_fwd	GGAGATATACATATGGACTATATCATGGAGCCGAGCATTTTTAACAC
DP120	NELFC_I57_fwd	GGAGATATACATATGATCATGGAGCCGAGCATTTTTAACAC
DP121	NELFC_3Ecluster_mut_fwd	GATTCCATTTTTACCGCAGCGGGCGCAACTCCAGCGTGGCTGG
DP122	NELFC_3Ecluster_mut_rev	CCAGCCACGCTGGAGTTGCGCCCGCTGCGGTAAAAATGGAATC
DP123	NELFC_QEKcluster_mut_fwd	GGAAGTCCAACGCTTTGCCGCAGCGGCAGGGCATGATGCTAGCCAGATCAC
DP124	NELFC_QEKcluster_mut_rev	GTGATCTGGCTAGCATCATGCCCTGCCGCTGCGGCAAAGCGTTGGACTTCC
DP125	NELFC_3Kcluster_mut_fwd	CCTATGCTGCTAGTGTTGTGGAAACGTGGGCAGCAAACGCACGTGTGTCTATTAACAAAGACGAACTGAAATCG
DP126	NELFC_3Kcluster_mut_rev	CGATTTCAGTTCGTCTTTGTTAATAGACACACGTGCGTTTGCTGCCCACGTTTCCACAACACTAGCAGCATAGG

22

ID	Name	Sequence 5' to 3'
DP127	NELFA1_EEQcluster _mut_fwd	GTTCACTGAACCTGGAACTGGCAGCAGCGAATCCGAACGT CCAAGAC
DP128	NELFA1_EEQcluster _mut_rev	GTCTTGGACGTTCGGATTCGCTGCTGCCAGTTCCAGGTTC AGTGAAC

1.2.3 Genes

Table 6: Synthetic genes used in this study

Name	organism	comment
NELF-A	*H. sapiens*	synthetic, codon optimized for *E. coli*
NELF-B	*H. sapiens*	synthetic, codon optimized for *E. coli*
NELF-C	*H. sapiens*	synthetic, codon optimized for *E. coli*

1.2.4 Nucleic acids for fluorescence anisotropy

Table 7: Nucleic acids used for fluorescence anisotropy experiments in this study

ID	Name	Type	Sequence 5' to 3'	Modification
DP 129	c-fos_RNA	ssRNA	CCGCAUCUGCAGCGAGCAUCUGAGA	5' 6-FAM
DP 130	junB_RNA	ssRNA	AGCGGCCAGGCCAGCCUCGGAGCCA	5' 6-FAM
DP 131	44%_RNA	ssRNA	ACCCCACAACUAAAAAAUCCCAACC	5' 6-FAM
DP 132	60%_RNA	ssRNA	AAGGGGAGCGGGGGAGGAUAAUAGG	5' 6-FAM
DP 133	72%_RNA	ssRNA	ACCACCCACCCACCCCACCGAACGC	5' 6-FAM
DP 134	c-fos_DNA	ssDNA	AAGACTGAGCCGGCGGCCGC	5' 6-FAM
DP 135	junB_DNA	ssDNA	AGGGAGCTGGGAGCTGGGGG	5' 6-FAM
DP 136	44%_DNA	ssDNA	ACCCCACAACTAAAAAATCCCAACC	5' 6-FAM
DP 137	60%_DNA	ssDNA	AAGGGGAGCGGGGGAGGATAATAGG	5' 6-FAM
DP 138	72%_DNA	ssDNA	ACCACCCACCCACCCCACCGAACGC	5' 6-FAM
DP 139	44%_rev_comp	ssDNA	GGTTGGGATTTTTTAGTTGTGGGGT	
DP 140	60%_rev_comp	ssDNA	CCTATTATCCTCCCCCGCTCCCCTT	
DP 141	72%_rev_comp	ssDNA	GCGTTCGGTGGGGTGGGTGGGTGGT	
DP 142	40%_hybrid	RNA:DNA hybrid	ACCCCACAACUAAAAAAUCCCAACC	5' 6-FAM [a]
DP 143	60%_hybrid	RNA:DNA hybrid	AAGGGGAGCGGGGGAGGAUAAUAGG	5' 6-FAM [a]
DP 144	72%_hybrid	RNA:DNA hybrid	ACCACCCACCCACCCCACCGAACGC	5' 6-FAM [a]

23

ID	Name	Type	Sequence 5' to 3'	Modification
DP 145	44%_dsDNA	dsDNA	ACCCCACAACTAAAAAATCCCAACC	5' 6-FAM [b]
DP 146	60%_dsDNA	dsDNA	AAGGGGAGCGGGGGGAGGATAATAGG	5' 6-FAM [b]
DP 147	72%_dsDNA	dsDNA	ACCACCCACCCACCCCACCGAACGC	5' 6-FAM [b]

[a] RNA strand is modified

[b] sense strand is modified

1.3 Reagents and consumables

Table 8: Reagents and consumables used in this study

Type	Source
Chemicals	Merck (Darmstadt, Germany), Roth (Karlsruhe, Germany), Sigma-Aldrich (Schnelldorf, Germany)
Enzymes, reagents and buffers for molecular cloning	Fermentas (St. Leon-Rot, Germany), NEB (Frankfurt am Main, Germany)
Oligonucleotides	ThermoScientific (Ulm, Germany), biomers (Ulm, Germany), IDT DNA (Leuven, Belgium)
Synthetic genes	Mr. Gene (Regensburg, Germany)
Commercial kits for DNA preparation	Qiagen (Hilden, Germany)
Crystallization reagents and tools	Hampton Research (Aliso Viejo, CA, USA), Qiagen (Hilden, Germany)

1.4 Growth media and additives

Table 9: Bacterial growth media used in this study

Medium	Description	Application
Lysogeny broth (LB)	1 % (w/v) Tryptone; 0.5 % (w/v) Yeast extract; 1 % (w/v) NaCl	*E. coli* culture
LB plates	1 % (w/v) Tryptone; 0.5 % (w/v) Yeast extract; 1 % (w/v) NaCl; 1.5 % (w/v) agar; 1x Antibiotic	*E. coli* culture on plates
X-Gal plates	1 % (w/v) Tryptone; 0.5 % (w/v) Yeast extract; 1 % (w/v) NaCl; 1.5 % (w/v) agar; 1x Antibiotic; 0.02 % X-Gal dissolved in N,N-Dimethylformamide (DMF)	*E. coli* culture, selection of pOPIN vectors containing the correct insert
SeMet medium	Obtained from AthenaES	Expression of seleno-methionine-substituted proteins in *E. coli*

Table 10: Growth media additives used in this study

Additive	Concentration	Application
Ampicillin	100 µg/ml	Selection of *E. coli*
Kanamycin	30 µg/ml	Selection of *E. coli*
Chloramphenicol	30 µg/ml	Selection of *E. coli*
Tetracycline	12.5 µg/ml	Selection of *E. coli*
IPTG	0.1 mM - 1.0 mM	Induction of protein expression in *E. coli*

1.5 Buffers and solutions

1.5.1 General buffers, solutions and dyes

Table 11: General buffers, solutions and dyes used in this study

Name	Composition/Source	Application
50x TAE	250 mM EDTA; 12.5 M Tris-acetate, pH = 8.0	Agarose gel electrophoresis
TE	1 mM Tris, pH 8.0; 0.1 mM EDTA	DNA storage
6x DNA-loading dye	1.5 g/L Bromphenol blue; 1.5 g/L Xylene cyanol; 50 % (v/v) Gylcerol (Fermentas)	Agarose gel electrophoresis
MOPS Electro-phoresis buffer	NuPAGE buffer (life technologie)	SDS-PAGE
5x SDS-loading buffer	10 % (w/v) SDS; 10 mM DTT; 20 % (v/v) glycerole; 0.2 M Tris-HCl, pH = 6.8; 0.05 % (w/v) Bromphenolblue	SDS-PAGE
Western blot transfer buffer	25 mM Tris; 192 mM glycine; 20 % EtOH	Western blot
Gel staining solution	Instant blue (Expedion)	Coomassie staining of PA gels
Thiosulfate solution	0.02 % sodium thiosulfate	Silver staining of PA gels
Silver nitrate solution	0.1 % silver nitrate; 0.02 % formaldehyde	Silver staining of PA gels
Developing solution	3 % sodium carbonate; 0.05 % formaldehyde	Silver staining of PA gels
100x PI	1.42 mg leupeptin; 6.85 mg pepstatin A; 850 mg PMSF; 1.685 mg benzamidine in 50 ml ethanol	protein purification
Denaturating buffer	8 M urea; 1 M Tris pH = 8.0	Resuspension of TCA-precipitated protein

Name	Composition/Source	Application
TFB-I	30 mM KAcetate; 50 mM MnCl$_2$; 100 mM RbCl; 10 mM CaCl$_2$; 15 % (v/v) glycerol	Prepare chemically competent *E. coli*
TFB-II	10 mM MOPS, pH = 7.0; 75 mM CaCl$_2$; 10 mM RbCl; 15 % (v/v) glycerol	Prepare chemically competent *E. coli*

1.5.2 Protein purification buffers

Table 12: Buffers used for protein purification in this study

Name	Composition	Application
Lysis buffer A	150 mM NaCl; 40 mM Na-HEPES pH 7.4 at 4°C; 10 mM imidazole; 2 mM DTT; 1x PI	NELF-AC purification
Lysis buffer B	250 mM NaCl; 40 mM Na-HEPES pH 7.4 at 4°C; 10 mM imidazole; 2 mM DTT; 1x PI	NELF-ABC purification
Lysis buffer C	500 mM NaCl; 40 mM Na-HEPES pH 7.4 at 4°C; 10 mM imidazole; 2 mM DTT; 1x PI	NELF-E purification
Dialysis buffer A	150 mM NaCl; 40 mM Na-HEPES pH 7.4 at 4°C; 2 mM DTT	NELF-AC purification
Dialysis buffer B	500 mM NaCl; 40 mM Na-HEPES pH 7.4 at 4°C; 2 mM DTT	NELF-E purification
Washing buffer MBP	300 mM NaCl; 40 mM Na-HEPES pH 7.4 at 4°C; 10 % glycerole; 2 mM DTT	NELF-ABC purification
Elution buffer MBP	300 mM NaCl; 40 mM Na-HEPES pH 7.4 at 4°C; 10 % glycerole; 2 mM DTT; 40 g/l maltose	NELF-ABC purification
Ion exchange low/ high salt buffer	100/1000 mM NaCl; 40 mM Na-HEPES pH 7.4 at 4°C; 2 mM DTT	Ion exchange chromatography
Size exclusion buffer A	150 mM NaCl; 10 mM Na-HEPES pH 7.4 at 4°C; 2 mM DTT	Size exclusion chromatography of NELF-AC for crystallization
Size exclusion buffer B	500 mM NaCl; 10 mM Na-HEPES pH 7.4 at 4°C; 2 mM DTT	Size exclusion chromatography of complete-NELF
Size exclusion buffer C	50 mM NaCl; 10 mM Na-HEPES pH 7.4 at 4°C; 2 mM DTT	Size exclusion chromatography of NELF-AC for fluorescence anisotropy

1.5.3 Fluorescence anisotropy buffers

Table 13: Buffers used for fluorescence anisotropy

Name	Composition	Application
Dilution buffer	50 mM NaCl; 40 mM Na-HEPES pH 7.4 at 4°C; 2 mM DTT	Serial dilution of protein
2.5x buffer	12.5 mM NaCl; 12.5 mM Na-HEPES pH 7.4; 7.5 mM $MgCl_2$; 2.5 mM DTT; 125 µg/ml BSA	Adjustment of final concentrations

1.5.4 Crosslinking buffers and solutions

Table 14: Buffers used for crosslinking

Name	Composition	Application
Crosslinking solution	disuccinimidyl suberate (DSS) 50 mM solution in DMSO prepared immediate before use, life technologies	Crosslinking
Quenching solution	1 M ammonium bicarbonate	Crosslinking

1.6 Crystallization screens

Table 15: 96-well high-throughput crystallization screens used in this study

Screen	Source
AJ1	in-house production[a]
AJ2	in-house production[a]
Complex screen	in-house production[a]
Complex screen 2	in-house production[a]
Crystal platform Magic 1	in-house production[a]
Crystal platform Magic 2	in-house production[a]
Morpheus	in-house production[a]
Wizars I II	in-house production[a]
Hampton research Index	Hampton
Qiagen Classics Suite	Qiagen
Qiagen Cryos Suite	Qiagen
Qiagen JCSG+ Suite	Qiagen
Qiagen PACT Suite	Qiagen
Qiagen PEGs Suite	Qiagen

[a] In-house production of the Crystallization Facility at the Max Planck Institute of Biochemistry (Martinsried, Germany).

2 Methods

2.1 Molecular cloning

Polymerase chain reaction

All inserts for molecular cloning were amplified by polymerase chain reaction (PCR) from codon optimized, synthetic genes (Mr. Gene) (NELF-A, -B and -C) or *H.sapiens* cDNA (NELF-E). PCR programs comprised 35 cycles (Biometra T3000 Thermocycler). Annealing temperature and elongation time were adjusted to the respective required conditions of the primers and template.

Primers were designed using the online NEB Tm-calculator (www.tmcalculator.neb.com) and optimized for an annealing temperature of 55°C in the complementary region. Primers used for molecular cloning contained an 5' overhang of 12 or 15 nt including the restriction site for conventional or ligation independent cloning (LIC), respectively, followed by the sequence complementary to the gene of interest. Tags and protease cleavage sites were introduced by in-frame cloning into a suitable vector.

For fragment amplification 2x Phusion High Fidelity PCR Master Mix (NEB) was used with a final primer concentration of 500 nM and 50 ng of pure DNA template in 50 µl final volume.

For colony PCR *Thermus aquaticus* (*Taq*) DNA polymerase (Fermentas) was used with final primer and Mg^{2+} concentrations of 640 nM and 2.5 mM, respectively, in 25 µl total volume according to the manufacturers instructions. Single *E. coli* colonies were picked to be used as template, resuspended in the complete reaction mix and additionally streaked on a LB-plate containing the appropriate antibiotic. Colonies containing the correct insert were later retrieved from this LB-plate.

Mutant genes were generated by amplifying two overlapping PCR products containing the mutant site in the overlapping region at the 3' and 5' end of the sense strand, respectively. In a second step both fragments were joined by fusion PCR.

PCR products were visualized by electrophoretic separation in a PerfectBlue Gelsystem electrophoresis chamber using 0.5-1 % agarose gel and staining with Sybr Safe diluted 1:10,000 (Invitrogen). Purification of PCR products from agarose gels was carried out using the QIAquick gel extraction kit (Qiagen).

Restriction and ligation

Vectors and inserts were digested with restriction endonucleases (NEB) according to the

manufacturers manual and purified with QIAquick PCR purification kit (Qiagen). 40 ng vector and a seven-fold molar excess of insert were ligated using T4 DNA ligase (NEB) in 20 µl reaction volume according to the manual. Cloning into pOPIN vectors was done by ligation independent cloning (LIC) using the InFusion Kit (Clontech) in 10 µl total volume according to the manufacturers manual.

Preparation and transformation of competent cells

Two strains of chemically competent *E. coli* cells were used for transformations. Ligation products and plasmids (for amplification) were transformed into XL1-Blue. For protein overexpression from the correspondent plasmids BL21-CodonPlus(DE3)-RIL were employed (Table 3).

To prepare chemically competent cells 400 ml of LB including the appropriate antibiotic were inoculated 1:100 with cells from an ON LB-culture, grown at 37°C until OD_{600} ~ 0.5 and cooled down on ice to stop growth. Subsequently cells were harvested by centrifugation for 10 min at 4°C and 4000 rpm, resuspended in 100 ml prechilled TFB-I on ice and centrifuged as before. The pellet was resuspended in 8 ml prechilled TFB-II on ice. 50 µl aliquots were flash frozen in liquid nitrogen and stored at -80°C.

100 ng of each pure plasmid, 4 µl ligated plasmid or 2.5 µl InFusion product, respectively, were transformed into chemically competent cells by heat shock. 50 µl cells were thawed on ice and incubated with DNA for 20 min on ice followed by 45 sec at 42°C and 2 min on ice. Subsequently 450 µl of LB were added and the cells shaken for 1 h at 37°C. Cells were spread on selective LB- plates and grown ON at 37°C.

Plasmid verification, preparation and storage

Bacterial colonies containing a plasmid with an insert were verified by colony-PCR as described and incubated ON in 10 ml LB. Plasmids were prepared from this ON cultures using Miniprep purification kits (Qiagen). To obtain glycerol stocks 500 µl of an ON culture were mixed with 500 µl glycerole and stored at -80°C.

2.2 Protein methods

2.2.1 General protein methods

Protein analysis methods

For sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) the NuPAGE

system (life technologies) was employed using 4-12 % gradient gels. Protein samples were mixed with 5x SDS-loading buffer to a final concentration of 1x and incubated for 3 min at 95°C. SDS-PAGE Molecular Weight Standard Broad Range (Bio-Rad) and PageRuler Prestained Protein Ladder (Fermentas) were used as molecular weight standards. Gels were developed in 1x MOPS buffer at 200 V until the dye reached the lower end of the gel and stained with instant blue (Expedeon).

In case of low protein concentration TCA-precipitation was performed before SDS-PAGE analysis. Trichloroacetic acid (TCA) was added to a final concentration of 10 %, incubated on ice for 20 min and centrifuged at 15000 rpm and 4°C for 15 min. The pellet was washed twice with prechilled (-20°C) acetone on ice, dried at 50°C and resuspended in 12 µl denaturing buffer.

Trace amounts of protein on a PA gel were detected by silver staining. The PA gel was washed in ddH$_2$O thoroughly and incubated in thiosulfate solution for 1 min. After washing 3x20 sec with ddH$_2$O the gel was incubated in silver nitrate solution for 20 min. After washing 3x20 sec with ddH$_2$O the gel was incubated in developing solution until the protein bands were visible and the reaction stopped with 5 % (v/v) acetic acid.

Protein concentration was determined with a NanoDrop 1000 spectrophotometer (Peqlab) using protein specific parameters regarding the molar attenuation coefficient ε and molecular weight.

Dynamic light scattering (DLS) was done with a Viscotek 802 DLS (Malvern Instruments) and the result analysed with the OmniSIZE software.

Limited proteolysis and fragment identification

Limited proteolysis was employed to identify stable fragments of proteins that were more likely to form crystals. 30 µg of protein in 100 µl total reaction volume were incubated at 37°C with varying amounts of protease (chymotrypsin or subtilisin) in order to ensure observable and complete degradation of flexible regions within 30 min. Samples of 12 µl volume were taken at varying points in time and the reaction stopped by immediate mixing with 3µl 5x SDS-loading dye and incubation at 95°C for 3 min. Degradation products were analysed by PAGE, transferred to a PVDF-membrane by Western blot (35 V, ON) and identified by Edman-Sequencing at the Max Planck Institute of Biochemistry core facility (Martinsried, Germany). Unknown protein bands were identified by MALDI-MS peptide mass fingerprinting at the Adolf Butenandt Institut, Zentrallabor für Proteinanalytik (ZfP) (Munich, Germany).

2.2.2 Purification of recombinant proteins

NELF-A$_{6-188}$C$_{36-590}$ and NELF-A$_{6-188}$C$_{183-590}$

The borders of NELF-A and NELF-C within the NELF-AC subcomplex were determined by limited proteolysis of human full-length NELF-AC complex with chymotrypsin and subtilisin followed by Edman sequencing. Human NELF-A and NELF-C were amplified from codon optimized DNA (Mr. Gene) and cloned into pET28a and pET21b vectors, between *NdeI* and *XhoI* or *NdeI* and *BamHI* restriction sites, respectively, resulting in N-terminally His$_6$-tagged NELF-A (6-188) and untagged NELF-C (36-590 or 183-590).

Plasmids encoding NELF-A (6-188) and NELF-C (36-590 or 183-590) were co-transformed into *E. coli* BL21 CodonPlus (DE3) RIL cells. Cells were grown in LB medium at 37°C until OD$_{600}$ ~0.6 and cooled on ice for 30 minutes. Protein expression was induced by the addition of 1 mM IPTG. After induction, cells were grown for an additional 16 hrs at 18°C. All purification steps were performed at 4°C. Cells were resuspended and lysed in lysis buffer A including 1x protease inhibitor. The lysate was applied to Ni-NTA agarose beads (Qiagen) and washed extensively with lysis buffer A containing 20 and 40 mM imidazole. Protein was eluted from the beads with lysis buffer A containing 200 mM imidazole. The eluted protein was mixed with 1 U thrombin/mg protein (Sigma) and dialyzed against dialysis buffer A for 16 hrs at 4°C. The protein was applied to Ni-NTA beads equilibrated in dialysis buffer A to remove uncleaved protein. The Ni-NTA flow through was applied to an anion exchange column (HiTrap Q-HP, 1 ml, GE Healthcare) equilibrated in ion exchange low salt buffer. Protein was eluted via a salt gradient from 100 mM (low salt) to 1 M NaCl (high salt) in ion exchange buffer. The protein was further purified by size exclusion chromatography with the use of a Superose 6 10/300 column (GE Healthcare) equilibrated in size exclusion buffer A. Peak fractions were pooled and concentrated by centrifugation in Amicon Ultra 4 ml concentrators (30 kDa and 10 kDa MWCO, respectively) (Millipore) to 6 mg/ml (NELF-A$_{6-188}$C$_{36-590}$) and 12 mg/ml (NELF-A$_{6-188}$C$_{183-590}$), respctively. Protein concentration was determined as described in 2.2.1. Protein was aliquoted, flash frozen, and stored at -80° C.

Selenomethionine-labeled protein NELF-A$_{6-188}$C$_{183-590}$

For production of selenomethionine-labeled protein, NELF-AC (6-188 and 183-590) plasmids were co-transformed into *E. coli* B834(DE3) cells. For protein expression, cells were grown in SelenoMet Medium (Table 9) supplemented with 40 µg/ml L-selenomethionine (SeMet). Selenomethionine-labeled protein was purified as above.

NELF-A$_{6-188}$C$_{36-590}$ mutant poteins

All NELF-A$_{6-188}$C$_{36-590}$ mutant poteins were purified like the wildtype protein.

NELF-ABC and ABCE

Human NELF-B (1-580) was cloned into a pOPIN-M-vector (Berrow et al., 2007) (provided by OPPF-UK) to produce a N-terminally His$_6$-MBP-tagged protein. Bicistronic expression plasmids containing full-length His$_6$-tagged NELF-A and NELF-C or His$_6$-tagged NELF-A (6-188) and NELF-C (36-590) were produced in the pET28a background between *NdeI* and *XhoI* sites. Full length NELF-E was cloned into a pOPIN-F-vector (Berrow et al., 2007) (provided by OPPF-UK) resulting in a protein with N-terminal His$_6$-tag. All protein expressions were performed as described above for the NELF-AC constructs. All purification steps were performed at 4°C and protease inhibitor was used during lysis only.

NELF-E was purified in lysis buffer C using Ni-NTA chromatography as described for NELF-AC. Protein was dialyzed against dialysis buffer B for 16 hrs at 4°C and simultaneously the His-tag was cleaved by 3C protease. Uncleaved protein and 3C protease were removed by a second Ni-NTA chromatography step. NELF-E was further purified by heparin affinity chromatography using a salt gradient from 100 mM to 1 M NaCl in ion exchange buffer.

The full-length and truncated NELF-ABC complexes were obtained by coexpression of full-length or truncated NELF-AC from a bicistronic vector together with full-length NELF-B. The complex was purified by Ni-NTA in lysis buffer B and amylose resin affinity chromatography (NEB) in washing buffer MBP. Protein was eluted from amylose resin with elution buffer MBP and diluted with water to achieve a final salt concentration of 150 mM NaCl. The complex was purified further by anion exchange chromatography (HiTrap Q-HP, 1 ml, GE Healthcare) as described for NELF-AC, and the salt concentration was adjusted to 500 mM NaCl. To remove the His$_6$-MBP -tag from NELF-B, the complex was incubated with 3C protease for 16 hrs at 4ºC. To obtain full-length and truncated NELF complexes, an excess of heparin-purified NELF-E was added to pure NELF-ABC prior to cleavage of the His$_6$-MBP- tag from NELF-B. Both the NELF-ABC and complete NELF complex were then applied to a Superose 6 10/300 size exclusion column equilibrated in size exclusion buffer B. Truncated NELF-ABC and complete NELF were concentrated by centrifugation to 2.1 mg/ml and 3.3 mg/ml, respectively, using Amicon Ultra 4 ml Centrifugal Filters (50 kDa MWCO) (Milipore), flash frozen, and stored at -80°C.

