Reconstitution of the Integrator complex and its interaction with paused transcription elongation complex of Pol II-DSIF-NELF

Dissertation

for the award of the degree

“Doctor rerum naturalium”

of the Georg-August-Universität Göttingen

in the doctoral program IMPRS Molecular Biology

of the Georg August University School of Science (GAUSS)

Göttingen, Germany

submitted by

Isaac Fianu

born in

Mobole, Ghana

Göttingen, April 2020

Thesis advisory committee

Prof. Dr. Patrick Cramer (Supervisor)	Department of Molecular Biology Max Planck Institute for Biophysical Chemistry, Göttingen
Prof. Dr. Kai Tittmann	Department of Molecular Enzymology Albrecht-von-Haller-Institute for Plant Sciences, University of Göttingen
Prof. Dr. Dirk Görlich	Department of Cellular Logistics Max Planck Institute for Biophysical Chemistry, Göttingen

Members of examination board

Prof. Dr. Patrick Cramer (1st reviewer)	Department of Molecular Biology Max Planck Institute for Biophysical Chemistry, Göttingen
Prof. Dr. Kai Tittmann (2nd reviewer)	Department of Molecular Enzymology Albrecht-von-Haller-Institute for Plant Sciences, University of Göttingen

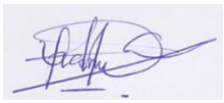
Extended members of examination board

Prof. Dr. Dirk Görlich	Department of Cellular Logistics Max Planck Institute for Biophysical Chemistry, Göttingen
Prof. Dr. Henning Urlaub	Research Group Bioanalytical Mass Spectrometry Max Planck Institute for Biophysical Chemistry, Göttingen
Dr. Alex Faesen	Research Group Biochemistry of Signal Dynamics Max Planck Institute for Biophysical Chemistry, Göttingen
Dr. Juliane Liepe	Research Group Quantitative and Systems Biology Max Planck Institute for Biophysical Chemistry, Göttingen

Proposed date of oral examination: 9th June, 2020

Affidavit

I, Isaac Fianu, hereby declare that my dissertation entitled ‘Reconstitution of the Integrator complex and its interaction with paused transcription elongation complex of Pol II-DSIF-NELF’ has been written independently and with no other sources and aids than quoted. This dissertation or parts thereof have not been submitted elsewhere for any academic award or qualification. The electronic version of this dissertation is congruent to the printed versions.



22nd April 2020 .....

Isaac Fianu

Acknowledgements

I wish to thank my supervisor, Prof. Dr. Patrick Cramer for the great opportunity to work with him on a very interesting and challenging project. For his believe in me through the period of my PhD thesis, for his guidance and finally for the great resources he provided for learning and research. Patrick remained interested in my work and provided very insightful ideas on how to approach the project. He also gave great advice on my personal development, time management and career choices.

I also want to thank my thesis advisory committee for their time and very useful inputs throughout the years. Their critical assessment of my reports thought me how to judiciously assess my own results and the results of others as a scientist. Furthermore, their kind words especially Kai, provided some momentum in times when I was discouraged and doubting my potential.

I am highly indebted to my collaborators on crosslinking mass spectrometry, Andreas Linden and Prof. Dr. Henning Urlaub who were instrumental to the progress of this project. Monika Raabe, Sabrine Köning Uwe Plessman and Annika Reinelt were extremely helpful in the mass spectrometric identification of proteins without which this project would not have advanced.

I am also grateful to Dr. Tino Pleiner and Prof. Dr. Dirk Görlich for efforts in attempted anti Integrator complex nanobody production.

I must extend sincere gratitude to my colleagues past and present who have helped me in many ways during my PhD. For creating a conducive working atmosphere, for open discussions and help with various equipments especially at the beginning of my PhD. Special thanks to members of Lab 213 and 214 especially Xiangyang Liu whom I shared workspace with.

I will also like to thank Dr. Seychelle Vos for sharing protocols, plasmids and very fruitful discussion on transcription elongation, pausing and the Integrator complex. Dr. Christian Dienemann for introduction to electron microscopy. Jürgen Wawrzinek and Ulrich Steuerwald for help with screening of crystallization conditions.

Furthermore, my sincere gratitude goes to Dr. Kristina Zumer, Dr. Livia Caizzi for introduction to human cell culture and fruitful discussions about transcription termination and the NELF complex. Dr. Annia Sawicka for sharing experience on cloning of challenging targets, Dr. Simon Neyer for introduction to Aektas and Franz Fischer for introduction to insect cell culture

I am grateful to Dr. Paulina Seweryn, Dr. Goran Kokic, Dr. Marc Böhning, Dr. Jana Schmitzova, Dr. Christian Dienemann, and Dr. Felix Wagner for critically proofreading parts of this thesis.

I will like to thank Dr. Sandra Schildbach, Dr. Suyang Zhang, Dr. Shintaro Aibara and Dr. Lucas Farnung for help with cryo-EM data processing and sample preparation discussions.
This work will be impossible without the relentless and excellent administrative and technical support of Kersten Maier, Kirsten Backs, Ute Neef, Petra Rus, Janine Blümel, Thomas Schulz, Mario Klein, Angelika Kruse and Manuela Wenzel. I say thank you very much.
I wish to express profound gratitude to Dr. Steffen Burkhadt and Kerstin Gruninger of the IMPRS Molecular Biology coordination office for great help with getting to start life in Göttingen and great support throughout my Masters and PhD.
During the period of my PhD, I met many great people some of which have become friends. They were of utmost support even with person problems outside work. I will like to thank Dr. S. Osman, Dr. F. Wagner Frau Dr. P. Seweryn, Dr. Marc Böhning Dr. King Faisal Yambire, Kingsley Kumashie and Julia Dziubek for their warm company during this time and for their help especially during my surgeries.
Finally, I am very thankful to my family back home who have been with me through the thick and thin of all my quests including this PhD. Their kind words and sometimes their problems provided motivation to keep going.

To Gladys Boni (Nanaa),
your Love, Kindness, and Altruism will never be forgotten

Table of Contents

Contents

Abstract

The Integrator complex (INT) is the latest and largest multi-subunit protein complex to be added to the list of factors involved in RNA Polymerase II (Pol II) transcription. INT consist of 15 subunits with an estimated molecular weight of 1.5 MDa. It is recruited to Pol II during initiation or early elongation of transcription. It plays important roles in transcription regulation during early elongation including premature termination of some messenger RNAs (mRNAs). It has also been shown to terminate small nuclear RNAs (snRNAs), enhancer RNAs and some viral RNAs. Despite the important roles of INT in transcription, there is currently no protocol to reconstitute INT for *in vitro* biochemical and structural studies. Here, I used the baculovirus and insect cell recombinant protein expression system to reconstitute INT. Most subunits of INT express poorly and form oligomers when purified independently and it was not possible to co-express all 15 subunits. I therefore divided INT into 4 subcomplexes based on identified inter-subunit interactions. Namely, a 7-subunits core-INT, a 3-subunit cleavage module (CM), a 3-subunit cleavage module interacting module (CMIM) and INTS1/12 heterodimer. Subsequently I established purification protocols for all these subcomplexes or their stable interacting domains as in the case of INTS1/12 heterodimer. Negative stain EM reveals that the core-INT (INTS2/3/5/6/7/8-DDX26B) is monomeric showing that the oligomerization of the constituent subunits is circumvented in the core-INT. Further cryo-EM reconstruction of core-INT reveals a low-resolution doughnut shape for this subcomplex. The CM (INTS4/9/11) has a trilobal shape as revealed by a cryo-EM reconstruction at ~20 Å.

For the first time, I reconstituted the full INT *in vitro* from the purified subcomplexes by amylose affinity pulldown. The interactions between the subcomplexes were assessed by crosslinking mass spectrometry (XL-MS). XL-MS reveals many crosslinks between core-INT, CMIM, and INTS1/12 heterodimer while the CM has very few crosslinks to the other subcomplexes. This result suggests that the CM may be a loosely associated module of INT.

Furthermore, I showed for the first time the *in vitro* binding of INT to the paused elongation complex (PEC) of Pol II, DSIF (DRB sensitivity-inducing factor) and NELF (negative elongation factor). I characterized inter complex protein-protein interactions between INT and PEC by XL-MS. It emerges that INT interacts mostly with NELF and Pol II but not DSIF according to the crosslinks observed. The INTS1, INTS6, and INTS12 subunits of INT appear to be involved in most of these interactions with PEC. This work has created the foundation for biochemical and structural characterization of INT as a complex and INT in the context of Pol II transcription regulation.

1 Introduction

1.1 General overview of transcription and RNA polymerases

Information pertaining to the phenotype of all organisms is stored in their deoxyribonucleic acid (DNA) macromolecule(s). The flow of this information from DNA to proteins via a ribonucleic acid (RNA) intermediate is simplified in the so-called central dogma of molecular biology coined by Francis Crick (Crick, 1970). The process of copying the information stored in DNA to RNA is called transcription and it is done by DNA-dependent RNA polymerases.

The simplest RNA polymerase is a single subunit enzyme encoded by some viruses and phages (Jeruzalmi & Steitz, 1998).

In prokaryotes, a single RNA polymerase (RNAP: used here for prokaryotic RNA polymerase only) composed of four polypeptides transcribes all genes (Hurwitz, 2005; Hurwitz et al., 1960; Stevens, 1960). In contrast to prokaryotes, eukaryotes employ three main RNA polymerases (Pols: for eukaryotic RNA polymerases) to transcribe their nuclear genomes namely Pol I, Pol II and Pol III (Roeder & Rutter, 1969). Plants have additional nuclear Pols, Pol IV and Pol V (Zhou & Law, 2015). Pol I has 14 subunits, Pol II has 12 subunits and Pol III has 17 subunits. The Pols share several subunits and some subunits are close homologues. (Cramer et al., 2000; Engel et al., 2013; Jasiak et al., 2006; Kuhn et al., 2007; Vannini & Cramer, 2012).

All the multi-subunit DNA-dependent RNA polymerases require additional factors for promoter DNA recognition and transcription initiation (Burgess et al., 1969; Sainsbury et al., 2015; Vannini & Cramer, 2012) (Table 1.1). The core structure of the three Pols is highly conserved from yeast to human but the Pols differ substantially in the transcription factors they require for RNA synthesis (Goodfellow & Zomerdijk, 2013) (Figure 1.1). The eukaryotic polymerases also share important structural features with the bacterial 4-subunit counterpart suggesting a conserved mechanism of catalysis (Allison et al., 1985; Cramer et al., 2001). The subunit composition of the three eukaryotic Pols in yeast is summarized in Table 1.1.

Soon after their discovery, it became clear that the three Pols have dedicated class of genes they transcribe. Pol I transcribes the precursor for the large ribosomal RNAs, Pol II transcribes all protein coding genes and some non-coding RNAs and Pol III produces transfer RNAs and 5S RNA (R. Weinmann et al., 1974; Roberto Weinmann & Roeder, 1974).

Table 1.1. Homology between the subunits of three Pols in yeast. Auxiliary factors important for transcription initiation by each Pol are also indicated. Table is modified from Vannini and Cramer, 2012.

Pol I	Pol II	Pol II	Function
Polymerase core			
A190	Rpb1	C160	Active center
AC135	Rpb2	C128	Active center
AC40	Rpb3	AC40	
Rpb5	Rpb5	Rpb5	
Rpb6	Rpb6	Rpb6	
Rpb8	Rpb8	Rpb8	
A12.2 N ribbon	Rpb9	C 11 N ribbon	RNA cleavage
Rpb10	Rpb10	Rpb10	
AC19	Rpb11	AC19	
Rpb12	Rpb12	Rpb12	
Polymerase stalk			
A14	Rpb4	C17	Initiation complex formation
A43	Rpb7	C25	Initiation complex formation
Transcription initiation factors in Pol II and their homologues in Pol I and III			
A49 and A34.5	TFIIF	C37 and C53	Initiation complex formation
A49	TFIIE	C31,34, 82	Open complex stabilization
TBP	TBP	TBP	Promoter recognition
	TAFs		Promoter recognition
Rrn7	TFIIB	Brf1	TBP and Pol binding

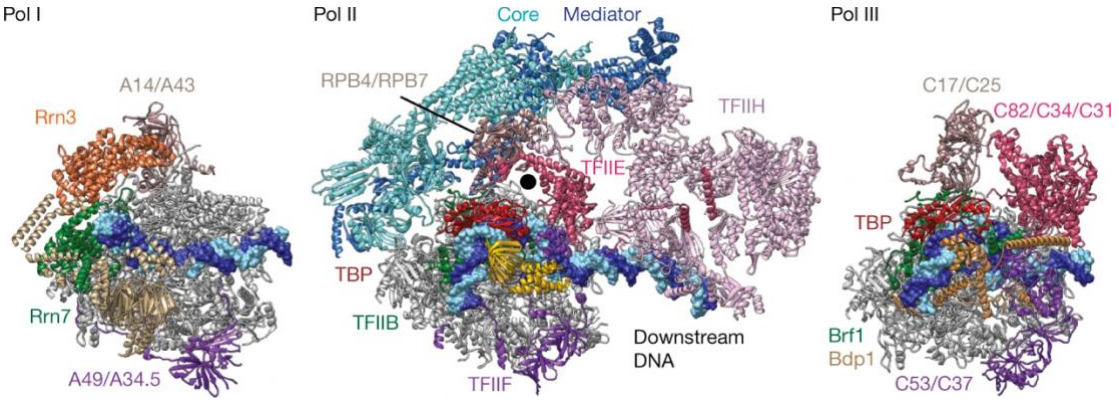


Figure 1.1. Structures of pre-initiation complexes of the three eukaryotic Pols. The conserved core of the Pols is depicted in grey. The transcription initiation factors decorating the Pols are indicated. The Pol II PIC has the coactivator complex, Mediator and the TFIIH complex which brings additional enzymatic activities and play regulatory functions. The figure was adapted from (Cramer, 2019).

1.2 Pol II transcription cycle

Pol II transcribes all protein-coding genes and a class of non-coding RNAs (ncRNA) in eukaryotic organisms (R. Weinmann et al., 1974). The ncRNAs include small nucleolar RNAs (snoRNA) that are important for ribosomal RNA (rRNA) modification and small nuclear RNAs (snRNAs) that are used in spliceosome assembly. Regulation of Pol II transcription is central to many cellular and biological processes including cell differentiation, organismal development and growth (Cramer, 2019; Sainsbury et al., 2015). Gene Transcription by Pol II is divided into three main phases: initiation, elongation, termination and recycling of Pol II (Hantsche & Cramer, 2016) (Figure 1.2). Each of these steps involve many auxiliary factors and are tightly regulated to achieve constitutive and activated expression of Pol II-transcribed genes.

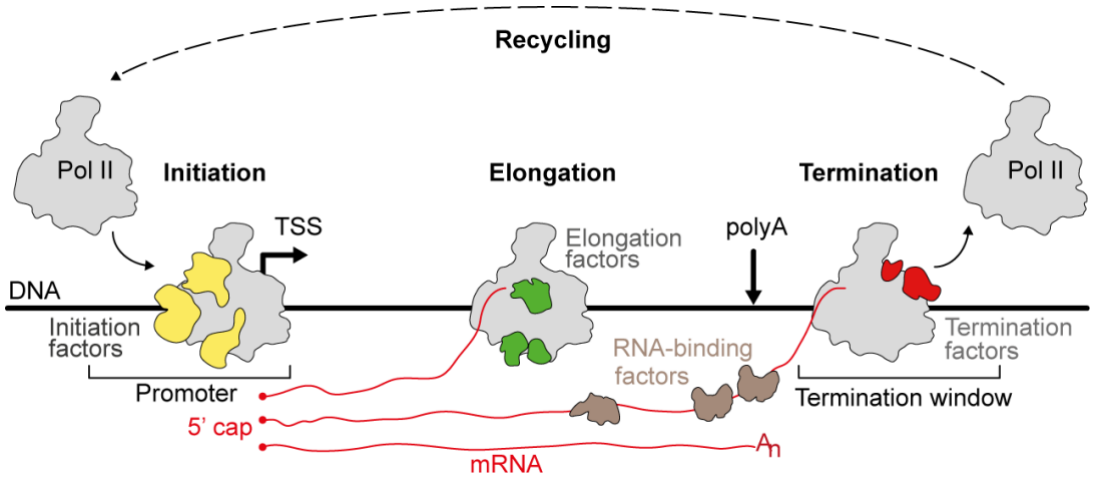


Figure 1.2. The Pol II transcription cycle. Pol II is shown in grey, initiation factors in yellow, elongation factors in green and termination factors in red. TSS stands for transcription start site. Figure is modified from Hantsche and Cramer, 2016.

1.2.1 Transcription initiation by Pol II

The DNA template at core promoters must be cleared of nucleosomes for Pol II to initiate transcription (Fuda et al., 2009; Lorch et al., 1987). This nucleosome barrier to transcription is removed by chromatin remodelers which create nucleosome-free DNA template at promoter regions (Lorch & Kornberg, 2017). Even in the absence of the nucleosome barrier, Pol II (like the other RNA polymerases) cannot initiate RNA synthesis by itself. This stimulated the hunt for Pol II initiation factors leading to the discovery of general transcription factors (GTFs) (Conaway & Conaway, 2019). In the classical view, Pol II assembles with the GTFs at nucleosome-free promoters in a sequential manner (Buratowski et al., 1989). This pioneering biochemical work paved way for detailed structural analysis of Pol II transcription initiation intermediates (Sainsbury et al., 2015). The TATA box binding protein (TBP) binds the minor groove of DNA at the promoter region and induces a 90 degrees bend (J. L. Kim et al., 1993). The mechanism of DNA binding by TBP confers no sequence specificity explaining how it is used by all three Pols in transcription initiation. The TBP-DNA complex is bound by the Pol II transcription factor IIA and IIB (TFIIA and TFIIB) and the Pol II-specific TBP-associated factors (TAFs). These factors stabilize the TBP-DNA complex by making contacts with both TBP and the promoter elements in the DNA and bestow specificity (Andel et al., 1999). TFIIB then recruits Pol II-TFIIF complex to form the core pre-initiation complex (cPIC). TFIIF is tightly associated with Pol II, is critical for promoter selectivity and stimulates RNA synthesis once the promoter DNA is opened (reviewed in Sainsbury et al., 2015). The 3D architecture of the core pre-initiation complex of TBP, TFIIA, TFIIB, Pol II and TFIIF is conserved between yeast and human (He et al., 2013; Mühlbacher et al., 2014; Sainsbury et al., 2015) and Figure 1.3. The transcription factor IIE (TFIIE) is recruited to the cPIC which in turn recruits the transcription factor IIH (TFIIH). The formation of an open initiation complex (OC) can happen spontaneously in some promoters (Dienemann et al., 2019) or requires ATPase activity of TFIIH (Schilbach et al., 2017) depending on physical properties of the DNA sequence such as meltability. The requirement of ATP for promoter melting is a major difference between Pol II initiation and other RNA polymerases.

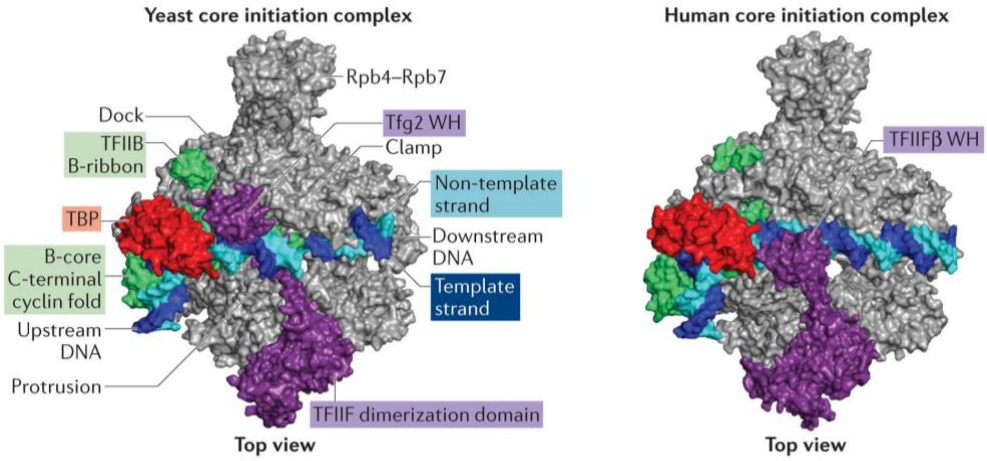


Figure 1.3. Structures of yeast and human core pre-initiation complexes. Shows the conservation of the architecture of the cPIC between these species. The factors are colored identically in each structure. Image was taken from Sainsbury et al., 2015.

The transcription co-activator complex Mediator, also regulates transcription initiation by Pol II (Kornberg, 2005). Mediator binds activating transcription factors bound to upstream enhancer elements of the promoter and at the same time interacts with TFIIB and TFIIH components of the Pol II initiation complex. This way it integrates upstream activation signals and transcription initiation to activate gene expression (Cramer, 2019). The Mediator complex and TFIIH have dissociable kinase modules, which phosphorylate the C-terminal domain (CTD) of the Pol II largest subunit, RPB1. The CTD of Pol II RPB1 is a conserved feature unique to Pol II. It consists of 26 (in yeast) and 52 (in human) repeats of the consensus sequence Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 (YSPTSPS). Residues of this repeat are subjected to reversible modifications such as phosphorylation (Tyr, Thr and Ser) and isomerization (Pro) throughout the transcription cycle. Phosphorylations are known to facilitate recruitment of factors to Pol II after initiation for regulation, co-transcriptional processing of the nascent RNA and termination (Dahmus, 1996; Eick & Geyer, 2013; Kwak & Lis, 2013; Shah et al., 2018; Zaborowska et al., 2016).

1.2.2 Pol II transcription elongation

The recruitment of TFIIE and TFIIH to the cPIC leads to promoter DNA melting and synthesis of RNA by Pol II (Sainsbury et al., 2015). TFIIH helicase XPB is furthermore required for promoter escape (Goodrich & Tjian, 1994; Moreland et al., 1999). After promoter escape, transcription is said to be in early elongation after the synthesis of about 12 nt long nascent RNA. The growing chain of newly synthesized RNA clashes with TFIIB leading to its displacement from the initiation complex (Sainsbury et al., 2013). The rate of elongation

observed in primer extension assays using purified Pol II is several folds lower than the *in vivo* synthesis rate of 100s of bases/min to an average of 2,300 bases/min (Conaway & Conaway, 2019; Gressel et al., 2017; Thummel et al., 1990). This suggests that, like the initiation phase, Pol II elongation requires additional factors to stimulate RNA synthesis. The requirement of transcription elongation and RNA processing factors as well as co-transcriptional processing of RNA makes the elongation phase of Pol II transcription an attractive regulatory step of gene expression (Kwak & Lis, 2013; Perales & Bentley, 2009). Several factors have been discovered that play a role in stimulating or repressing Pol II transcription. Transcription factor IIS (TFIIS) was the first Pol II elongation stimulating factor to be discovered (Sekimizu et al., 1976). TFIIS, like the GreA and GreB factors in the bacterial transcription system (Borukhov et al., 2005), stimulates endoribonuclease activity of Pol II when it is arrested and backtracked during transcription (Cheung & Cramer, 2011; Ehara et al., 2017; Izban & Luse, 1992; Reines et al., 1992). TFIIS-like subunits A12.2 and C11 in the Pol I and Pol III systems respectively stimulate RNA cleavage in these Pols (Kuhn et al., 2007; Riva et al., 1998; Vannini & Cramer, 2012).

Selective inhibition of transcription elongation by the purine nucleoside analog 5,6-dichloro-l-3-ribofuranosylbenzimidazole (DRB) (Chodosh et al., 1989) led to the discovery of the DRB sensitivity inducing factor (DSIF) (Wada et al., 1998). DSIF is a heterodimer of SPT4 and SPT5 and conserved from yeast to human (Swanson & Winston, 1992; Wada et al., 1998). DSIF binds to the upstream DNA and makes contacts with the exiting RNA while interacting with Pol II to enhance its processivity during elongation (Bernecky et al., 2017; Ehara et al., 2017).

A key regulatory feature of Pol II early elongation is promoter-proximal pausing (PPP) where Pol II accumulates about 12-65 nt downstream of the transcription start site (Gilmour & Lis, 1986). PPP is a conserved regulatory mechanism from *C. elegans* to human (but not in yeast) and have been observed in many protein-coding genes (Core & Adelman, 2019). The accumulation of paused Pol II prevents new initiation events. This makes PPP an effective way to repress transcription (Gressel et al., 2017). Many factors contribute in establishing PPP including DNA and RNA sequence as well as protein factors (Core & Adelman, 2019; Kwak & Lis, 2013). The negative elongation factor (NELF) composed of 4 polypeptides, NELF -A, -B, -C/-D and -E is important for establishing PPP (Narita et al., 2003; Yamaguchi et al., 1999). The NELF complex arrests mobile domains of Pol II necessary for escaping the paused state. It additionally induces tilting of the DNA-RNA hybrid in the Pol II active site into a non-canonic state which is not compatible with nucleotide addition. And sterically prevents binding of TFIIS thereby maintaining the PPP (Vos, Farnung, Urlaub, et al., 2018). NELF is a

metazoan-specific factor needed to regulated the metazoan-specific process of PPP (Narita et al., 2003).

Transition from the promoter-proximally paused state to processive elongation or activated transcription requires specific phosphorylation of the Pol II CTD (Core & Adelman, 2019; Dahmus, 1996; Kwak & Lis, 2013). The positive elongation factor b (P-TEFb) was identified as the main CTD kinase responsible for this transition and it is inhibited by DRB (Marshall & Price, 1995). Recent structural and biochemical studies shows that P-TEFb kinase, CDK9 phosphorylates the CTD of Pol II, the CTD linker domain, NELF and DSIF to stimulate activated transcription (Vos, Farnung, Boehning, et al., 2018; Vos, Farnung, Urlaub, et al., 2018). NELF phosphorylation weakens its interaction with Pol II and allows it to be competed off by the PAF complex. Also, phosphorylation of the CTD linker domain recruits the histone chaperone SPT6 to the activated elongation complex (Vos, Farnung, Boehning, et al., 2018). This study confirms a wide variety of *in vivo* studies that highlight P-TEFb, DSIF and NELF as key players of transcription regulation at the step of PPP (Core & Adelman, 2019; Kwak & Lis, 2013). Furthermore, additional factors such as histone methyltransferases SET1 and SET2 may be recruited to allow transcription through nucleosome bound regions within the gene body (Cramer, 2019; Kwak & Lis, 2013).

1.2.3 Transcription termination by Pol II

Among the eukaryotic Pols, RNA Pol II transcribes the most diverse groups of genes. It transcribes genes coding to very short enhancer RNAs (eRNAs) to protein-coding genes that can be larger than 100 kb. This suggests Pol II must be very processive in the latter case, which in turn requires robust mechanism(s) for termination (Proudfoot, 2016). Pol II termination relies on DNA encoded sequence features that induce slowing down/pausing of the polymerase at the end of a transcription unit. The binding of termination factors are thought to play major roles in inducing Pol II pausing at the end of the transcriptional unit (Nojima et al., 2015). Pol II CTD modifications such as phosphorylation of Ser2 and Tyr1 of the heptad repeat installed earlier in the transcription cycle play key roles in recruiting terminations factors (Larochelle et al., 2018; Mayer et al., 2012; Shah et al., 2018; Zaborowska et al., 2016). For mRNAs, transcription termination and 3’ processing is done by the multi-subunit cleavage and polyadenylation specific factor, cleavage stimulating factor and the poly(A) polymerase referred to as CPF for simplicity (Mandel et al., 2006; Y. Zhang et al., 2019). The CPF is a modular complex conserved from yeast to human that can recognize the polyA site (PAS), cleave and polyadenylate the nascent RNA. (Casañal et al., 2017; Mandel et al., 2006; Sun et al., 2020).

The CPF subunit Pcf11 possesses a CTD interaction domain, which binds Ser2- phosphorylated CTD demonstrating how the CTD phosphorylation recruits the CPF (Licatalosi et al., 2002; Mayer et al., 2012).

Substantial effort has been dedicated into characterizing the CPF but it is not clear how the elongating Pol II is cleared of the gene during termination. Two models are currently used to describe Pol II termination. The allosteric model suggests that the Pol II elongation complex undergo conformational changes upon transcribing through the PAS which leads to termination. This model is supported by the observation that *in vitro* reconstituted transcription can terminate without pre-mRNA cleavage and without an exonuclease (Zhang et al., 2015). The allosteric model is challenged by an alternative model called the 'torpedo' model. The torpedo model proposes a mechanism of Pol II termination in which pre-mRNA cleavage at the PAS by the CPF generate an uncapped 5' ended RNA. This free 5' end recruits a 5' - 3' exonuclease which degrades the exposed RNA and dismantles the polymerase when it catches-up with it (Porrua et al., 2016; Proudfoot, 2016). This model is supported by observation in both yeast and mammals that knockdown of Rat1 and XRN2 (the 'torpedo' nucleases) respectively causes major Pol II termination defects (Fong et al., 2015; M. Kim et al., 2004). A recent genome-wide study in yeast further highlighted requirement of Rat1 for global mRNA termination in support of the torpedo model but also observed rearrangements of Pol II elongation complex that might support the allosteric model. The authors therefore suggested that the two models are not mutually exclusive and likely a combination of both is at play. (Baejen et al., 2017).

In *Saccharomyces cerevisiae*, Pol II transcribed ncRNAs are terminated by an alternative termination pathway, the Nrd1-Nab3-Sen1 pathway (NNS pathway) (Porrua et al., 2016). Nrd1 binds phosphorylated Pol II CTD similar to Pcf11 of the CPF and interacts with Nab3. Nrd1 preferentially binds to CTD phosphorylated on Ser5 instead of Ser2 by Pcf11 of the CPF (Mayer et al., 2012; Vasiljeva et al., 2008). Nrd1 and Nab3 bind specific sequence elements in both the DNA and the nascent RNA at the termination site of snRNAs and snoRNAs and recruit Sen1 for termination (Creamer et al., 2011). Sen1 is an ATP dependent helicase similar to the Rho factor in the bacterial RNAP termination. Sen1 uses ATP hydrolysis to dismantle the elongation complex (Porrua et al., 2016; Porrua & Libri, 2013).

The NNS pathway is not conserved in fission yeast where the CPF complex is employed in the termination of Pol II-transcribed ncRNA (Larochelle et al., 2018). It is also not present in metazoans. Senataxin, the mammalian homologue of Sen1 is involved in resolution of non-template DNA-RNA hybrids (R-loops) formed by paused Pol II (Skourti-Stathaki et al., 2011) and upon DNA damage (Cohen et al., 2018).

Since Pol II-transcribed snRNAs do not have PAS prerequisite for CPF cleavage and the NNS pathway is not conserved in metazoans, it was enigmatic how this class of Pol II-transcribed genes is terminated in multicellular organisms until the discovery of the Integrator complex (INT).

1.3 The Integrator complex, discovery and subunit composition

INT was discovered by serendipity when Baillat and colleagues identified a set of 12 polypeptides in a search for proteins associated with DSS1 (deleted in split hand/split foot protein 1). These polypeptides turned out to interact with the CTD of Pol II. In a preliminary amino acid sequence analysis of these polypeptides, they discovered that two of them share conserved domains with CPSF73 and CPSF100 subunits of the CPF hinting at a potential role of the newly discovered Pol II associated complex in RNA processing. Indeed, mutation of a conserved active site residue in the CPSF73-like polypeptide (later names INTS11) resulted in misprocessing of snRNAs consistent with their suspicion. This set of 12 polypeptides was then characterized biochemically as a complex, shown to be metazoan-specific, and named the Integrator complex (INT) for integrating CTD phosphorylation signal and 3' processing of snRNAs (Baillat et al., 2005). The 12 polypeptides were named INTS1 to INTS12 according to their apparent molecular weight on SDS-PAGE. Based on its elution profile on gel filtration chromatography, it was estimated that INT is over 1 MDa in size (Baillat et al., 2005; Baillat & Wagner, 2015). Further genetic screen in *Drosophila* identified two additional proteins, which associates with INT and are important for snRNA 3' processing namely Asunder and VWA9. Asunder and VWA9 were therefore named INTS13 and INTS14 respectively (Jiandong Chen et al., 2012a). Subsequent, proteomic analysis of human INT confirmed INTS13, INTS14 as well as DDX26B are bona fide members of the complex. DDX26B is also called INTS6-like (INTS6L) because it shares some sequence similarities with INTS6 (Malovannaya et al., 2011). Figure 1.4 is a cartoon representing INT subunits scaled to their sizes.

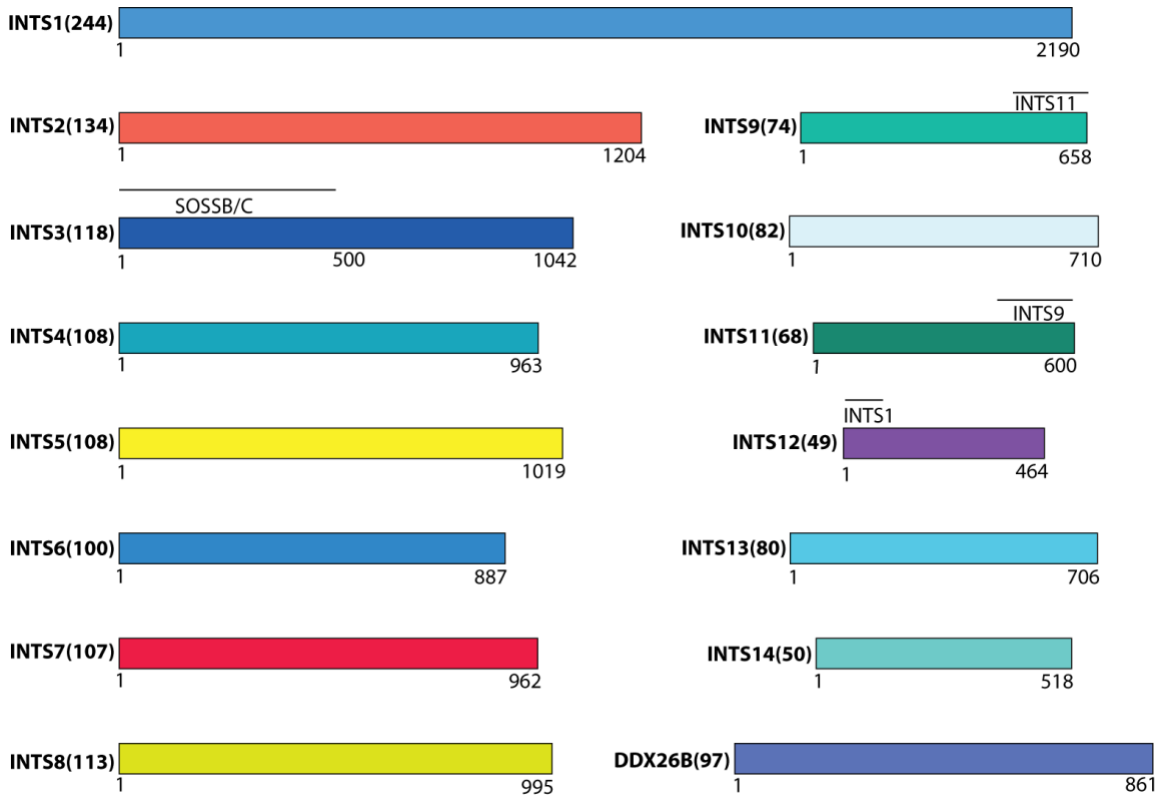


Figure 1.4 Subunit composition of INT. INT subunits are depicted as rectangle scaled to their sizes (number of amino acids). Interacting domains of INTS9, INTS11, and INTS12 are indicated. The part of INTS3 that interacts with SOSS-B and SOSS-C is also indicated. The molecular weight in kDa of each subunit is indicated in brackets.

The INT subunits are predicted to be alpha helical with protein-protein interactions folds such as ARM, HEAT, TPR and vWA (vWFA) (Baillat & Wagner, 2015; Gómez-Orte et al., 2019) (Figure 4.1). Apart from INTS9 and INTS11, which are highly similar to CPSF100 and CPSF73 respectively, none of the other subunits share similarity with any known protein (Baillat & Wagner, 2015). Further characterization of INTS9 and INTS11 showed that they form a heterodimer and this heterodimerization is important for snRNA 3' processing (Albrecht & Wagner, 2012). Molecular characterization reveals that the very C-terminal regions of INTS9 and INTS11 are important for INTS9/11 heterodimer formation (Wu et al., 2017).

INTS12 has a PHD domain known for binding methylated histones. This subunit was shown to interact with INTS1 via an N-terminal microdomain in the *Drosophila* INT (Jiandong Chen et al., 2013). Yeast two-hybrid analysis revealed an interaction between *Drosophila* IntS10 and IntS14 (Jiandong Chen et al., 2012a). INTS3 was shown to be part of the SOSS complex involved in sensing DNA double strand breaks. Careful analysis of proteins co-immunoprecipitated with the SOSS complex showed that INTS6 interacts with this complex (Ren et al., 2014; Skaar et al., 2009, 2015).

1.3.1 Pol II transcription of snRNA genes and the role of INT

Pol II-transcribed snRNA genes have a relatively simple structure composed of a conserved distal sequence element (DSE) which act as an upstream enhancer, a proximal sequence element (PSE) which is equivalent to a core promoter in protein-coding genes, an intronless gene body and a 3' box where pre-snRNAs are cleaved for termination (Guiro & Murphy, 2017) (Figure 1.5a). The DSE and PSE are conserved elements in the Pol III-transcribed 7SK and U6 snRNAs (Jawdekar & Henry, 2008). The 7SK and U6 snRNA genes have in addition to DSE and PSE a TATA box, which is important for the selective recruitment of Pol III to these genes. Addition of a TATA box downstream of the PSE of a Pol II-transcribed snRNA switches its transcription to Pol III. Analogously, mutation of the TATA box downstream of the PSE within the Pol III-transcribed U6 snRNA gene switches its transcription to Pol II (Lobo & Hernandez, 1989; Mattaj et al., 1988). These observations show that the promoter elements of snRNA genes determine which Pol transcribes them and therefore insure their accurate termination since Pol II and Pol III termination are substantially different. However, changing the U1 snRNA promoter to the promoter of a Pol II-transcribed protein-coding gene resulted in severe misprocessing of the resultant pre-snRNA, even though the 3'box was not changed (Neuman de Vegvar et al., 1986). It can therefore be assumed that the promoter elements of the Pol II-transcribed snRNAs encode information for the specific recruitment of factors needed for their 3' processing beyond the DSE and PSE binding factors. Pol II transcription initiation from an snRNA promoter requires the Pol II-specific GTFs except that only a subset of TBP associated factors (TAFs 5-9, 11 and 13) are recruited (Guiro & Murphy, 2017; Zaborowska et al., 2012). Beyond the recruitment of gene-specific factors, specific Pol II CTD modifications installed during snRNA specific initiation complex assembly have important roles in snRNA gene transcription. In line with this notion, phosphorylation of Ser-7 of the CTD heptad repeat has been shown to be important for transcription of U2 snRNA but not protein-coding genes (Egloff et al., 2007). This implies that, this modification among others may recruit specific factors for elongation and 3'processing of Pol II-transcribed snRNA genes. As expected, Egloff and colleagues showed that CTD phosphorylated on Ser-7 of the consensus repeat recruits the protein phosphatase RPAP2 to the promoter region of the U2 snRNA gene which facilitates recruitment of a subset of INT subunits including INTS1, INTS4, INTS5, INTS6 and INTS7.

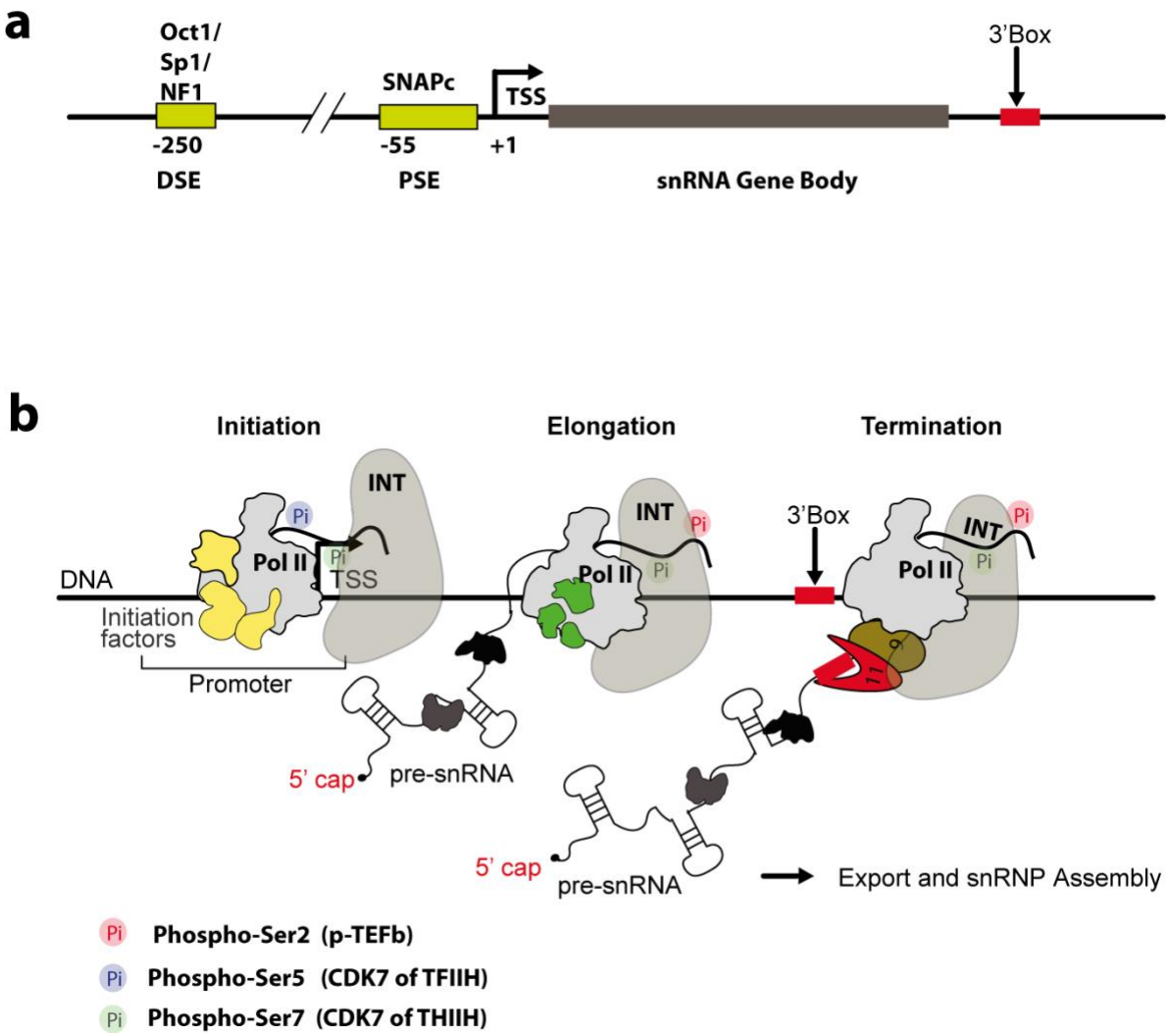


Figure 1.5 snRNA gene organization and transcription. (**a**) Organization of a typical Pol II-transcribed snRNA gene. Distal sequence elements (DSE) are bound by snRNA gene-specific activators such as Oct1, Sp1 or NF1. SNAPc an snRNA-specific Pol II transcription initiation factor binds to the proximal sequence element (PSE) and interacts with activators. TSS stands for transcription start site and the 3'box is indicated with a red filled box. (**b**) A model depicting the Pol II transcription cycle on a typical snRNA gene highlighting the involvement of INT. Phosphorylation status of the Pol II CTD is indicated and the kinases installing such modifications are indicated below the cartoon. INTS11 (red) is shown cleaving the pre-snRNA at the 3' box during termination.