2.2.3 Protein interaction studies

Popin-M vector containing full length His-MBP-NELF-B was transformed into *E. coli* BL21 CodonPlus (DE3) RIL cells. NELF-B was expressed and purified by Ni-NTA chromatogaphy as described above. His-MBP-NELF-B was bound to amylose resin and incubated for 16 hrs at 4°C with NELF-AC or NELF-C constructs NELF-A$_{6-188}$C$_{36-590}$, NELF-A$_{6-188}$C$_{183-590}$, NELF-C$_{36-190}$ and NELF-C$_{183-590}$ which have been purified by Ni-NTA as described above for NELF-A$_{6-188}$C$_{36-590}$ and NELF-A$_{6-188}$C$_{183-590}$. After incubation the resin was washed extensively with washing buffer MBP, eluted with elution buffer MBP and the eluate analyzed by PAGE.

2.3 X-Ray crystallography

2.3.1 Crystallization screens of NELF-A$_{6-188}$C$_{36-590}$

Screens to find initial crystallization conditions for NELF-A$_{6-188}$C$_{36-590}$ were conducted at the crystallization facility of the Max Planck Institute of Biochemistry (Martinsried, Germany) using 96-well high throughput screens from Hampton, Qiagen and in-house productions (Table 15). Screens were set up using a nanoliter crystallization robot (Phoenix) by mixing each 100 nl protein and reservoir solution. Extensive screening yielded conditions that were used as starting point. Initial crystals were improved manually in 24-well hanging drop plates (VDX plates with sealant, Hampton, 0.5 ml reservoir volume) by varying pH, concentrations of salt and precipitant, temperature from 4°C – 20°C and drop size from 0.5 – 4 µl. Thin needle shaped crystals grew within 3-5 days in 1.1 – 1.2 M (NH$_4$)$_2$SO$_4$, 0.2 M NaCl and 100 mM Na-HEPES pH 7.4.

2.3.2 Crystallization screens and optimization of NELF-A$_{6-188}$C$_{183-590}$

Screens to find crystallization conditions for NELF-A$_{6-188}$C$_{183-590}$ were conducted at the crystallization facility of the Max Planck Institute of Biochemistry (Martinsried, Germany) using 96-well high throughput screens from Hampton, Qiagen and in-house production (Table 15). Screens were set up using a nanoliter crystallization robot (Phoenix) by mixing each 100 nl protein and reservoir solution. Extensive screening yielded a suitable crystallization condition to produce well-diffracting crystals. Initial crystals were optimized manually in 15-well EasyXtal hanging drop vapor diffusion plates (Qiagen, 0.2 ml reservoir volume) by varying pH, concentrations of salt and precipitant, temperature from 4°C – 20°C and drop size between 0.5 and 4 µl.

Optimized native and selenomethionine-labeled NELF-AC crystals were grown by hanging-drop vapor diffusion and obtained by mixing 1μl NELF-AC protein (12 mg/ml) with 1 μl reservoir solution containing 14-14.5 % (w/v) PEG 3350 and 200 mM sodium malonate pH 6.8-7.0. Tetrahedral NELF-AC crystals grew within 3-5 days. Crystals were cryo-protected in mother liquor containing 25 % (w/v) glucose, and flash frozen in liquid nitrogen.

2.3.3 Data collection and processing

Diffraction data for native NELF-A$_{6-188}$C$_{183-590}$ crystals were collected under cryo conditions (100 K) in 0.1° increments at beamline X06DA of the Swiss Light Source in Villigen (Switzerland) using a wavelength of 1.0000 Å and a Pilatus 2M-F detector (Table 17) (Broennimann et al., 2006).

Raw data were processed and scaled with XDS. The structure was solved by single isomorphous replacement with anomalous scattering (SIRAS) using diffraction data from an isomorphous crystal of SeMet-labeled protein. Location of 13 selenomethionine sites, calculation of initial phases and density modification were performed with the SHELX suite (Sheldrick, 2008). An initial model was built with Buccaneer (Cowtan, 2006). The model was iteratively built with COOT (Emsley and Cowtan, 2004) and refined with REFMAC (Vagin et al., 2004) and phenix.refine (Afonine et al., 2005) until the R-factors converged (Table 17).

2.4   Identification of P-TEFb *in-vivo* phosphorylation sites on NELF-A$_{6-188}$C$_{183-590}$

All experiments described in chapter II 2.4 were conducted by Seychelle M. Vos (NELF-AC phosphorylation by P-TEFb) and Henning Urlaub (Phosphopeptide enrichment and LC-MSMS analysis), both members of the MPI for Biophysical Chemistry, Goettingen.

NELF-AC phosphorylation by P-TEFb

Human Cdk9 (1-372) and Cyclin T1 (1-272) were co-expressed in High 5 insect cells and purified as described (Schulze-Gahmen et al., 2013). NELF-A$_{6-188}$C$_{183-590}$ (10-200 μM) and P-TEFb (0.4-2 μM) were incubated for 4-16 hrs at 30 ºC in a buffer containing 3-10 mM ATP, 3-15 mM MgCl$_2$, 1 mM DTT, 30 mM Na-HEPES pH 7.4, 4 % (v/v) glycerol, and 100 mM NaCl. The kinase-treated NELF-AC protein (100-500 pmol) was analyzed on a 4-12

% Bis-Tris acrylamide gel and submitted for mass spectrometry.

Phosphopeptide enrichment and LC-MSMS analysis

Phosphopeptides derived after in-gel digest of the sample were enriched as described previously (Oellerich et al., 2009). Enriched phosphopeptides were analyzed on a LC-coupled Q-Exactive HF mass spectrometer (ThermoFisherScientific) under standard chromatography conditions as described (Oellerich et al., 2009). The MS raw files were processed by MaxQuant (Cox and Mann, 2008) (version 1.5.2.8) and MS/MS spectra were searched against Uniprot human database with Andromeda (Cox et al., 2011) search engine. Allowed variable modifications included phosphorylation of serine, threonine, and tyrosine, methionine oxidation, and carbamidomethlyation of cysteine.

2.5   Fluorescence anisotropy

All experiments described in chapter II 2.5 were planned and conducted in cooperation with Seychelle M. Vos.

Preparation of mutant proteins

WT and mutant NELF-A$_{6-188}$C$_{183-590}$ proteins (Table 4) were expressed and purified as described above (II 2.2.2). For the final size exclusion step, the column was equilibrated in size exclusion buffer C. Peak fractions were pooled, concentrated by centrifugation to 30 mg/ml, aliquoted, flash frozen, and stored at -80ºC.

Nucleic acids

5' 6-FAM labeled ssDNA, ssRNA and dsDNA were obtained from Integrated DNA Technologies and dissolved in water to 100 µM. Sequences of artificial ssDNA and dsDNA with 44 %, 60 % and 72 % GC content are listed in Table 7 (corresponding sequences for ssRNA). Natural ssDNA sequences correspond to sequences of exposed coding (non-template) strands at the *c-fos* gene ('DP 134', bps 87-96 downstream of the TSS (Fivaz et al., 2000)) and the *junB* gene ('DP 135', bps 45-54 downstream of the TSS (Aida et al., 2006)) during promoter-proximal pausing +/-5 bps (Figure 14E). Natural ssRNA sequences correspond to 25 nt of nascent mRNA sequence predicted to be proximal to the RNA exit pore on the Pol II surface at *c-fos* ('DP 129', bps 53-77 relative to TSS) and *junB* ('DP 130', bps 13-37 relative to TSS) during promoter-proximal pausing (Figure 14E). To produce RNA-DNA hybrids, labeled ssRNA template was mixed at a 1:1 molar ratio with reverse complementary ssDNA (Thermo Fisher Scientific), heated to 95° for 10 min

and cooled down at a rate of 0.1°/min.

Assays

NELF- A$_{6-188}$C$_{183-590}$ was serially diluted in 2-fold steps in dilution buffer (Table 13). Nucleic acid (2.4µl, 10nM final concentration) and NELF-AC (12µL, 100-0.1µM final concentration) were mixed on ice and incubated for 10 minutes. The assay was brought to a final volume of 24 µl 2.5x buffer (Table 13) and incubated for 20 min at RT in the dark (final conditions: 30 mM NaCl, 3 mM MgCl$_2$, 10 mM Na-HEPES 7.4, 2 mM DTT and 50 µg/ml BSA). 20 µl of each solution was transferred to a Greiner 384 Flat Bottom Black Small volume plate.

Measurement and evaluation

Fluorescence anisotropy was measured at 30°C with an Infinite® M1000Pro reader (Tecan) with an excitation wavelength of 470 nm (±5 nm), an emission wavelength of 518 nm (±20 nm) and a gain of 72. All experiments were done in triplicate and analyzed with GraphPad Prism Version 6. Binding curves were fit with a single site quadratic binding equation:

$$y = \left(\frac{Bmax * \left([x] + [L] + Kd, app - \sqrt{\left([x] + [L] + Kd, app \right)^2 - 4 \left([x] * [L] \right)} \right)}{2 * [L]} \right)$$

where Bmax is the maximum specific binding, L is the concentration of nucleic acid, x is the concentration of NELF-A$_{6-188}$C$_{183-590}$, Kd,app is the apparent disassociation constant for NELF- A$_{6-188}$C$_{183-590}$ and nucleic acid. Error bars (Figure 14, Figure 15, Figure 18, Figure 26, Figure 27) are representative of the standard deviation from three experimental replicates.

2.6 Crosslinking and mass spectrometry

Crosslinking with disuccinimidyl suberate (DSS)

Truncated NELF-ABC complex and complete NELF complex were incubated with 1.0 mM and 1.1 mM DSS H12/D12 (Creative Molecules), respectively, for 30 min at 30°C. The crosslinking reaction was quenched by adding ammonium bicarbonate to a final concentration of 100 mM and incubation for 20 min at 30°C (Table 14).

Mass spectrometry identification of lysine-lysine crosslinking sites

All experiments described in this paragraph were performed by Tomasz Zimniak and Franz Herzog, both members of the Gene Center, LMU Munich.

The chemical cross-links on NELF complexes were identified by mass spectrometry as described previously (Herzog et al., 2012). Briefly, cross-linked complexes were reduced with 5 mM TCEP (Thermo Scientific) at 35°C for 15 min and subsequently treated with 10 mM iodoacetamide (Sigma-Aldrich) for 30 min at room temperature in the dark. Digestion with lysyl enodpeptidase (Wako) was performed at 35°C, 6M Urea for 2 h (at enzyme-substrate ratio of 1:50 w/w) and was followed by a second digestion with trypsin (Promega) at 35°C overnight (also at 1:50 ratio w/w). Digestion was stopped by the addition of 1 % (v/v) trifluoroacetic acid (TFA). Acidified peptides were purified using C18 columns (Sep-Pak, Waters). The eluate was dried by vacuum centrifugation and reconstituted in water/acetonitrile/TFA, 75:25:0.1. Cross-linked peptides were enriched on a Superdex Peptide PC 3.2/30 column (300 × 3.2 mm) at a flow rate of 25 µl min−1 and water/acetonitrile/TFA, 75:25:0.1 as a mobile phase. Fractions of 100 µl were collected, dried, and reconstituted in 2 % acetonitrile and 0.2 % FA, and further analyzed by liquid chromatography coupled to tandem mass spectrometry using a hybrid LTQ Orbitrap Elite (Thermo Scientific) instrument. Cross-linked peptides were identified using xQuest (Walzthoeni et al., 2012). False discovery rates (FDRs) were estimated by using xProphet (Walzthoeni et al., 2012). and results were filtered according to the following parameters: FDR = 0.05, min delta score = 0.90, MS1 tolerance window of -4 to 4 ppm, ld-score > 22.

2.7   Bioinformatic tools

ClustalW (Larkin et al., 2007) was used to produce multiple sequence alignments. Secondary structure predictions and analysis were done with Hhpred (Soding, 2005) and psipred (Buchan et al., 2013). DNA and protein sequences were obtained from the NCBI´s database and *www.uniprot.org*, respectively. Protein specific parameters were calculated using the ProtParam software (Wilkins et al., 1999). For protein domain predictions we used the NCBI´s conserved domain database (Marchler-Bauer et al., 2011). To predict cluster of surface amino acids suitable for surface entropy reduction we employed the SERp server (http://services.mbi.ucla.edu/SER/) (Goldschmidt et al., 2007). The DALI server (Holm and Rosenstrom, 2010) was used for structure similarity searches. Evaluation of binding affinity data obtained by fluorescence anisotropy was conducted by Prism 6.0 (see II 2.5). Pymol was used to visualize and investigate molecular models and to produce images. Crystallography software tools were used as described in II 2.3.3.

III   Results and discussion

1   Structure and function of a truncated NELF-AC subcomplex and architecture of complete NELF

Data presented in chapter III 1 have been obtained during this thesis and were submitted for publication (see page IV).

1.1   NELF subcomplex NELF-AC

In a long-standing effort to obtain structural information on the intrinsically flexible NELF complex, we delineated regions in human NELF subunits that form soluble subcomplexes amenable to structural analysis (Figure 6, Figure 24, Figure 25, Table 16, II 2.2.2, III 2.1). Bacterial co-expression of NELF subunit variants revealed that the N-terminal region of NELF-A could be co-purified with NELF-C. Limited proteolysis and co-expression analysis with truncated protein variants showed that the N-terminal residues 6-188 of human NELF-A and residues 183-590 of human NELF-C formed a stable subcomplex ('NELF-AC'). Purified NELF-AC could be crystallized by vapor diffusion, and the X-ray structure was solved by SIRAS (Figure 7, II 2.3). The structure contained one NELF-AC heterodimer in the asymmetric unit and was refined to a free *R*-factor of 25.6 % at 2.8 Å resolution (Table 17). The structure shows very good stereochemistry and lacks only the mobile NELF-A residues 183-188, and NELF-C residues 183-185, 401-402, 445-448, 523, and 564-572. In the final model, 98.2 % of residues are in preferred Ramachandran regions and 1.8 % of residues are in additionally allowed regions.
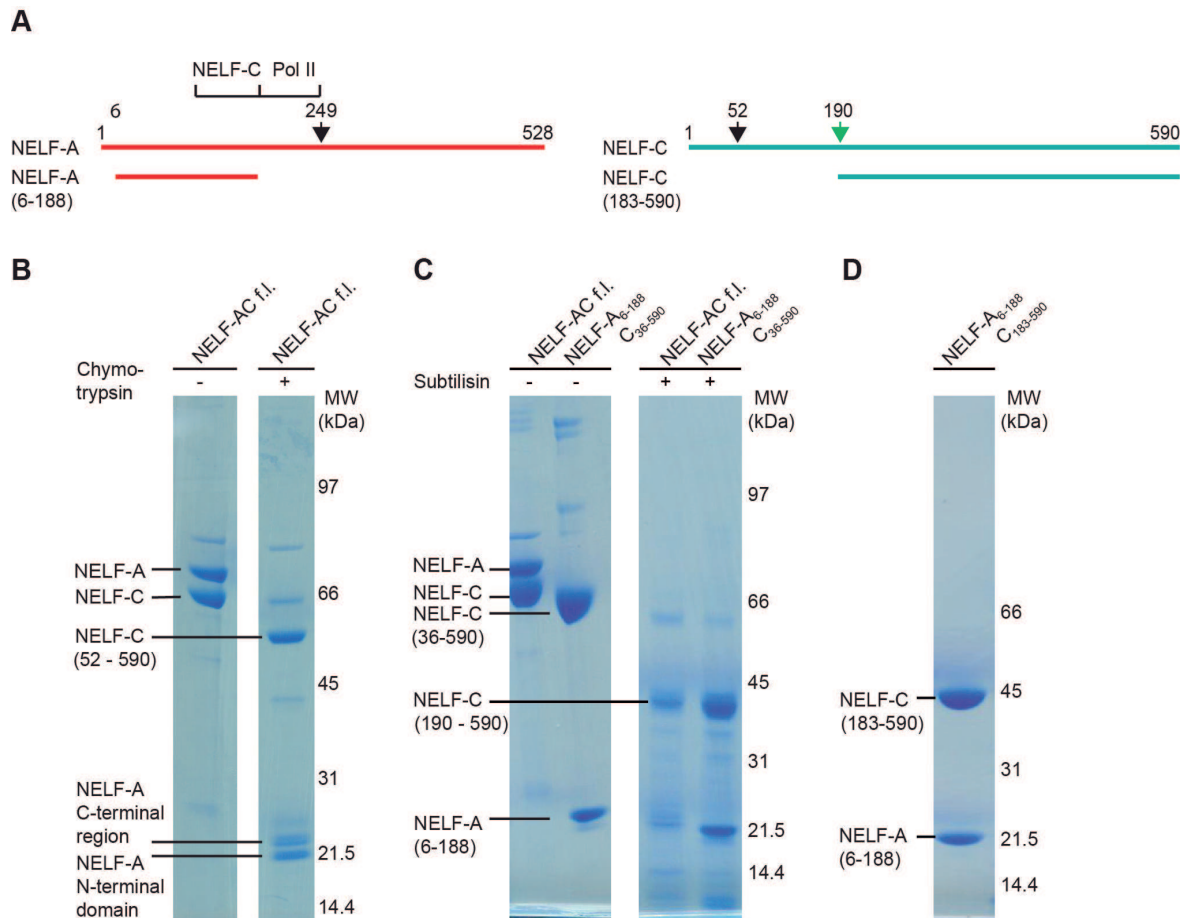
Figure 6: Iterative truncation of full-length NELF-AC yields a variant amenable to crystallization

(A) Crystallized variant and previously identified functional regions in human NELF-A and NELF-C. Cleavage sites of chymotrypsin and subtilisin are indicated by black and green arrows, respectively. 'NELF-C' delineates the previously identified NELF-C-binding region in NELF-A (Narita et al., 2003), whereas 'Pol II' marks the region in NELF-A that associates with Pol II (Narita et al., 2003).

(B) Partial digestion of pure full-length NELF-AC with chymotrypsin yields three stable degradation products that were identified as NELF-A N-terminal domain (residues 6-188), NELF-A C-terminal region (residues 248-~485) and NELF-C (residues 52-590). The resulting truncated construct NELF-A$_{6-188}$C$_{36-590}$ did not yield diffracting crystals. Shown are SDS-PAGE analyses.

(C) Partial digestion of pure full-length NELF-AC and truncated NELF-A$_{6-188}$C$_{36-590}$ with subtilisin yield the same stable degradation products for NELF-A (residues 6-188) and NELF-C (residues 190-590).

(D) The resulting truncated variant NELF-A$_{6-188}$C$_{183-590}$ ('NELF-AC') was successfully used for crystallization.

39

Table 16: Solubility of bacterially expessed NELF variants

Protein variant	Solubility
NELF-AC	(+), (1)
NELF-AD	(++), (1)
NELF-AC$_{36-590}$	(++), (1)
NELF-A$_{6-188}$C$_{36-590}$	(+++), (2)
NELF-A$_{6-188}$ C$_{183-590}$	(+++)
NELF-ABC	(++), (1)
NELF-A$_{6-188}$BC$_{36-590}$	(++), (3)
NELF-ABCE	(+), (1), (3)
NELF-A$_{6-188}$BC$_{36-590}$E	(++), (2), (3)

Variants are full-length proteins if not otherwise specified. (+) = low solubility, (++) = medium solubility, (+++) = high solubility, (1) = aggregation, (2) = slight aggregation, (3) = stable at high salt concentrations only (500 mM NaCl).
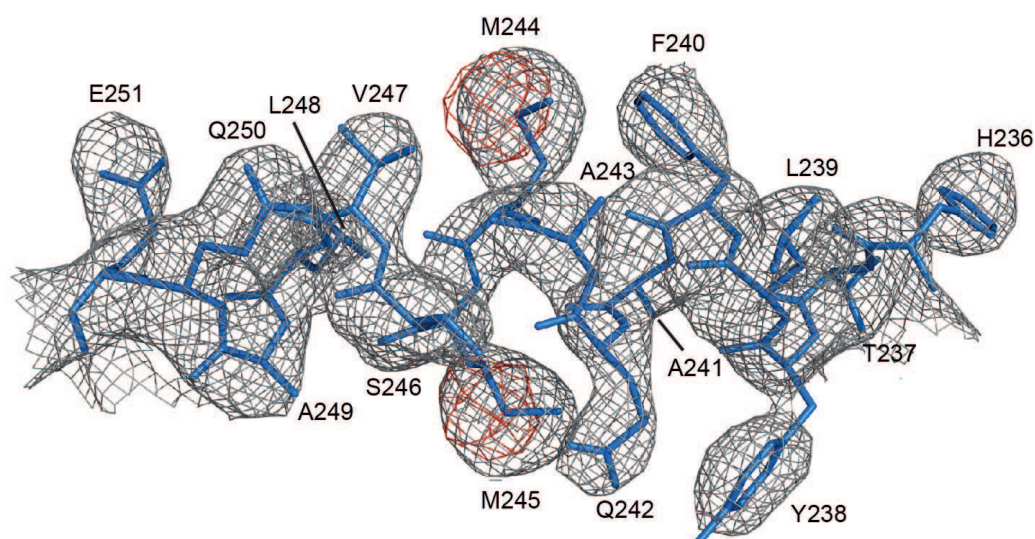
Figure 7: Exemplary region of the electron density map

The final $2F_o$-F_c electron density map was contoured at 1.5 σ (grey) and the anomalous difference Fourier electron density for the selenomethionine-labeled crystal was contoured at 4.0 σ. The final model for NELF-C helix α4´ (Figure 8, Figure 9) is superimposed in stick representation, showing the position of selenium atoms in selenomethionine residues (red) used for phasing.

Table 17: X-ray diffraction and refinement statistics

	Native	SeMet
Data collection [a]		
Space group	I213	I213
Cell dimensions		
$a=b=c$ (Å)	185.07	184.45
Wavelength (Å)	1.00000	0.97910
Resolution (Å)	12.3-2.75 (2.82-2.75) [b]	14.54-3.25 (3.33-3.25)
R_{sym} (%)	9 (271)	9 (130)
$I / \sigma I$	32.2 (1.9)	24.0 (2.2)
Completeness (%)	100 (100)	100 (100)
Redundancy	39.8 (40.7)	20.6 (15.8)
CC $_{(1/2)}$ [c] (%)	100 (73.9)	100 (78.4)
Figure of merit for SeMet sites		0.323
Refinement		
Resolution (Å)	2.75	
No. reflections observed	1,093,935	657,156
No. Reflections unique	27,492	31,923
R_{work} / R_{free} (%)	23.7 / 25.6	
No. atoms		
Protein	4434	
Ligand/ion	2	
Water	13	
B-factors (Å^2)		
Protein	110.8 (NELF-A)	
	108.4 (NELF-C)	
Ligand/ion	103.5	
Water	77.4	
R.m.s deviations		
Bond lengths (Å)	0.003	
Bond angles (°)	0.662	

[a] Diffraction data were collected at beamline X06DA of the Swiss Light Source, Switzerland and processed with XDS (Kabsch, 2010).

[b] Values in parentheses are for the highest-resolution shells

[c] $CC_{1/2}$ = percentage of correlation between intensities from random half-datasets (Karplus and Diederichs, 2012).

## 1.2 Unusual structure of the human NELF-AC subcomplex

The structure of NELF-AC reveals a novel fold and an extended interface between the two NELF subunits (Figure 8, Figure 9, Figure 10B and C). NELF-C adopts a horseshoe-shaped structure (Figure 9). NELF-C consists of 22 α-helices (α1´-α22´) and a small two-stranded β-sheet (β1´-β2´, residues 367-379) that protrudes from the surface. The C-terminal half of NELF-C (helices α14´-α19´) forms three HEAT repeats (H1-H3). The HEAT repeat region shows structural similarity (Holm and Rosenstrom, 2010) to the C-terminal repeat domain (CTD)-interacting domain (CID) (Meinhart and Cramer, 2004) and the polyadenylation factor symplekin (Xiang et al., 2010). Despite the presence of a CID-like fold, NELF-AC did not show significant binding to CTD diheptad peptides carrying phosphorylations at CTD residues serine-2 or serine-2 and serine-5 (not shown). Subunit NELF-A forms a highly conserved helical 'N-terminal domain' (helices α1-α5, residues 6-110) that resembles (Holm and Rosenstrom, 2010) the fold of the HIV integrase-binding domain in human PC4 and SFRS1-interacting protein (Figure 10A) (Cherepanov et al., 2005). This domain is followed by an 'extended region' in NELF-A that forms four additional helices (helices α6-α9, residues 111-182) arrayed around the NELF-C horseshoe (Figure 8, Figure 9).

Both NELF-A regions interact extensively with NELF-C through hydrophobic and polar contacts. Two invariant tryptophan side chains (W24 and W89) on the NELF-A N-terminal domain insert into largely conserved hydrophobic pockets of NELF-C (Figure 8, Figure 10B). The extended region of NELF-A is essential for NELF-C interaction (Narita et al., 2003) and contacts the N- and C-terminal regions of NELF-C with its helices α6 and α9, respectively. NELF-A helices α7 and α8 are buried in the central surface of the NELF-C horseshoe (Figure 10C). Overall, the heterodimer interface buries a large surface area (3,690 $Å^2$), explaining the high stability of the complex in 2 M sodium chloride (not shown).

Figure 8: Conservation of human NELF-A and NELF-C

Alignment of NELF-A and NELF-C regions present in the structure from *Homo sapiens* (*H.s.*), *Drosophila melanogaster* (*D.m.*) and *Dictyostelium discoideum* (*D.d.*). Invariant and conserved residues are highlighted in green and yellow, respectively. Barrels above the alignment represent α-helices, arrows β-sheets. HEAT-repeats H1-H3 are marked with black lines above the alignment. Residues residing in the heterodimeric interface and hydrophobic core residues are marked by black and red squares, respectively. Red triangles label residues involved in nucleic acid interaction, blue and orange dots mark previously known phosphorylation sites located at the protein surface and residues identified to be phosphorylated by P-TEFb, respectively. The "N-terminal domain" and "extended region" of NELF-A are indicated. Sequence alignments were carried out with ClustalW2 (Larkin et al., 2007) followed by manual editing and rendered with JALVIEW (Waterhouse et al., 2009).
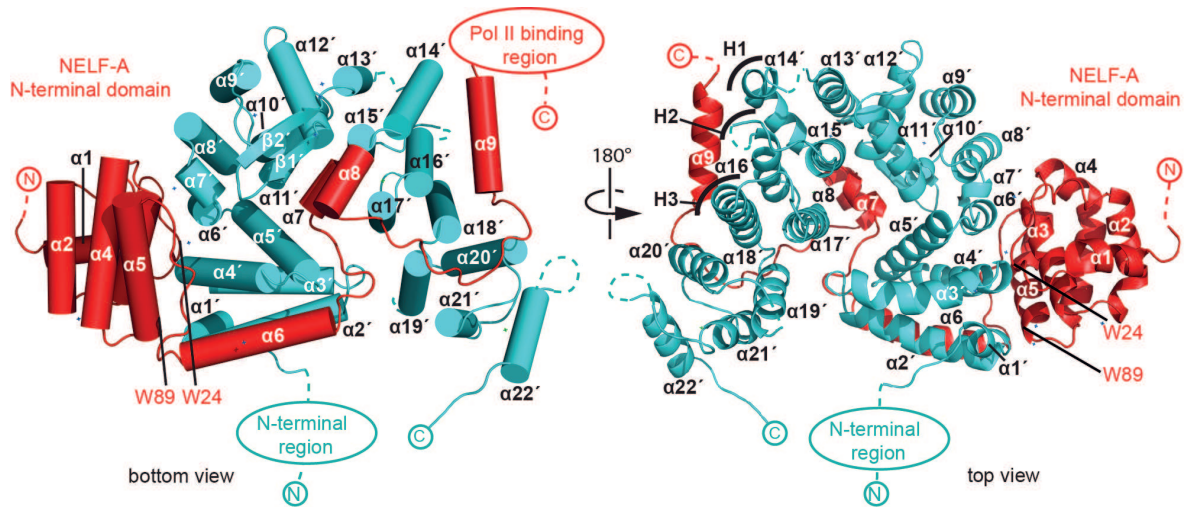
Figure 9: Crystal structure of human NELF-AC complex.

Ribbon model of NELF-AC with NELF-A in red and NELF-C in cyan. N- and C-termini, mobile regions, and truncated regions are indicated by dashed lines. The two views are related by a 180° rotation around the vertical axis.
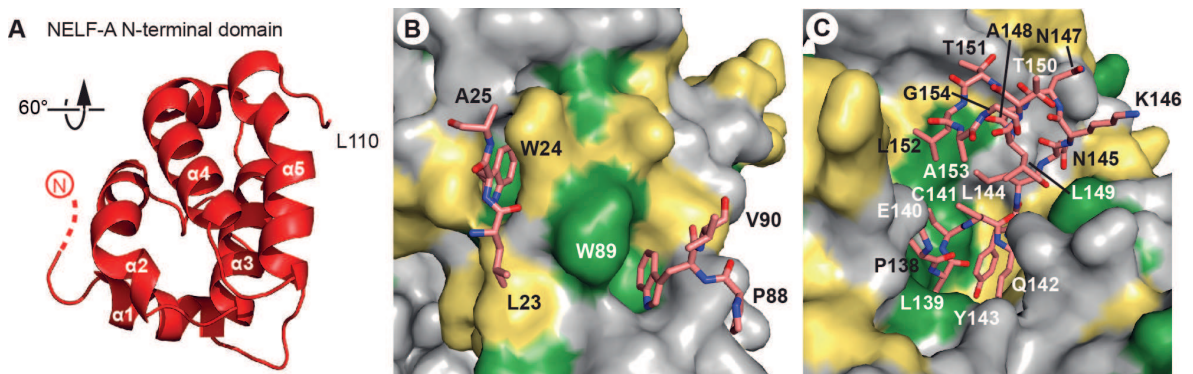
Figure 10: Details of NELF-AC structure and subunit interaction

(A) NELF-A N-terminal domain enlarged and rotated 60° around the horizontal axis relative to "bottom view" (Figure 9).

(B) Detailed view of invariant NELF-A residues W24 and W89 and surrounding residues (stick model) interacting with the NELF-C surface. NELF-C surface conservation colored according to Figure 8. The view is rotated by 90° around the vertical axis relative to "bottom view" (Figure 9).
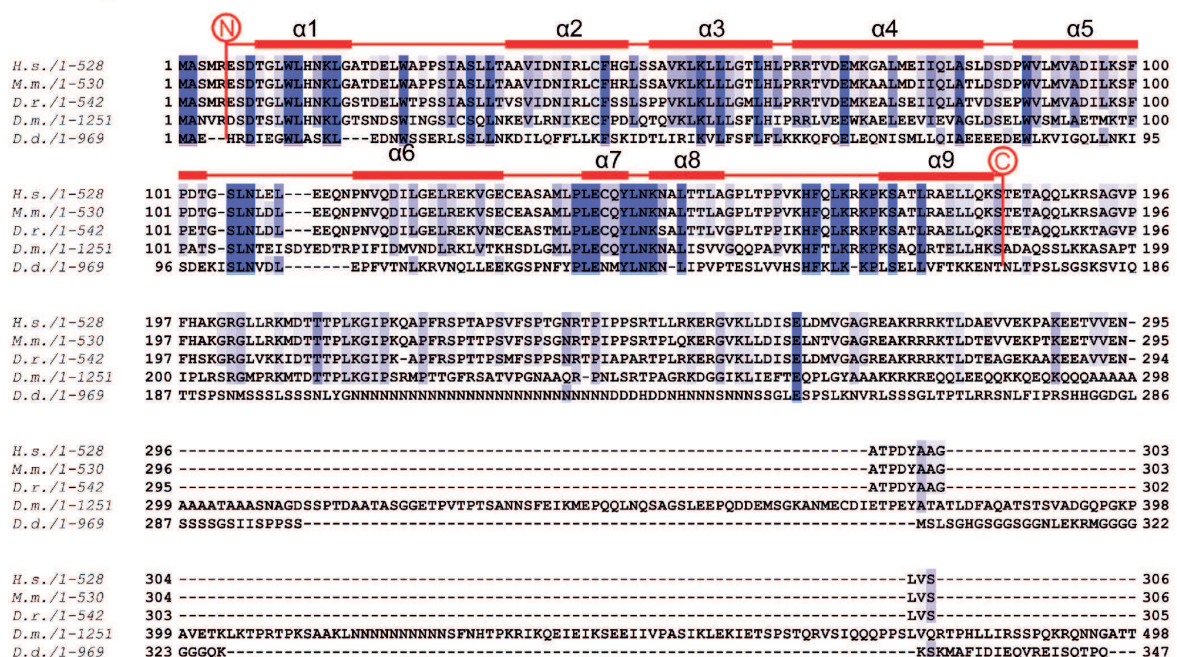
(C) Detailed view of NELF-A helices α7 and α8 (stick model, residues 138–154) surrounded by NELF-C. NELF-C surface conservation is colored according to Figure 8. The view is rotated 60° around the horizontal axis relative to "bottom view" (Figure 9).

44

1.3 The NELF-AC core is highly conserved

The crystallized regions of human NELF-AC share considerable homology among metazoans, particularly at residues forming the hydrophobic cores and the interface between NELF-A and NELF-C (Figure 8). The extent of conservation is evident when human and *Drosophila melanogaster* are compared, which share 55 % identity for NELF-A and 50 % identity for NELF-C. Most regions outside of the crystallized core appear to have diverged in metazoans, suggesting a conserved role for the central NELF-AC core (Figure 11, Figure 12). Thus, the human NELF-AC structure is an excellent model for NELF-AC complexes in other multicellular species.

Surprisingly, a part of the NELF-AC complex also exists in the single cell slime mold *Dictyostelium discoideum*. The hypothetical *Dictyostelium* proteins *DDB_G0286295* and *DDB_G0268678* share sequence similarity with NELF-A and the crystallized C-terminal region of NELF-C (Figure 8). The conservation of many residues in the hydrophobic core and heterodimer interface indicates that the NELF-AC subcomplex exists in this single cell organism. A putative *Dictyostelium* homolog is also found for a region of human NELF-B comprising residues 1-410 (*DDB_G0284195*) (Chang et al., 2012), but not for NELF-E, suggesting that *Dictyostelium* NELF may be composed of a three-subunit NELF-ABC complex that contains the conserved NELF-AC core.

Full-length NELF-A

45

```
H.s./1-528    307 ------------TQKLGSLNNEPALPSTSYLPSTPS------------------------------------------------------------------ 330
M.m./1-530    307 ------------TQKLGSLNSEPTLPSTSYLPSTPS------------------------------------------------------------------ 330
D.r./1-542    306 ------------TQKLGALNNESALPSTSYLPATPS------------------------------------------------------------------ 329
D.m./1-1251   499 SAGTTTTTVGNTTIKMEKLDIKPMIRATGVSPSTSAGTTTTLLTPQQLRQAANPLANLPNNISVKITSAKAKAAAAAAATSSGSQTQSQQPQTLQVQQAP 598
D.d./1-969    348 -----------------KKRRQSIDPQPQQSQPSQQQQQQQ------------------------------------------------------QQQQ 375


H.s./1-528    331 ---------------VVPASS---------YIPSSETPPAPSS---REAS------------------------------RPPEE---PSAPSPT-------- 365
M.m./1-530    331 ---------------VVPASS---------YIPSSETPPAPPS---REAS------------------------------RPPEE---PSAPSPT-------- 365
D.r./1-542    330 ---------------MVPSSS---------YIPSSETQPANAGGSGRETTN---------------------------RQSEESAAPSAAAAS-------- 371
D.m./1-1251   599 QQSHPPLLINSSTPVILASSPSAQRAKALALPSSSTSATTTTTIPSQAIKTMPLSQLKTATNSGPVIISQTIIQPAKRAQQQAGTSSAAAASQQQQQQQS 698
D.d./1-969    376 QQQQQQQQSEKDNNNNNNNNNNNNNNNNNNNNNNSYLPGINESNDQSQTSPQQSPHS--------------------QHLQQQQQLKEQQQAAQVLV 452


H.s./1-528    366 -----------------LPAQFKQR-------------------------------------APMYNSG-----------------------LS 382
M.m./1-530    366 -----------------LPTQFKQR-------------------------------------APMYNSG-----------------------LS 382
D.r./1-542    372 -----------------LPGQFKQR-------------------------------------PPMYNAST----------------------AS 389
D.m./1-1251   699 QQQLLHQQIQQQTQQQAQLPQQPQQQQTQYILATPQQQPQQIQQQQQQQQPILPTLTSFSHSRPMPQTTTLYQATTPSGSGQTPTKILLKTSGSSSVVMT 798
D.d./1-969    453 QQQQSKEQQQKEQQMRAQQQQQQQQQQAIQQQQLKEQQLREQQQQQQQQQLREQQLREQQLREQQFREQQFREQQQFQQQNRHN----------YQNEQ 541


H.s./1-528    383 PATP-TPAAPTSPLTPTTPP-----------------------------------------------AVAPTTQTPP------------ 411
M.m./1-530    383 PATP-APAAPTSPLTPTTPP-----------------------------------------------AVTPTAQTPP------------ 411
D.r./1-542    390 PTAPTSPTTPTSTPTNNGPPPAAT-------------------------------------------AAQPDMPTQP------------- 423
D.m./1-1251   799 PLRQAQPQQATAVVSSNPPPLVATSAAVVSPGQTTLNIQNVQLPNRPVTIQPASQAAQQQHMQAQLQQQQPHPQHTIVANTTATQQPKLSQVLMQPTAAV 898
D.d./1-969    542 DNSQQQQQQQYQPLQPPPVIPQNYQQPSLQTLQAAQAIQSMQSMHQQQQQQQQQQQQQQQQQQQQQQHPQLSIQQQALLQQQQLQQQQQQQQQQQQHQQQQQ 641


H.s./1-528        ---------------------------------------------------------------------------------------------------
M.m./1-530        ---------------------------------------------------------------------------------------------------
D.r./1-542        ---------------------------------------------------------------------------------------------------
D.m./1-1251   899 GTSGVSALNVSPTSGKNKTIILTQKGVILRNIGGDMYQQIPISNVGNMQGLGGTTLMTTTAGPPSLVKTTPSSGVQLQQQQSGKQILPTLIPTSSLGGQH 998
D.d./1-969    642 QQHQQQQTQQAITSPLLNASPPMTGKRKKDDQESLASQKKKEKQEKKDQKEREKEREKQDKEREKQEKKDQKEREKREKKEQKEKEKKEKKSSQRANKKA 741


H.s./1-528    412 ----------------------------------------------------------VAMVAPQTQAPA--QQQPKKN----------------- 430
M.m./1-530    412 ----------------------------------------------------------VAMVAPQTQAPAPVQQQPKKN----------------- 432
D.r./1-542    424 ----------------------------------------------------------PTTAAQPAPTPAPPQTQPKKN---------------- 444
D.m./1-1251   999 VIVQQQQPTNVIGNSQQQTIIRPVMTNVGGGLTTLPQGLTLIQRPGQQPQLVQVQAAPGSTQRTIITQSNTAAAASQQPRQQQQQILVQHKPAPTLQQR 1098
D.d./1-969    742 KTGDSDNAGSGDSPLMSSPTIPSHDLSGLSTLMNIVDSQQGGNSNNQQSQLNQNRSPTQQPLIQQQIQQQFQQLQQQQQLQQQQQQQQQQQQQ-------- 833


H.s./1-528    431 --------------------------------------------------LSLTREQMFAAQEMFKTANKVTRPEKALILGFMAGSRENPCQEQGD 476
M.m./1-530    433 --------------------------------------------------LSLTREQMFAAQEMFKTANKVTRPEKALILGFMAGSRENPCPEQGD 478
D.r./1-542    445 --------------------------------------------------LSLTREQMYAAQEMFKTANKVTRPEKALILGFMAGSRENPCPEQGD 490
D.m./1-1251  1099 LVTSTTSGGQGQQGNPNAGLPRTVQVQVQAQQQQQPQQQQATQQQSQQAPQRRGLSLSNEHVHKAHEMFRKANRVSRPDKALILGFMAGLRENPRPNNEN 1198
D.d./1-969    834 --------------------------------------------QQHNPMQQPQIHQQIQQQFQQHQHHHQHQHQHQHQQQQQQQQQYQQHNQQQHNQNQ 889


H.s./1-528    477 VIQIKLSEHTEDLPKADGQGSTTMLVDT-VFEMNYATGQWTRFKKYKPMTNVS------------------------- 528
M.m./1-530    479 VIQIKLSEHTEDLPKADGQGSTTMLVDT-VFEMNYATGQWTRFKKYKPMTNVS------------------------- 530
D.r./1-542    491 IIQIKLSEHTEVLPKADGTGSTTMLVDT-VFEMNYSTGQWTRLKKYKPITNVS------------------------- 542
D.m./1-1251  1199 VLVIKLGETEEKVQQDNG-HTALCLVES-HIRLDYNTGEWKTFQNYRLQDQSAAS------------------------ 1251
D.d./1-969    890 QHNQSQHNQNQQNQSYPFDPQYYLQRSTQQQTQSPPAHLNYQSPQYYNQQKQEGLGIIEQQRRIIEQSGQPYFGNYNNNK 969
```
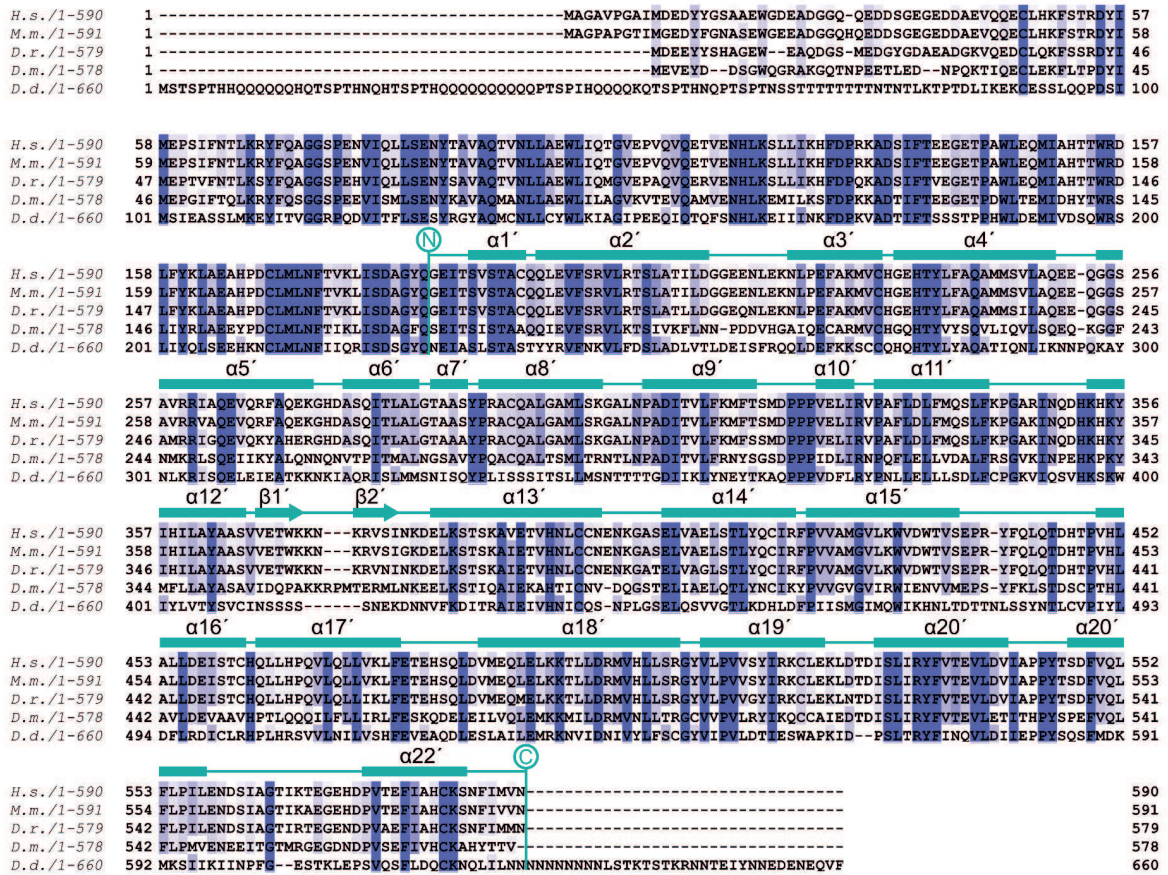
Figure 11: Multiple sequence alignment of full-length NELF-A demonstrating the comparatively high conservation of the crystallized region

The alignment compares full-length NELF-A from *Homo sapiens* (*H.s.*), *Mus musculus* (*M.m.*), *Danio rerio* (*D.r.*), *Drosophila melanogaster* (*D.m.*) and *Dictyostelium discoideum* (*D.d.*). Residues are colored according to percent conservation with brighter colors representing higher conservation. Barrels above the alignment represent α-helices, arrows β-sheets and are colored according to Figure 8. N- and C-terminal borders of solved crystal structure are indicated. Sequence alignment was done with ClustalW2 (Larkin et al., 2007) followed by manual editing and rendered with JALVIEW (Waterhouse et al., 2009).

Full-length NELF-C

Figure 12: Multiple sequence alignment of full-length NELF-C demonstrating the low conservation of the N-terminal region

The alignment compares full-length NELF-C from *Homo sapiens* (*H.s.*), *Mus musculus* (*M.m.*), *Danio rerio* (*D.r.*), *Drosophila melanogaster* (*D.m.*) and *Dictyostelium discoideum* (*D.d.*). Residues are colored according to percent conservation with brighter colors representing higher conservation. Barrels above the alignment represent α-helices, arrows β-sheets and are colored according to Figure 8. N- and C-terminal borders of solved crystal structure are indicated. Sequence alignment was done with ClustalW2 (Larkin et al., 2007) followed by manual editing and rendered with JALVIEW (Waterhouse et al., 2009).