Interestingly, the catalytic subunits of INT (INTS11) was not present at the promoter regions. Phosphorylation of Ser-2 of the Pol II CTD by the P-TEFb kinase CDK9 has been shown to be required for termination of snRNAs (Albrecht & Wagner, 2012). In summary, phosphorylation of Ser-7 by CDK7 (THIIH kinase) during initiation and Ser-2 by CDK9 in early elongation are the two marks required to efficiently recruit the full INT to allow for correct Pol II mediated snRNA transcription (Egloff et al., 2010) (Figure 1.5b).

Termination of Pol II-transcribed snRNA genes requires the enzymatic activity of INTS11 (Baillat et al., 2005) and is dependent on the Pol II elongation factors NELF and DSIF. Knockdown of NELF resulted in polyadenylation of the Pol II-transcribed snRNAs which are not polyadenylated (Yamamoto et al., 2014). The exact roles these Pol II elongation factors play in termination of snRNAs is not clear, although it can be speculated that they induce pausing (slow down) of Pol II to allow 3' box recognition and RNA cleavage by INT. It has also been shown that recruitment of NELF in the 3' region of snRNAs blocks the recruitment of the mRNA-specific CPF (Yamamoto et al., 2014).

1.3.2 INT beyond snRNA transcription

Regulation of pre-mRNA synthesis

The discovery of a role for NELF and DSIF in snRNA 3' processing was inspired by the observation that INT co-immunoprecipitated with SPT5 (DSIF) and NELF -E suggesting interaction between these complexes (Yamamoto et al., 2014). Interaction between INT and these transcription elongation factors suggests that INT might have a broader role in Pol II transcription. The first evidence in this direction came from Gardini and coworkers who observed that INT is recruited to immediate early genes (IEG) upon epidermal growth factor (EGF) stimulation (Gardini et al., 2014). These genes, similar to *Drosophila* Hsp70, are strongly regulated by promoter-proximal pausing (PPP). They showed that knockdown of INTS1 and INTS11 abrogated escape of paused Pol II into productive elongation upon EGF stimulation. Mechanistically, INT controls release of paused Pol II by recruiting the super elongation complex which contains the positive transcription elongation factor (P-TEFb). P-TEFb then releases paused Pol II by phosphorylating Pol II, NELF and DSIF (Gardini et al., 2014; Vos, Farnung, Urlaub, et al., 2018).

Furthermore, it was observed that more than 2000 protein-coding genes are differentially expressed when INTS3 and INTS11 are knocked down, suggesting a broader role of INT in regulating expression of protein-coding genes (Stadelmayer et al., 2014). Stadelmayer and colleagues observed some genes that are upregulated upon INTS11 knockdown in the absence of stimulus (resting cells) contrary to downregulation of IEGs reported by Gardini and colleagues. This suggests a dual function of INT at protein-coding genes in a stimulus-dependent manner. Additionally, recent studies demonstrated INTS11-dependent premature termination of paused Pol II within a class of INT regulated genes (Elrod et al., 2019; Tatomer et al., 2019). This study is well supported by the observation that INT regulated pre-mRNAs are rich in 3'box-like sequences which might be recognized and cleaved by INT to attenuate these genes (Stadelmayer et al., 2014).

Taken together, these *in vivo* studies suggest a model where INT is a critical switch that represses promoter proximally paused Pol II to downregulate certain genes via endonuclease activity of INT11 or recruits the super elongation complex containing P-TEFb that releases Pol II from PPP and activates other genes.

Termination of other RNAs

Apart from termination of snRNAs, INT has been shown to play important roles in termination of other classes of genes. Knockdown of INTS3 and INTS9 was shown to induce aberrant 3' processing of replication-dependent histone mRNAs (Skaar et al., 2015). These class of mRNAs are intronless and non-polyadenylated. They require a unique complex of U7 snRNP, the cleavage module of the CPF and other accessory factors for their 3' formation (Marzluff et al., 2008; Sun et al., 2020).

Additionally, INT is recruited to the promoter region of enhancer RNAs (eRNA) in a stimulus-dependent manner where its catalytic activity was shown to be important for their termination (Lai et al., 2015). The *Herpesvirus saimiri* also require INT for the biogenesis of a microRNA important for downregulating host gene expression. This microRNA possesses a 3' sequence highly analogous to the 3' box of snRNAs (Xie et al., 2015).

In summary, it emerges that INT is needed for termination of genes that are relatively short compared to pre-mRNAs and this activity depends on a 3' box-like sequence and catalytic activity of INTS11.

1.4. Aims and scope of this work

I have described in section 1.3 above the multi-subunit INT which is added recently to the list of Pol II transcription regulators in metazoans. There has been substantial amount of *in vivo* work dedicated to understanding the specific roles of INT in Pol II transcription and regulation summarized in the aforementioned section. While these studies have been important in disentangling the *in vivo* functions of INT, there are critical questions that can only be addressed by *in vitro* biochemical reconstitution approach using purified components. For example, to obtain structural and biochemical details on the architecture of INT and its roles in Pol II transcription regulation, purification of INT is absolutely necessary. There seems to be redundancy in the kinases that phosphorylate the Pol II CTD (Zaborowska et al., 2016). To unequivocally identify which CTD phosphorylation is important for INT recruitment to Pol II transcriptional complexes, an *in vitro* biochemical assay using purified components is necessary. Furthermore, INT has been shown to interact with phosphorylated Pol II CTD,

NELF, DSIF and the super elongation complex to orchestrate its functions in Pol II transcription. Identification of the specific INT subunits that interact with these key components also requires *in vitro* biochemical reconstitution. While there are well-established protocols for the purification of Pol II, DSIF and NELF (Vos, Farnung, Urlaub, et al., 2018), protocols for production of INT in good amounts and purity for biochemical studies is still lacking. Miniscule amounts of INT has been obtained from endogenous sources but these are normally contaminated with Pol II and other interacting complexes of INT (Baillat et al., 2005; Stadelmayer et al., 2014; Yamamoto et al., 2014). This difficulty in obtaining homogenous INT in good quantity hampers biochemical and structural characterization of INT. To avert the problems of low yield and heterogeneity, recombinant reconstitution of INT was pursued in this study. The Integrator complex (INT) comprises 15 polypeptides with an estimated molecular weight of ~1.5 MDa assuming a single copy of each subunit (Baillat et al., 2005; Jiandong Chen et al., 2012). The large size of this complex poses a major challenge for recombinant expression. Furthermore, it is not known how the subunits are interacting within INT, as in which subunits might need each other for co-translational folding. This poses the challenge of first identifying interacting subunits and then establishing an expression and purification protocol for the full INT.

The baculovirus-insect cell recombinant protein expression system has been used successfully to produce large multi-subunit protein complexes (Berger et al., 2004; Gradia et al., 2017; Schilbach et al., 2017; Zhang et al., 2013).

In this study, I sought to reconstitute INT recombinantly using the baculovirus-insect cell expression system and its interaction with Pol II paused elongation complex (PEC) using the following steps

1. Identify interacting subunits of INT that form stable subcomplexes for co-expression and purification using a co-infection assay in insect cells coupled to XL-MS
2. Reconstitute the full INT from the identified subcomplex and obtain an inter-subunit interaction network using XL-MS
3. Study the interaction between INT and PEC *in vitro* and identify the proteins involved in the interaction between the two complexes using XL-MS.

2 Materials

Table 2.1. Sources of cDNA of INT subunits and GenBank Accession codes

Plasmid ID	Selection	Gene ID	GenBank Accession	Website
Assembled from cNDA fragments (IDT)		INTS1	NM_001080453	
Assembled from cNDA fragments (IDT)		INTS2	AL136800.1	
pcDNA5D FRT/TO Flag INTS3	Ampicillin	INTS3	NM_023015.3	https://mrcppureagents.dundee.ac.uk/reagents-view-cdna-clones/607541
HsCD00329509		INTS4	BC009995	https://plasmid.med.harvard.edu/PLASMID/GetCloneDetail.do?cloneid=329509&species=
HsCD00343174 replaced with HsCD00438603	Spectinomy-cin	INTS5	BC060841	https://plasmid.med.harvard.edu/PLASMID/SearchClone.do?&searchType=PlasmID+Clone+ID&searchString=HsCD00343174
HsCD00338395	Ampicillin	INTS6	BC039829	https://plasmid.med.harvard.edu/PLASMID/SearchClone.do?&searchType=PlasmID+Clone+ID&searchString=HsCD00338395

HsCD00327522	Ampicillin	INTS7	BC030716	https://plasmid.med.harvard.edu/PLASMID/SearchClone.do?&searchType=PlasmID+Clone+ID&searchString=HsCD00327522
HsCD00342288	Ampicillin	INTS8	BC136754	https://plasmid.med.harvard.edu/PLASMID/SearchClone.do?&searchType=PlasmID+Clone+ID&searchString=HsCD00342288
HsCD00337647	Chloramphe-nicol	INTS9	BC025267	https://plasmid.med.harvard.edu/PLASMID/SearchClone.do?&searchType=PlasmID+Clone+ID&searchString=HsCD00337647
HsCD00321882	Chloramphe-nicol	INTS10(79-710)	BC006209	https://plasmid.med.harvard.edu/PLASMID/SearchClone.do?&searchType=PlasmID+Clone+ID&searchString=HsCD00321882
HsCD00326484	Chloramphe-nicol	INTS11 (CPSF3L)	BC007978	https://plasmid.med.harvard.edu/PLASMID/SearchClone.do?&searchType=PlasmID+Clone+ID&searchString=HsCD00326484
HsCD00334255	Chloramphe-nicol	INTS12	BC014442	https://plasmid.med.harvard.edu/PLASMID/SearchClone.do?&searchType=PlasmID+Clone+

				ID&searchString=HsCD00334255
HsCD00323796	Chloramphe-nicol	INTS13/ ASUN	BC008368	https://plasmid.med.harvard.edu/PLASMID/GetCloneDetail.do?cloneid=323796&species=
HsCD00323194	Chloramphe-nicol	INTS14/ VWA9	BC007991	https://plasmid.med.harvard.edu/PLASMID/GetCloneDetail.do?cloneid=323194&species=
Assembled from cNDA fragments (IDT**)**		INTS6L/ DDX26B	BC140715.1	

Table 3.2. Vectors, tags and resistance marker

Name	**Affinity tag**	**Resistance marker**
438-A	No tag	Ampicillin
438-B	6xHis	Ampicillin
438-C	6xHis-MBP	Ampicillin

Tables 3.3. Bacteria strains and their genotypes

Strain	**Genotype**	**Provider**
XL1-Blue	*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1lac [F' proAB lacIqZΔM15 Tn10 (Tetr)]*	Agilent
NEB Stable	Not available	NEB
DH10EMBacy	*F- mcrA Δ(mrr-hsdRMS-mcrBC) ϕ80dlacZΔM15 ΔlacX74 endA1 recA1 deoR Δ(ara, leu)7697 araD139 galU galK λ- rpsL nupG / bMON14272‡ yfp+/ pMON7124*	Geneva Biotech

Table 3.4. Composition of bacterial growth media

Media	Composition
LB broth	1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl (+/- antibiotics or additives based on specific needs)
LB Agar	1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, 0.1 mg/ml Ampicillin sulphate 2% agar (+/- antibiotics or additives based on specific needs)

Table 3.5. Insect cell lines

Species	Cell line	Growth Medium	Provider
Spodoptera frugiperda	SF21 *(IPLB-Sf21-AE)*	Gibco® Sf-900™ III SFM	Thermo Fisher Scientific
Spodoptera frugiperda	SF21	Gibco® Sf-900™ III SFM	Thermo Fisher Scientific
Trichoplusia ni	Hi5 *(BTI-TN-5B1-4)*	ESF921™	Thermo Fisher Scientific

Table 3.6. Chemicals and their supplies

Chemical (full name)	Supplier
Acetic acid	Merck
Acetone	Merck
Agarose	Invitrogen
BME (β-mercaptoethanol)	Roth
BS3 (bissulfosuccinimydyl suberate)	Thermo Fisher Scientific
DTT (Dithiothreitol)	Roth
EDC (1-ethyl-3-(3-dimethylaminopropyl)carbodiimide)	Thermo Fisher Scientific
EDTA (Ethylenediaminetetraacetic acid)	Roth
Ethanol, Isopropanol	Merck
Glycerol	Roth
HCl	Merck
HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid)	Roth
Imidazole	Roth
IPTG(Isopropyl β-d-1-thiogalactopyranoside)	Roth

KCl	Merck
KOH	Merck
Maltose	Roth
$MgCl_2$	Merck
$MnCl_2$	Merck
NaCl	Merck
Sucrose	Merck
SYBRsafe	Invitrogen
Tris (trisaminomethane)	VWR
TCEP (tris(2-carboxyethyl) phosphine)	Roth
x-gal (5-bromo-4-chloro-3-indolyl-β-D-galactoside)	Roth
$ZnCl_2$	Merck

Table 3.7. Buffers and their composition/suppliers

Buffer	**Composition/Supplier**	**Application**
10x TAE	50 mM EDTA pH 8.0 at 20°C, 2.5 M Tris-acetate	Agarose gel electrophoresis
10x PBS	137 mM NaCl, 2.7 mM KCl, 10 mM Na2HPO4 pH 7.4, 1.76 mM KH2PO4 pH 7.4	Western blotting
20x MOPS ((3-(N-morpholino)propanesulfonic acid)) Buffer	Thermo Fisher Scientific	LDS-PAGE
20x MES (2-ethanesulfonic acid)	Thermo Fisher Scientific	LDS-PAGE
4x LDS (lithium dodecyl sulfate) sample buffer	Thermo Fisher Scientific	LDS-PAGE
6*x* DNA loading dye	NEB	Agarose gel electrophoresis
Cutsmart buffer	NEB	Restriction endonuclease digestion with PmeI and SspI

NEB Buffer 3.1	NEB	Restriction endonuclease digestion with SwaI

Table 3.8. Antibodies used in this study

Antibody	Host organism	Supplier/catalogue Number
Anti INTS2	Rabbit	Abcam/ab178334
Anti INTS3	Rabbit	Abcam/ab70451
Anti INTS5	Rabbit	Abcam/ab74405
Anti Histag-HRP	Mouse	Mitenyl Biotech/130-092-785
Anti MBP-HRP	Mouse	Abcam/ab49923
Donkey anti Rabbit	donkey	Abcam/ab150075

Table 3.9. Overview of key primers used in cloning INT subunits

Oligo ID/Name	SEQUENCE	Purpose
E44_INTS1_LicV1_For	TACTTCCAATCCAATATGAACCGGGCCAAGC	LIC cloning of wt-INTS1
E45_INTS1_LicV1_Rev	TTATCCACTTCCAATTCACATCACGGCCTCCATATG	LIC cloning of wt-INTS1
E46_INTS1_Seq_1	GCCAGTTCTCGTCCTCCTC	Sequencing of INTS1
E47_INTS1_Seq_2	GAGCCTCATGTACCTGGCC	Sequencing of INTS1
E48_INTS1_Seq_3	CAACCACTTCATGCTGTGCA	Sequencing of INTS1
E49_INTS1_Seq_4	CCATGGAGCTTGCTGACCAC	Sequencing of INTS1
E50_INTS1_sqe_5	GACCTGGTACAGTCCAGCG	Sequencing of INTS1

E51_INTS1_Seq_6	GTCGGAGTCTCAGGACCAGG	Sequencing of INTS1
E52_INTS1_seq_7	CTCGGTGGCAGAGCTCC	Sequencing of INTS1
E53_INTS1_seq8	GCCTCCTAGTGGACTGGCTG	Sequencing of INTS1
E54_INTS1_seq_9	CACCTCAACTTCCAGGAGTTCC	Sequencing of INTS1
IF_43_Int1_U_438C_1F	ACTTCCAATCCAAUATGAACCGGGCCAAGC	U excision cloning of INTS1
IF_44_INTS1_U_438C_1R	ACAGGCGTGTGCUGG	U excision cloning of INTS1
IF_45_INTS1_U_438C_2F	AGCACACGCCTGUGGAG	U excision cloning of INTS1
IF_46_INTS1_U_438C_2R	ACTTCCAATTCACAUCACGGCCTCCATATGC	U excision cloning of INTS1
IF_47_INTS1_U_438C_3F	ATGTGAATTGGAAGUGGATAACGGATCCG	U excision cloning of INTS1
IF_48_INTS1_U_438C_3R	ACAGCTTGTCTGUAAGCGGATG	U excision cloning of INTS1
IF_49_INTS1_U_438C_4F	ACAGACAAGCTGUGACCGTC	U excision cloning of INTS1
IF_50_INTS1_U_438C_4R	ATTGGATTGGAAGUACAGGTTTTCCTCG	U excision cloning of INTS1

IF_51_Int1_U_438C_1F	AAGGAGATATAGTUATGAACCGGGCCAAGC	U excision cloning of INTS1
IF_52_c10xHis_438_U_excision	AACTATATCTCCTUCTTAAAGGGATCCGCGCCC	U excision cloning of INTS1
IF_53_INTS1_U_c10xHis_438	AAGTAGAGGTTCAUCACGGCCTCCATATGC	U excision cloning of INTS1
IF_54_c10xHis_438_U_excision	ATGAACCTCTACTUCCAATCCGGCTC	U excision cloning of INTS1
IF65_INTS1_1-194_LicV1_R	TTATCCACTTCCAATTTAGTCCTCCTCCTCCGTGAGG	Truncation of INTS1
IF66_INTS1_1-1010_LicV1_R	TTATCCACTTCCAATTTACTCCTTCTCCTCCCCGTCC	Truncation of INTS1
IF67_INTS1(M-C)_LicV1_F	TACTTCCAATCCAATATGGAGGAGGATGTGGGGG	Truncation of INTS1
E55_INTS2_LicV 1_For	TACTTCCAATCCAATATGAAGGATCAACAAACAGTAATAATGACTG	LIC cloning of INTS2
E56_INTS2_LicV1_Rev	TTATCCACTTCCAATTTAAATTCCACTAACACTCATATTTATTATTTCAATTACTGTTC	LIC cloning of INTS2
E57_INTS2_seq_1	GAGACACAAGAACCAGGCACC	Sequencing of INTS2
E58_INTS2_seq_2	GAGAGTCCAGTATATTTGGAGGAAGC	Sequencing of INTS2
E59_INTS2_seq_3	CTGCAGTTGATGACGAGCCG	Sequencing of INTS2
E60_INTS2_seq_4	CGAAGACTTTAGCTGCCATGC	Sequencing of INTS2