1.4   NELF-AC binds single-stranded nucleic acids

Analysis of the NELF-AC surface showed that one face of the NELF-AC complex contains four positively charged patches (Figure 13, bottom view). Patch 1 consists of NELF-A residues R65 and R66 and NELF-C residues R291 and K315. Patch 2 encompasses NELF-C residues K372, K373, and K374, and patch 3 contains NELF-C residues K384 and K388. Patch 4 is composed of NELF-A residues K146, K161, K168, and R175, and NELF-C residues R419 and R506. These patches are well conserved among metazoa, and are partially conserved in *Dictyostelium* (Figure 8). In addition to the four positive

patches, the NELF-AC surface contains a polar conserved surface (patch 5) that is formed by NELF-A residues K166, R167, K170, L174, E177, K181, and S182, and residues E491, K494, D498, D526, S528, R531, Y532, T535 and E536 that protrude from NELF-C helices α18´ and α20´ (Figure 13, side view).
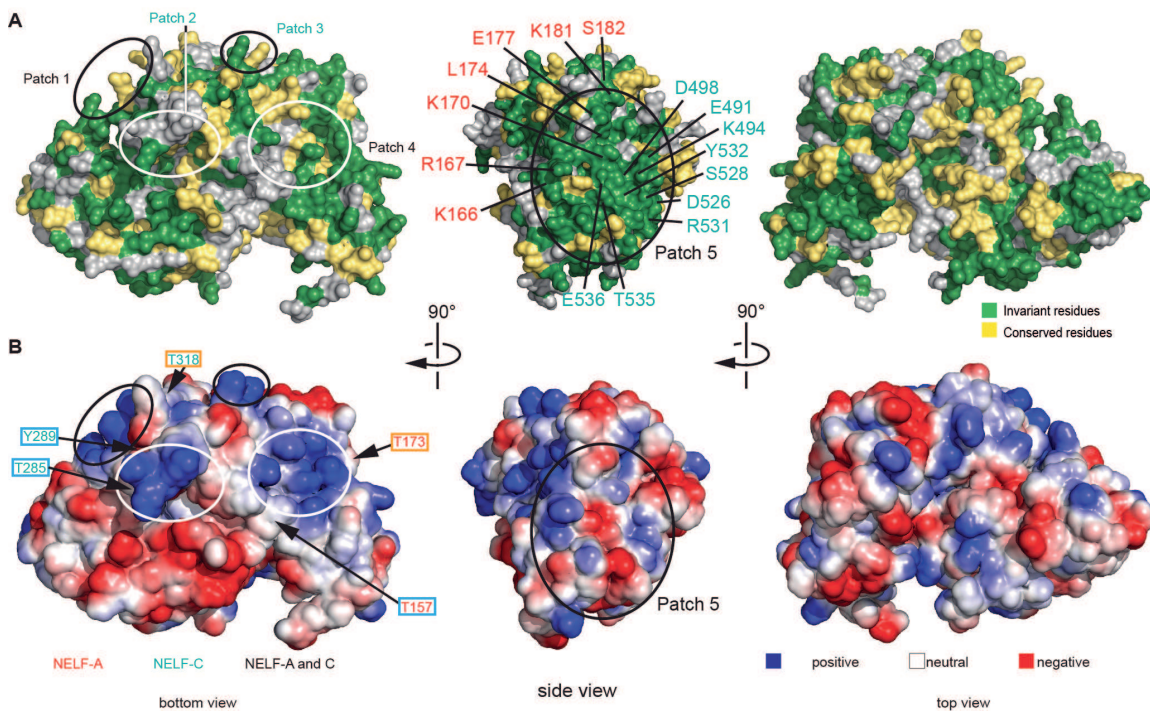
Figure 13: Surface properties of NELF-AC

Three views of the solvent-accessible surface related by 90° rotations around a vertical axis are shown.

(A) Surface conservation. Residues that are invariant from human to *Drosophila* are in green, conserved residues in yellow (Figure 8). Surface areas involved in nucleic acid binding (patches 1–4) (III 1.4) are highlighted. Colors of labels according to color code of protein features belong to (Figure 9). A conserved polar surface area (patch 5) is formed by the C-terminal region of NELF-A and NELF-C helices α18´ and α20´.

(B) Electrostatic surface potential generated with ABPS. Blue, red, and white areas indicate positive, negative and neutral charge, respectively. Surface areas involved in nucleic acid binding (patches 1–4), exposed phosphorylation sites mapped *in vivo* (T157, T285 and Y289) and sites phosphorylated by P-TEFb *in vitro* (T173 and T318) are highlighted (III 1.5). Colors of labels according to color code of protein features belong to (Figure 9). Phosphorylation sites mapped *in vivo* and *in vitro* are framed by blue and orange boxes, respectively (Figure 8).

The positively charged patches of NELF-AC suggested that the subcomplex may associate with nucleic acid. To investigate this idea, we used fluorescence anisotropy titration assays (Figure 14, II 2.5). We tested NELF-AC binding to fluorescently labeled, synthetic 25-nt single-stranded (ss) DNA and ssRNA random sequence (Table 7). Indeed, we detected binding of NELF-AC to both ssDNA and ssRNA. Regression analysis of the

binding curves revealed that K_d values in the micromolar range (Figure 14A). In contrast, NELF-AC did not associate with nucleic acid duplexes composed of DNA or DNA-RNA hybrids (not shown).

To investigate whether the positively charged patches were involved in nucleic acid binding, we generated NELF-AC variants in which lysine and arginine residues in the patches were substituted with methionine and glutamine, respectively. Indeed, single-stranded nucleic acid binding was strongly impaired in variants with mutations in three or four of the positively charged patches (Figure 14B, C). We also tested whether single-stranded nucleic acids corresponding to known Pol II *in vivo* pause sites could associate with NELF-AC (Figure 14E). A ssRNA oligonucleotide with a sequence corresponding to a promoter-proximal transcript from the *junB* gene bound NELF-AC with a K_d of ~8.0±0.9 µM, whereas ssDNA corresponding to the non-template strand in this region bound more weakly (Aida et al., 2006) (Figure 14D). Furthermore, ssRNA and ssDNA derived from the *c-fos* promoter-proximal region sequences (Fivaz et al., 2000) also bound NELF-AC, albeit with a preference for DNA (Figure 14D). Taken together, NELF-AC binds single-stranded nucleic acids via positively charged patches, and both the strength of binding and the preference for RNA or DNA are sequence-dependent.
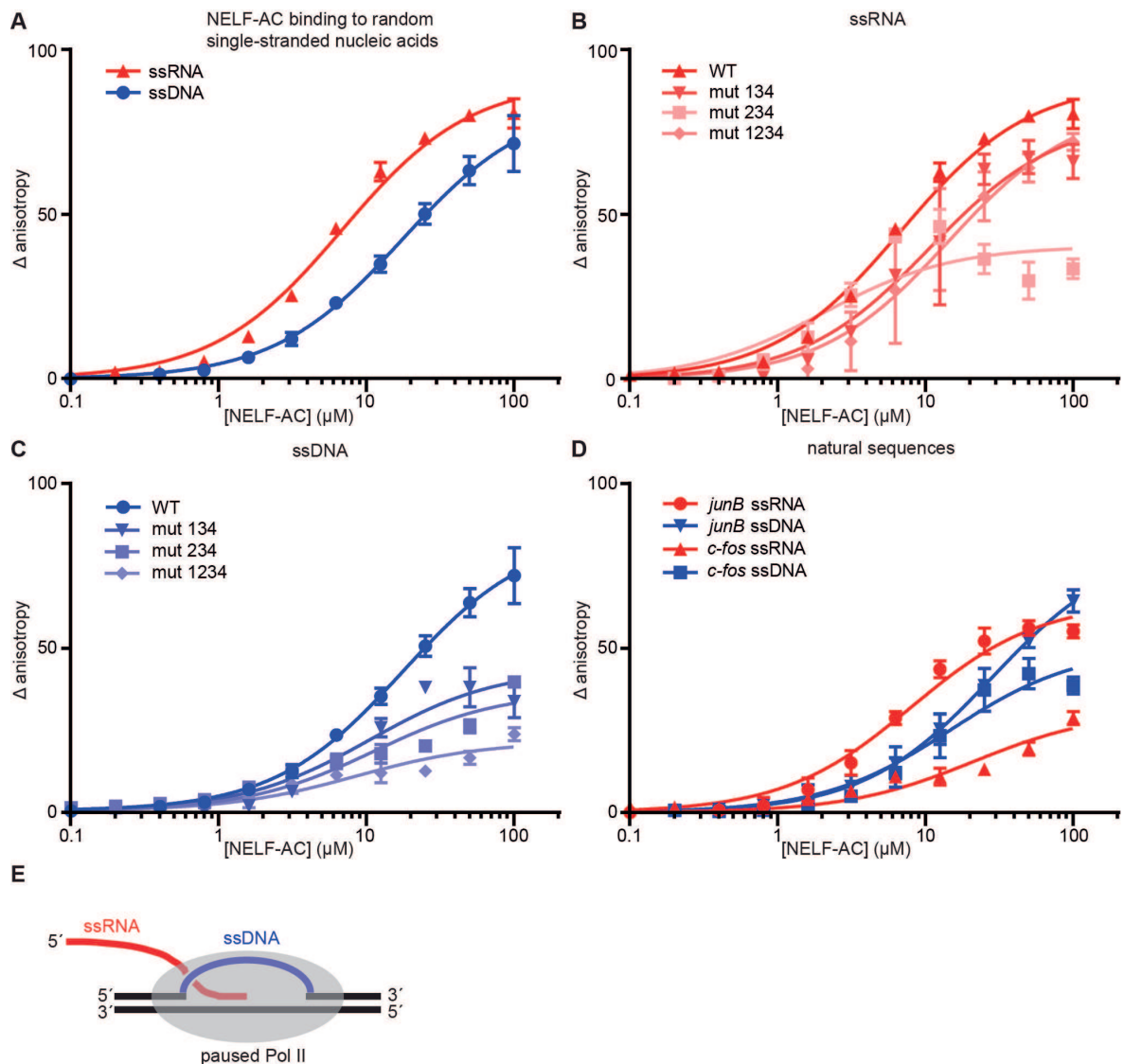
Figure 14: NELF-AC binds single-stranded nucleic acids

(A) Binding of wild type (WT) NELF-AC to 10 nM fluorescently labeled ssRNA or ssDNA with 60 % GC content as monitored by the change in relative fluorescence anisotropy. Error bars reflect the standard deviation from three experimental replicates.

(B, C) Binding of WT NELF-AC and variants containing mutations in surface patches (Figure 13) to the same ssRNA (B) or ssDNA (C) used in panel A. Numbers indicate mutated patches present in NELF-AC variants.

(D) Binding of WT NELF-AC to 10 nM of fluorescently labeled ssRNA and ssDNA derived from natural sequences of promoter-proximal regions of paused genes *junB* and *c-fos* (II 2.5, Table 7) as monitored by the change in relative fluorescence anisotropy.

(E) Schematic of the presence of single-stranded nucleic acids (ssRNA, ssDNA) in the promoter-proximally paused Pol II elongation complex.

1.5   NELF-AC phosphorylation counteracts nucleic acid binding

The above results suggested that NELF acts at least partially during Pol II pausing by binding single-stranded nucleic acids in the promoter-proximal transcription elongation complex. Because it is known that release of paused Pol II involves NELF phosphorylation (Chiba et al., 2010; Fujinaga et al., 2004), we hypothesized that NELF phosphorylation counteracts nucleic acid binding by NELF-AC. In support of this, known phosphorylation sites mapped to the nucleic acid-binding face of NELF-AC. The human NELF-A residue T157 (Q9H3P2) and NELF-C residues T285, Y289, and S301 (Q8IXH7) can be phosphorylated *in vivo* (*www.phosphosite.org*) and T157, T285 and Y289 are located close to the nucleic acid-binding patches 1 and 4 in our structure. We therefore generated protein variants with the phosphomimetic mutations T157D, T285D and Y289E, and tested their affinity for ssRNA and ssDNA. Double mutants exhibited poor solubility. All three variants significantly impaired binding to ssRNA and ssDNA (Figure 15), arguing that phosphorylation of NELF-AC counteracts nucleic acid binding, and providing a possible explanation for how paused Pol II may be released upon NELF phosphorylation.
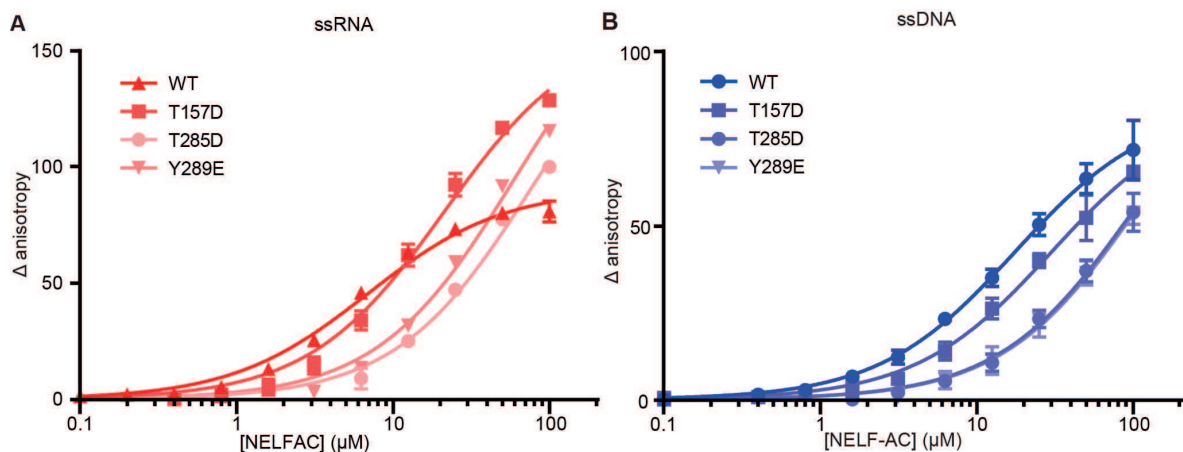

Figure 15: *In vivo* phosphorylation of NELF-AC counteracts binding of single-stranded nucleic acids

(A, B) Binding of WT NELF-AC and variants carrying phosphomimetic mutations that mimic phosphorylations previously identified *in vivo* to 10 nM of fluorescently labeled ssRNA (A) or ssDNA (B) with 60 % GC content as monitored by changes in relative fluorescence anisotropy (same as Figure 14A-D). Error bars reflect the standard deviation from three experimental replicates.

To test whether P-TEFb can phosphorylate NELF-AC *in vitro*, and to investigate whether phosphorylation may impair nucleic acid binding of NELF-AC, we prepared recombinant P-TEFb by co-expressing its subunits CDK9 and cyclin-T1 in insect cells (II 2.4). We incubated NELF-AC with purified P-TEFb and ATP, and subjected the two NELF subunits to phosphopeptide analysis by mass spectrometry (Figure 16, Figure 17, II 2.4). We

indeed detected two P-TEFb dependent phosphosites corresponding to NELF-A residue T173 (patch 4) and NELF-C residue T318 (patch 1). To ensure complete phosphorylation and to avoid complications caused by inherent P-TEFb nucleic acid binding activity, we cloned and purified NELF-AC variants with phosphomimetic mutations of NELF-A T173D and NELF-C T318D. Binding experiments showed strongly decreased affinity for ssRNA and ssDNA compared to wild type NELF-AC (Figure 18). Thus nucleic acid binding by NELF-AC can be impaired by P-TEFb-dependent phosphorylation of its nucleic acid-binding face.
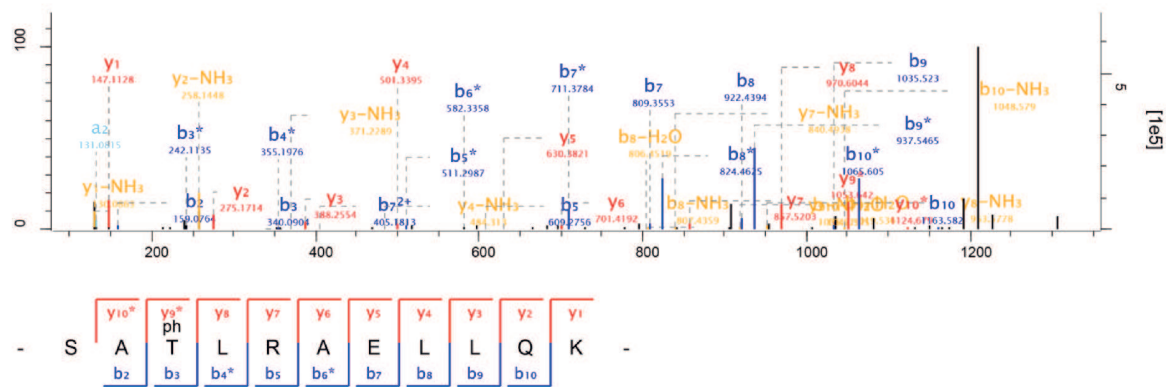
Figure 16: Identification of pT-173 in NELF-C by LC-MSMS

Production ion spectrum of phosphorylated peptide SATLRAELLQK (position 171 – 181) in NELF-A. The b- and y-type fragmention ions are depicted within the spectrum and in the corresponding peptide sequence. b-type- and/or y-fragment ions marked with an asterisk (*) are those that reveal a loss of H_3PO_4 (98 Da) in the spectrum and thus contain the phosphorylated amino acid. The b_3-ion (Ser-Ala-Thr) but not the b_2-ion (Ser-Ala) shows unambigiously a loss of H_3PO_4 so that Thr-173 is the phosphorylated amino acid. Scan number is the according to Excalibur software (ThermoFisherScientific) and score according to MaxQuant software (Cox and Mann, 2008).
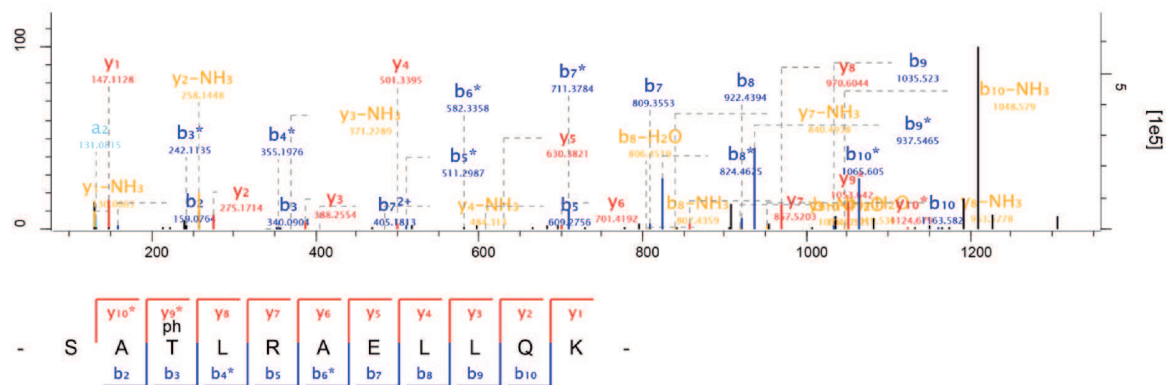
Figure 17: Identification of pT-318 in NELF-AC by LC-MSMS

Production ion spectrum of phosphorylated peptide MFTSMDPPPVELIR (position 316 – 329) in NELF-C. The b- and y-type fragment ions are depicted within the spectrum and in the corresponding peptide sequence. b-type- and/or y-fragment ions marked with an asterisk (*) are those that reveal a loss of H_3PO_4 (98 Da) in the spectrum, and thus contain the phosphorylated amino acid. The b_3-ion (Met_{ox}-Phe-Thr) shows unambiguously a loss of H_3PO_4 so that Thr-318 is the phosphorylated amino acid. Scan number is the according to Excalibur software (ThermoFisherScientific) and score according to MaxQuant software (Cox and Mann, 2008).
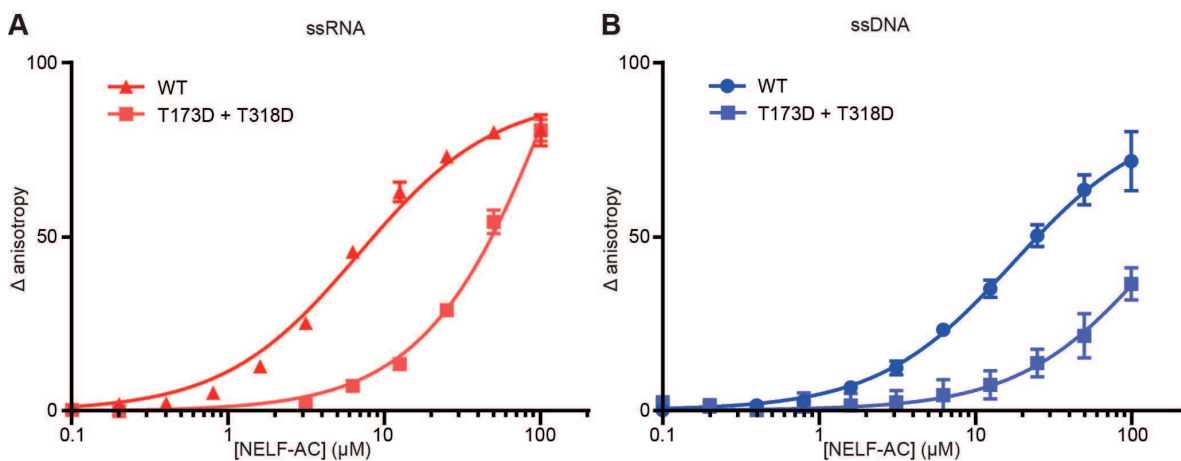
Figure 18: Phosphorylation of NELF-AC by P-TEFb counteracts binding of single-stranded nucleic acids

(A, B) Binding of WT NELF-AC and variants carrying phosphomimetic mutations that mimic phosphorylations made by P-TEFb *in vitro* to 10 nM of fluorescently labeled ssRNA (A) or ssDNA (B) with 60 % GC content and monitored binding by changes in relative fluorescence anisotropy (same as Figure 14A-D). Error bars reflect the standard deviation from three experimental replicates.

1.6   Complete NELF has an accessible nucleic acid-binding face

To investigate whether the nucleic acid-binding face on NELF-AC is accessible in the four-subunit NELF complex, we prepared the entire NELF complex in recombinant form (II 2.2.2). We co-expressed full-length NELF-A, NELF-C, and NELF-B carrying a solubility-enhancing maltose-binding protein (MBP) tag in *E. coli*. The resulting NELF-ABC complex was partially purified, and supplemented with independently expressed and purified NELF-E. Because the resulting NELF complex was prone to aggregation, we also prepared a truncated version that lacked the C-terminal region of NELF-A (residues 189-528) and the non-conserved N-terminal tail of NELF-C (residues 1-35) (Figure 11, Figure 12, Figure 19A, II 2.2.2). Similarly, we prepared a three-subunit truncated NELF-ABC complex lacking the NELF-E subunit.
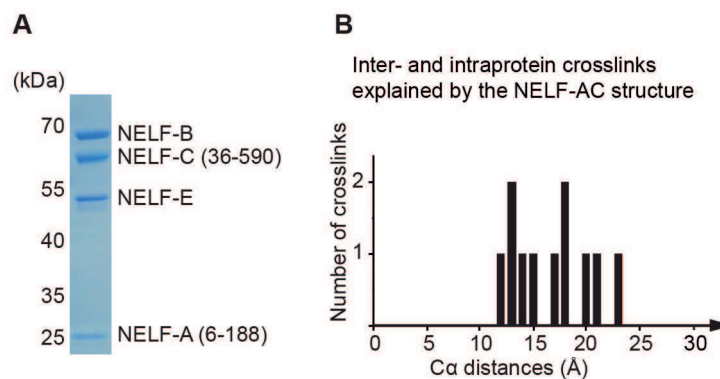
Figure 19: Controls of complete NELF crosslinking
(A) SDS-PAGE of pure truncated four-subunit NELF used for crosslinking analysis.
(B) Crosslinking of truncated four-subunit NELF complex. Cα-atom distances of lysine-lysine crosslinks located within the known NELF-AC structure.

We crosslinked purified truncated three- and four-subunit NELF complexes DSS and detected lysine-lysine crosslinks by mass spectrometry as described (Herzog et al., 2012). For the four-subunit NELF, we obtained a total of 158 unique high-confidence lysine-lysine crosslinks, including 70 inter-subunit and 88 intra-subunit crosslinks (Figure 20A). Our NELF-AC crystal structure explained 11 inter- and intra-subunit crosslinks, with Cα distances below the maximum allowed distance of 30 Å (Figure 19B). This provided a positive control and argued that the structure of NELF-AC is preserved within the complete NELF complex.
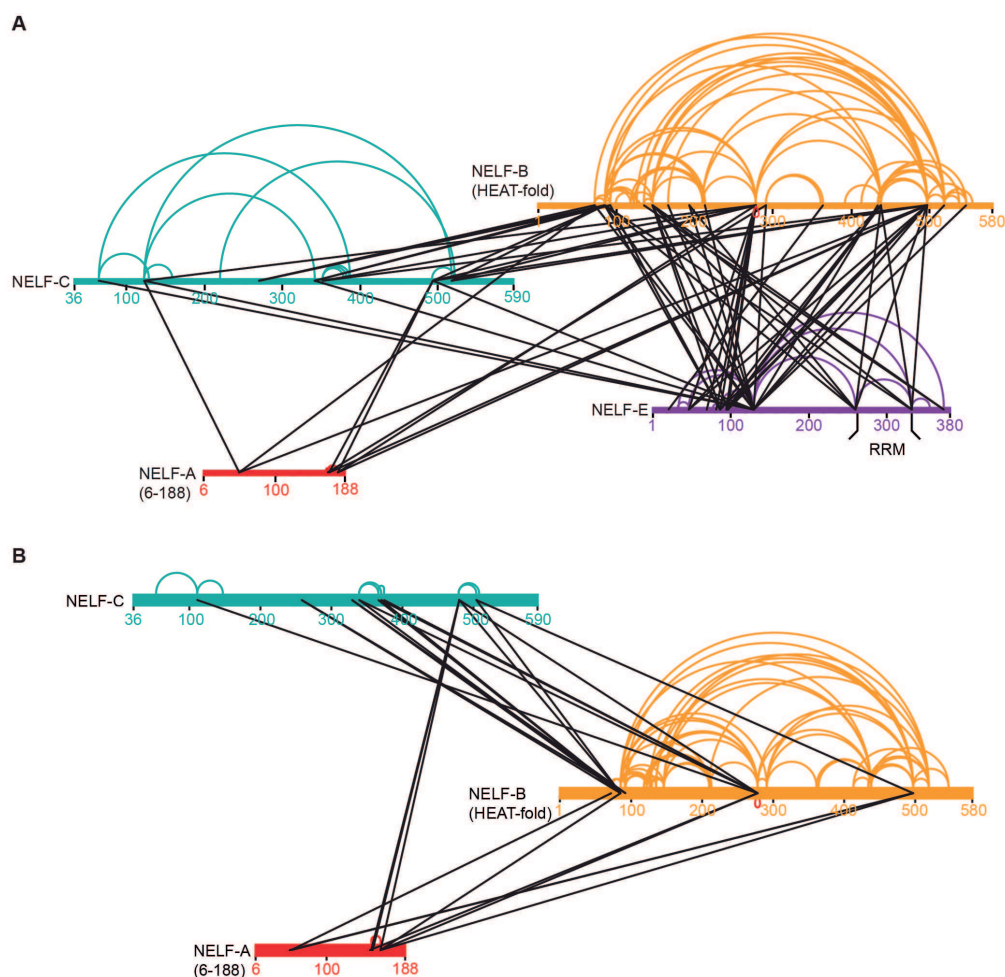
Figure 20: Interactions detected by crosslinking

Inter-protein and intra-protein crosslinks are colored in black and in the colors of the corresponding NELF subunits, respectively. The crosslinking interaction map has been created using the xiNET program (Combe et al., 2015). NELF-A, NELF-B, NELF-C and NELF-E are colored as in Figure 22.

(A) Crosslinking interaction map of a truncated 4-subunit NELF complex as determined by lysine crosslinking followed by mass spectrometric identification of the crosslinked sites. A red loop at NELF-B K278 indicates a self-crosslink.

(B) Crosslinking interaction map of a truncated three-subunit NELF-ABC as determined by lysine crosslinking followed by mass spectrometric identification of the crosslinked sites. A red loop at NELF-B K278 indicates a self-crosslink.

The lysine residues in the nucleic acid-binding patches were devoid of crosslinks. Patches 3 and 4 formed monolinks with the crosslinking reagent, indicating they are accessible to solvent (Figure 22). Independent crosslinking analysis of the 3-subunit NELF-ABC complex confirmed and complemented these results (Figure 20B). Taken together, the nucleic acid-binding face of NELF-AC remains accessible for nucleic acid binding in the complete NELF complex.

1.7 Location of NELF-B and NELF-E

The crosslinking analysis also revealed the topology of the complete NELF. To position NELF-B with respect to the NELF-AC structure, we first prepared a model of NELF-B. The program I-TASSER (Yang et al., 2015) predicted that NELF-B forms a HEAT repeat fold (Figure 22) (C-score = -2.31, best template structure is 1B3U – human PP2A). The model is strongly supported by 54 intra-crosslinks, which suggest a strong curvature of the HEAT repeat fold (Figure 20A), as observed for a HEAT repeat protein folding around an interaction partner (Cingolani et al., 1999).

Three distinct regions of NELF-B crosslinked to NELF-AC; the N-terminal region (residues K72, K85 and K92), the central region (K278 and K292), and the C-terminal region (K497) (Figure 20A, Figure 22). These three regions of NELF-B crosslink to the side opposite of the NELF-AC nucleic acid-binding face, indicating that NELF-B and nucleic acids bind to opposite faces of NELF-AC. The central region of NELF-B also crosslinks to the N-terminal region of NELF-C that is not present in our crystal structure (K125). *In vitro,* the N-terminal region of NELF-C bound NELF-B (Figure 21), suggesting that NELF-B embraces the N-terminal region of NELF-C. Crosslinks detected with the truncated four-subunit NELF complex were corroborated by independent XL-MS experiments performed with the three-subunit NELF-ABC complex (Figure 20B).
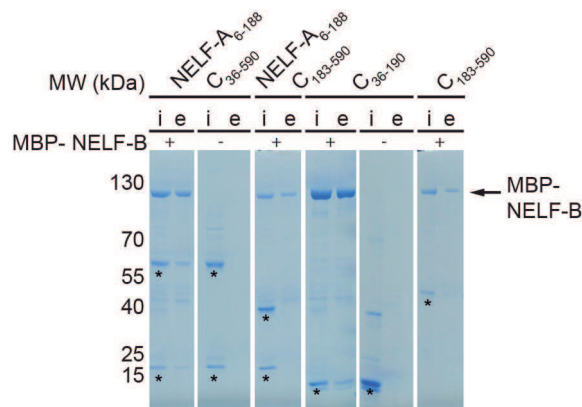
Figure 21: Interaction between NELF-AC and NELF-B
Pull-down assays demonstrating that the NELF-C N-terminal region (residues 36-190) is sufficient for interaction with NELF-B. MBP-NELF-B was bound to amylose resin and incubated with different NELF-AC constructs. Only NELF-A$_{6-188}$C$_{36-590}$ and NELF-C$_{36-190}$ could bind to NELF-B. Asterisks mark NELF-A and NELF-C constructs added to NELF-B as described above the lanes. i = input, e = eluate.

Finally, the crosslinking indicated that NELF-E is located at the periphery of NELF-ABC. NELF-E crosslinked to NELF-B, except for one residue (K130) within a mobile region that crosslinked to multiple distant sites on NELF-C (Figure 20). The extensive crosslinking between NELF-E and NELF-B is consistent with biochemical interaction data (Narita et al., 2003). A predicted N-terminal helix of NELF-E (residues 7-36) formed a crosslink to NELF-B residue K364 (Figure 22), placing this helix near the central region in NELF-B. The RRM domain of NELF-E (Rao et al., 2006) (residues 257-335, PDB-code 2JX2) crosslinks to numerous locations on NELF-B with its residues K260 and K332 on one side of the RRM (Figure 20, Figure 22). This crosslinking pattern indicates that the RRM domain remains mobile within NELF. Residue K326 of NELF-E is located in the RNA-binding β-sheet and was monolinked, arguing that the RNA-binding face of NELF-E is accessible within NELF. Taken together, our analysis revealed the topology of complete NELF and showed that both nucleic acid-binding sites of NELF are accessible.
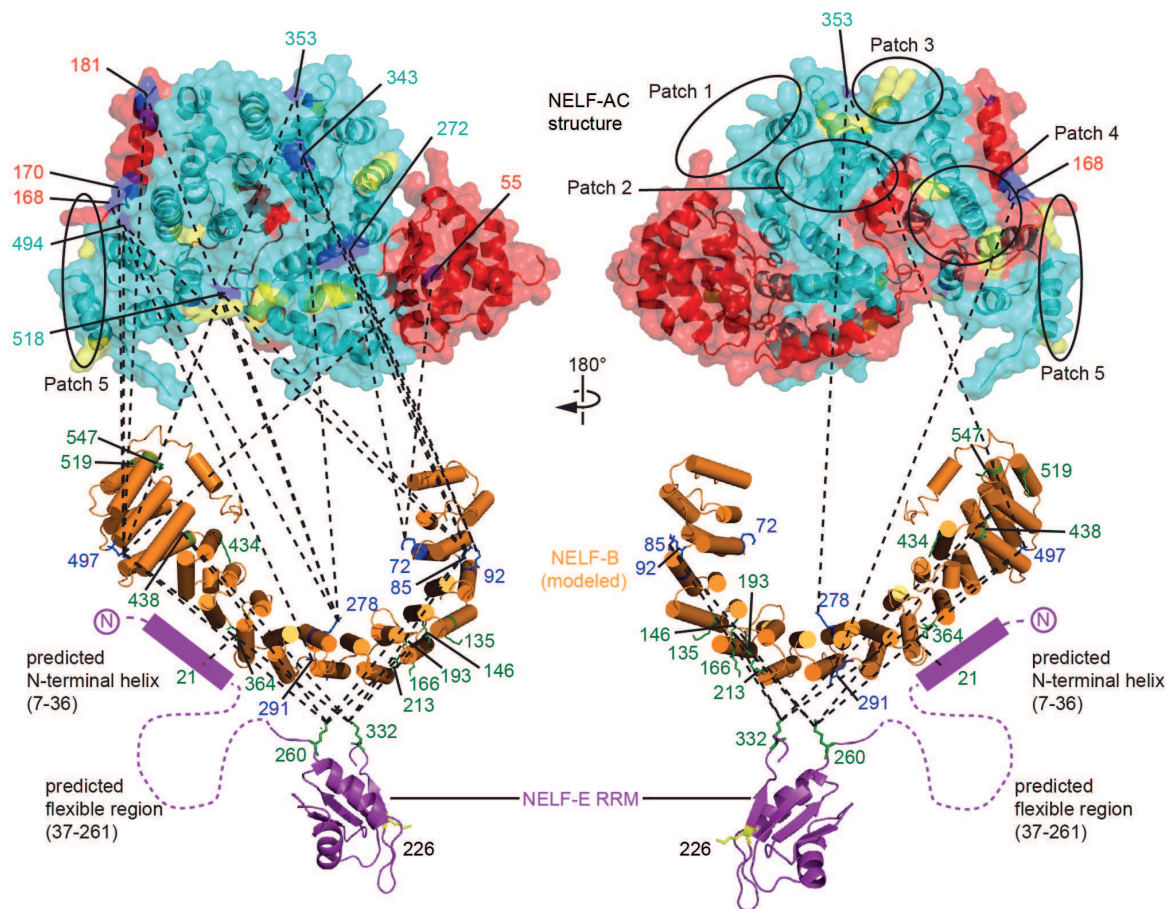
Figure 22: Molecular architecture of complete NELF

Structural overview of observed crosslinks between NELF-A$_{6-188}$ (red), NELF-C$_{183-590}$ (cyan), NELF-B (orange), and NELF-E (magenta). NELF-B was modeled with I-TASSER (Yang et al., 2015). Two views related by a 180º rotation around a vertical axis are shown. The NELF-E RRM-domain structure was previously solved (Rao et al., 2006) (2JX2). A predicted (Buchan et al., 2013) N-terminal helix (7-36) and a predicted flexible region (37-261) of NELF-E are indicated. Lysines on NELF-AC surface forming crosslinks to NELF-B are colored blue and labeled as in Figure 9. Residues of NELF-B that form crosslinks with both NELF-AC and NELF-E are blue, residues solely forming crosslinks between NELF-B and NELF-E are green. NELF-AC patches (1-4) responsible for nucleic acid interaction are circled and the conserved patch 5 is indicated (Figure 13, III 1.4). NELF-AC lysine residues forming monolinks with DSS are in yellow.

1.8 Discussion

Understanding the mechanisms of promoter-proximal pausing and release is essential for understanding gene regulation and ultimately requires structural information of Pol II elongation complexes bound by DSIF, NELF, and P-TEFb. To this end, structures of the involved multi-protein components are required. Structural information is already available for Pol II elongation complexes (Martinez-Rucobo and Cramer, 2013), DSIF (Klein et al., 2011; Martinez-Rucobo et al., 2011), and P-TEFb (Baumli et al., 2012; Baumli et al., 2008; Schulze-Gahmen et al., 2014; Schulze-Gahmen et al., 2013; Tahirov et al., 2010), but is lacking for NELF, except for the RRM domain of NELF-E (Rao et al., 2006; Rao et al., 2008). To close this gap, we report here the overall architecture of NELF and the crystal structure of its conserved core subcomplex NELF-AC. In addition, we report a single-stranded nucleic acid-binding function of NELF-AC that is repressed by P-TEFb-dependent phosphorylation of the NELF-AC surface. Our results are generally relevant for understanding NELF function and the mechanisms of Pol II pausing because the NELF-AC structure and functionally important residues are conserved.

Based on our data and available literature (Chiba et al., 2010; Yamaguchi et al., 1999a) we propose that Pol II pausing involves binding of NELF-AC to nascent RNA, and that Pol II release involves phosphorylation of the RNA-binding face of NELF-AC by P-TEFb, thereby impairing RNA binding. We also suggest a topological model for NELF interaction with the promoter-proximally paused Pol II-DSIF transcription elongation complex (Figure 23). In this model, the positively charged face of NELF-AC interacts with exiting RNA because in a Pol II-DSIF elongation complex ssRNA is accessible, whereas non-template ssDNA is most likely not (Martinez-Rucobo et al., 2011). The NELF-A region immediately following the region in our structure (residues 189-248) is flexible (proteolysis data not shown) and may interact with the Pol II clamp (Yamaguchi et al., 2007). The adjacent, highly conserved surface patch 5 on NELF-AC (Figure 13, side view) may contribute to Pol II binding or may bind DSIF or another factor, such as P-TEFb. NELF-B and NELF-E are likely oriented away from the Pol II surface, but NELF-E is flexible and can bind nascent RNA when it is approximately 70 nucleotides long (Missra and Gilmour, 2010; Pagano et al., 2014; Rao et al., 2008). The comparatively weak affinity of NELF-AC for nucleic acids would not contradict our model since upon binding to the Pol II clamp, NELF-AC is positioned close to nascent RNA effecting an increases binding efficiency. Further, for reversible protein-nucleic acid interactions the affinity of individual binding sites would be expected to be rather weak so nucleic acids can be released easily.
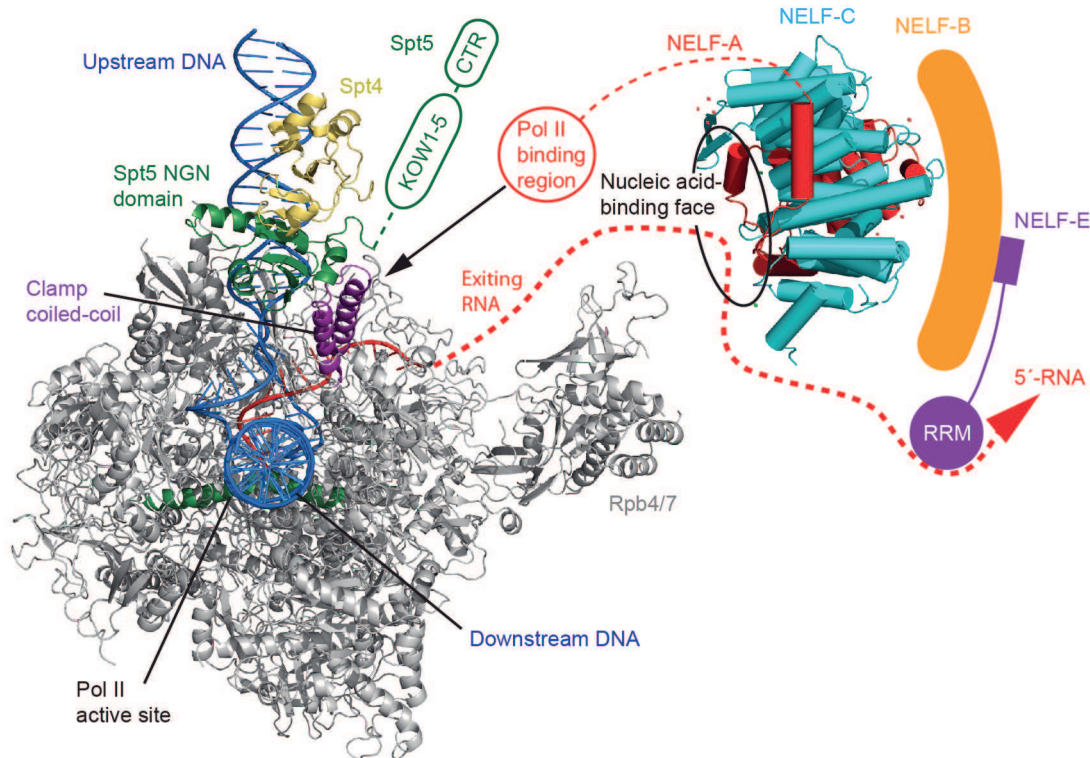
Figure 23: Topology of a promoter-proximally paused Pol II transcription elongation complex containing DSIF and NELF

NELF (Figure 22) was positioned next to the previously published Pol II-DSIF elongation complex model (Martinez-Rucobo et al., 2011). We used a single scale to allow for a comparison of the relative sizes of the complexes. The view of NELF-AC is rotated 90° relative to the 'bottom view' in Figure 9 and Figure 13 around the vertical axis. The positively charged, nucleic acid-binding face of NELF-AC was oriented towards RNA exiting Pol II. Note the Pol II-binding region of NELF-A that may contact the Pol II clamp helices (violet) is not present in the NELF-AC structure. The presumed interaction between the NELF-A Pol II-binding region and the Pol II clamp domain positions the NELF complex close to the RNA exit tunnel of Pol II.

It is known that the extent of Pol II pausing strongly differs between different genes (Muse et al., 2007). Such gene specificity may be explained by differences in promoter-proximal DNA regions. How can DNA sequence influence pausing? First, certain sequences may lead to DNA-RNA hybrids that favor Pol II pausing by slowing down the elongation rate, similar to DNA sequences that influence pausing of bacterial RNA polymerase (Greive and von Hippel, 2005; Larson et al., 2014; Vvedenskaya et al., 2014). Second, nascent RNA may bind to NELF with different affinities, directly influencing the efficiency of NELF recruitment to pausing sites. Indeed we observed that nucleic acid binding of NELF-AC is strongly sequence-dependent. It is also known that DNA regions differ in their GC content (Ginno et al., 2012) and in *Drosophila* there is a known sequence motif that is associated with pausing (Hendrix et al., 2008). Third, nucleosome stabilities vary with DNA sequence and nucleosomes are known to influence Pol II elongation (Gilchrist et al., 2010; Gilchrist

et al., 2008; Mayer et al., 2015).

Our results are essential for deciphering the molecular basis of Pol II pausing and release and its sequence dependency and thus gene specificity. We note that nucleic acid binding alone may explain recruitment of NELF to certain genes and its association with promoter-proximal regions, but is insufficient to explain Pol II pausing, which would additionally require a change in the elongation behavior of the polymerase. This may involve a conformational change in the enzyme that may be triggered or stabilized by NELF binding to the Pol II surface. Analysis of this intricate mechanism awaits the formation of functional complexes comprising Pol II, DSIF, and NELF, and their structural analysis.

2   Further analysis of the NELF subcomplex NELF-AC

Results described in this chapter have been obtained during this thesis. They support and broaden the results described in chapter III 1 but were not published.

2.1   Crystallization experiments with NELF-A$_{6-188}$C$_{36-590}$

A first phase of limited proteolysis using chymotrypsin and subsequent truncation of mobile and unconserved regions resulted in a highly soluble and compact construct consisting of NELF-A (6-188) and NELF-C (36-590) ('NELF-A$_{6-188}$C$_{36-590}$') (Figure 6B, C). Validity of this result was supported by bioinformatic analysis predicting NELF-C to consist of a single and entirely alpha helically structured TH1 domain comprising the conserved region 36-590 (Figure 3A, Figure 12) (Buchan et al., 2013; Marchler-Bauer et al., 2011). Pure NELF-A$_{6-188}$C$_{36-590}$ formed trimers as determined by size exclusion chromatography (Figure 24A) and dynamic light scattering (~250 kDa extimated size, monomer: 85.8 kDa) with an RSD of 24.5 %. Thin needle shaped crystals were obtained after extensive screening (II 2.3.1, Figure 25A). However, the small size and fragility of the crystals prevented us from collecting diffraction data. To improve crystal quality we varied the termini of the construct NELF-A$_{6-188}$C$_{36-590}$ with the objective to remove flexible terminal sequences not identified before or add terminal amino acids important for crystal contacts (Table 4 'DP33-DP47') (Derewenda, 2004). The four termini of NELF-A$_{6-188}$C$_{36-590}$ were varied individually with the remaining three termini invariant resulting in NELF-A constructs with N-termini beginning at S9, T20, G29 or A36 and C-termini ending at S182, S193 or R202. NELF-C N- and C-termini were defined at Q30, S52, D55 or I57 and N559, T568, D573 or N585, respectively. Yet none of these constructs could improve crystal quality.

Our second approach to enhance the quality of NELF-A$_{6-188}$C$_{36-590}$ crystals was based on surface entropy reduction (Cooper et al., 2007; Derewenda, 2004; Derewenda and Vekilov, 2006). Clusters of flexible and polar side chains, in particular glutamate and lysine, comprise a high entropy and consecutively account for protein-protein repulsion preventing crystallization. Mutation to alanine thus can increase chances of successful crystallization. Based on results of the "surface entropy reduction prediction server" (Goldschmidt et al., 2007) four high-entropy clusters were modified: NELF-A E111, E112, Q113; NELF-C E138, E139, E141; NELF-C Q270, E271, K272; NELF-C K371, K372, K374. Clusters were mutated to alanine individually and in combinations (Table 4, 'DP48 - DP53'). We found that all mutations did not affect protein solubility. However, crystal quality could not be increased.

In a second phase of limited proteolysis we found that subtilisin cleaves the N-terminal region of NELF-C (36-189) resulting in the new construct NELF-A$_{6-188}$C$_{183-590}$ ('NELF-AC') (Figure 6C and D, III 1.1). Pure NELF-A$_{6-188}$C$_{183-590}$ was homogeneous and showed no signs of aggregation as demonstrated by size exclusion chromatography (Figure 24B). Using DLS the particle size was calculated to 60 kDa (65.9 kDa in theory) with an RSD of 19.1 %.

In combination with results from interaction assays (Figure 21) we concluded that NELF-C consists of at least two compact and independently folding regions 36-189 and 190-590. The NELF-C N-terminal region accounts for aggregation of NELF-A$_{6-188}$C$_{36-590}$, possibly due to the lack of its binding partner NELF-B. Thus crystallization of the NELF-C N-terminal region presumably requires previous complex formation with NELF-B.