E61_INTS2_seq_5	GCTCAGGATAGTGCAGCTGTCC	Sequencing of INTS2
113_INTS2_Frag1_F	GCGGATCCTTTATTAAGTACTTCCAATCC	CPEC cloning of INTS2
114_INTS2_Frag1_R	CAGTATATAGTAGAGCACCAAAAGCTGTG	CPEC cloning of INTS2
115_INTS2_Frag2_F	ACAGCACAGCTTTTGGTGC	CPEC cloning of INTS2
116_INTS2_Frag2_R	CGGACCGGTTATCCACTTCC	CPEC cloning of INTS2
E62_INTS3_LicV1_For	TACTTCCAATCCAATATGGAGTTGCAGAAGGGAAAAGG	LIC cloning of INTS3
E63_INTS3_Licv1_Rev	TTATCCACTTCCAATTCAGTCACTGTCAGAGCCCAC	Lic cloning of INTS3
E64_INTS3_seq_1	CCTTCTGTCGCAGGGCC	Sequencing of INTS3
E65_INTS3_Seq_2	CAGCCAAAAATATCTGGTTGGCAG	Sequencing of INTS3
E66_INTS3_seq_3	CAACTTCTATCCACCATTGGAGGG	Sequencing of INTS3
E67_INTS3_seq_4	CTACCCAGCTGGGCGATCTG	Sequencing of INTS3
IF96_INTS3_471_RTH_R	GTGTGCCAAGACCCGTTTCTC	Truncation of INTS3
IF97_INTS3_800_RTH_R	TAGGCTCTGAATGAGTATGTTGAGAACTGAG	Truncation of INTS3
IF98_INTS3_438c_Stop_RTH	TGAATTGGAAGTGGATAACGGATCCG	Truncation of INTS3
IF99_TEV_438c_RTH	ATTGGATTGGAAGTACAGGTTTTCCTCG	Truncation of INTS3
IF100_INTS3_472_RTH_F	CTAGCTCCCCTGTTTGACAACCC	Truncation of INTS3

IF101_INTS3_801_RTH_F	GACTGGGAGACCTTTGAGCAG	Truncation of INTS3
E68_INTS4_LicV1_For	TACTTCCAATCCAATATGGCGGCGCACCTTAAG	LIC cloning of INTS4
E69_INTS4_LicV1_Rev	TTATCCACTTCCAATTTAGCGCCGTGCAGGTTTG	LIC cloning of INTS4
E70_INTS4_Seq_1	CAGTCCTCTTTCATGGAGCTGC	Sequencing of INTS4
E71_INTS4_Seq_2	GTGCCTGCAGTTACTTGGCAATC	Sequencing of INTS4
E72_INTS4_Seq_3	TGCCTTGATTTCCTAGTTGACATGTTC	Sequencing of INTS4
E73_INTS4_seq_4	GGAAAAGTTGTGGAATGTAGCTGCC	Sequencing of INTS4
E80_INTS5_LicV1_For	TACTTCCAATCCAATATGTCCGCGCTGTGCG	LIC cloning of INTS5
E81_INTS5_LicV1_Rev	TTATCCACTTCCAATCTACGTCCCCTGTCGAAGG	LIC cloning of INTS5
E39_INTS5_seq_1	GGCCACAGGAGAGAACCC	Sequencing of INTS5
E40_INTS5_seq_2	GGATACCTCTGTTCAGCATTCTCC	Sequencing of INTS5
E41_INTS5_seq_3	CTGGTTCATCACCGGGGAG	Sequencing of INTS5
E42_INTS5_seq_4	GGGACAATGAGACTCTCTCAGTTG	Sequencing of INTS5
E43_INTS5_seq_5	CGGCCCTGGGTAATATGCATG	Sequencing of INTS5
E82_INTS6_LicV1_For	TACTTCCAATCCAATATGCCCATCTTACTGTTCCTGATAGAC	LIC cloning of INTS6
E83_INTS6_LicV1_Rev	TTATCCACTTCCAATTTAATTGCTATTAATATGGTTGATCTGATTGGC	LIC cloning of INTS6
E35_INTS6_seq_1	CAGGGGAAGGATCTGGTCC	Sequencing of INTS6

E36_INTS6_Seq_2	GTGTCTGGAGTCCTTGGTGC	Sequencing of INTS6
E37_INTS6_seq_3	GATGGGAGCACCTAACCTAATAGC	Sequencing of INTS6
E38_INTS6_Seq_4	CCACCTGCACCTACAACTC	Sequencing of INTS6
E90_INTS6_Swa1_mutation_F	GCTTTTGATTTATTAAATTTGAATAGATTAGTAACTGGC	Mutation of SwaI site in INTS6 CDNA
E91_INTS6_Swa1_mutation_R	GCCAGTTACTAATCTATTCAAATTTAATAAATCAAAAGC	Mutation of SwaI site in INTS6 CDNA
E74_INTS7_LicV1_For	TACTTCCAATCCAATATGGCGTCAAACTCAACTAAGTCTTTCC	LIC cloning of INTS7
E75_INTS7_LicV1_Rev	TTATCCACTTCCAATTTAAAACCGTGTGTAGGCATTGCG	LIC cloning of INTS7
E76_INTS7_Seq_1	GCAAGCAGAGTGAAAGTGTGC	Sequencing of INTS7
E77_INTS7_seq_2	GAGGAAGAATGCTCATCATAGTATTCGTC	Sequencing of INTS7
E78_INTS7_Seq_3	CCTGGAATCCCTACTGGTACTTTGTAG	Sequencing of INTS7
E79_INTS7_Seq_4	GACCTCCAGAGGTGTGGTCG	Sequencing of INTS7
E84_INTS8_LicV1_For	TACTTCCAATCCAATATGAGCGCGGAGGCG	LIC cloning of INTS8
E85_INTS8_LicV1_rev	TTATCCACTTCCAATTTAAAAGTAAAGTTTTGCCATTGCTTGGAG	LIC cloning of INTS8
E31_INTS8_Seq_1	CATAGCACACCTGGCACTGC	Sequencing of INTS8
E32_INTS8_seq_2	GGCCATGGAACCAGGC	Sequencing of INTS8
E33_INTS8_Seq_3	GATCCCCTAGAGTAAATCTGTGC	Sequencing of INTS8

E34_INTS8_Seq_4	CTGTAGTGTGTCCAGTCAGCAC	Sequencing of INTS8
E10_INTS9_licV1_14_For	TACTTCCAATCCAATATGAAACTGTATTGCCTGTCAGGG	LIC cloning of INTS9
E11_INTS9_licV1_14_rev	TTATCCACTTCCAATTCATCAGAACTTCTGTAAGAATTTGAGGACAAGG	LIC cloning of INTS9
E22_INTS9_Seq_1	CCACAAGGAGGCAGACTGAG	Sequencing of INTS9
E23_INTS9_seq_2	GCAGGCTTCTCATGGAAGAGC	Sequencing of INTS9
E24_INTS9_seq_3	GGTCCACTTCATGGAGCTCTG	Sequencing of INTS9
E14_Int10_LicV1_For	TACTTCCAATCCAATGAAATCAGCATTATTACATCAGCATTAAGGAAC	LIC cloning of INTS10 truncation
E94_INTS10_LicV1_F_flC	TACTTCCAATCCAATATGTCTGCCCAGGGGGAC	LIC cloning of INTS10
E15_INTS10_Lic_V1_Rev	TTATCCACTTCCAATTCAGGTCAGAGTCTGAAGGAGC	LIC cloning of INTS9
E28_INTS10_Seq_1	GACCAGCATGGTGGAATACACC	Sequencing of INTS10
E29_INTS10_seq_2	GAATGGCAGATGGATAAAGGAAGACG	Sequencing of INTS10
E30_INTS10_seq_3	GATTTGATGTGCTACATGGTACTCC	Sequencing of INTS10
C13_INTS10_CPEC_for	GAAATCAGCATTATTACATCAGCATTAAGG	CPEC cloning of INTS10
C14_INTS10_CPEC_rev	CGGACCGGTTATCCACTTCCAATTCAGGTCAGAGTCTGAAGGAGC	CPEC cloning of INTS10
E5_Int11_licV1_14C_For	TACTTCCAATCCAATATGCCTGAGATCAGAGTCACG	LIC cloning of INTS11
E6_INTS11_licV1_14C_rev	TTATCCACTTCCAATTCATCAGCTGGGGGCCTG	LIC cloning of INTS11

E25_INTS11_seq_1	CAGCACCTTCCCACCACG	Sequencing of INTS11
E26_INTS11_Seq_2	GTACGCCACGACCATCCG	Sequencing of INTS11
E27_INTS11_seq_3	CAGCATCCCCGTAGGCATC	Sequencing of INTS11
E86_INTS12_LicV1_For	TACTTCCAATCCAATATGGCTGCTACTGTGAACTTGG	LIC cloning of INTS12
E87_INTS12_LicV1_Rev	TTATCCACTTCCAATTTACTTCTTGAGTTTCTTTTGGGCAGC	LIC cloning of INTS12
E88_INTS12_Seq_1	CCCTCGCCTGGTGTGG	Sequencing of INTS12
E89_INTS12_Seq_2	GTTCTCTTAAACGCTAGAAAAGTTGTCTCTTG	Sequencing of INTS12
IF_89_1B_LV1_R_RTH	ATTGGAAGTGGATAACGGATCCG	Truncation of INTS12
IF88_12-194_RTH	TTACTGGGGTTTATGACAATCTCGGTG	Truncation of INTS12
117_INTS13_LicV1_F	TACTTCCAATCCAATATGAAGATTTTTTCTGAATCTCATAAAACAGTG	LIC cloning of INTS13
118_INTS13_LicV1_R	TTATCCACTTCCAATTCACTGCCGGCTGGCTTTTC	LIC cloning of INTS13
119_INTS13_Seq_1	GTGTTAGCATGCTGTTCTTCCTTC	Sequencing of INTS13
120_INTS13_Seq_2	GTCCCCGGTTTTAACCAGTG	Sequencing of INTS13
121_INTS13_Seq_3	GCAAGCGGTAGTTCCATTAGCC	Sequencing of INTS13
122_INTS14_LicV1_F	TACTTCCAATCCAATATGCCGACAGTGGTGG	LIC cloning of INTS14
123_INTS14_LicV1_R	TTATCCACTTCCAATTCAAATTCTTTCAGTGCTGCTCC	LIC cloning of INTS14
124_INTS14_Seq_1	GACCAGATGTCTGGACAGAAC	Sequencing of INTS14

125_INTS14_Seq_2	CCCCAGGCCAGAACC	Sequencing of INTS14
E128_DDX26B_CPEC_1	GCGGATCCTTTATTAAGTACTTCCAATCCAATATGCCCATCCTGCTGTTCC	CPEC cloning of DDX26B
E129_DDX26B_CPEC_2	GCAACTGGAACACTATGAAGGGAATCTTC	CPEC cloning of DDX26B
E130_DDX26B_CPEC_3	CTCGCAGTGGAGAAGCCAATGTCTTCAGATATTCCTGATAGTTACCCATTTGTGCAACTGGAACACTATGAAGGG	CPEC cloning of DDX26B
E131_DDX26B_CPEC_4	GACATTGGCTTCTCCACTGCGAGAGATTGATCCAGACCAACCCAAAAGACTGCATACTTTTGGCAATCCGTTC	CPEC cloning of DDX26B
E132_DDX26B_CPEC_5	CTTTTGGCAATCCGTTCAAACAAGATAAG	CPEC cloning of DDX26B
E133__DDX26B_CPEC_6	CGGACCGGTTATCCACTTCCAATCTAACATGATGATCTGCTGTTGATGTGAC	CPEC cloning of DDX26B
E134_DDX26b_Seq_1	GACAACTATGCCATGGCTGAGC	Sequencing of DDX26B
E135_DDX26B_Seq_2	GGTCGCTCCTACTGTGTGAG	Sequencing of DDX26B
E136_DDX26B_Seq_3	GAAATCACAGGGGAAACTGCAC	Sequencing of DDX26B
E137_DDX26B_Seq_4	GTCTGACGACTTCACAAGTCTCAGC	Sequencing of DDX26B
E138_DDX26B_LicV1_F	TACTTCCAATCCAATATGCCCATCCTGCTGTTCC	LIC cloning of DDX26B
E139_DDX26B_LicV1_R	TTATCCACTTCCAATCTAACATGATGATCTGCTGTTGATGTGAC	LIC cloning of DDX26B

Note: Primers are written in 5’ to 3’ direction, CPEC – circular polymerase extention cloning. LIC – Ligation independent cloning.

Table 3.10. Expression constructs

A comprehensive list of final expression vector and all intermediates generated during this study can be provided upon request.

3 Methods

3.1 General methods for cloning

3.1.1 Polymerase Chain reactions (PCR)

The open reading frames (ORFs) of Integrator subunits were amplified from purchased plasmid DNA using PCR except INTS1, INTS2 and DDX26B. INTS1, INTS2 and DDX26B were assembled from DNA fragments (Table 2.1). PCR primers were design targeting the 5' and 3' ends of the ORFs with lengths between 10 and 20 nt and a melting temperature of 58 – 60 ^{o}C. Additional DNA overhangs were added to the primers for compatibility with the ligation independent cloning (LIC) into the MacroLab 438 series of vectors (Gradia et al., 2017). A typical PCR reaction was carried out using the Phusion® High-Fidelity DNA polymerase (NEB) and contained 1 - 10 ng of plasmid DNA depending on the size, ~3% (v/v) DMSO and all other reagents were added in the recommended amounts of the manufacturer's protocol. Normally, 25 - 30 thermocycles were carried out in a TProfessional TRIO Thermocycler® (Biometra).
Parameters such as the primer annealing temperature and primer extension time were set for each PCR according to primers and the size of the gene of interest. In some cases, it was necessary to test a gradient of annealing temperatures for amplification. Under standard conditions an average synthesis rate of 1,000 nt/min was assumed for Phusion® polymerase (NEB). To reduce background after PCR in subsequent cloning steps, the plasmid DNA (which is methylated) in the product was removed with DpnI (NEB) digestion for 3 hrs or overnight. The products were then analyzed on agarose gel. Products of interest were excised from the gel and purified using the QIAquick gel extraction kit (QIAGEN) according to the manufacturer's instructions or in some cases using the QIAGEN PCR purification kit.

3.1.2 Round-the-horn PCR for mutagenesis

The ORFs of INTS3 and INTS6 in the purchased plasmids had the restriction sites of PmeI and SwaI respectively. Since these restriction enzymes are used to combine vectors into polyprotein expression plasmids, it was necessary to mutate these positions. For some INT subunits disordered regions were removed or different truncation variants were cloned (INTS1, INTS3 and INTS12).
For these purposes, the 'round-the-horn' PCR (RTH-PCR) technique was applied ((Liu & Naismith, 2008)). This method entails the amplification of an entire template vector and re-ligation of the ends. This provides the flexibility to add a DNA sequence as an overhang, make

a point mutation in the annealing region of the primer, or design primers to remove a DNA fragment in the vector. Genes of interest were first cloned into a Macrolab 438 vector of choice and primers were designed to change a base in the sequence of an undesired restriction site or remove a whole region. For efficient ligation of the amplified vector DNA, one of primers in a primer pair for RTH-PCR was purchased pre-phosphorylated at the 5' end. PCR conditions were as for a standard PCR described above with primer extension time adjusted depending on the size of the vector. Successful amplification was assessed by analyzing 5 % of the RTH-PCR product on agarose gel. DpnI digestion was then used to remove the template vector and the amplified vector was purified using the QIAGEN PCR purification kit (QIAGEN). An amount of 100 - 200 ng of the purified vector was then ligated using Quick Ligase (NEB) following the manufacturer's instructions. An aliquot of 10 µl of the ligation reaction was transformed into XL1 blue chemically competent bacteria.

3.1.3 Restriction endonuclease digestion of DNA

Linear vectors and inserts for LIC were generated by restriction endonuclease digestion. A typical reaction included 1 - 2 µg of DNA, 5 µl of the recommended 10x buffer of the restriction enzyme used (NEB) and 10000 - 20000 U of the restriction enzyme in a final volume of 50 µl. The reaction was incubated for 4 - 5 hrs or overnight at 25 ^{o}C for SwaI and 37 ^{o}C for SspI and PmeI. The restriction digest of inserts was always purified from agarose gel using the QIAGEN gel extraction kit while the linearized vectors were mostly purified using the QIAGEN quick PCR purification kit.

3.1.4 Agarose gel electrophoresis

Agarose gel electrophoresis was applied to visualize DNA (plasmid or PCR products) and for preparative isolation of DNA after restriction digestion and PCR.

The products of PCR or restriction digestions were analyzed by separation on 1 - 2 % (w/v) agarose in 1x TAE buffer depending of the expected product sizes. SYBR™ Safe DNA gel stain (Invitrogen) was added (1:20 v/v) to the molten agarose upon casting for subsequent visualization of DNA samples under ultraviolet (UV) light. DNA samples were supplemented with DNA loading buffer to a final concentration of 1x and loaded on the gel covered with 1x TAE buffer. Commercial DNA ladders were used as a size standard. The electrophoresis was run at 120 V for 20 - 45 min or until sufficient separation of the DNA fragments was achieved. DNA bands were visualized with the GEL iX20 Imager system (Intas) for analytical gels or excised for gel extraction in case of preparative gels.

3.1.5 DNA Extraction from agarose gel and PCR purification

After agarose gel electrophoresis, the desired DNA fragments were visualized by UV-illumination with the BST-20G-D2E BlueLED BioTransilluminator (Biostep), excised with a sterile razor blade and the gel pieces were weighed.

The DNA was then extracted from the gel using the QIAquick Gel Extraction kit following the manufacturer's protocol. Alternatively, only about 5% of the PCR product or restriction digest reaction of SwaI and SspI is analyzed on agarose gel. The products were then purified using the QIAGEN PCR purification kit following the manufacturer's protocol. The DNA was always eluted in 20 - 40 µl of double distil water (ddH_2O).

3.1.6 Strategy for cloning of INT subunits in MacroLab 438 series vectors

The baculovirus - insect cell system was employed for the expression of proteins in this study. The 438 series of MacroLab vectors (modified from pFastBac, Addgene #55218 and #55220) allows for the cloning of multiple gene cassettes together and further permits the integration of these genes into the baculoviral genome for subsequent baculovirus production. There are several types of these vectors allowing one to exploit different tagging options. The three main ones employed in this study were 438-A (No tag), 438-B (amino (N) terminus 6xHis tag cleavable by TEV protease) and 438-C (N-terminus 6xHis followed by maltose binding protein (6xHis-MBP) and a TEV protease cleavage site). Each of these vectors have a polyhydrin (polH) promoter, an SV40 polyadenylation signal and a SspI restriction site in between them. A gene of interest can be cloned into the SspI restriction site creating an expression cassette with a polH promoter and an SV40 terminator. Figure 3.1 provides a schematic summary of the LIC cloning method.

The empty vectors were linearized by digestion with SspI. The linearized vectors were purified using the QIAGEN PCR purification kit. To generate sticky overhangs on the linearized vectors to make them amenable to ligation independent cloning (LIC), they were treated with LIC certified T4 DNA polymerase (purified in our laboratory) in the present of dGTP. The lack of dNTPs in the reaction activated the 3‘ - 5‘ exonuclease activity of the T4 DNA polymerase which then digests one strand of the DNA in the 3’ to 5’ direction until a dCMP nucleotide is encountered and then gets arrested in a futile cycle of addition and removal of dGMP nucleotide due to the presence of dGTP. This generates a linear vector with sticky ends compatible with LIC cloning (Gradia et al., 2017).

The primers for PCR amplification of ORFs of the INT subunits were designed to contain adaptor sequences called the LICV1 sequences. When the PCR products are treated with the

T4 DNA polymerase in the presence of dCTP, a sticky overhang complementary to the sticky overhang of the T4 DNA polymerase treated linearized vector is created.