Initial crystallization screens with NELF-A$_{6-188}$C$_{183-590}$ (Table 15) identified aplenty of predominantly polyethylene glycole (PEG) containing crystallization conditions (not shown) providing a basis for optimization and consecutive structure solution as described in III 1.1 (Figure 25B, C).
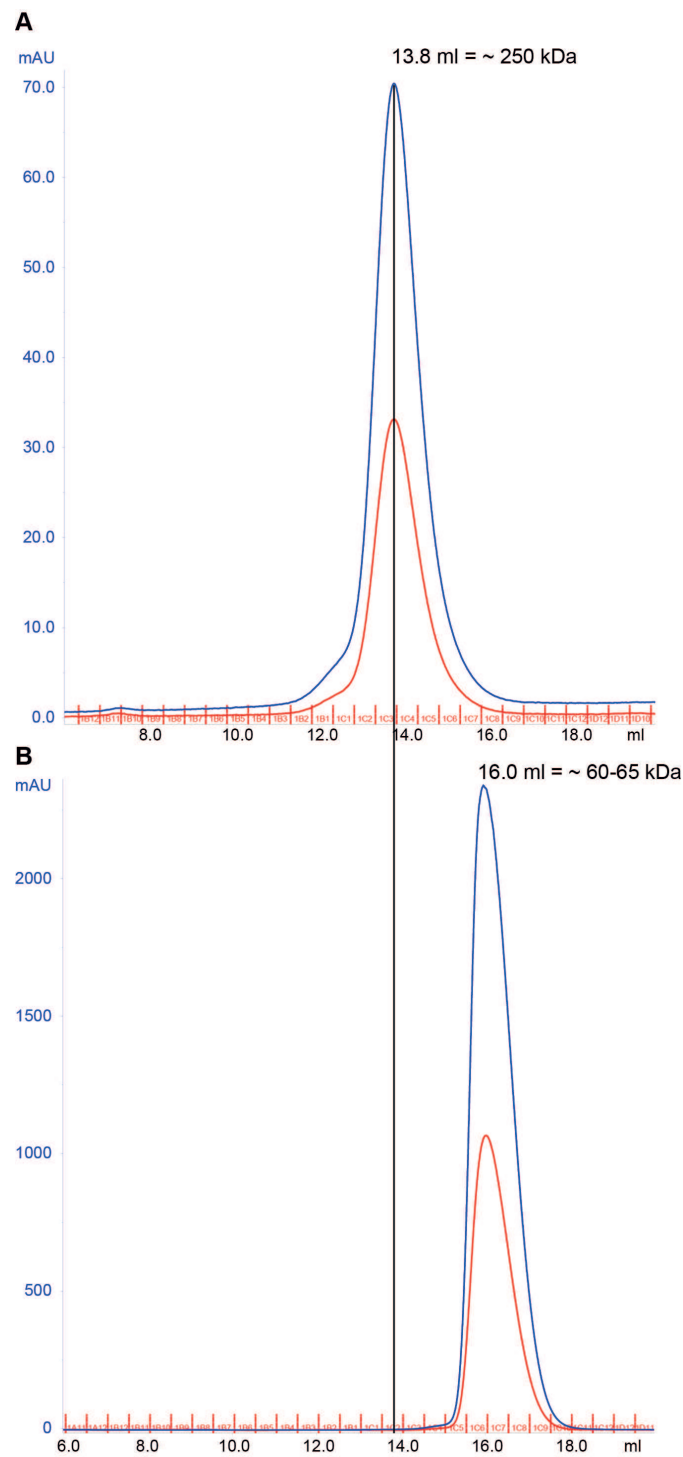
Figure 24: Size-exclusion analysis of truncated NELF-AC subcomplexes.

Size exclusion chromatography of trimeric NELF-A$_{6-188}$C$_{36-590}$ (A) and monomeric NELF-A$_{6-188}$C$_{183-590}$ (B). Truncation of the NELF-C N-terminal region (36-182) results in a substantial shift towards a higher retention volume (13.8 and 16.0 ml, respectively) and hence lower apparent molecular weight (250 and 65 kDa, respectively). For both size-exclusion chromatographies a Superose 6 10/300 column equilibrated in size exclusion buffer A was used. The blue and red curves indicate absorption in milli absorption units (mAU) at 280 nm and 256 nm, respectively.

63
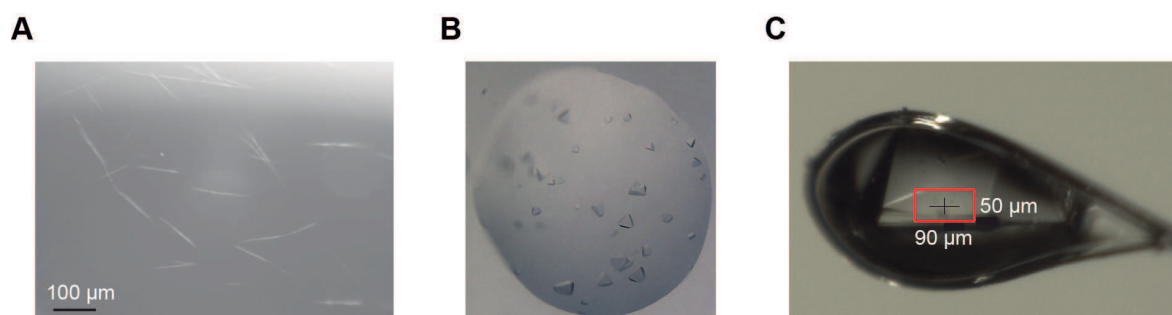
Figure 25: Crystallization of NELF-AC variants

(A) Needle shaped crystals of NELF-A$_{6-188}$C$_{36-590}$ WT grown in 1.15 M (NH$_4$)$_2$SO$_4$, 0.2 M NaCl, 0.1 M Na-HEPES pH 7.4. Despite extensive protein modification quality improvement was not possible.

(B) Crystallization screen Hampton Research Index HT, condition H3 (0.2 M Sodium malonate pH 7.0 and 20 % w/v Polyethylene glycol 3,350). This condition was used as starting point for successful optimization.

(C) Optimized NELF-A$_{6-188}$C$_{183-590}$ crystal that was used to solve the structure, mounted in a cryo-loop. The size of the tetraedric shaped crystal was ~300x200 µM. The crystal was grown in 0.2 M Sodium malonate pH 7.0 and 14.25 % w/v Polyethylene glycol 3,350 (II 2.3.2).

2.2 In depth analysis of NELF-AC interactions with nucleic acids

Additional fluorescence anisotropy experiments measuring the interaction between NELF-A$_{6-188}$C$_{183-590}$ wildtype or mutant protein and nucleic acids enabled us to further explicate and differentiate the results obtained in chapters III 1.4 and III 1.5.

NELF-AC preferentially binds nuleic acids with high GC content

We tested ssDNA and ssRNA with 44 %, 60 % and 72 % GC content as described in III 1.4 (II 2.5, Table 7). We found a low affinity of NELF-AC for both 44 % ssDNA and ssRNA. Binding assays using nucleic acids with higher GC content (60 % and 72 %) yielded no clear pattern. However, by trend nucleic acids with higher GC content exhibited a higher affinity to NELF-AC (Figure 26). Affinites of natural sequences supported this results (Figure 14D): *In vivo c-fos* ssRNA sequence (60 % GC content) showed a lower affinity to NELF-AC compared to *in vivo junB* ssRNA, *junB* ssDNA and *c-fos* ssDNA (76 %, 75 % and 75 % GC content, respectively). Summarized, NELF-AC tends to bind to nucleic acids with high GC content more tightly than to nucleic acids with low GC content.
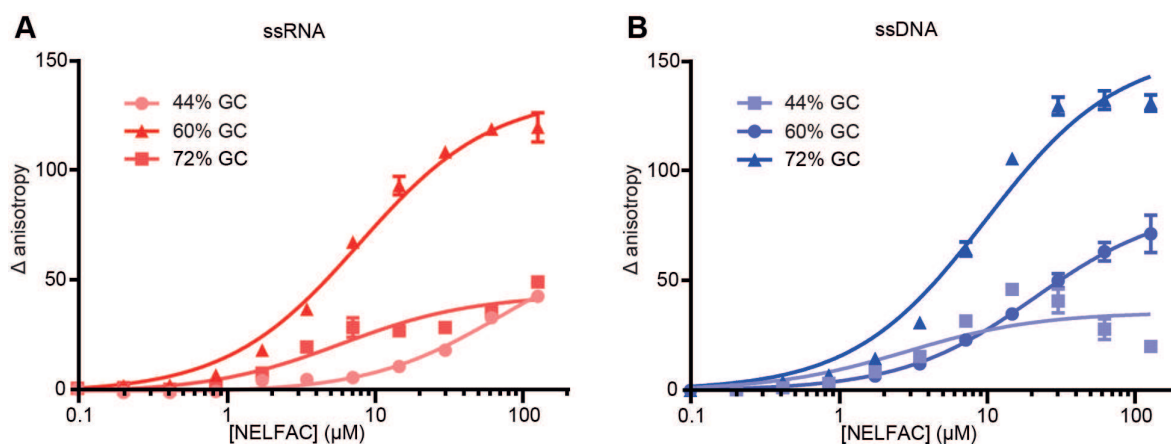
Figure 26: NELF-AC preferentially binds single stranded nucleic acids with high GC content
(A, B) Binding of WT NELF-AC to 10 nM of fluorescently labeled ssRNA (A) or ssDNA (B) with 44 %, 60 % or 72 % GC content as monitored by the changes in relative fluorescence anisotropy (same as Figure 14A-D). Error bars reflect the standard deviation from three experimental replicates.

Surface patch 1 is essential for interaction with nucleic acids

We asked whether specific surface patches contribute to nucleic acid binding in particular. Hence, in addition to the combined mutation of three or four surface patches (Figure 14B and C, III 1.4) we generated NELF-AC variants containing individual mutations of patches 1, 2, 3 or 4 (Figure 13). We then tested their affinity for fluorescently labeled, synthetic 25-nt single-stranded ssDNA and ssRNA random sequence with 60 % GC content and ssDNA random sequence with 72 % GC content (III 1.4, Figure 27, Table 7). We found that mutation of patch 1 strongly reduced binding to all three nucleic acids tested. Patch 1 thus appears to be crucial for binding nucleic acids. In contrast, mutating patch 2 had little effect on the affinity for both types of nucleic acid in general. Binding assays with NELF-AC variants carrying mutations in patches 3 and 4 produced ambiguous results, only interaction with ssDNA but not ssRNA was affected. Taken together patch 1 is an essential binding site for nucleic acids while patch 2 contributes to binding of nucleic acids to a minor extent and rather stabilizes the interaction. Contribution of patches 3 and 4 to nucleic acid binding remains unclear.
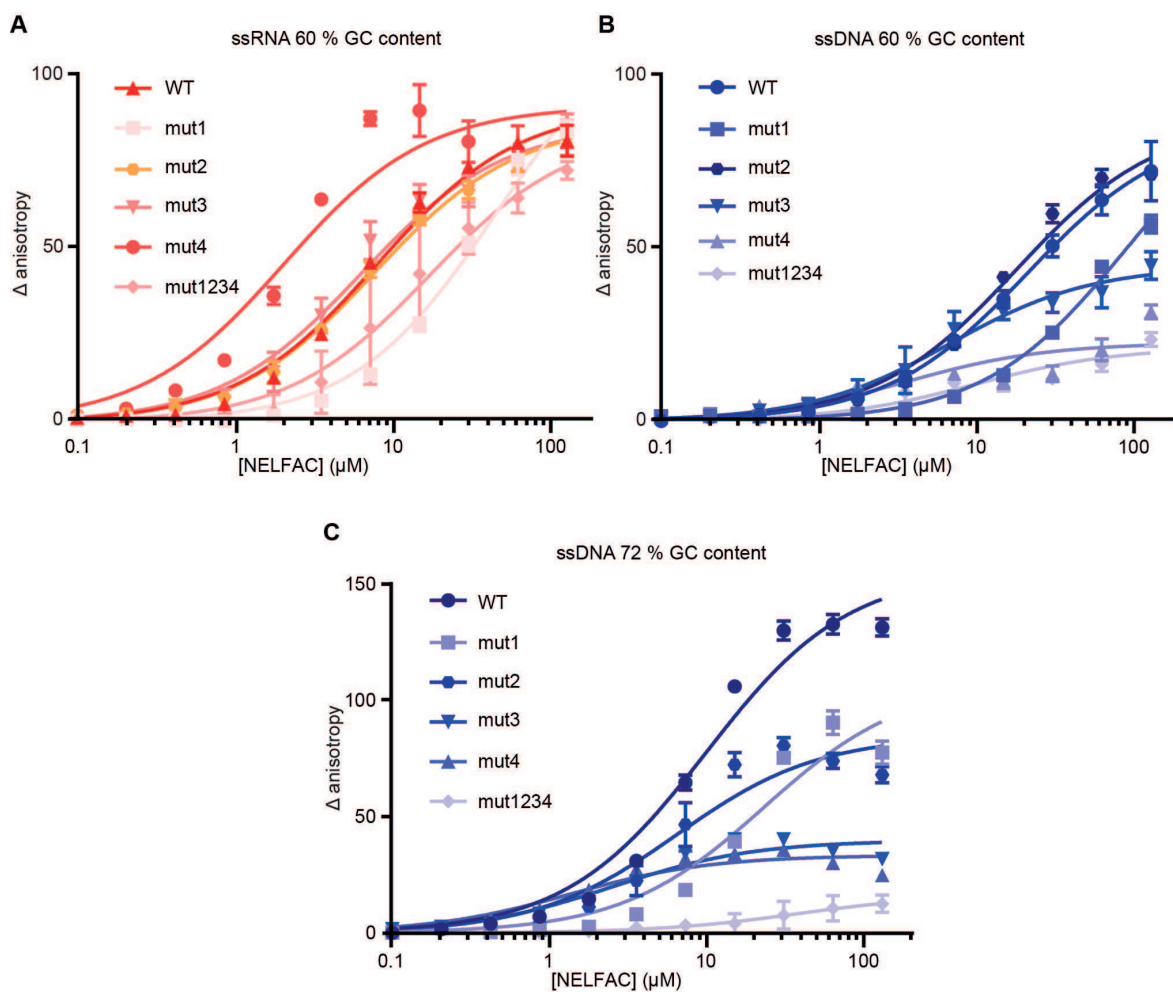
Figure 27: Effect of mutating individual suface patches on NELF-AC affinity for nucleic acids

Binding of WT NELF-AC and variants containing one or four mutations in surface patches to 10 nM of fluorescently labeled ssRNA (A), ssDNA (B) with 60 % and ssDNA (C) with 72 % GC content as monitored by the changes in relative fluorescence anisotropy (same as Figure 14A-D). Numbers indicate mutated patches present in NELF-AC variants (Figure 13). Error bars reflect the standard deviation from three experimental replicates.

2.3 Discussion

Detailed analysis of interaction between NELF- $A_{6-188}C_{183-590}$ and nucleic acids revealed a differentiated pattern. The observed significance of patch 1 for nucleic acid binding becomes momentous in view of its vicinity to the phosphorylation sites NELF-C T285, Y289 and T318. Phosphorylation abrogates the nucleic acid binding capacity of a nearby patch by neutralizing the positive charge. Indeed, phosphomimetic mutations T285D and Y289E reduce nucleic acid affinity much stronger than NELF-A T157D, which is located proximal to the positive patch 4 (Figure 15). Thus our results explain the efficient regulation of nucleic acid binding by phosphorylation. Further, multiple phosphorylation sites that are presumably controlled by different kinases, provide the possibility for dynamic regulation of promoter-proximal pausing.

The observed tendency of NELF-AC to stronger bind nucleic acids with high GC content can contribute to explaining the establishment of promoter-poximal pausing. Gene regions circumjacent and downstream of the TSS exhibit a significantly above-average GC content (Ginno et al., 2013; Ginno et al., 2012). A preference of NELF-AC for nucleic acids with high GC content sequences would ensure a tight grip and prevent unspecific interactions outside promoter-proximal pausing regions.

IV   Conclusion and outlook

Since NELF has been discovered in 1999 (Yamaguchi et al., 1999a) numerous studies revealing the biochemistry (Li et al., 2013; Missra and Gilmour, 2010; Narita et al., 2003; Wu et al., 2005) and genomic function (Fujita and Schlegel, 2010; Gilchrist et al., 2008; Williams et al., 2015; Zeitlinger et al., 2007) were published. The structure of the common and widespread (Koonin and Makarova, 2013) NELF-E RRM domain has been solved (Rao et al., 2006; Rao et al., 2008) and two functional regions of NELF-A, the NELF-C and the Pol-II interaction domain, have been mapped (Narita et al., 2003). The general importance of NELF for transcriptional regulation has recently been recognized (Yamaguchi et al., 2013). Yet, no specific structural information on NELF was available to date.

In this work we could solve the high-resolution structure of a highly conserved NELF-AC subcomplex and identify a binding site for single stranded nucleic acids previously unknown. NELF-AC affinity for nucleic acids strongly depends on sequence and GC content and can be regulated through phosphorylation by P-TEFb and likely other kinases. Analysis of the holo-NELF architecture showed that both single stranded nucleic acid binding faces are located at opposite ends of the NELF complex and freely accessible. These results allowed to propose a model describing the topology of a pausing complex and the molecular basis of NELF action.

Our model now needs to be reviewed, verified and expanded. It provides a starting point for future structural and mechanistic studies to deeper understand the molecular background of promoter-proximal pausing. However, important aspects of the pausing mechanism such as the change in Pol II conformation upon NELF binds to Pol II and resulting consequences for the transcription rate cannot be explained yet and require further research.

The next steps to better understand NELF are in-depth analysis of the function of the NELF-AC subcomplex and the structural characterization of further parts of NELF. In continuation of this work two major points should be addressed. First, the specificity and biological relevance of the newly identified NELF-AC nucleic acid binding capacity. Despite the relatively weak interaction (K_d ~10 µM), binding is not unspecific as proven by the strong sequence dependency. One way to verify the significance of nucleic acid binding *in vivo* could be to knock out NELF-A and -C genes in cell culture and subsequently replenish NELF-A and -C carrying mutations in positive patches. The resulting cells can then be analyzed globally for differences in transcription and

specifically for defects in promoter-proximal pausing at well-known paused genes like *hsp70* or *junB*. Consecutively the specificity of NELF-AC for single stranded nucleic acids regarding type, sequence and secondary structure needs to be determined *in vivo*. Although *in vivo* interaction with ssDNA is less probable than with ssRNA, as discussed above, our results were ambiguous. PAR-CLIP to discover RNA-bound NELF (Hafner et al., 2010), chromatin immunoprecipitation of DNA-bound protein and related methods would clarify this question. Subsequently, the preference of NELF-AC for specific sequences, sequence motifs or sequence characteristics like the GC content should be defined more accurately by bioinformatic analysis of *in vivo* experiments and *in vitro* validation of data using systematic evoluation of ligands by exponential enrichment (SELEX) as demonstrated for NELF-E RRM (Pagano et al., 2014). Once a tightly interacting nucleic acid has been identified, crystallization of NELF-AC and its nucleic acid ligand is within the realms of possibility. Structural elucidation of a NELF-AC-nucleic acid complex would render a deeper insight into the interaction between the two molecules possible. Since NELF-AC was crystallized in a variety of PEG-containing conditions (III 2.1) this provides a promising starting point.

Another important question emanating from this work is for the function and interaction partners of the highly conserved NELF-AC surface area 'patch 5' (Figure 13). This site is outstanding in the context of the generally well conserved NELF surface and likely fundamental for NELF function. One approach could be comparative pull-down assays using affinity-tagged NELF-AC subcomplexes with and without mutations in patch 5 and NELF-depleted nuclear extracts as demonstrated previously (Narita et al., 2003) or alternatively *in vivo* over-expression of affinity-tagged NELF-AC followed by pull-down.

Structural studies on other NELF subunits as well are necessary to clarify how NELF contributes to establish promoter-proximal pausing. Crystallization of NELF-AC has revealed a novel function and its regulation as well as a highly conserved surface area previously unknown. Similarly, high-resolution structures of other NELF regions might lead to unexpected new discoveries. Regarding its stucture and function NELF-B is poorly analyzed. Considering the high sequence conservation of NELF-B (Figure 3A) and its central position within the NELF complex (Figure 22) (Narita et al., 2003), resolving the structure would be an important step. A promising approach might be co-expression of NELF-B carrying a solubility-enhancing tag together with the NELF-C N-terminal region (36-189) (Figure 21) followed by determination of a minimal complex using repetitive limited proteolysis and truncation as applied successfully for crystallizing NELF-AC. Results obtained during this thesis indicate a flexible linker between NELF-C N- and C-

terminal region (36-189 and 190-590) (Figure 6). For that reason successful crystallization of a NELF-ABC complex presumably requires genetic engeneering of this linker region. Determining the position of the NELF-E RRM on the NELF-B surface and in the context of complete NELF would allow to better estimate the minimal length of a nascent RNA bound to the RRM and the significance of NELF-E in promoter-proximal pausing (Missra and Gilmour, 2010). The flexible linker region between the predicted NELF-E N-terminal helix (7-36) and the RRM would allow the RRM to be dynamic within NELF if it was not fixed on the NELF-B surface. Defining the position of the N-terminal helix on NELF-B more precisely thus is important, too. This could be achieved by co-crystallization of NELF-B and the NELF-E region 7-36.

Structure and function of the partially conserved NELF-A C-terminal region (249-528) are unknown (Figure 3A, Figure 11). This part of NELF-A has been analysed in the context of the bachelor thesis of Denis Höfler. All results are described in this work. In brief, we found that NELF-A C-terminal region can be separated into two stable and soluble subregions (250-369, 349-528) that can readily be expressed and purified. However, successful crystallization requires further construct optimization.

The final question is how NELF action pauses elongating RNA polymerase. In order to find the answer to this question, the interactions of NELF with other protein complexes need to be analyzed. A first step would be to identify the NELF binding site on the Pol II surface. The inherently unstable and likely unfolded Pol II-interaction region of NELF-A (Narita et al., 2003) (189-248) presumably requires an interaction partner to adept a defined conformation. Crosslinking of a Pol II-NELF complex coupled with mass-spectrometry could locate the position of NELF on Pol II. As mentioned above (III 1.8) detailed structural studies with a ternary pausing complex of Pol II-DSIF-NELF including a nucleic acid scaffold are requisite to obtain an overview of all molecular aspects of pausing. Cryo-electron microscopy studies of large complexes combined with crystallographic data from subcomplexes or crosslinking were proven to be a powerful approach to such problems (He et al., 2013; Klein et al., 2011; Martinez-Rucobo et al., 2015). With the crystallization of large functional complexes possibly being intricate, cryo-electron microscopy studies of a complete pausing complex accompanied by x-ray analysis of subcomplexes would be the method of choice to answer the molecular enigmas of the pausing complex.

References

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

Afonine, P.V., Grosse-Kunstleve, R.W., and Adams, P.D. (2005). A robust bulk-solvent correction and anisotropic scaling procedure. Acta crystallographica Section D, Biological crystallography *61*, 850-855.

Aida, M., Chen, Y., Nakajima, K., Yamaguchi, Y., Wada, T., and Handa, H. (2006). Transcriptional pausing caused by NELF plays a dual role in regulating immediate-early expression of the junB gene. Mol Cell Biol *26*, 6094-6104.

Aiyar, S.E., Sun, J.L., Blair, A.L., Moskaluk, C.A., Lu, Y.Z., Ye, Q.N., Yamaguchi, Y., Mukherjee, A., Ren, D.M., Handa, H.*, et al.* (2004). Attenuation of estrogen receptor alpha-mediated transcription through estrogen-stimulated recruitment of a negative elongation factor. Genes Dev *18*, 2134-2146.

Amir-Zilberstein, L., Ainbinder, E., Toube, L., Yamaguchi, Y., Handa, H., and Dikstein, R. (2007). Differential regulation of NF-kappaB by elongation factors is determined by core promoter type. Mol Cell Biol *27*, 5246-5259.

Amleh, A., Nair, S.J., Sun, J., Sutherland, A., Hasty, P., and Li, R. (2009). Mouse cofactor of BRCA1 (Cobra1) is required for early embryogenesis. PLoS One *4*, e5034.

Andrecka, J., Lewis, R., Bruckner, F., Lehmann, E., Cramer, P., and Michaelis, J. (2008). Single-molecule tracking of mRNA exiting from RNA polymerase II. Proc Natl Acad Sci U S A *105*, 135-140.

Andrulis, E.D., Guzman, E., Doring, P., Werner, J., and Lis, J.T. (2000). High-resolution localization of Drosophila Spt5 and Spt6 at heat shock genes in vivo: roles in promoter proximal pausing and transcription elongation. Genes Dev *14*, 2635-2649.

Armache, K.J., Kettenberger, H., and Cramer, P. (2003). Architecture of initiation-competent 12-subunit RNA polymerase II. Proc Natl Acad Sci U S A *100*, 6964-6968.

Armache, K.J., Mitterweger, S., Meinhart, A., and Cramer, P. (2005). Structures of complete RNA polymerase II and its subcomplex, Rpb4/7. J Biol Chem *280*, 7131-7134.

Artsimovitch, I., and Landick, R. (2002). The transcriptional regulator RfaH stimulates RNA chain synthesis after recruitment to elongation complexes by the exposed nontemplate DNA strand. Cell *109*, 193-203.

Bai, L., Shundrovsky, A., and Wang, M.D. (2004). Sequence-dependent kinetic model for transcription elongation by RNA polymerase. J Mol Biol *344*, 335-349.

Barboric, M., Nissen, R.M., Kanazawa, S., Jabrane-Ferrat, N., and Peterlin, B.M. (2001). NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II. Mol Cell *8*, 327-337.

Barboric, M., Yik, J.H., Czudnochowski, N., Yang, Z., Chen, R., Contreras, X., Geyer, M., Matija Peterlin, B., and Zhou, Q. (2007). Tat competes with HEXIM1 to increase the active pool of P-TEFb for HIV-1 transcription. Nucleic Acids Res *35*, 2003-2012.

Baugh, L.R., Demodena, J., and Sternberg, P.W. (2009). RNA Pol II accumulates at promoters of growth genes during developmental arrest. Science *324*, 92-94.

Baumli, S., Hole, A.J., Wang, L.Z., Noble, M.E., and Endicott, J.A. (2012). The CDK9 tail determines the reaction pathway of positive transcription elongation factor b. Structure *20*, 1788-1795.

Baumli, S., Lolli, G., Lowe, E.D., Troiani, S., Rusconi, L., Bullock, A.N., Debreczeni, J.E., Knapp, S., and Johnson, L.N. (2008). The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation. Embo J *27*, 1907-1918.

Belogurov, G.A., Vassylyeva, M.N., Svetlov, V., Klyuyev, S., Grishin, N.V., Vassylyev, D.G., and Artsimovitch, I. (2007). Structural basis for converting a general transcription factor into an operon-specific virulence regulator. Mol Cell *26*, 117-129.

Bernecky, C., Grob, P., Ebmeier, C.C., Nogales, E., and Taatjes, D.J. (2011). Molecular architecture of the human Mediator-RNA polymerase II-TFIIF assembly. PLoS biology *9*, e1000603.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K.*, et al.* (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315-326.

Berrow, N.S., Alderton, D., Sainsbury, S., Nettleship, J., Assenberg, R., Rahman, N., Stuart, D.I., and Owens, R.J. (2007). A versatile ligation-independent cloning method suitable for high-throughput expression screening applications. Nucleic Acids Res *35*, e45.

Boettiger, A.N., and Levine, M. (2009). Synchronous and stochastic patterns of gene activation in the Drosophila embryo. Science *325*, 471-473.

Broennimann, C., Eikenberry, E.F., Henrich, B., Horisberger, R., Huelsen, G., Pohl, E., Schmitt, B., Schulze-Briese, C., Suzuki, M., Tomizaki, T*., et al.* (2006). The PILATUS 1M detector. Journal of synchrotron radiation *13*, 120-130.

Buchan, D.W., Minneci, F., Nugent, T.C., Bryson, K., and Jones, D.T. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res *41*, W349-357.

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

Bushnell, D.A., Westover, K.D., Davis, R.E., and Kornberg, R.D. (2004). Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. Science *303*, 983-988.

Chang, G.S., Noegel, A.A., Mavrich, T.N., Muller, R., Tomsho, L., Ward, E., Felder, M., Jiang, C., Eichinger, L., Glockner, G*., et al.* (2012). Unusual combinatorial involvement of poly-A/T tracts in organizing genes and chromatin in Dictyostelium. Genome Res *22*, 1098-1106.

Chen, H.T., Warfield, L., and Hahn, S. (2007). The positions of TFIIF and TFIIE in the RNA polymerase II transcription preinitiation complex. Nat Struct Mol Biol *14*, 696-703.

Chen, R., Liu, M., Li, H., Xue, Y., Ramey, W.N., He, N., Ai, N., Luo, H., Zhu, Y., Zhou, N*., et al.* (2008). PP2B and PP1alpha cooperatively disrupt 7SK snRNP to release P-TEFb for transcription in response to Ca2+ signaling. Genes Dev *22*, 1356-1368.

Chen, R., Yang, Z., and Zhou, Q. (2004). Phosphorylated positive transcription elongation factor b (P-TEFb) is tagged for inhibition through association with 7SK snRNA. J Biol Chem *279*, 4153-4160.

Chen, Y., Yamaguchi, Y., Tsugeno, Y., Yamamoto, J., Yamada, T., Nakamura, M., Hisatake, K., and Handa, H. (2009). DSIF, the Paf1 complex, and Tat-SF1 have nonredundant, cooperative roles in RNA polymerase II elongation. Genes Dev *23*, 2765-2777.

Cheng, B., and Price, D.H. (2007). Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. J Biol Chem *282*, 21901-21912.

Cherepanov, P., Sun, Z.Y., Rahman, S., Maertens, G., Wagner, G., and Engelman, A. (2005). Solution structure of the HIV-1 integrase-binding domain in LEDGF/p75. Nat Struct Mol Biol *12*, 526-532.

Cheung, A.C., and Cramer, P. (2011). Structural basis of RNA polymerase II backtracking, arrest and reactivation. Nature *471*, 249-253.

Cheung, A.C., Sainsbury, S., and Cramer, P. (2011). Structural basis of initial RNA polymerase II transcription. Embo J *30*, 4755-4763.

Chiba, K., Yamamoto, J., Yamaguchi, Y., and Handa, H. (2010). Promoter-proximal pausing and its release: molecular mechanisms and physiological functions. Exp Cell Res *316*, 2723-2730.

Cingolani, G., Petosa, C., Weis, K., and Mueller, C. (1999). Structure of importin-b bound to the IBB domain of importin-a. Nature *399*, 221-229.

Combe, C.W., Fischer, L., and Rappsilber, J. (2015). xiNET: cross-link network maps with residue resolution. Molecular & cellular proteomics : MCP *14*, 1137-1147.

Cooper, D.R., Boczek, T., Grelewska, K., Pinkowska, M., Sikorska, M., Zawadzki, M., and Derewenda, Z. (2007). Protein crystallization by surface entropy reduction: optimization of the SER strategy. Acta crystallographica Section D, Biological crystallography *63*, 636-645.

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

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta crystallographica Section D, Biological crystallography *62*, 1002-1011.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature biotechnology *26*, 1367-1372.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V., and Mann, M. (2011). Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J Proteome Res *10*, 1794-1805.

Cramer, P. (2002a). Common structural features of nucleic acid polymerases. Bioessays *24*, 724-729.

Cramer, P. (2002b). Multisubunit RNA polymerases. Current opinion in structural biology *12*, 89-97.

Cramer, P., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. Science *292*, 1863-1876.

Derewenda, Z.S. (2004). The use of recombinant methods and molecular engineering in protein crystallization. Methods (San Diego, Calif) *34*, 354-363.

Derewenda, Z.S., and Vekilov, P.G. (2006). Entropy and surface engineering in protein crystallization. Acta crystallographica Section D, Biological crystallography *62*, 116-124.

Edwards, A.M., Kane, C.M., Young, R.A., and Kornberg, R.D. (1991). Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter in vitro. J Biol Chem *266*, 71-75.

Emsley, P., and Cowtan, K. (2004). Coot: model-building tools for molecular graphics. Acta crystallographica Section D, Biological crystallography *60*, 2126-2132.

Fivaz, J., Bassi, M.C., Pinaud, S., and Mirkovitch, J. (2000). RNA polymerase II promoter-proximal pausing upregulates c-fos gene expression. Gene *255*, 185-194.

Forget, D., Langelier, M.F., Therien, C., Trinh, V., and Coulombe, B. (2004). Photo-cross-linking of a purified preinitiation complex reveals central roles for the RNA polymerase II mobile clamp and TFIIE in initiation mechanisms. Mol Cell Biol *24*, 1122-1131.

Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature *461*, 186-192.

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

Fujita, T., and Schlegel, W. (2010). Promoter-proximal pausing of RNA polymerase II: an opportunity to regulate gene transcription. J Recept Signal Transduct Res *30*, 31-42.

Gargano, B., Amente, S., Majello, B., and Lania, L. (2007). P-TEFb is a crucial co-factor for Myc transactivation. Cell Cycle *6*, 2031-2037.

Giardina, C., Perez-Riba, M., and Lis, J.T. (1992). Promoter melting and TFIID complexes on Drosophila genes in vivo. Genes Dev *6*, 2190-2200.

Gilchrist, D.A., Dos Santos, G., Fargo, D.C., Xie, B., Gao, Y., Li, L., and Adelman, K. (2010). Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. Cell *143*, 540-551.

Gilchrist, D.A., Nechaev, S., Lee, C., Ghosh, S.K., Collins, J.B., Li, L., Gilmour, D.S., and Adelman, K. (2008). NELF-mediated stalling of Pol II can enhance gene expression by blocking promoter-proximal nucleosome assembly. Genes Dev *22*, 1921-1933.

Gilmour, D.S. (2009). Promoter proximal pausing on genes in metazoans. Chromosoma *118*, 1-10.

Gilmour, D.S., and Lis, J.T. (1986). RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. Mol Cell Biol *6*, 3984-3989.

Ginno, P.A., Lim, Y.W., Lott, P.L., Korf, I., and Chedin, F. (2013). GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination. Genome Res *23*, 1590-1600.

Ginno, P.A., Lott, P.L., Christensen, H.C., Korf, I., and Chedin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell *45*, 814-825.

Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A., and Kornberg, R.D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. Science *292*, 1876-1882.

Goldschmidt, L., Cooper, D.R., Derewenda, Z.S., and Eisenberg, D. (2007). Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. Protein Sci *16*, 1569-1576.

Gonatopoulos-Pournatzis, T., and Cowling, V.H. (2014). Cap-binding complex (CBC). Biochem J *457*, 231-242.

Greive, S.J., and von Hippel, P.H. (2005). Thinking quantitatively about transcriptional regulation. Nat Rev Mol Cell Biol *6*, 221-232.

Grohmann, D., Nagy, J., Chakraborty, A., Klose, D., Fielden, D., Ebright, R.H., Michaelis, J., and Werner, F. (2011). The initiation factor TFE and the elongation factor Spt4/5 compete for the RNAP clamp during transcription initiation and elongation. Mol Cell *43*, 263-274.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Guo, M., Xu, F., Yamada, J., Egelhofer, T., Gao, Y., Hartzog, G.A., Teng, M., and Niu, L. (2008). Core structure of the yeast spt4-spt5 complex: a conserved module for regulation of transcription elongation. Structure *16*, 1649-1658.

Haag, J.R., and Pikaard, C.S. (2011). Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. Nat Rev Mol Cell Biol *12*, 483-492.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M.*, et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell *141*, 129-141.

Hahn, S., and Young, E.T. (2011). Transcriptional regulation in Saccharomyces cerevisiae: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. Genetics *189*, 705-736.

Hargreaves, D.C., Horng, T., and Medzhitov, R. (2009). Control of inducible gene expression by signal-dependent transcriptional elongation. Cell *138*, 129-145.

He, N., Liu, M., Hsu, J., Xue, Y., Chou, S., Burlingame, A., Krogan, N.J., Alber, T., and Zhou, Q. (2010). HIV-1 Tat and host AFF4 recruit two transcription elongation factors into a bifunctional complex for coordinated activation of HIV-1 transcription. Mol Cell *38*, 428-438.

He, Y., Fang, J., Taatjes, D.J., and Nogales, E. (2013). Structural visualization of key steps in human transcription initiation. Nature *495*, 481-486.

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

Hendrix, D.A., Hong, J.W., Zeitlinger, J., Rokhsar, D.S., and Levine, M.S. (2008). Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. Proc Natl Acad Sci U S A *105*, 7762-7767.

Henriques, T., Gilchrist, D.A., Nechaev, S., Bern, M., Muse, G.W., Burkholder, A., Fargo, D.C., and Adelman, K. (2013). Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. Mol Cell *52*, 517-528.

Herzog, F., Kahraman, A., Boehringer, D., Mak, R., Bracher, A., Walzthoeni, T., Leitner, A., Beck, M., Hartl, F.U., Ban, N.*, et al.* (2012). Structural probing of a protein phosphatase 2A network by chemical cross-linking and mass spectrometry. Science *337*, 1348-1352.

Hirtreiter, A., Damsma, G.E., Cheung, A.C., Klose, D., Grohmann, D., Vojnic, E., Martin, A.C., Cramer, P., and Werner, F. (2010). Spt4/5 stimulates transcription elongation through the RNA polymerase clamp coiled-coil motif. Nucleic Acids Res *38*, 4040-4051.

Hocine, S., Singer, R.H., and Grunwald, D. (2010). RNA processing and export. Cold Spring Harb Perspect Biol *2*, a000752.

Holm, L., and Rosenstrom, P. (2010). Dali server: conservation mapping in 3D. Nucleic Acids Res *38*, W545-549.

Holstege, F.C., van der Vliet, P.C., and Timmers, H.T. (1996). Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors IIE and IIH. Embo J *15*, 1666-1677.

Hsu, C.L., and Stevens, A. (1993). Yeast cells lacking 5'-->3' exoribonuclease 1 contain mRNA species that are poly(A) deficient and partially lack the 5' cap structure. Mol Cell Biol *13*, 4826-4835.

Iida, M., Iizuka, N., Tsunedomi, R., Tsutsui, M., Yoshida, S., Maeda, Y., Tokuhisa, Y., Sakamoto, K., Yoshimura, K., Tamesa, T.*, et al.* (2012). Overexpression of the RD RNA binding protein in hepatitis C virus-related hepatocellular carcinoma. Oncol Rep *28*, 728-734.

Iyer, V., and Struhl, K. (1995). Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. Embo J *14*, 2570-2579.

Jang, M.K., Mochizuki, K., Zhou, M., Jeong, H.S., Brady, J.N., and Ozato, K. (2005). The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. Mol Cell *19*, 523-534.

Jeronimo, C., Forget, D., Bouchard, A., Li, Q., Chua, G., Poitras, C., Therien, C., Bergeron, D., Bourassa, S., Greenblatt, J.*, et al.* (2007). Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. Mol Cell *27*, 262-274.

Jimeno-Gonzalez, S., Ceballos-Chavez, M., and Reyes, J.C. (2015). A positioned +1 nucleosome enhances promoter-proximal pausing. Nucleic Acids Res *43*, 3068-3078.

Jonkers, I., Kwak, H., and Lis, J.T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. Elife *3*, e02407.

Jonkers, I., and Lis, J.T. (2015). Getting up to speed with transcription elongation by RNA polymerase II. Nat Rev Mol Cell Biol *16*, 167-177.

Kabsch, W. (2010). Xds. Acta crystallographica Section D, Biological crystallography *66*, 125-132.

Kaplan, N., Moore, I.K., Fondufe-Mittendorf, Y., Gossett, A.J., Tillo, D., Field, Y., LeProust, E.M., Hughes, T.R., Lieb, J.D., Widom, J.*, et al.* (2009). The DNA-encoded nucleosome organization of a eukaryotic genome. Nature *458*, 362-366.

Karn, J., and Stoltzfus, C.M. (2012). Transcriptional and posttranscriptional regulation of HIV-1 gene expression. Cold Spring Harbor perspectives in medicine *2*, a006916.

Karplus, P.A., and Diederichs, K. (2012). Linking crystallographic model and data quality. Science *336*, 1030-1033.

Kassube, S.A., Fang, J., Grob, P., Yakovchuk, P., Goodrich, J.A., and Nogales, E. (2013). Structural insights into transcriptional repression by noncoding RNAs that bind to human Pol II. J Mol Biol *425*, 3639-3648.

Keegan, B.R., Feldman, J.L., Lee, D.H., Koos, D.S., Ho, R.K., Stainier, D.Y., and Yelon, D. (2002). The elongation factors Pandora/Spt6 and Foggy/Spt5 promote transcription in the zebrafish embryo. Development *129*, 1623-1632.

Kerzendorfer, C., Hannes, F., Colnaghi, R., Abramowicz, I., Carpenter, G., Vermeesch, J.R., and O'Driscoll, M. (2012). Characterizing the functional consequences of haploinsufficiency of NELF-A (WHSC2) and SLBP identifies novel cellular phenotypes in Wolf-Hirschhorn syndrome. Hum Mol Genet *21*, 2181-2193.

Kettenberger, H., Armache, K.J., and Cramer, P. (2004). Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. Mol Cell *16*, 955-965.

Kim, H.J., Jeong, S.H., Heo, J.H., Jeong, S.J., Kim, S.T., Youn, H.D., Han, J.W., Lee, H.W., and Cho, E.J. (2004). mRNA capping enzyme activity is coupled to an early transcription elongation. Mol Cell Biol *24*, 6184-6193.

Kim, T.K., Ebright, R.H., and Reinberg, D. (2000). Mechanism of ATP-dependent promoter melting by transcription factor IIH. Science *288*, 1418-1422.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993). Crystal structure of a yeast TBP/TATA-box complex. Nature *365*, 512-520.

Kininis, M., Isaacs, G.D., Core, L.J., Hah, N., and Kraus, W.L. (2009). Postrecruitment regulation of RNA polymerase II directs rapid signaling responses at the promoters of estrogen target genes. Mol Cell Biol *29*, 1123-1133.

Klein, B.J., Bose, D., Baker, K.J., Yusoff, Z.M., Zhang, X., and Murakami, K.S. (2011). RNA polymerase and transcription elongation factor Spt4/5 complex structure. Proc Natl Acad Sci U S A *108*, 546-550.

Klimas, N., Koneru, A.O., and Fletcher, M.A. (2008). Overview of HIV. Psychosomatic medicine *70*, 523-530.

Koonin, E.V., and Makarova, K.S. (2013). CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA Biol *10*, 679-686.

Kostrewa, D., Zeller, M.E., Armache, K.J., Seizl, M., Leike, K., Thomm, M., and Cramer, P. (2009). RNA polymerase II-TFIIB structure and mechanism of transcription initiation. Nature *462*, 323-330.

Krueger, B.J., Varzavand, K., Cooper, J.J., and Price, D.H. (2010). The mechanism of release of P-TEFb and HEXIM1 from the 7SK snRNP by viral and cellular activators includes a conformational change in 7SK. PLoS One *5*, e12335.

Krumm, A., Meulia, T., Brunvand, M., and Groudine, M. (1992). The block to transcriptional elongation within the human c-myc gene is determined in the promoter-proximal region. Genes Dev *6*, 2201-2213.

Kuehner, J.N., Pearson, E.L., and Moore, C. (2011). Unravelling the means to an end: RNA polymerase II transcription termination. Nat Rev Mol Cell Biol *12*, 283-294.

Kulaeva, O.I., Hsieh, F.K., Chang, H.W., Luse, D.S., and Studitsky, V.M. (2013). Mechanism of transcription through a nucleosome by RNA polymerase II. Biochim Biophys Acta *1829*, 76-83.

Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013). Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. Science *339*, 950-953.

Kwak, H., and Lis, J.T. (2013). Control of transcriptional elongation. Annu Rev Genet *47*, 483-508.

Kyrpides, N.C., Woese, C.R., and Ouzounis, C.A. (1996). KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. Trends in biochemical sciences *21*, 425-426.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R.*, et al.* (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947-2948.

Larson, M.H., Mooney, R.A., Peters, J.M., Windgassen, T., Nayak, D., Gross, C.A., Block, S.M., Greenleaf, W.J., Landick, R., and Weissman, J.S. (2014). A pause sequence enriched at translation start sites drives transcription dynamics in vivo. Science *344*, 1042-1047.

Li, B., Carey, M., and Workman, J.L. (2007). The role of chromatin during transcription. Cell *128*, 707-719.

Li, J., and Gilmour, D.S. (2011). Promoter proximal pausing and the control of gene expression. Curr Opin Genet Dev *21*, 231-235.

Li, J., Liu, Y., Rhee, H.S., Ghosh, S.K., Bai, L., Pugh, B.F., and Gilmour, D.S. (2013). Kinetic competition between elongation rate and binding of NELF controls promoter-proximal pausing. Mol Cell *50*, 711-722.

Li, Q., Price, J.P., Byers, S.A., Cheng, D., Peng, J., and Price, D.H. (2005). Analysis of the large inactive P-TEFb complex indicates that it contains one 7SK molecule, a dimer of HEXIM1 or HEXIM2, and two P-TEFb molecules containing Cdk9 phosphorylated at threonine 186. J Biol Chem *280*, 28819-28826.

Lis, J.T., Mason, P., Peng, J., Price, D.H., and Werner, J. (2000). P-TEFb kinase recruitment and function at heat shock loci. Genes Dev *14*, 792-803.

Liu, X., Bushnell, D.A., Silva, D.A., Huang, X., and Kornberg, R.D. (2011). Initiation complex structure and promoter proofreading. Science *333*, 633-637.

Liu, Y., Warfield, L., Zhang, C., Luo, J., Allen, J., Lang, W.H., Ranish, J., Shokat, K.M., and Hahn, S. (2009). Phosphorylation of the transcription elongation factor Spt5 by yeast Bur1 kinase stimulates recruitment of the PAF complex. Mol Cell Biol *29*, 4852-4863.

Luecke, H.F., and Yamamoto, K.R. (2005). The glucocorticoid receptor blocks P-TEFb recruitment by NFkappaB to effect promoter-specific transcriptional repression. Genes Dev *19*, 1116-1127.

Mandal, S.S., Chu, C., Wada, T., Handa, H., Shatkin, A.J., and Reinberg, D. (2004). Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. Proc Natl Acad Sci U S A *101*, 7572-7577.

Marchler-Bauer, A., Lu, S., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R.*, et al.* (2011). CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res *39*, D225-229.

Margeat, E., Kapanidis, A.N., Tinnefeld, P., Wang, Y., Mukhopadhyay, J., Ebright, R.H., and Weiss, S. (2006). Direct observation of abortive initiation and promoter escape within single immobilized transcription complexes. Biophysical journal *90*, 1419-1431.

Markert, A., Grimm, M., Martinez, J., Wiesner, J., Meyerhans, A., Meyuhas, O., Sickmann, A., and Fischer, U. (2008). The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes. EMBO reports *9*, 569-575.

Marshall, N.F., Peng, J., Xie, Z., and Price, D.H. (1996). Control of RNA polymerase II elongation potential by a novel carboxyl-terminal domain kinase. J Biol Chem *271*, 27176-27183.

Marshall, N.F., and Price, D.H. (1992). Control of formation of two distinct classes of RNA polymerase II elongation complexes. Mol Cell Biol *12*, 2078-2090.

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

Martinez-Rucobo, F.W., and Cramer, P. (2013). Structural basis of transcription elongation. Biochim Biophys Acta *1829*, 9-19.

Martinez-Rucobo, F.W., Kohler, R., van de Waterbeemd, M., Heck, A.J., Hemann, M., Herzog, F., Stark, H., and Cramer, P. (2015). Molecular Basis of Transcription-Coupled Pre-mRNA Capping. Mol Cell *58*, 1079-1089.

Martinez-Rucobo, F.W., Sainsbury, S., Cheung, A.C., and Cramer, P. (2011). Architecture of the RNA polymerase-Spt4/5 complex and basis of universal transcription processivity. Embo J *30*, 1302-1310.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C.*, et al.* (2008). Nucleosome organization in the Drosophila genome. Nature *453*, 358-362.

Mayer, A., di Iulio, J., Maleri, S., Eser, U., Vierstra, J., Reynolds, A., Sandstrom, R., Stamatoyannopoulos, J.A., and Churchman, L.S. (2015). Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. Cell *161*, 541-554.

Mayer, A., Schreieck, A., Lidschreiber, M., Leike, K., Martin, D.E., and Cramer, P. (2012). The spt5 C-terminal region recruits yeast 3' RNA cleavage factor I. Mol Cell Biol *32*, 1321-1331.

McChesney, P.A., Aiyar, S.E., Lee, O.J., Zaika, A., Moskaluk, C., Li, R., and El-Rifai, W. (2006). Cofactor of BRCA1: a novel transcription factor regulator in upper gastrointestinal adenocarcinomas. Cancer Res *66*, 1346-1353.

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

Missra, A., and Gilmour, D.S. (2010). Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. Proc Natl Acad Sci U S A *107*, 11301-11306.

Muhlbacher, W., Sainsbury, S., Hemann, M., Hantsche, M., Neyer, S., Herzog, F., and Cramer, P. (2014). Conserved architecture of the core RNA polymerase II initiation complex. Nat Commun *5*, 4310.

Munoz, M.J., de la Mata, M., and Kornblihtt, A.R. (2010). The carboxy terminal domain of RNA polymerase II and alternative splicing. Trends in biochemical sciences *35*, 497-504.

Muse, G.W., Gilchrist, D.A., Nechaev, S., Shah, R., Parker, J.S., Grissom, S.F., Zeitlinger, J., and Adelman, K. (2007). RNA polymerase is poised for activation across the genome. Nature genetics *39*, 1507-1511.