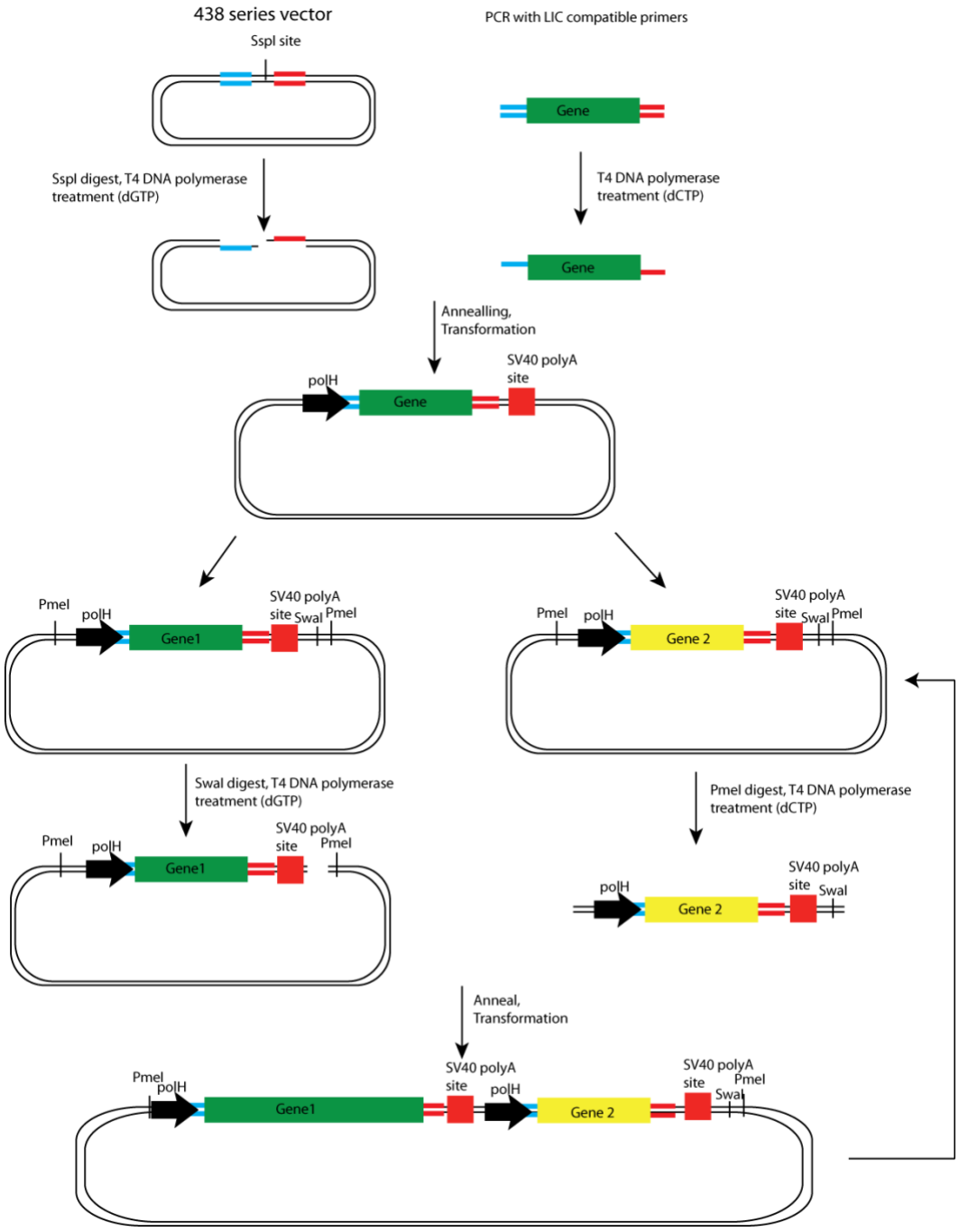


Figure 3.1. Schematic Summary of ligation independent cloning using the MacroLab system. The key steps are illustrated including the restriction sites of the enzymes; SspI, SwaI and PmeI. The polH (polyhedrin) promoter and the SV40 (simian virus 40) termination signal are also shown. Scheme was created based on information in Gradia et al., 2017.

The T4 polymerase treated linearized vector and insert were annealed at room temperature and transformed into XL1 blue chemically competent *E. coli.* The nick in the annealed product is repaired in *E. coli*. 3 - 5 single colonies were picked and cultured overnight, plasmid DNA isolated and analyzed by restriction digest using PmeI for 1 - 2 hrs followed by agarose gel electrophoresis. One or two positive clones (containing the bands corresponding to the right size of the vector and the insert) were sequenced by sanger sequencing (seqlab) for final validation. This approach was used for all subunits of INT except INTS1, INTS2, INTS10 and DDX26B.

3.1.7 Uracil excision cloning of INTS1 into MacroLab 438 series vectors

Uracil excision PCR provides a robust way to assemble several DNA fragments into one construct (Bitinaite & Nichols, 2009). It is particularly useful for cloning large DNA fragments that are challenging to amplify by PCR. This method was used to assemble INTS1 into the 438 series vector.

To achieve this, the ORF of INTS1 was divided into 4 DNA fragments and purchased from IDT (Integrated DNA Technologies). Each fragment was amplified with PCR primers containing 6 nt overlap to the adjoining fragment and a deoxyuridine placed at the boundary of the overlapping regions. The PCR reactions were performed using Phusion U hot start DNA polymerase (Thermo Fisher Scientific) which can permit extension from the deoxyuridine containing primer. The first and last fragments had overlaps with the polH promoter and SV40 polyadenylation signal-containing ends of the linear vector. The 438 vectors were linearized by PCR amplification from the SspI site. The first primer was designed to contain a 6 nt overlap with the first fragment of the INTS1 (N-terminus) with a deoxyuridine at the boundary whiles the right primer had a 6 nt overlap with the fourth fragment of INTS1 (C-terminus) with a deoxyuridine in the boundary region. This arrangement ensured that the ORF is correctly assembled in the 438 vector which will guarantee expression of the correct protein. The PCR products of the fragments and the vector backbone were treated with the USER enzyme mix (Thermo Fisher Scientific) according the manufacturer's protocol and transformed into XL1 blue *E. coli.* The plasmid DNA was than extracted from overnight cultures prepared from single colonies in LB medium supplemented with ampicillin. Successful assembly of INTS1 ORF into a 438 vector was assessed by restriction analysis with PmeI and sanger sequencing using sequencing primers which cover the full length of the INTS1 ORF.

3.1.8 Circular polymerase extension cloning (CPEC) for assembly of INTS2, INTS10 and DDX26B from DNA fragments into MacroLab 438 vectors

To obtain the ORFs for INTS2 and DDX26B, synthetic DNA fragments for the full ORFs for each gene were purchased and assembled into the 438 series vectors using Circular Polymerase Extension Cloning (CPEC)(Quan & Tian, 2009) (Figure 3.2).

Each of INTS2 and DDX26B ORFs were divided into 2 DNA fragments. The first fragment for each gene corresponding to the N-terminal half of the protein had a 32 nt sequence from the 438 series vector when it is linearized with SspI. This created a 60 ^{o}C melting temperature homology region between the first fragment of each gene and the upstream region of the vector (Figure 3.2). Similar homology regions were created between the first and second fragments, the second and the downstream end of the linearized vector. To assemble the full-length gene from its fragments into a vector, a 50 µl PCR mix was prepared containing 1x Q5 DNA polymerase buffer (NEB), 150 ng each of the INTS2 or DDX26B fragments, 200 ng of SspI linearized 438 vector, 200 µM dNTPs (NEB), and 0.02 U/µl of Q5 DNA polymerase (NEB). After initial denaturation for 30 s at 98 ^{o}C, the reaction was cooled 55 ^{o}C for annealing and extension was done at 72 ^{o}C for 4 mins. This was repeated for 26 cycles in a thermos cycler. At the end of the cycles, 40 µl of the reaction was transformed into chemically competent XL1 blue *E. coli.* Plasmid DNA was isolated from the single colonies that was cultured overnight in LB supplemented with ampicillin. Assembly of ORFs were assessed by restriction digestion with PmeI and positive clones were sequenced using sequencing primers covering the full length of the ORF. For INTS10, I purchased a construct with a truncated cDNA missing amino acid 1 to 78 and it was cloned into a 438 vector by the standard LIC cloning. To generated a CPEC compatible fragment for this deletion mutant of INTS10, the shortened cDNA was amplified from the vector with primers that created a 60 ^{o}C -melting-temperature overlap between the C-terminal region and the downstream end of the vector. A DNA fragment corresponding to amino acids 1 to 89 of INTS10 which encompasses the missing region, an additional 32 nt overlap with the upstream region of the SspI linearized 438 vector and a 33 nt overlap with the rest of the INTS10 ORF was purchased from IDT. A CPEC reaction was set up including the linearized vector, and the two fragments of INTS10. Successful cloning of the full length INTS10 was verified by restriction digestion and Sanger sequencing from a single colony as described for INTS2 and DDX26B.

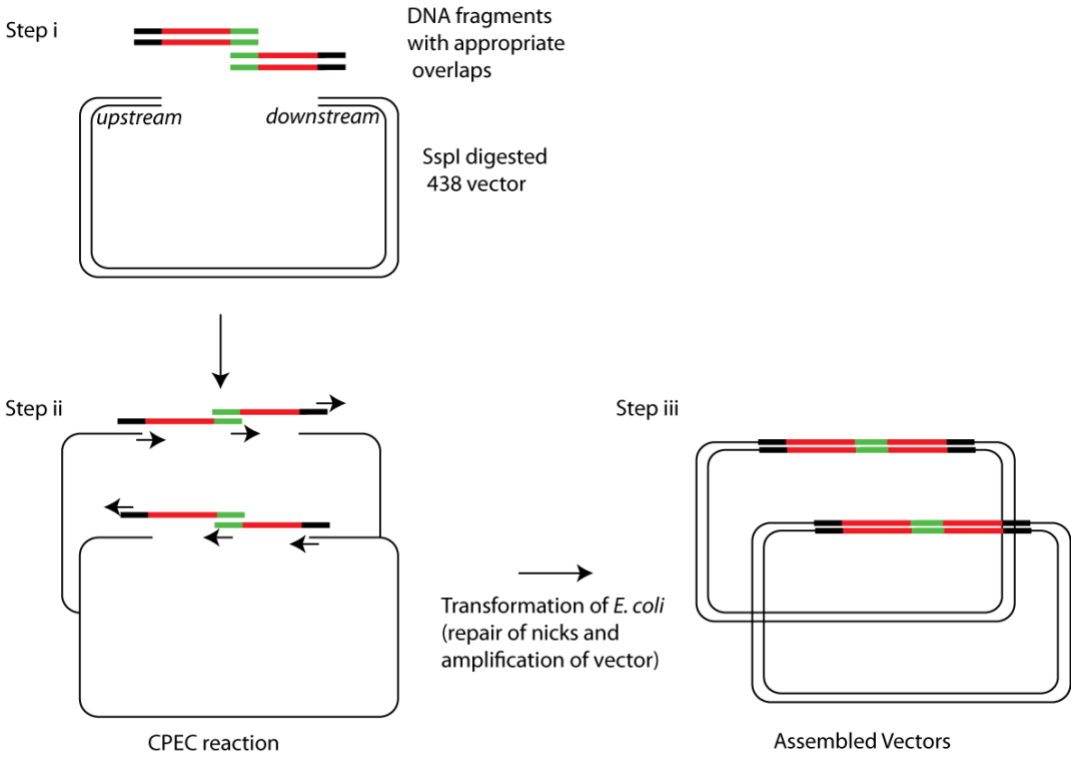


Figure 3.2. Schematic illustration of CPEC cloning strategy. DNA fragments of gene of interest is illustrated by a red lines. Homology regions between two DNA fragments of gene of interest is colored green and homology region between cDNA fragments and the linearized vector are shown in black. In step ii, arrows show direction of extension by DNA polymerase in the PCR reaction.

3.1.9 Cloning of multiple genes into one construct for co-expression

The MacroLab 438 vectors have a restriction site of PmeI flanking a cloned expression cassette and a SwaI site downstream of the SV40 polyadenylation signal and before the PmeI restriction site (Figure 3.1). It is possible to combine two vectors containing expression cassettes into a polyprotein co-expression vector. To achieve this, the 'acceptor vector' was linearized by digestion with SwaI and sticky ssDNA overhangs were generated by treating the linear 'acceptor vector' with T4 DNA ligase in the presence of dGTP. The expression cassette from the 'donor vector' was cut out by digestion with PmeI. The DNA fragment was separated on agarose gel and fragment of interest was purified by gel extraction. ssDNA overhangs complementary to the overhangs of the linearized T4 DNA polymerase treated 'acceptor vector' were generated by treating the gel purified expression cassette with T4 DNA polymerase in the presence of dCTP. The two fragments were annealed and transformed into *E. coli* for repair of the nick in the annealed fragments creating a multi protein expression vector in which each gene has its own promoter and termination sequence. This approach allowed for the creation of co-expression vectors containing up to 8 subunits of INT and ~34,000 bp in size.

3.1.10 Cloning of a heteropentameric (INTS3/5/6/8-DDX26B) and heteroheptameric (INTS2/3/5/6/7/8-DDX26B) subcomplexes

Cloning of co-expression constructs larger that 30,000 bp in size was problematic. Due to the size of this construct, the transformation efficiency of the standard XL1 blue chemically competent cells used was extremely low. Additionally, the cells turn to remove one or two subunits from the constructs (frequently the ORF for INTS5). To avert these problems, the 'NEB stable' chemically competent cells were employed for cloning this construct and generally any construct bigger than 25,000 base pairs in size. Briefly, LIC qualified T4 DNA polymerase-treated SwaI digest of vector containing ORFs for INTS3, INTS6 and DDX26B was annealed with LIC qualified T4 DNA polymerase-treated PmeI digested insert of INTS5/8 heterodimer to create the heteropentameric subcomplex. Variants of this construct where INTS3 or INTS6 was tagged were created.

Constructs expressing INTS2, INTS5, INTS7 and INTS8 were also generated by the standard LIC cloning with affinity tag on INTS5 or INTS7 or no affinity tag. An expression constructs harbouring ORFs for INTS2, INTS3, INTS5, INTS6, INTS7, INTS8 and DDX26B (core-INT) was created by combining INTS2-INTS5-INTS7- INTS8 expression vector and INTS3-INTS6-DDX26B expression cassette using LIC cloning with the NEB stable cell lines. Variants of this construct with no affinity tagged subunit or 6xHis-MBP affinity tag on INTS5 or INTS6 or INTS7 were created.

3.1.11 Transformation of chemically competent cells

In order to amplify empty cloning vectors, or for LIC cloning of ORFs into vectors, or for the assembly of multiple expression cassettes into a polyprotein expression construct, the ability of chemically competent *E. coli* to take up, repair, maintain and amplify plasmid DNA was exploited. To transformed chemically competent *E. coli* (XL1 Blue or NEB stable), 1 ng of vector plasmid DNA or 10 µl of annealed T4 DNA polymerase-treated vector and insert was added to 100 µl of freshly thawed cells and incubated on ice for 20 - 30 min. Cells were heat-shock at 42 ^{o}C for 45 s and then kept on ice. Transformed cells were recovered in 1 ml of LB media (or the provided media in case of NEB stable cells) at 37 ^{o}C for 1 hr shaking at 1000 rpm on an Eppendorf thermomixer. For antibiotic selection of transformed cells, recovered cells were collected by spinning at 3000 rpm in an Eppendorf table top centrifuge for 3 min and plated onto LB agar supplemented with ampicillin (or appropriate antibiotic) and incubated overnight at 37 °C. Single colonies were picked from the overnight plate and grown overnight

in LB broth supplemented with ampicillin. Plasmid was isolated from each clonal culture using the QIAGEN miniprep kit following the manufacturer's protocol.

3.1.12 Insertion of expression cassettes into baculovirus shuttle vectors (bacmids)

The DH10EMBacY strain of bacteria contain two accessory plasmids, pMON7124 and bMON14272. The pMON7124 contain among other DNA sequences a tetracycline resistance gene and the bacterial Tn7 transposon required for transposition of a DNA sequence flanked by TN7 recognition site (TN7L and TN7R) (Waddell & Craig, 1989). The bMON14272 is a modified version of the baculovirus shuttle vector which has a mini Tn7 attachment site (mini – attTn7) inserted inside a LacZα gene, a kanamycin resistance marker and all genomic regions of *Autographa californica nuclear polyhedrosis virus* (*AcNPV*, a species of baculoviridae) that are necessary for viral DNA stability and propagation in infected host cells. The chiA gene encoding chitinase and v-cath gene encoding cysteinase were removed for enhanced protein production. The shuttle vector has additionally all sequences needed for the maintenance and amplification of the bacmid in bacteria. It contains an eYFP gene under polH promoter for tracking the protein expression upon transfection of the host cells (Berger et al., 2004; Luckow et al., 1993). The MacroLab 438 vectors used in this study have a Tn7L and Tn7R sequences flanking the region which includes cloned expression cassettes and gentamycin resistance marker. This allows transposition of the expression cassette(s) in the 438 series into the mini – attTn7 in bMON14272 when transformed into DH10EMBacY *E. coli*. Successful transposition results in destruction of the LacZα gene preventing the breakdown of x-gal and resulting in white colored colonies. The insertion of an expression cassette was monitored by blue - white colony selection by plating transformed cells on LB plates supplemented with 100 ug/ml x-gal, 1mM IPTG and 50 µg/ml gentamycin. To achieve this, 1 µg (2 µg for constructs bigger that 25,000 base pairs) of expression vector containing expression cassette(s) of interest was transformed into DH10EMBacY electrocompetent cells. Cells were incubated on ice for 15 min after DNA was added and then transferred into a BIORAD Gene Pulser®/Micropulser™ electroporation cuvettes (0.1 cm gap) (BioRad). For electroporation, one pulse (25 µF, 1.8 kV) was applied. Transformed cells were immediately recovered in 1 ml LB media and incubated in 13 ml tubes (Sarstedt) at 37 ^{o}C for 5 h or overnight shaking at 150 rpm. An amount of 50 – 100 µl of recovered cells was plated on LB agar plates supplemented with gentamycin, x-gal and IPTG for blue-white selection. Few white colonies were re-streaked on fresh LB plates supplemented with gentamycin, x-gal and IPTG to avoid false positives.

3.1.13 Isolation of bacmid DNA by alkaline lysis and isopropanol precipitation

To isolate bacmid, 5 ml (or 10 ml in the case of construct > 30 kpb) overnight culture was made from selected and re-streaked white colonies in LB media supplemented with 50 µg/ml gentamycin. Cells were harvested by centrifugation at 4000 rpm for 10 min in Eppendorf 5810 R centrifuge. Cell pellet was resuspended in 250 µl of buffer P1, lysed by adding 250 µl of buffer P2 and the lysate was neutralized by adding 350 µl of buffer N3 (QIAprep SpinMiniprep kit). Cell debris were removed by centrifugation (10 min, 21,000g) at 4 ^{o}C. The supernatant was transferred into a fresh 1.5 ml Eppendorf tube and mixed with 700 µl of isopropanol. For efficient precipitation, the mixture was incubated overnight at -20 ^{o}C or for 1 - 2 hours at -80 ^{o}C. DNA was pelleted by centrifugation (21,0000g, 30 min) at 4 ^{o}C. The DNA pellet was washed with 500 µl of 70% (v/v) ethanol and spun at 21,000g for 10 min. Finally, 30 µl of 70% (v/v) ethanol was added on top of the DNA and stored at -20 ^{o}C until use.

3.2 Insect cell culture

The baculovirus insect cell recombinant protein expression system was used for the production of proteins in this study because of its documented success in the production of multiprotein complexes (Bieniossek et al., 2013; Trowitzsch et al., 2010; Zhang et al., 2013). The SF9 cells were used for the production of initial baculoviruses (V_0) harboring protein of interest. The SF21 cells were used to amplify the V_0 to V_1 and for expression tests and the Hi5 cells were used for large scale expression of proteins by infecting them with baculoviruses generated using SF9 and SF21. The cells were maintained in a shaking incubator (60 rpm, 27 ^{o}C) with limited exposure to light. Stocks were maintained at cell density of $0.5x10^6$ – $1x10^6$ cells/ml. Cell density and size were monitored daily by measuring in a CASY cell counter and analyzer (OMNI Life Science) and diluted back to the appropriate density. For SF9 and SF21, the final volume of culture was always kept at 10% of the volume of the cell culture flask whiles Hi5 were up to 20% of the volume of the flask to allow adequate aeration. Cell cultures were kept under sterile conditions with aseptic techniques to avoid microbial contamination.

3.2.1 Transfection of SF9 with bacmid for V_0 production

The 30 µl ethanol covering the bacmid DNA (from section 3.1.13) was removed and the DNA pellet was dried in the sterile hood for 5 -10 min to remove residual ethanol. The DNA was dissolved in 20 µl sterile dH_20 by adding it on top of it without pipetting as this may break the DNA. For complete resuspension, the mixture was incubated for 10 - 20 minutes until pellet was fully dissolved. A master mix for transfection containing 10 µl Xtreme Gene 9 transfection agent (Roche) and 100 µL SF9 media (Gibco® Sf-900™ III SFM) was prepared for each

bacmid transfection. Once the DNA was completely dissolved, 200 µl Sf9 media and 100 µl of transfection agent master mix was added to each bacmid and incubated for 60 minutes.
Two wells of a 6-well plate were seeded with 3 ml of SF9 cells at a cell density of $1x10^6$ cells/ml for each bacmid. The plate was incubated at 27 ºC for at least 20 min before the end of the 60 min incubation of the bacmid and the transfection agent. A volume of 150 µl of the transfection master mix of each bacmid was added on top of the cells in each well (two wells for each bacmid). For each construct, bacmid was prepared from two independent verified white colonies. Plates were incubated at 27 ºC and after 48 hrs, checked for successful transfection under the fluorescent microscope (the infected cells became fluorescent due the presence of eYFP). The media on top of the cells (containing the baculovirus) were harvested after 72 hrs and stored at 4 ºC as the V_0 baculovirus.

3.2.2 Production of V_1 baculoviruses

In order to generate a highly infective baculovirus solution for large scale protein expression, the V_0 baculoviruses were amplified in SF21 cells. To this end, 0.5 - 1 ml of V_0 baculovirus was added to 25 ml of SF21 cells at a concentration of $1x10^6$ cells/ml. Cells were monitored daily by measuring their density and size (diameter) using the CASY cell counter and analyzer. The cell diameter usually increases as cells swell-up upon infection. The cells were maintained at a density of $1x10^6$ cells/ml until growth was arrested by the infection. The V_1 was harvested 24 - 48 hrs after the day of arrest which usually coincided with a drop-in cell viability to below 90%. To harvest cells, the culture was transferred into a sterile 50 ml falcon and spun at 250xg for 15 min. The supernatant containing the V_1 baculovirus was collected and stored at 4 ºC while the cell pellet was used for Ni or Amylose affinity pull-down to monitor protein expression.