Nag, A., Narsinh, K., and Martinson, H.G. (2007). The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. Nat Struct Mol Biol *14*, 662-669.

Napolitano, G., Lania, L., and Majello, B. (2014). RNA polymerase II CTD modifications: how many tales from a single tail. Journal of cellular physiology *229*, 538-544.

Narita, T., Yamaguchi, Y., Yano, K., Sugimoto, S., Chanarat, S., Wada, T., Kim, D.K., Hasegawa, J., Omori, M., Inukai, N.*, et al.* (2003). Human transcription elongation factor NELF: identification of novel subunits and reconstitution of the functionally active complex. Mol Cell Biol *23*, 1863-1873.

Narita, T., Yung, T.M., Yamamoto, J., Tsuboi, Y., Tanabe, H., Tanaka, K., Yamaguchi, Y., and Handa, H. (2007). NELF interacts with CBC and participates in 3' end processing of replication-dependent histone mRNAs. Mol Cell *26*, 349-365.

Natarajan, M., Schiralli Lester, G.M., Lee, C., Missra, A., Wasserman, G.A., Steffen, M., Gilmour, D.S., and Henderson, A.J. (2013). Negative elongation factor (NELF) coordinates RNA polymerase II pausing, premature termination, and chromatin remodeling to regulate HIV transcription. J Biol Chem *288*, 25995-26003.

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

Nechaev, S., Fargo, D.C., dos Santos, G., Liu, L., Gao, Y., and Adelman, K. (2010). Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science *327*, 335-338.

O'Brien, S.K., Cao, H., Nathans, R., Ali, A., and Rana, T.M. (2010). P-TEFb kinase complex phosphorylates histone H1 to regulate expression of cellular and HIV-1 genes. J Biol Chem *285*, 29713-29720.

Oellerich, T., Gronborg, M., Neumann, K., Hsiao, H.H., Urlaub, H., and Wienands, J. (2009). SLP-65 phosphorylation dynamics reveals a functional basis for signal integration by receptor-proximal adaptor proteins. Molecular & cellular proteomics : MCP *8*, 1738-1750.

Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. Genes Dev *10*, 2657-2683.

Ott, M., Geyer, M., and Zhou, Q. (2011). The control of HIV transcription: keeping RNA polymerase II on track. Cell Host Microbe *10*, 426-435.

Oven, I., Brdickova, N., Kohoutek, J., Vaupotic, T., Narat, M., and Peterlin, B.M. (2007). AIRE recruits P-TEFb for transcriptional elongation of target genes in medullary thymic epithelial cells. Mol Cell Biol *27*, 8815-8823.

Pagano, J.M., Kwak, H., Waters, C.T., Sprouse, R.O., White, B.S., Ozer, A., Szeto, K., Shalloway, D., Craighead, H.G., and Lis, J.T. (2014). Defining NELF-E RNA binding in HIV-1 and promoter-proximal pause regions. PLoS Genet *10*, e1004090.

Palangat, M., Renner, D.B., Price, D.H., and Landick, R. (2005). A negative elongation factor for human RNA polymerase II inhibits the anti-arrest transcript-cleavage factor TFIIS. Proc Natl Acad Sci U S A *102*, 15036-15041.

Palermo, R.D., Webb, H.M., and West, M.J. (2011). RNA polymerase II stalling promotes nucleosome occlusion and pTEFb recruitment to drive immortalization by Epstein-Barr virus. PLoS Pathog *7*, e1002334.

Perales, R., and Bentley, D. (2009). "Cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. Mol Cell *36*, 178-191.

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

Plaschka, C., Lariviere, L., Wenzeck, L., Seizl, M., Hemann, M., Tegunov, D., Petrotchenko, E.V., Borchers, C.H., Baumeister, W., Herzog, F.*, et al.* (2015). Architecture of the RNA polymerase II-Mediator core initiation complex. Nature *518*, 376-380.

Plet, A., Eick, D., and Blanchard, J.M. (1995). Elongation and premature termination of transcripts initiated from c-fos and c-myc promoters show dissimilar patterns. Oncogene *10*, 319-328.

Price, D.H. (2000). P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II. Mol Cell Biol *20*, 2629-2634.

Proudfoot, N.J., Furger, A., and Dye, M.J. (2002). Integrating mRNA processing with transcription. Cell *108*, 501-512.

Qiu, H., Hu, C., Wong, C.M., and Hinnebusch, A.G. (2006). The Spt4p subunit of yeast DSIF stimulates association of the Paf1 complex with elongating RNA polymerase II. Mol Cell Biol *26*, 3135-3148.

Rao, J.N., Neumann, L., Wenzel, S., Schweimer, K., Rosch, P., and Wohrl, B.M. (2006). Structural studies on the RNA-recognition motif of NELF E, a cellular negative transcription elongation factor involved in the regulation of HIV transcription. Biochem J *400*, 449-456.

Rao, J.N., Schweimer, K., Wenzel, S., Wohrl, B.M., and Rosch, P. (2008). NELF-E RRM undergoes major structural changes in flexible protein regions on target RNA binding. Biochemistry *47*, 3756-3761.

Rasmussen, E.B., and Lis, J.T. (1993). In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. Proc Natl Acad Sci U S A *90*, 7923-7927.

Renner, D.B., Yamaguchi, Y., Wada, T., Handa, H., and Price, D.H. (2001). A highly purified RNA polymerase II elongation control system. J Biol Chem *276*, 42601-42609.

Roeder, R.G., and Rutter, W.J. (1970). Specific nucleolar and nucleoplasmic RNA polymerases. Proc Natl Acad Sci U S A *65*, 675-682.

Rohr, O., Marban, C., Aunis, D., and Schaeffer, E. (2003). Regulation of HIV-1 gene transcription: from lymphocytes to microglial cells. Journal of leukocyte biology *74*, 736-749.

Rougvie, A.E., and Lis, J.T. (1988). The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. Cell *54*, 795-804.

Rougvie, A.E., and Lis, J.T. (1990). Postinitiation transcriptional control in Drosophila melanogaster. Mol Cell Biol *10*, 6041-6045.

Sadowski, C.L., Henry, R.W., Lobo, S.M., and Hernandez, N. (1993). Targeting TBP to a non-TATA box cis-regulatory element: a TBP-containing complex activates transcription from snRNA promoters through the PSE. Genes Dev *7*, 1535-1548.

Sainsbury, S., Bernecky, C., and Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. Nat Rev Mol Cell Biol *16*, 129-143.

Sainsbury, S., Niesser, J., and Cramer, P. (2013). Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. Nature *493*, 437-440.

Saunders, A., Core, L.J., and Lis, J.T. (2006). Breaking barriers to transcription elongation. Nat Rev Mol Cell Biol *7*, 557-567.

Schneider, E.E., Albert, T., Wolf, D.A., and Eick, D. (1999). Regulation of c-myc and immunoglobulin kappa gene transcription by promoter-proximal pausing of RNA polymerase II. Curr Top Microbiol Immunol *246*, 225-231.

Schones, D.E., Cui, K., Cuddapah, S., Roh, T.Y., Barski, A., Wang, Z., Wei, G., and Zhao, K. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887-898.

Schulze-Gahmen, U., Lu, H., Zhou, Q., and Alber, T. (2014). AFF4 binding to Tat-P-TEFb indirectly stimulates TAR recognition of super elongation complexes at the HIV promoter. Elife *3*, e02375.

Schulze-Gahmen, U., Upton, H., Birnberg, A., Bao, K., Chou, S., Krogan, N.J., Zhou, Q., and Alber, T. (2013). The AFF4 scaffold binds human P-TEFb adjacent to HIV Tat. Elife *2*, e00327.

Sedore, S.C., Byers, S.A., Biglione, S., Price, J.P., Maury, W.J., and Price, D.H. (2007). Manipulation of P-TEFb control machinery by HIV: recruitment of P-TEFb from the large form by Tat and binding of HEXIM1 to TAR. Nucleic Acids Res *35*, 4347-4358.

Shandilya, J., and Roberts, S.G. (2012). The transcription cycle in eukaryotes: from productive initiation to RNA polymerase II recycling. Biochim Biophys Acta *1819*, 391-400.

Sheldrick, G.M. (2008). A short history of SHELX. Acta crystallographica Section A, Foundations of crystallography *64*, 112-122.

Shilatifard, A. (2006). Chromatin modifications by methylation and ubiquitination: implications in the regulation of gene expression. Annu Rev Biochem *75*, 243-269.

Shuman, S. (2001). Structure, mechanism, and evolution of the mRNA capping apparatus. Progress in nucleic acid research and molecular biology *66*, 1-40.

Sikorski, T.W., and Buratowski, S. (2009). The basal initiation machinery: beyond the general transcription factors. Current opinion in cell biology *21*, 344-351.

Siliciano, R.F., and Greene, W.C. (2011). HIV latency. Cold Spring Harbor perspectives in medicine *1*, a007096.

Sobhian, B., Laguette, N., Yatim, A., Nakamura, M., Levy, Y., Kiernan, R., and Benkirane, M. (2010). HIV-1 Tat assembles a multifunctional transcription elongation complex and stably associates with the 7SK snRNP. Mol Cell *38*, 439-451.

Soding, J. (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics *21*, 951-960.

Squazzo, S.L., Costa, P.J., Lindstrom, D.L., Kumer, K.E., Simic, R., Jennings, J.L., Link, A.J., Arndt, K.M., and Hartzog, G.A. (2002). The Paf1 complex physically and functionally associates with transcription elongation factors in vivo. Embo J *21*, 1764-1774.

St Amour, C.V., Sanso, M., Bosken, C.A., Lee, K.M., Larochelle, S., Zhang, C., Shokat, K.M., Geyer, M., and Fisher, R.P. (2012). Separate domains of fission yeast Cdk9 (P-TEFb) are required for capping enzyme recruitment and primed (Ser7-phosphorylated) Rpb1 carboxyl-terminal domain substrate recognition. Mol Cell Biol *32*, 2372-2383.

Sun, J., Watkins, G., Blair, A.L., Moskaluk, C., Ghosh, S., Jiang, W.G., and Li, R. (2008). Deregulation of cofactor of BRCA1 expression in breast cancer cells. J Cell Biochem *103*, 1798-1807.

Sydow, J.F., Brueckner, F., Cheung, A.C., Damsma, G.E., Dengl, S., Lehmann, E., Vassylyev, D., and Cramer, P. (2009). Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. Mol Cell *34*, 710-721.

Tahirov, T.H., Babayeva, N.D., Varzavand, K., Cooper, J.J., Sedore, S.C., and Price, D.H. (2010). Crystal structure of HIV-1 Tat complexed with human P-TEFb. Nature *465*, 747-751.

Tarun, S.Z., Jr., and Sachs, A.B. (1996). Association of the yeast poly(A) tail binding protein with translation initiation factor eIF-4G. Embo J *15*, 7168-7177.

te Velthuis, A.J. (2014). Common and unique features of viral RNA-dependent polymerases. Cellular and molecular life sciences : CMLS *71*, 4403-4420.

Thomas, M.C., and Chiang, C.M. (2006). The general transcription machinery and general cofactors. Critical reviews in biochemistry and molecular biology *41*, 105-178.

Toth, Z., Brulois, K.F., Wong, L.Y., Lee, H.R., Chung, B., and Jung, J.U. (2012). Negative elongation factor-mediated suppression of RNA polymerase II elongation of Kaposi's sarcoma-associated herpesvirus lytic gene expression. J Virol *86*, 9696-9707.

Tsai, F.T., and Sigler, P.B. (2000). Structural basis of preinitiation complex assembly on human pol II promoters. Embo J *19*, 25-36.

Vagin, A.A., Steiner, R.A., Lebedev, A.A., Potterton, L., McNicholas, S., Long, F., and Murshudov, G.N. (2004). REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. Acta Crystallogr D *60*, 2184-2195.

Valouev, A., Johnson, S.M., Boyd, S.D., Smith, C.L., Fire, A.Z., and Sidow, A. (2011). Determinants of nucleosome organization in primary human cells. Nature *474*, 516-520.

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

Venters, B.J., and Pugh, B.F. (2009). A canonical promoter organization of the transcription machinery and its regulators in the Saccharomyces genome. Genome Res *19*, 360-371.

Vvedenskaya, I.O., Vahedian-Movahed, H., Bird, J.G., Knoblauch, J.G., Goldman, S.R., Zhang, Y., Ebright, R.H., and Nickels, B.E. (2014). Interactions between RNA polymerase and the "core recognition element" counteract pausing. Science *344*, 1285-1289.

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G.A., Winston, F.*, et al.* (1998). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. Genes Dev *12*, 343-356.

Walzthoeni, T., Claassen, M., Leitner, A., Herzog, F., Bohn, S., Forster, F., Beck, M., and Aebersold, R. (2012). False discovery rate estimation for cross-linked peptides identified by mass spectrometry. Nat Methods *9*, 901-903.

Wang, D., Severinov, K., and Landick, R. (1997). Preferential interaction of the his pause RNA hairpin with RNA polymerase beta subunit residues 904-950 correlates with strong transcriptional pausing. Proc Natl Acad Sci U S A *94*, 8433-8438.

Wang, X., Hang, S., Prazak, L., and Gergen, J.P. (2010). NELF potentiates gene transcription in the Drosophila embryo. PLoS One *5*, e11498.

Watanabe, T., Hayashi, K., Tanaka, A., Furumoto, T., Hanaoka, F., and Ohkuma, Y. (2003). The carboxy terminus of the small subunit of TFIIE regulates the transition from transcription initiation to elongation by RNA polymerase II. Mol Cell Biol *23*, 2914-2926.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Weinzierl, R.O. (2011). The Bridge Helix of RNA polymerase acts as a central nanomechanical switchboard for coordinating catalysis and substrate movement. Archaea (Vancouver, BC) *2011*, 608385.

Wen, Y., and Shatkin, A.J. (1999). Transcription elongation factor hSPT5 stimulates mRNA capping. Genes Dev *13*, 1774-1779.

Wenzel, S., Martins, B.M., Rosch, P., and Wohrl, B.M. (2010). Crystal structure of the human transcription elongation factor DSIF hSpt4 subunit in complex with the hSpt5 dimerization interface. Biochem J *425*, 373-380.

Werner, F. (2007). Structure and function of archaeal RNA polymerases. Mol Microbiol *65*, 1395-1404.

Werner, F. (2012). A nexus for gene expression-molecular mechanisms of Spt5 and NusG in the three domains of life. J Mol Biol *417*, 13-27.

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

White, R.J., and Jackson, S.P. (1992). Mechanism of TATA-binding protein recruitment to a TATA-less class III promoter. Cell *71*, 1041-1053.

Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C., Williams, K.L., Appel, R.D., and Hochstrasser, D.F. (1999). Protein identification and analysis tools in the ExPASy server. Methods in molecular biology (Clifton, NJ) *112*, 531-552.

Williams, L.H., Fromm, G., Gokey, N.G., Henriques, T., Muse, G.W., Burkholder, A., Fargo, D.C., Hu, G., and Adelman, K. (2015). Pausing of RNA polymerase II regulates mammalian developmental potential through control of signaling networks. Mol Cell *58*, 311-322.

Wright, T.J., Costa, J.L., Naranjo, C., Francis-West, P., and Altherr, M.R. (1999). Comparative analysis of a novel gene from the Wolf-Hirschhorn/Pitt-Rogers-Danks syndrome critical region. Genomics *59*, 203-212.

Wu, C.H., Lee, C., Fan, R., Smith, M.J., Yamaguchi, Y., Handa, H., and Gilmour, D.S. (2005). Molecular characterization of Drosophila NELF. Nucleic Acids Res *33*, 1269-1279.

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

Xiang, K., Nagaike, T., Xiang, S., Kilic, T., Beh, M.M., Manley, J.L., and Tong, L. (2010). Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. Nature *467*, 729-733.

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

Yamaguchi, Y., Filipovska, J., Yano, K., Furuya, A., Inukai, N., Narita, T., Wada, T., Sugimoto, S., Konarska, M.M., and Handa, H. (2001). Stimulation of RNA polymerase II elongation by hepatitis delta antigen. Science *293*, 124-127.

Yamaguchi, Y., Inukai, N., Narita, T., Wada, T., and Handa, H. (2002). Evidence that negative elongation factor represses transcription elongation through binding to a DRB sensitivity-inducing factor/RNA polymerase II complex and RNA. Mol Cell Biol *22*, 2918-2927.

Yamaguchi, Y., Mura, T., Chanarat, S., Okamoto, S., and Handa, H. (2007). Hepatitis delta antigen binds to the clamp of RNA polymerase II and affects transcriptional fidelity. Genes Cells *12*, 863-875.

Yamaguchi, Y., Shibata, H., and Handa, H. (2013). Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond. Biochim Biophys Acta *1829*, 98-104.

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

Yamaguchi, Y., Wada, T., Watanabe, D., Takagi, T., Hasegawa, J., and Handa, H. (1999b). Structure and function of the human transcription elongation factor DSIF. J Biol Chem *274*, 8085-8092.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nat Methods *12*, 7-8.

Yang, Z., Yik, J.H., Chen, R., He, N., Jang, M.K., Ozato, K., and Zhou, Q. (2005). Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. Mol Cell *19*, 535-545.

Ye, Q., Hu, Y.F., Zhong, H., Nye, A.C., Belmont, A.S., and Li, R. (2001). BRCA1-induced large-scale chromatin unfolding and allele-specific effects of cancer-predisposing mutations. The Journal of cell biology *155*, 911-921.

Yik, J.H., Chen, R., Nishimura, R., Jennings, J.L., Link, A.J., and Zhou, Q. (2003). Inhibition of P-TEFb (CDK9/Cyclin T) kinase and RNA polymerase II transcription by the coordinated actions of HEXIM1 and 7SK snRNA. Mol Cell *12*, 971-982.

Yudkovsky, N., Ranish, J.A., and Hahn, S. (2000). A transcription reinitiation intermediate that is stabilized by activator. Nature *408*, 225-229.

Yue, Z., Maldonado, E., Pillutla, R., Cho, H., Reinberg, D., and Shatkin, A.J. (1997). Mammalian capping enzyme complements mutant Saccharomyces cerevisiae lacking mRNA guanylyltransferase and selectively binds the elongating form of RNA polymerase II. Proc Natl Acad Sci U S A *94*, 12898-12903.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nature genetics *39*, 1512-1516.

Zhang, Z., Klatt, A., Gilmour, D.S., and Henderson, A.J. (2007). Negative elongation factor NELF represses human immunodeficiency virus transcription by pausing the RNA polymerase II complex. J Biol Chem *282*, 16981-16988.

Zhou, H., Liu, Q., Gao, Y., Teng, M., and Niu, L. (2009a). Crystal structure of NusG N-terminal (NGN) domain from Methanocaldococcus jannaschii and its interaction with rpoE". Proteins *76*, 787-793.

Zhou, K., Kuo, W.H., Fillingham, J., and Greenblatt, J.F. (2009b). Control of transcriptional elongation and cotranscriptional histone modification by the yeast BUR kinase substrate Spt5. Proc Natl Acad Sci U S A *106*, 6956-6961.

Zhou, Q., and Yik, J.H. (2006). The Yin and Yang of P-TEFb regulation: implications for human immunodeficiency virus gene expression and global control of cell growth and differentiation. Microbiology and molecular biology reviews : MMBR *70*, 646-659.

Zhu, Y., Pe'ery, T., Peng, J., Ramanathan, Y., Marshall, N., Marshall, T., Amendt, B., Mathews, M.B., and Price, D.H. (1997). Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. Genes Dev *11*, 2622-2632.

List of Figures

List of Tables

Abbrevations

5' 6-FAM	6-carboxyfluorescein covalently linked to the 5'-end
°C	degree celcius
Å	Ångstrom
µg	microgramm
µl	microliter
ε	molar attenuation coefficient
aa	amino acid
Amp	ampicillin
bp(s)	base pair(s)
BSA	bovine serum albumine
Ca	calcium
CBC	cap binding complex
CDK	Cyclin-dependent kinase
cDNA	complementary DNA
Cl	chloride
CID	CTD interacting domain
CPSF	cleavage and polyadenylation specificity factor
comp	complementary
CstF	cleavage stimulation factor
CTD	C-terminal domain of pol II subunit rpb1
CTR	C-terminal region of Spt5
CV	column volumes
Da	dalton
ddH_2O	bidest water
D.me.	*Drosophila melanogaster*
DMF	dimethylformamide
DMSO	dimethylsulfoxid
DNA	deoxyribonucleic acid
DLS	dynamic light scattering
DRB	5,6-Dichloro-1-β-D-ribofuranosylbenzimidazole
dsDNA	double-stranded DNA
DSIF	DRB sensitivity inducing factor
DSS	disuccinimidyl suberate
DTT	dithiothreitol

E. coli	*Escherichia coli*
EDTA	ethylenediaminetetraacetic acid
EtOH	ethanol
fl	full length
fwd	forward
GTF	general TF
H.s.	*Homo sapiens*
H.sapiens.	*Homo sapiens*
HCl	hydrochloric acid
HDAg	hepatitis delta antigen
HEPES	2-[4-(2-hydroxyethyl)-1-piperazinyl]-ethanesulfonic acid
HIV	human immumodeficiency virus
h/hrs	hour/hours
hsf	heat shock factor
IPTG	isopropyl β-D-1-thiogalactopyranoside
IRES	internal ribosome entry site
Kacetate	potassium acetate
Kan	kanamycine
kDa	kilodalton
KOW	Kyrpides-Ouzounis-Woese
l	liter
LB	lysogeny broth
LIC	ligation independent cloning
LTR	long terminal repeat
Met	methionine
M	molar (mol/liter)
mAU	milli absorption units
MALDI-MS	matrix-assisted laser desorption/ionization and mass spectrometry
me	methyl
min	minute
ml	milliliter
mM	millimolar
Mn	manganese
MOPS	3-(N-morpholino)propanesulfonic acid

mRNA	messenger RNA
miRNA	micro RNA
mut	mutated
MW	molecular weight
MWCO	molecular weight cut-off
Na	sodium
NCBI	National Center for Biotechnology Information
NEB	New England Biolabs
NELF	negative elongation factor
NFR	nucleosome free region
N-His$_6$	N-terminal hexa-histidine tag
Ni-NTA	nickel-nitrilotriacetic acid
ng	nanogramm
nl	nanoliter
nM	nanomolar
nt	nucleotide
OD$_{600}$	optical density at 600 nm
ON	over night
P-TEFb	positive transcription elongation factor
PA	polyacrylamide
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PEG	polyethylene glycol
PI	protease inhibitor
PIC	pre-initiation complex
PMSF	phenylmethylsulfonylfluorid
Pol	DNA-dependent RNA polymerase
ppp	promoter-proximal pausing
PVDF	polyvinylidenfluorid
Rb	rubidium
rev	reverse
RNA	ribonucleic acid
rpm	revolutions per minute
rRNA	ribosomal RNA
RRM	RNA recognition motif

RSD	routed-square deviation
RT	room temperature
sec/secs	second/seconds
SeMet	selenomethionine
SIRAS	single isomorphous replacement with anomalous scattering
SDS	sodium dodecyl sulphate
SELEX	systematic evoluation of ligands by exponential enrichment
SLS	Swiss Light Source
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleic protein
ssDNA	single-stranded DNA
ssRNA	single-stranded RNA
TAE	tris-acetate-EDTA
TAR	trans-activation response element
Taq	*Thermus aquaticus*
TBP	TATA-binding protein
TCA	trichloroacetic acid
TE	tris-EDTA
Temp	temperature
TFII	transcription factor of pol II
TFB-I/II	transformation buffer I/II
Tris	tris(hydroxymethyl)aminomethane
tRNA	translator RNA
TSS	transcription start site
v/v	volume per volume
WT	wildtype
w/v	weight per volume
X-Gal	5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside
ZfP	Zentrallabor für Proteinanalytik

Nucleotides

A	adenine
C	cytosine
G	guanine
T	thymine
U	uracil

Amino acids

A	alanine
D	aspartate
His	histidine
K	lysine
Q	glutamine
R	arginine
S	serine
T	threonine
Y	tyrosine