3.3 Methods for protein production, purification and analysis

All assays involving proteins were done on ice or at 4 ºC unless otherwise stated. Cell lysis was done by sonication using a BRANSON Digital sonifier. Protein concentration was determined using UV absorbance values at 280 nm and the predicted molar extinction coefficient, ε using the Lambert-Beer law. ε was calculated for each protein or complex from the amino acid sequence using the online tool at https://web.expasy.org/protparam/.

3.3.1 Pulldown Assay for protein expression and interaction test

Amylose affinity or Nickel affinity pull-down was used to assess the expression and interaction between subunits and subcomplexes of the Integrator complex. Cell pellets from harvested V_1

baculoviruses or expression tests were suspended in 1 ml of lysis buffer containing 20 mM HEPES 7.4, 150 mM NaCl, 10% glycerol, 1 mM DTT or BME or TCEP (30 mM imidazole was added in the case of Ni affinity pulldown). Cells were lysed by sonication with 10% continuous pulse for 10s. Cell debris were removed by centrifugation at 15000 rpm for 30 min in an Eppendorf table top centrifuge operated at 4 ^{o}C. An input sample was taken and the rest of the supernatant was applied to 150 µl of pre-equilibrated amylose resin (NEB) or Ni-NTA agarose (QIAGEN) (washed once with water and three times in lysis buffer). The beads were incubated on a rotating wheel (12rpm) at 4 ^{o}C for 30 and 20 min to allow binding to amylose resin or Ni-NTA agarose respectively. After centrifugation at 2000 rpm for 1 min, the supernatant containing unbound proteins was discarded and the beads were washed 3 - 4 times with 1 ml of lysis buffer. Bound protein was then eluted with 100 - 200 µl of elution buffer (100 mM Maltose in lysis buffer for amylose affinity and 500 mM Imidazole in lysis buffer for Ni affinity pulldown). About 10 - 20% of eluted fraction and the aliquots of 0.2 µl of the input were analysed by LDS-PAGE.

3.3.2 LDS-PAGE Electrophoresis

Protein biochemistry results were analysed using denaturing polyacrylamide gel electrophoresis (PAGE). To this end, appropriate fractions of expression tests or protein-protein interaction assays or fractions of protein purification procedures were denatured by adding 1 - 2x of LDS sample buffer (Thermo Fisher Scientific) which denatures proteins and confers the same negative charge on them. An appropriate amount of the denatured proteins was resolved on a NuPAGE 4 - 12 % bis tris precast gel (Thermo Fisher Scientific) using 1x MOPS buffer (for better resolution of proteins bigger than 80 kDA) or 1x MES buffer for proteins with molecular weight smaller than 70 kDa. Gels were run at 180 - 200 V for 45 - 60 min and stained with Instant Blue (Expedeon) for visualisation.

3.3.3 Western blotting

The INT subunits have narrow size distribution with most of them having a molecular weight between 100 and 150 kDA (INTS2, INTS3, INTS4, INTS5, INTS6, INTS8 and DDX26B). This makes LDS-PAGE insufficient when analysing the expression or interactions of sub complexes involving multiple of these subunits. Western blot was applied for the detection of specific INT subunits using their specific antibodies. Protein bands from LDS-PAGE were transferred onto a nitrocellulose membrane using Trans-Blot Turbo Transfer System™ from BioRad. Membrane was blocked by incubating with 2% milk in PBS-T (1x PBS and 0.1% Tween20) on a shaker at RT for 1 hr. Specific antibody for protein of interest was added to the

blocked membrane and incubated at 4 ºC on shaker overnight or at room temperature for 2 - 3 hrs. Unbound antibody was discarded and the membrane was washed 3 times with 1x PBS-T. The membrane was then incubated with specific secondary antibody with reactivity to the primary antibody used. Incubation with the secondary antibody was done for 1 - 2 hrs at room temperature. The secondary antibodies were conjugated with horseradish peroxidase (HRP). Free secondary antibody was discarded and the membrane was washed 3 times with 1x PBS-T. The membrane was covered with equal amounts (0.5 - 1 ml) of each of peroxide and an enhancer solution (Thermo Fisher Scientific). The blot was then imaged for chemiluminescence using an Intas imager.

3.3.4 Identification of interacting subunits by systematic co-expression of subunits

3.3.4.1 INTS5/8 heterodimer

The ORF of INTS8, INTS10 and INTS12 were cloned into 438-A vector and INTS5 was cloned into 438-C which contain an N-terminus 6xHis-MBP tag. The four vectors were further cloned by LIC resulting in a four-subunits co-expression construct containing 6xhis-MBP-INTS5, INTS8, INTS10 and INTS12. For protein expression, V_0 and V_1 baculoviruses were produced for this construct (Section 3.2). The cell pellet from the V_1 baculovirus was used for amylose affinity pulldown (Section 3.3.1) followed my mass spectrometric identification of proteins in the elution fraction of the pulldown.

3.3.4.2 INTS/14 heterodimer

The ORFs of INTS14 was cloned into 438-A vector and INTS13 was cloned into 438-C which contain an N-terminus 6xHis-MBP tag. The vectors were combined by LIC into a two-subunits expression construct containing 6xhis-MBP-INTS13, and INTS14. V_0 and V_1 baculoviruses were produced for this construct. The cell pellet from the V_1 baculovirus was used for amylose affinity pulldown followed my mass spectrometric identification of proteins in the elution fraction of the pulldown.

The experiment described above (sections 3.3.4.1 and 3.3.4.2) was done using various combinations of INT subunits/subcomplexes varying the tagged subunit in order to test their direct protein-protein interaction

3.3.5 Co-infection and partial purification of full Integrator complex from three baculoviruses

3.3.5.1 Co-infection

Based on the information accumulated from systematic co-expression and the available literature (Section 1.3), three polyprotein expression constructs were made each containing

interacting subunits of INT where possible. Subunits, INTS1, INTS3, INTS6 and INTS12 were in one vector with a 6xHis-MBP affinity tag on INTS1. Subunits, INTS5, INTS8, INTS10, INTS13, INTS14 and DDX26B were also in one vector with a 6xHis-MBP affinity tag on INTS5. DDX26B was added to this construct because it was convenient and faster to clone into it. The third vector contained INTS2, INTS4, INTS7, INTS9 and INTS11 with a 6xHis-MBP on INTS11. INTS2 and INTS7 did not interact with the INTS4/9/11 heterotrimer but were clone together for convenience. V_0 and V_1 baculoviruses were produced for each expression construct and the expression of at least the affinity tagged subunits and its associated interaction partners were assessed via amylose affinity pulldown and LDS-PAGE analysis. This was important in order to know that all baculoviruses are viable. Protein co-expression from each baculoviruses was induced by co-infecting 600ml of Hi5 cells at $1x10^6$ cells/ml with 100 µl of each baculovirus. Cell viability, density and diameter were monitored in 24 – hour interval and cells were diluted when density exceeded $1x10^6$ cells/ml. Cultures were harvested after 72 hours, the supernatant was discarded and the cell pellet was resuspended in 60 ml of resuspension buffer (20 mM HEPES pH 7.4, 300 mM NaCl, 5% glycerol, 1 mM DTT, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine), flash frozen in liquid nitrogen and stored at -80 ^oC.

3.3.5.2 Partial purification of full INT.

The term “partial purification” is used here because three 6xHis-MBP affinity tagged subunits of INT were involved in the purification and the stoichiometry of the resultant complex cannot be guaranteed. Pellets were quickly thawed in water bath at room temperature and transferred on ice. Lysis was done by sonication (30% amplitude for 4 min 0.4 s pulse on and 0.6 s pulse off). After sonication, lysate was cleared by centrifugation at 87,207xg for 30 min at 4 ^oC and the supernatant was transferred into an ultracentrifuge tube and further spun for 1 hr at 45000 rpm at 4 ^oC using a Ti45 rotor in an ultracentrifuge (Beckman Coulter). The supernatant was then filtered through a 0.8 um syringe filter for additional clearance.

A 20ml of 50% amylose slurry (NEB) was washed 3x with ddH_2O water and then another 3x with lysis buffer. Each wash and equilibration step were done in a 50 ml falcon tube and all centrifugations were done at 190xg for 5 min. The cleared lysate was incubated with the equilibrated slurry for 60 min rotating at 4 rpm at 4 °C to bind the complexes through the tagged proteins. After binding, beads were washed 2x with 30 ml lysis buffer and then 2x with low salt buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 5% glycerol, 1 mM DTT). After the washing steps, bound protein was eluted 4x with 20ml of 100 mM maltose in low salt buffer and 5 µl of each elution fraction and 5 µl of 1 in 30 dilution of the unbound and wash fractions were

analyzed on LDS-PAGE. Elution fractions containing protein of interest were pulled and applied to a 5 ml HiTrap Q column equilibrated in low salt buffer. The column was washed with 50 ml (10 column volumes) of the low salt buffer and then eluted with a gradient from 0 to 100% high salt buffer (20 mM HEPES pH 7.4, 850 mM NaCl, 5% glycerol, 1 mM DTT) over 18 column volumes. A 10 µl sample from each fraction of the elution peak was analysed by LDS-PAGE. Peak fractions were pulled and concentrated and a 100 µg portion was removed and crosslinked with 1 mM BS3 for 30 min on ice. The crosslinking reaction was quenched by adding 50 mM NH_4HCO_3 and the proteins were precipitated using 4 volumes of acetone. The precipitates were dissolved in 4 M urea/50 mM NH_4HCO_3 followed by identification of crosslinked peptides by mass spectrometry (XL-MS). An aliquot of the partially purified complex (~100 µg) was used for analytical gel filtration by applying it to a Superose 6 Increase 3.2/300 column (GE life sciences) equilibrated in low salt buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 5% glycerol, 1 mM DTT). A sample from the main peaks of the gel filtration run (0.05 - 0.1 mg/ml) was stabilized by crosslinking with 0.1% glutaraldehyde for 10 min on ice. The crosslinking reaction was quenched by adding 100 mM Tris - HCl (pH 8) and and 10 mM aspartate. The stabilized complex was analysed by negative stain electron microscopy.

3.3.6 Identification of interaction partners of INTS3/6-DDX26B heterotrimer by co-infection with different subcomplexes/subunits of INT

The XL-MS results from section 3.3.5 show a strong interaction between INTS3 and DDX26B leading to the identification of the heterotrimeric subcomplex of INTS3, INTS6 and DDX26B. V_1 baculoviruses expressing INTS3/6-DDX26B (no affinity tag) as well as 6xHis-MBP tagged INTS1, 6xHis-MBP tagged INTS12, INTS2/7 heterodimer with a 6xHis-MBP tag on INTS7, INTS5/8 heterodimer with a 6xHis-MBP tag on INTS5, INTS4/9/11 heterotrimer with 6xHis-MBP tag on INTS11 and INTS10/13/14 heterotrimer with 6xHis-MBP tag on INTS13 were produced. To identify which subcomplexes/subunits have direct protein–protein interaction with INTS3/6-DDX26B heterotrimer, 25 ml of Hi5 cells at $1x10^6$ cells/ml were co-infected with 12.5 µl of the V_1 baculovirus expressing INTS3/6-DDX26B and 12.5 µl of another V_1 baculovirus expressing the tagged version of the subunits/subcomplexes of INT mentioned above. Cell density, diameter and viability in each culture were measured after 24 hours and the cultures were diluted appropriately to bring cell density to $1x10^6$ cells/ml and cultures were grown for 48 hrs. After measuring cell density, $25x10^6$ cells were harvested from each culture by centrifugation at 385xg for 15 min at 4 ^oC. Cells were resuspended in a lysis buffer containing 20 mM HEPES 7.4, 150 mM NaCl, 10% glycerol, 1 mM TCEP 10 µM $ZnCl_2$, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine and used

for pulldown as described in section 3.3.1 with 100 µl of 50 % amylose resin. A sample of the elution (10 µl) and 0.5 µl of input for each sample were analyzed by LDS-PAGE and western blot with a rabbit anti INTS3 antibody (Abcam).

3.3.7 Sucrose Density Gradient Centrifugation

Sucrose density centrifugation is a gentle method for separating protein complexes according to their sizes in their native state. It is particularly useful for fragile complexes that might fall-apart under gel filtration conditions and for huge complexes that elute close to the void volume and cannot be separated from aggregated proteins by gel filtration. This method was used to separate the various subcomplexes of INT as well as complexes between INT and Pol II elongation complexes. A light sucrose solution (15 % w/v) was made containing sucrose and all other components of the protein buffer of choice as well as a heavy solution prepared similarly but contained 30% (w/v) sucrose. A 15 - 30% linear sucrose density gradient was prepared by layering 2 ml of light solution (15% sucrose) on top of 2 ml of heavy solution. The two layers were mixed into a linear gradient using BioComp 108 Gradient master. 50 - 300 µg of protein or complex (in 50 - 200 µl) was loaded on top of the gradient after appropriate volume was removed from the top to accommodate the sample. The gradient was then spun at 32,000 rpm for 16 hours at 4 °C. The gradient was fractionated into 200 µl fractions from the top and stored on ice. Sample from selected fractions were analysis by LDS-PAGE.

3.4 Expression and purification of Proteins

3.4.1 Expression and Purification of INTS4/9/11 heterotrimer

Expression

The ORF of INTS4 and INTS9 were cloned into the MacroLab 438-A vector and INTS11 was cloned into 438-C which has an N-terminus 6xHis-MBP tag. The three vectors were combined by LIC into a three subunits polyprotein expression construct containing 6xhis-MBP-INTS11, INTS4 and INTS9. V_0 and V_1 baculoviruses were produced for this construct and expression of all three proteins was validated by pulldown experiment using the cell pellet from the V_1 baculovirus. For protein expression, 2x 600 ml of Hi5 insect cells at $1x10^6$ cells/ml was infected with 300 µl of the V_1 baculovirus in a 3 l flask. Cell density, viability and size were monitored in 24 hrs intervals and cells were diluted when the density went higher than $1x10^6$ cells/ml. Normally, the cells divided once in 24 hours and then got arrested by the baculovirus indicated by increase in cell diameter and stationary growth. Cultures were harvested when viability drops below 85% (usually within 72 hrs) by centrifugation at 238xg for 30 min. The supernatant was discarded and the cell pellet from 1 l culture was resuspended in 35 ml of lysis buffer (20

mM HEPES pH 7.4, 300 mM NaCl, 30 mM imidazole 10% glycerol, 10 mM BME 10 µM $ZnCl_2$, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine). The harvested cells were flash frozen in liquid nitrogen and store at -80 ^oC until purification.

Purification

The frozen pellets were thawed in a water bath at 25 ^oC and lysed by sonication with 30% amplitude for 2 min with 0.6 s pulse on and 0.4 s pulse off. The lysate was cleared by centrifugation at 87,207xg for 1 hr and filtered with a 0.8 µm syringe filter. The cleared lysate was applied to a pre-equilibrated 5 ml HisTrap column at a flow rate of 1.5 ml/min and the column was washed with 100 ml of wash buffer (20 mM HEPES pH 7.4, 300 mM NaCl, 30 mM imidazole 10% glycerol, 10 mM BME 10 µM $ZnCl_2$). A self-packed amylose column with a total bed volume of 15 ml pre-equilibrated in wash buffer was connected in tandem to the 5 ml HisTrap column and the bound protein was eluted from the HisTrap column unto the amylose column using Ni elution buffer containing 20 mM HEPES pH 7.4, 300 mM NaCl, 10% glycerol, 10 mM BME 10 µM $ZnCl_2$, 250 mM Imidazole. The HisTrap column was detached and the amylose column was washed with 100 ml of wash buffer before eluting with wash buffer supplemented with 100 mM maltose. The fractions containing the heterotrimeric complex were pulled and treated overnight with catalytic amounts of 6xHis-TEV protease and lambda phosphatase (home made in our laboratory) in the presence of 1 mM $MnCl_2$ to remove the affinity tag and remove post-translational phosphorylations from the insect cells. The protein was applied again to an equilibrated 5 ml HisTrap in a reverse nickel affinity step to remove the TEV protease, the affinity tag and proteins with affinity tag. The unbound protein was collected and concentrated using a 100 kDa cut-off Amicon ultra centrifugal filter (MERCK Millipore). The protein was purified on a superpose 6 increase gel filtration column equilibrated in gel filtration buffer containing 20 mM HEPES pH 7.4, 300 mM NaCl, 10% glycerol, 10 mM BME 10 µM $ZnCl_2$. The elution peak was concentrated and stored is 5 µl aliquots at -80 ^oC. In order to test interaction with other subunits/subcomplexes via *in vitro* pulldown experiments, INTS4/9/11 heterotrimer was also purified without removing the affinity tag.

3.4.2 Expression and Purification of INTS10/13/14 heterotrimer

Expression

The ORF of INTS10 and INTS14 were cloned into the MacroLab 438-A vector and INTS13 was cloned into 438-C carrying a 6xHis-MBP tag or 438-B carrying a 6xHis tag. The three vectors were combined by LIC into a three subunits polyprotein expression construct containing

6xhis-MBP-INTS13, INTS10 and INTS14 or 6xhis-INTS13, INTS10 and INTS14. V_0 and V_1 baculoviruses were produced for these constructs and expression of all three proteins was validated by pulldown experiment using the cell pellet from the V_1 baculovirus. Protein expression and harvesting of cultures was as described for the INTS4/9/11 heterotrimer (section 3.4.1). The cell pellet from 1 l culture was resuspended in 35 ml of lysis buffer (20 mM HEPES pH 7.4, 200 mM NaCl, 10% glycerol, 10 mM BME 30 µM imidazole, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine). The harvested cells were flash frozen in liquid nitrogen and store at -80 ^{o}C until purification.

Purification

Frozen cell pellets expressing the recombinant subcomplex was thawed in a water bath at 25 ^{o}C. Cell lysis and clearance of the lysate was as described for the INTS4/9/11 heterotrimer (section 3.4.1). The cleared lysate was applied to a pre-equilibrated 5 ml HisTrap column at a flow rate of 1.5 ml/min and contaminating proteins were removed by washing the column with 100 ml of the lysis buffer. A self-packed amylose column with a total bed volume of 15 ml pre-equilibrated in lysis buffer was connected in tandem to the 5 ml HisTrap column and the bound protein was eluted from the HisTrap column onto the amylose column using Ni elution buffer containing 20 mM HEPES pH 7.4, 200 mM NaCl, 10% glycerol, 10 mM BME, 250 mM Imidazole. The HisTrap column was detached and the amylose column was washed with 100 ml of lysis buffer before eluting with amylose elution buffer containing 20 mM HEPES pH 7.4, 200 mM NaCl, 10% glycerol, 10 mM BME, 100 mM maltose. For purification from baculovirus expressing 6xhis-INTS13, INTS10 and INTS14, protein was eluted from the HisTrap column with a gradient (0-100%) of Ni elution buffer containing 20 mM HEPES pH 7.4, 200 mM NaCl, 10% glycerol, 10 mM BME, 500 mM Imidazole. Fractions containing the heterotrimeric complex were pulled and treated with catalytic amounts 6xHis-tagged TEV protease and lambda phosphatase (home made in our laboratory) in the presence of 1 mM $MnCl_2$ to remove the affinity tag and potential post-translational phosphorylations. The protein was applied again to an equilibrated 5 ml HisTrap in a reverse nickel affinity step to remove the TEV protease, the affinity tag and proteins with affinity tag. The unbound protein was collected, concentrated using a 100 kDa cut-off Amicon ultra centrifugal filter and applied to a Superose 6 Increase gel filtration column equilibrated in gel filtration buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 10 mM BME. The elution peak was concentrated and stored in 5 µl aliquots at -80 ^{o}C. In order to test interaction with other subunits/subcomplexes via affinity pull-down experiments, INTS10/13/14 heterotrimer was also purified without removing the 6xHis-MBP affinity tag.

3.4.3 Expression and purification of heteropentameric subcomplex (INTS3/5/6/8/-DDX26B)

Expression

A construct expressing the pentameric subcomplex of INTS3, INTS5, INTS6, INTS8 and DDX26B (INTS3/5/6/8-DDX26B) was created using LIC cloning (Section 3.1.10) with 6xHi-MBP affinity tag on INTS6. Baculovirus expressing this subcomplex was produced (V_0 and V_1) and expression was confirmed via V_1 amylose affinity pulldown. For protein expression, Hi5 insect cells at a density of $1x10^6$ cell/ml were infected with the V_1 baculovirus at 1:2000 (v/v) baculovirus to culture volume ratio. Cultures were monitored in 24-hr interval and diluted accordingly to ensure cell density did not exceed $1x10^6$ cell/ml. Cultures were harvested after 48 hours and cell pellets were suspended in a lysis buffer (35 ml/l of culture) containing 25 mM Tris-HCl pH 8, 200 mM NaCl, 10% glycerol, 10 mM BME 10 µM $ZnCl_2$, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine. Resuspended cells were flash frozen in liquid nitrogen and stored at -80 ^oC until purification.

Purification

Lysis of cells expressing the heteropentamer and clearance of the lysate to remove cell debris was as described for the INTS4/9/11 heterotrimer (section 3.4.1). The cleared lysate was applied to a preequilibrated 25 ml bed volume self-packed amylose column at a flow rate of 1.0 ml/min and the column was washed with 100 ml of wash buffer (25 mM Tris–HCl pH 8, 0.2 M NaCl, 10 mM BME, 10 µM ZnSO4, 10% glycerol). An equilibrated 5 ml HiTrap Q column was then connected to the amylose column and the bound protein complex was eluted from the amylose column using wash buffer supplemented with 100 mM maltose. The amylose column was disconnected and the HiTrap Q column was washed with 50 ml of wash buffer prior to elution with a gradient (0-100%) of high salt buffer (25 mM Tris–HCl pH 8, 1 M NaCl, 10 mM BME, 10 µM ZnSO4, 10% glycerol). The peak fractions were pulled, concentrated (100 kDa cut-off Amicon Ultra Centrifugal filter form Millipore) and applied to a Sephacryl S-300 gel filtration column equilibrated in gel filtration buffer (20 mM HEPES pH 8, 0.2 M NaCl, 10 mM BME, 10 µM ZnSO4, 10% glycerol). An aliquot from the elution peak was used for negative stain electron microscopy after fixation with 0.1% glutaraldehyde and the rest was used for pulldown assays.

3.4.4 Expression and purification of core-INT (INTS2/3/5/6/7/8-DDX26B)

Expression

Cloning of the core-INT subcomplex was described (Section 3.1.10) and expression was exactly as described for the heteropentameric subcomplex (INTS3/5/6/8-DDX26B) in section 3.4.3. Harvested Hi5 cells expressing the core-INT were resuspended in lysis buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 1 mM TCEP 10 µM $ZnCl_2$, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine) and stored at -80 ^oC until purification.

Purification

The purification strategy used was the same as for the heteropentameric subcomplex (INTS3/5/6/8-DDX26B) described above in section 3.4.3. Except that HEPES pH 7.4 was used in place of Tris-HCl pH 8 and the final gel filtration step was omitted. The protein was either purified further by a sucrose density gradient or was used directly for downstream experiments after concentration.

3.4.5 Reconstitution of 13-subunits subcomplex of INT and full INT

Freshly purified core-INT with 6xhis-MBP tag on INTS6 (100 µg) was immobilized on pre-equilibrated amylose beads by incubating for 1 hr at 4 ^oC on a rotating wheel (10 rpm). Beads were washed 3x with 800 µl of binding buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 1 mM TCEP 10 µM $ZnCl_2$) to remove unbound core-INT and divided into two parts of 100 µl. Purified INTS4/9/11 and INTS10/13/14 were added in 3 molar excess to one part of the immobilized core-INT and incubated for 1 hour at 4 ^oC on a rotating wheel (10 rpm). Beads were washed 4x with binding buffer and eluted with 40 µl of binding buffer containing 100mM maltose. 10 µl of LDS sample buffer was added to the elution and analyzed by LDS-PAGE.

The reconstitution of the full Integrator complex for crosslinking mass spectrometry was done similarly with the addition of partially purified INTS1 and INTS12 starting with 500 µg of core-INT. The crosslinking reaction was done as described before (Section 3.3.5.2) except that 3 mM BS3 was used and the reaction was incubated at room temperature.

3.4.6 Expression and purification of INTS1 and INTS12 interacting domains

Expression

The ORFs for INTS1 and INTS12 were cloned in 438-B and 438-A respectively by LIC. The truncation of these proteins was done using ‘round-the-horn’ PCR (see Section 3.1.2). The truncated variants in their respective 438 vectors were combined into a polyprotein expression vector harboring 6xHis-INTS1(1-294) and INTS12(1-194) by LIC. The production of V_0 and

V_1 baculoviri was done in SF9 and SF21 cells respectively and protein expression was done in Hi5 insect cells. The Hi5 cells were monitored and the cells were harvested after 72 hrs and cell pellets were resuspended in 20 mM HEPES pH 7.4, 150 mM NaCl, 1 mM TCEP, 10 µM $ZnCl_2$, 10% glycerol (35 ml/liter of hi5) and flash frozen and stored at -80 ^oC.

Purification

The cell pellets from 1.4 l of Hi5 cells expressing the INTS1(1-294)/INTS12(1-194) heterodimer were thawed in a water bath at room temperature and lysed by sonication with 30% amplitude for 2 min with 0.6 s pulse on and 0.4 s pulse off in a sonicator. The lysate was cleared by centrifugation at 87,207xg for 1 hr and filtered with a 0.8 µm syringe filter. The cleared lysate was applied to a preequilibrated 5 ml HisTrap column at a flow rate of 1.5 ml/min and the column was washed with 50 ml of a wash buffer (20 mM HEPES pH 7.4, 0.4 M NaCl, 1 mM TCEP, 10 µM $ZnCl_2$, 10% glycerol) followed by 25 ml of lysis buffer (20 mM HEPES pH 7.4, 0.15 M NaCl, 1 mM TCEP, 10 µM $ZnCl_2$, 10% glycerol). An equilibrated 5 ml HiTrap Q column was then connected in tandem to the HisTrap column and the bound protein complex was eluted from the HisTrap column using Ni elution buffer (lysis buffer supplemented with 500 mM Imidazole). The protein did not bind to the HiTrap Q column but this step was essential for removing some contaminating proteins that were bound by this column. The fractions from the HisTrap elution containing the heterodimeric complex were pulled and treated overnight with 6xHis-TEV and dialyzed against lysis buffer to remove the affinity tag and reduce imidazole concentration respectively. The protein was applied again to an equilibrated 5 ml HisTrap in a reverse nickel affinity step to remove the TEV protease, the affinity tag and proteins with the affinity tag (undigested). The protein still bound the column but could be separated from the TEV protease in a different peak. The fractions containing the complex of interest were collected, concentrated using a 30 kDa cut-off Amicon ultra Centrifugal filter and applied to a hiload 16/600pg Superdex 200 gel filtration column equilibrated in gel filtration buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 10 mM TCEP 10 µM $ZnCl_2$. The elution peak was concentrated and stored in 5 µl aliquots at -80 ^oC.

3.4.7 Expression and purification of NELF and INTS3

Expression

The cDNA encoding the full length INTS3 was cloned into 438-C expression vector. This expression vector was combined by LIC with an expression vector harbouring the expression cassettes of the full length NELF complex comprising NELF -A, NELF -B, 6xHis tagged NELF

-C and NELF -E or a NELF complex variants lacking NELF -E or NELF tentacle deletion mutants in which NELF -A (1-188) and/or NELF -E (1-138), were combined with full length NELF -B and C. All NELF constructs were kindly provided by Seychelle Vos and described (Vos, Farnung, Urlaub, et al., 2018). Expression of INTS3-NELF or its variants was done as described (Vos, Farnung, Urlaub, et al., 2018). Harvested Hi5 cells expressing INTS3-NELF was resuspended in 40 ml of lysis buffer (20 mM HEPES pH 7.4, 300 mM NaCl, 10% glycerol, 30 mM imidazole, 1 mM DTT, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, and 0.33 mg/ml benzamidine), flash-frozen in liquid nitrogen and stored at –80°C until purification.

Purification

Frozen pellets were quickly thawed in a water bath at room temperature and 1 mg DNase1 (Sigma-Aldrich) was added. Cells were lysed by sonication (30% amplitude, 2 minutes, 0.6s on, 0.4s off). Lysate was cleared by centrifugation for 60 minutes (87,207xg, 4 ^{o}C). Clarified lysate was filtered with a 0.8 µm syringe filter. A 5 ml HisTrap column, 5 ml HiTrap Q column and a self-packed 15 ml amylose resin were equilibrated in a low salt buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 30 mM imidazole, 1 mM DTT).

Filtered lysate was applied to the HisTrap column via a peristaltic pump/sample pump. The column was washed with 50 ml lysis buffer, followed by 25 ml low salt buffer. The nickel column was connected to the equilibrated self-packed amylose and the HiTrap Q columns (HisTrap - amylose - HiTrap Q) and bound proteins were eluted from the HisTrap column using Ni elution buffer (250 mM imidazole in low salt buffer). The columns were washed with 150 ml of low salt buffer and the 5ml HiTrap Q (with bound free NELF) was detached before the amylose column was eluted with amylose elution buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 30 mM imidazole, 1 mM DTT, 100 mM maltose). The HiTrap Q was eluted separately with a gradient (0-100%) of high salt buffer (20 mM HEPES pH 7.4, 850 mM NaCl, 10% glycerol, 30 mM imidazole, 1 mM DTT). Peak fractions were analyzed by LDS-PAGE (10 µl, 4-12% Bis-Tris gel, MES buffer). The fractions from the amylose elution, containing INTS3 and under-stoichiometric NELF complex as well as the HiTrap Q elution with only NELF complex were separately treated with catalytic amounts of His6-TEV protease and lambda protein phosphatase to remove affinity tags and potential post-translational phosphorylations of the proteins respectively.

The samples (after amylose and HiTrap Q elution) were applied separately to 5 mL HisTrap column equilibrated in low salt buffer with a peristaltic pump and the flow through was

collected. The bound proteins were eluted with the Ni elution buffer from the columns and fractions from the flow through as well as the elution were analyzed by LDS-PAGE.

The flow through fractions from both samples were concentrated with a 100K cut-off Amicon Ultra filter (MERCK Millipore) to 2-4 ml. The sample containing the NELF complex was applied to a hiload 16/600pg Superdex 200 gel filtration column equilibrated in gel filtration buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 1 mM DTT). The elution peak was analyzed by LDS-PAGE (10 µl, 4-12% Bis-Tris gel, MES buffer). The peak fractions were pooled and concentrated aliquoted and stored at -80 ^{o}C.

Protein from the amylose elution which contains mostly INTS3 was also applied to the hiload 16/600pg Superdex 200 gel filtration column equilibrated in gel filtration buffer. The elution peaks were analyzed by SDS-PAGE (10 µL, 4-12% Bis-Tris gel, MES buffer). Appropriate fractions containing INTS3 – NELF complex as well as just INTS3 were pulled separately and concentrated. Proteins were aliquoted, flash-frozen in liquid nitrogen and stored at -80 ^{o}C.

3.4.8 Purification of Mammalian RNA Pol II

Mammallian RNA Pol II was purified from *S. scrofa* thymus (obtained from E. Wolf, Ludwig Maximilian University of Munich) essentially as described (Bernecky et al., 2016). Gdown containing fractions from Uno Q ion exchange step were discarded before a final size-exclusion step was done using a Sephacryl S-300 16/60 column preequilibrated into Pol II buffer (10 mM HEPES pH 7.25, 150 mM NaCl, 10% glycerol, 10 µM $ZnCl_2$, and 1 mM DTT). Peak fractions containing Pol II were concentrated in 100K cut-off Amicon Ultra Centrifugal Filter (Merck Millipore), aliquoted, flash-frozen in liquid nitrogen, and stored at -80 ^{o}C.

3.4.9 Purification of human DSIF

Cloning and expression of full length DSIF (SPT4 and SPT5) was described (Bernecky et al., 2017) and purification was done according to the protocol by Vos and colleagues (Vos, Farnung, Urlaub, et al., 2018).

3.5 Formation and characterization of complexes between INT (INTS3) and RNA Pol II paused elongation complex

3.5.1 Formation of INTS3 – Paused elongation complex (INTS3-PEC)

INTS3 - PEC was formed essentially as described for the PEC (Vos, Farnung, Urlaub, et al., 2018) using the same nucleic acid scaffolds. Briefly 88 pmol final Pol II, 200 pmol of annealed RNA-template DNA, 300 pmol non-template DNA were assembled as described. INTS3, DSIF and NELF were added in a fourfold molar excess relative to Pol II in a final buffer containing 20 mM HEPES pH 7.4, 100 mM NaCl, 3 mM $MgCl_2$, 1 mM DTT, 4% glycerol. The complex was incubated for 30 min at 30 ^oC and applied to a Superose 6 Increase 3.2/300 column equilibrated in the complex buffer. Peak fractions were analyzed by LDS-PAGE and appropriate fractions from the peak was pulled and crosslinked with 2 mM BS3 for XL-MS. Similar approach was used for the preparation of cryo-EM grids except that each fraction from the elution peak was treated independently and few cryo grids were froze from each fraction from the elution peak.

3.5.2 Formation of INT – PEC

To test whether the reconstituted INT interacts with the recently described pause elongation complex (PEC), PEC was formed as previously described with the following modifications (Vos, Farnung, Urlaub, et al., 2018). Briefly 30 pmol final Pol II, 100 pmol of annealed RNA-template DNA, 200 pmol non-template DNA were assembled as described. ~ 30 pmols of freshly purified core-INT, 100 pmol of INTS10/13/14 (CMIM), 100 pmol of INTS4/9/11(CM) and 200 pmols of INTS1(1-294)/INTS12(1-194) heterodimer were pre-incubated together for 1 hr at 4 ^oC and added to the Pol II and nucleic acids complex and incubated for 10 mins before DSIF and NELF were added (200 pmols each). The reaction was incubated for additional 30 min at 30 ^oC in a final buffer contain 20 mM HEPES pH 7.4, 100 mM NaCl, 4% glycerol, 3 mM $MgCl_2$, 1 mM TCEP. The complex was separated on a 15-30% sucrose density gradient spinning at 32000 rpm using an SW60 rotor in an ultra-centrifuge (Beckmann Coulter). The gradient was manually fractionated into 200 µl from the top. Each gradient fraction was concentrated to 40 µl using a 0.5 ml 100 kDa cut-off Amicon Ultra Centrifugal filter and 10 µl of LDS sample buffer was added for LDS-PAGE analysis.

For crosslinking mass spectrometry analysis of INT-PEC complex, PEC was formed as described using the same nucleic acid scaffolds (Vos, Farnung, Urlaub, et al., 2018). Briefly 117 pmol final Pol II, 300 pmol of annealed RNA–template DNA, 400 pmol non-template DNA were assembled as described. DSIF and NELF were added in a fourfold molar excess relative

to Pol II in a final buffer containing 20 mM HEPES pH 7.4, 100 mM NaCl, 4% glycerol, 3 mM $MgCl_2$, 1 mM TCEP. The sample was incubated for 30 min at 30 ºC and applied to a Superose 6 Increase 3.2/300 column equilibrated in the complex buffer. Peak fractions were analyzed by SDS–PAGE. In parallel, INT was assembled using 100 pmol of freshly purified core-INT, 100 pmol each of INTS4/9/11 (CM) and INTS10/13/14 (CMIM) and 200 pmol of INTS1(1-294)/INTS12(1-194) and incubated on ice for 1 hour in 20 mM HEPES 7.4, 150 mM NaCl, 10% glycerol, 1 mM TCEP 10 µM $ZnCl_2$. The assembled INT was mixed with the peak fractions of PEC and incubated for 30 min at 30 ºC. The complex was crosslinked with 3 mM BS3 for 30 min at 30 ºC. The crosslinking reaction was quenched with 50 mM Tris pH 7.5. The crosslinked complex was precipitated with 4 volumes of ice-cold acetone and 0.2 volumes of CH_3COONa and incubated at -20 ºC overnight. Protein precipitates were isolated by centrifugation at 15000 rpm for 30 min in an Eppendorf table top centrifuge and dissolved in 4 M urea, 50 mM NH_4HCO_3 for mass spectrometric analysis.

3.5.3 Analysis of complexes by XL-MS

Enzymatic digestion of crosslinked complexes and enrichment of crosslinked peptide pairs

Crosslinked proteins (resuspended in 4 M urea/50 mM ammonium bicarbonate) were reduced with 10 mM dithiothreitol (DTT) for one hour at room temperature (RT). Alkylation was performed by adding iodoacetamide (IAA) to a final concentration of 40 mM, incubated 30 min in the dark at RT. After dilution to 1 M urea with 50 mM ammonium bicarbonate (pH 8.0), crosslinked protein complexes were digested with trypsin (Promega) in a 1:50 enzyme-to-protein ratio at 37 ºC overnight. Peptides were acidified with trifluoroacetic acid (TFA) to a final concentration of 0.5% (v/v), desalted with MicroSpin Colums (Harvard Apparatus) following manufacturer's instructions and vacuum dried. Dried peptides were dissolved in 50 µl 30% acetonitrile (ACN)/0.1% TFA and peptide size exclusion (pSEC, Superdex Peptide 3.2/300 column) was performed to enrich for crosslinked peptides at a flow rate of 50 µl/min. Fractions of 50 µl were collected. Fractions containing the crosslinked peptides (1-1.5 ml) were vacuum dried and dissolved in 4% ACN/0.05% TFA (v/v) for subsequent LC-MS/MS analysis.

In-gel digestion

Cross-linked proteins were separated by SDS-PAGE using a 4-12% gradient gel (NuPAGE, Invitrogen) for 45 min at 200 V. After Coomassie staining, gel lanes were cut into 23 slices and chopped into small pieces. Proteins were reduced and alkylated with 10 mM DTT and 55 mM IAA, respectively. 0.3 µg trypsin (Sigma Aldrich) was used for digestion, incubated at 37°C

overnight. Peptides were extracted with 100% ACN and 5% formic acid, vacuum dried, dissolved in 2% ACN/0.05% TFA and subjected to LC-MS/MS analysis.

LC-MS/MS analysis

Crosslinked peptides derived from pSEC or in-gel digestion were analysed as technical duplicates on an Orbitrap Fusion, an Orbitrap Fusion Lumos Tribrid or an Q Exactive HF-X Mass Spectrometer (Thermo Scientific), depending on the experimental design, coupled to a Dionex UltiMate 3000 UHPLC system (Thermo Scientific) equipped with an in house-packed C18 column (ReproSil-Pur 120 C18-AQ, 1.9 µm pore size, 75 µm inner diameter, 30 cm length, Dr. Maisch GmbH). Samples were separated applying the following 58 min gradient: mobile phase A consisted of 0.1% formic acid (FA, v/v), mobile phase B of 80% ACN/0.08% FA (v/v). The gradient started at 5% B, increasing to 8-20% B within 3 min, followed by 8-20% to 46% B within 43 min, depending on the experimental design, then keeping B constant at 90% for 6 min. After each gradient the column was again equilibrated to 5% B for 6 min. The flow rate was set to 300 nl/min. MS1 spectra were acquired with a resolution of 120,000 in the orbitrap (OT) covering a mass range of 380-1580 m/z. Injection time was set to 60 ms and automatic gain control (AGC) target to 5×105. Dynamic exclusion covered 10-30 s. Only precursors with a charge state of 3-8 were included. MS2 spectra were recorded with a resolution of 30,000 in OT, injection time was set to 128 ms, AGC target to 5×104 and the isolation window to 1.6 m/z. Fragmentation was enforced by higher-energy collisional dissociation (HCD) at 30%.

Data analysis

For crosslinked peptide searches with pLink 1 (v. 1.23, pFind group (Chen et al., 2019; Yang et al., 2012)), .raw files were converted to mgf format using ProteomeDiscoverer 1.4 (Thermo Scientific, signal-to-noise ratio 1.5, 1000-10000 Da precursor mass). Within pLink 1 BS3 was used as crosslinker and trypsin as digestion enzyme with a maximum of two missed cleavage sites. Carbamidomethylation of cysteines was set as fixed modification, oxidation of methionines as variable modification. Searches were conducted in combinatorial mode with a precursor mass tolerance of 5 Da and a fragment ion mass tolerance of 20 ppm. The used database contained all proteins within the complexes. FDR was set to 0.01. Results were filtered by applying a precursor mass accuracy of ±10 ppm. Spectra of all technical duplicates were combined.

For crosslinked peptide searches with pLink 2 (v. 2.3, pFind group (Chen et al., 2019)), .raw files were used as input. The same settings were applied as described for pLink 1 searches, with the following changes: BS3 or EDC-DE were used as crosslinker, depending on the experimental design. Trypsin was used as protease with a maximum of three missed cleavage

sites. Precursor tolerance was set to 10 ppm. FDR was set to 0.01 or 0.05 in separate or global mode, depending on the experimental design. Protein interaction networks were generated by xiNET (Combe et al., 2015).

3.6 Electron Microscopy

3.6.1 Negative stain electron microscopy

Negative stain electron microscopy was used to assess the oligomerization behaviour of the subcomplexes of INT prior to cryo-electron microscopy (cryo-EM). To this end, 4 µl of protein (~50 nM) was incubated on the film-side of a glow-discharged carbon support copper mesh negative stain grid (S160-4; Plano) for 30 - 180 s depending on protein concentration. The grid was washed for 30 s in a drop of ddH_2O and then stained 3 times for 20 s each in separate drops of 2% (w/v) uranyl acetate solution (Electron Microscopy Sciences). The stained grids were incubated for 60 - 120 s and then blotted with a Whatman filter paper leaving a thin film of stain and then allowed to air-dry. Negative stained grids were images in a Philips CM120 transmission electron microscope operated at 120 kV at a magnification of 37,000x or 74,000x (for the cleavage module) after a defocus of 2.3 µm was applied.

3.6.2 Cryo - electron microscopy

Samples for cryo-EM analysis were gently fixed by crosslinking with 0.1% glutarylaldehyde on ice for 10 min. The crosslinking reaction was quenched by addition of 100 mM NH_4HCO3 (for core-INT) or 8 mM aspartate and 2 mM lysine (INTS4/9/11). The crosslinked samples were dialyzed for 4 - 6 hrs into a buffer containing 20 mM HEPES 7.4, 150 mM NaCl, 1 mM TCEP 10 µM $ZnCl_2$ using a 20 kDa MWCO Slide-A-Lyzer MINI dialysis units. Samples prepared by sucrose density gradient (core-INT), were dialyzed for 10 hrs to overnight and concentrated before cryo-grid preparation. The dialyzed samples (4 µl with final concentration between 100 and 300 nM) were applied to a glow- discharged UltrAuFoil 2/2 grids (Quantifoil). After a 10 s incubation, the grids were blotted for 5-8 s with a blot force of 5 and plunge frozen in liquid ethane. Freezing was done using a Mark IV vitrobot (FEI) operated at 4 ^oC and 100 % humidity.

3.6.3 Data collection and processing

Core-INT

Cryo-grids were pre-screened for good particle distribution using a Glacios transmission electron microscope (FEI) operated at 200 keV. Cryo-EM data was automatically collected on the same microscope using EPU and a Falcon 3EC direct electron detector (FEI) operated in linear mode. Images were recorded for 1.52 seconds at a nominal magnification of 120,000x corresponding to 1.23 Å/pixel with an electron dose rate of 41.58 e-/px/s. The resulting total dose of 41.78 e-/Å2 was fractionated into 30 frames. A total of 1900 images were collected with a defocus range of 1.5-3 µm. Movie frames were subjected to motion correction and contrast transfer function (CTF) estimation in Warp (Tegunov & Cramer, 2019). Particles were automatically picked in WARP yielding about 250,000 particles. Calculation of 2D class averages, generation of initial model and 3D volume calculations were done in cryoSPARC2 (Punjani et al., 2017).

4 Results

4.1 Reconstitution of the Integrator complex

4.1.1 Sequence Analysis of Integrator complex subunits

There is currently limited tertiary and quaternary structural information on INT. A crystal structure of the C-terminal interacting domains of INTS9 and INTS11 has been solved covering about 100 amino acids from each protein (Wu et al., 2017). And a crystal structure of SOSS-A/INTS3 (1-500) in complex with SOSSB1 and SOSSC in the SOSS1 complex has also been reported (Ren et al., 2014). In the absence of experimental data, bioinformatic analysis of the primary amino acid sequence may be helpful in providing preliminary insight into the secondary and tertiary structure of a protein. To this end, I run protein domain/family predictions using the domain/protein family recognition tool 'InterPro' (https://www.ebi.ac.uk/interpro/about/interpro/) for all subunits of INT from their amino acid sequence. The default settings were used for the predictions. Domain predictions were in agreement with previously observed predictions when a different domain prediction tool was used (Gómez-Orte et al., 2019). The INTS9 and INTS11 subunits of INT have 40% sequence identity respectively with CPSF100 and CPSF73 of the cleavage and polyadenylation specific factor (Baillat et al., 2005). These subunits are also predicted to have the metallo-beta-lactamase and beta CASP domains (Figure 4.1). Other domains predicted included a domain of unknown function (DUF3677) in the N-terminus of INTS1, 8 HEAT (Huntingtin, elongation factor 3, the A subunit of protein phosphatase 2A (PP2A) and the signaling kinase TOR1) repeats in INTS4, A van Willebrand Factor Associated (vWFA) domain in INTS6, INTS14 and DDX26B, an armadillo like (ARM) fold in INTS7, 4 TPR (Tetratricopeptide repeat) repeats in INTS8 and a PHD like zinc finger in INTS12. The majority of the predicted folds, ARM, TPR, vWFA, HEAT, are involved in protein-protein interaction and signaling (Tewari et al., 2010; Yoshimura & Hirano, 2016; Zeytuni & Zarivach, 2012a) and might be important for the inter-subunit interaction within INT and interactions with other proteins in cellular processes that INT is involved in. There were no known domains/protein families recognized in INTS2, INTS3, INTS5, INTS10 and INTS13.

Intrinsically disordered proteins and domains are very prevalent in eukaryotic proteins especially those involved in gene transcription (Dyson & Wright, 2005; Mitchell & Tjian, 1989).

Given the low number of recognizable domains found from domain predictions, the disorder prediction tool, disorder2 (Ward et al., 2004) was used to check for intrinsically disordered regions in the amino acid sequences of the INT subunits (Figure 4.1). The protein binding propensities of the predicted disordered regions were also predicted. There were widespread disordered regions predicted in various subunits with high confidence. Most of the predicted disordered regions have high protein binding propensity except a stretch of about 100 amino acids in the C-terminal regions of INTS6 and DDX26B (also called INTS6-like). In the exception of a small N-terminal domain and the PHD domain, INTS12 is predicted to be disordered and lacking in interaction with other proteins. These predictions might be useful when it is necessary to truncate a protein to improve expression and solubility or for crystallization.

4.1.2 Co-expression of known subcomplexes

It is not known how the subunits of INT are interacting to form the complex. For recombinant production, it is critical to identify which subunits are interacting and might need each other for co-translational folding (Hardesty et al., 1999; Schilbach et al., 2017). The first thing that I attempted was to identify the interacting subunits of INT. The INTS9 and INTS11 subunits have been shown to form a stable heterodimer *in vivo* (INTS9/11 heterodimer) which hosts the endoribonuclease activity of the complex (Albrecht & Wagner, 2012; Wu et al., 2017). My preliminary co-expression test of these subunits show they interact to form INTS9/11 heterodimer. The expression of INTS9/11 heterodimer was very poor in both *E. coli* and baculovirus - insect cell expression systems suggesting an additional subunit might be needed to improve the expression and solubility of these subunits or a different expression condition is needed. INT was identified as a target of the protein phosphatase 2A using affinity purification and crosslinking coupled to mass spectrometry (Herzog et al., 2012a). In their work, Herzog and colleagues identified chemical crosslinks between INTS4 and the heterodimer of INTS9 and INTS11. They also observed crosslinks between INTS2 and INTS7 (Herzog et al., 2012a; Solis-Mezarino & Herzog, 2017). This suggests that, INTS4 have direct protein - protein interaction with INTS9/11 heterodimer.

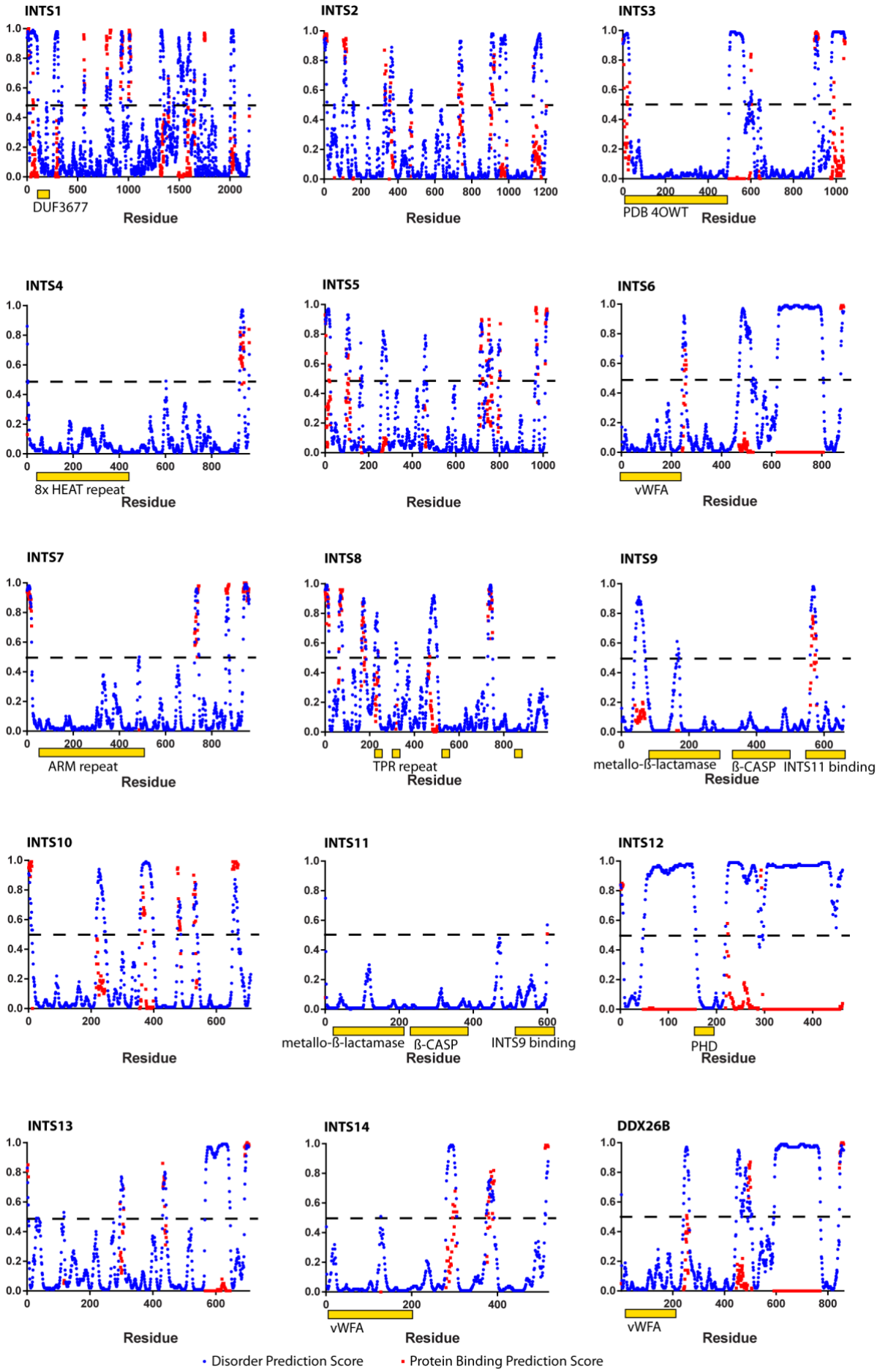

INTS1
DUF3677
Residue
INTS2
Residue
INTS3
PDB 4OWT
Residue
INTS4
8x HEAT repeat
Residue
INTS5
Residue
INTS6
vWFA
Residue
INTS7
ARM repeat
Residue
INTS8
TPR repeat
Residue
INTS9
metallo-ß-lactamase
ß-CASP
INTS11 binding
Residue
INTS10
Residue
INTS11
metallo-ß-lactamase
ß-CASP
INTS9 binding
Residue
INTS12
PHD
Residue
INTS13
Residue
INTS14
vWFA
Residue
DDX26B
vWFA
Residue
Disorder Prediction Score
Protein Binding Prediction Score

Figure 4.1. Domain and disorder prediction of INT subunits. Disorder predictions (blue) and protein binding prediction (red) scores are plotted as a function of amino acid residue numbers for each subunit of INT indicated on the top left corner of each plot. The dashed lines represent the cutoff of 0.5 score. Predicted domains from InterPro for each subunit, crystallized domain (of INTS3) and interacting C-terminal regions of INTS9 and INTS11 are indicated with a yellow filled rectangle. Abbreviations: DUF (domain of unknown function), HEAT (Huntingtin, Elongation factor 3, protein phosphatase 2A, and the yeast kinase TOR1), ARM (armadillo-like repeats), vWFA (von Willebrand Factor type A domain), β-CASP (CPSF73, Artemis, SNM1 PSO2 domain), PHD (plant homeodomain finger) and TPR (tetratricopeptide repeats).

To test this, I constructed a baculovirus co-expressing INTS4, INTS9 and 6xHis-MBP-INTS11. Results from amylose affinity pulldown of expression test in Hi5 insect cells showed the three proteins interact and form a stable heterotrimeric subcomplex (Figure 4.2a). Also, co-expressing INTS4 with the INTS9/11 heterodimer drastically improved their expression in insect cells and solubility under the same conditions used for INTS9/11 heterodimer. This shows that INTS4 is needed for proper folding and assembly of INTS9/11 heterodimer. INTS4 alone however is soluble, expressible and can be purified in high purity and yield (Figure S5c and d). This heterotrimeric subcomplex was characterized around the same time by Albretch and colleagues and named as the cleavage module of INT due to its similarity with the cleavage module of the CPSF complex involved in the 3' processing of mRNAs (Albrecht et al., 2018). Also, based on the observed chemical crosslink between INTS2 and INTS7 (Solis-Mezarino & Herzog, 2017), the interaction between these two subunits could be established by co-expression and co-purification using affinity tag on INTS7 (Figure 4.2b and Figure S4).

INTS3 also known as SOSS-A is known to be a part of the SOSS complex involved in single stranded DNA sensing (Huang et al., 2009; Li et al., 2009; Ren et al., 2014; Skaar et al., 2009). INTS6 is the only subunit of INT that interacted with the SOSS1 (NABP2) complex (Skaar et al., 2015). I suspected that INTS6 might have direct protein-protein interaction with INTS3. Co-expression of INTS3 and INTS6 and amylose affinity pulldown using 6xHis-MBP affinity tag on INTS3 showed the two subunits interact as they co-purified (Figure 4.2c). Taken together, the available literature helped to establish preliminary inter-subunit interactions within the INT, namely INTS4/9/11 heterotrimer (also known as the cleavage module), INTS2/7 and INTS3/6 heterodimers.

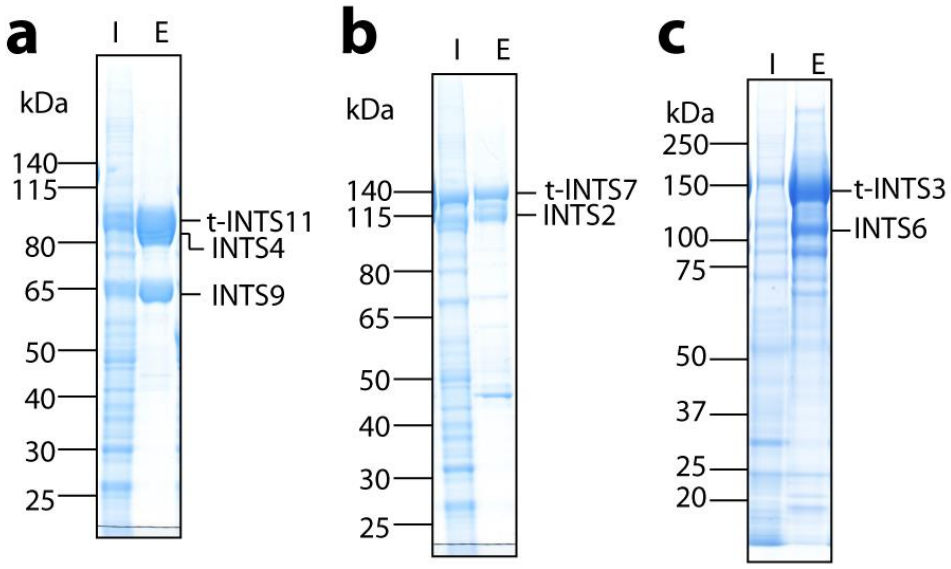
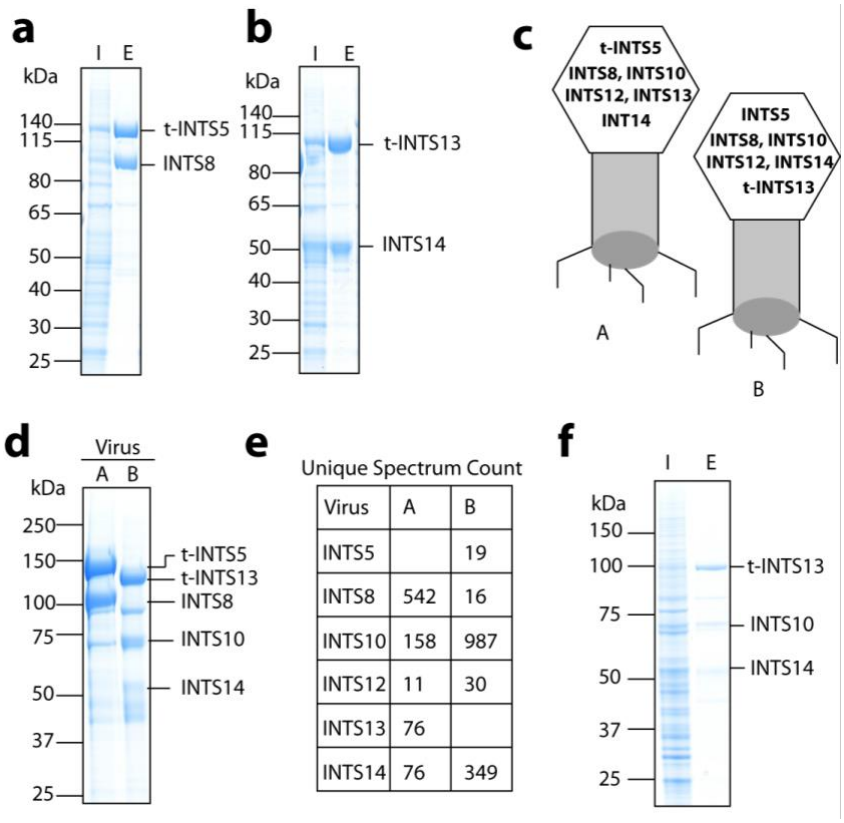


Figure 4.2. Co-expression test of known INT subcomplexes. LDS-PAGE analysis of expression and interaction tests of various subcomplexes of INT. Letter 'I' represent an input sample from the soluble fraction of the lysate of infected SF21 or Hi5 cells and E is 5 - 10% of the elution of an amylose affinity pulldown from infected SF21 or Hi5 cells expressing INTS4/9/11 heterotrimer (**a**), INTS2/7 heterodimer (**b**), and INTS3/6 heterodimer (**c**). 't' stands for 6xHis-MBP affinity tag. The identity of proteins was confirmed by mass spectrometry.

4.1.3 Identification of novel subcomplex of INT by systematic co-expression of subunits

To further expand on the known subcomplexes, I randomly cloned the rest of the INT subunits for co-expression. Amylose affinity pulldown analysis from insect cells infected with a baculovirus co-expressing INTS5, INTS8, INTS10 and INTS12 with affinity tag on INTS5 showed INTS5 and INTS8 are interacting partners (Figure 4.3a). INTS10 and INTS12 had weak interaction with INTS5/8 heterodimer as shown on the LDS-PAGE and very few peptides from these subunits were identified by mass spectrometry (results not shown). I also identified INTS13 and INTS14 as a stable heterodimer by co-expressing them with affinity tag on INTS13 (Figure 4.3b). Furthermore, I constructed baculoviruses co-expressing INTS5, INTS8, INTS10, INTS12, INTS13 and INTS14 with affinity tag on either INTS5 or INTS13. Proteins associated with INTS5 and INTS13 in Hi5 expression tests of the two baculoviruses were assessed by amylose affinity pulldown, LDS-PAGE and mass spectrometric identification of peptides in the elution fractions. Trace amounts of INTS10, INTS12, INTS13 and INTS14 were co-purified with INTS5/8 heterodimer when construct with affinity tag on INTS5 was analyzed (virus A Figure 4.3 c, d and e). However, when the virus with affinity tagged INTS13 was analyzed, there was significant enrichment of INTS10 in the elution fraction alongside INTS13 and

INTS14 (Figure 4.2 d and e) suggesting INTS10 has a strong physical interaction with either INTS13 or INTS14 or both in the Integrator complex. Yeast-two-hybrid analysis of *Drosophila* INTS14 shows it interacts with INTS10 *in vivo* (Jiandong Chen et al., 2012b).



Virus	A	B
INTS5		19
INTS8	542	16
INTS10	158	987
INTS12	11	30
INTS13	76	
INTS14	76	349

Figure 4.3. systematic co-expression of INT subunits to identify subcomplexes. LDS-PAGE analysis of expression test and amylose affinity pulldown of a baculovirus expressing INTS5, INTS8, INTS10 and INT12 (**a**) and INTS13 and INTS14 (**b**). (**c**) A cartoon representation of baculoviruses expressing 6xH-MBP-INTS5, INTS8, INTS10, INTS12, INTS13 and INTS14 (virus A) and INTS5, INTS8, INTS10, INTS12, 6xHis-MBP-INTS13 and INTS14 (virus B). (**d**) LDS-PAGE analysis of the elution fractions of amylose affinity pulldown from SF21 cells expressing virus A and B described in (**c**). (**e**) Mass spectrometric identification of peptides in the elution fractions of the expression tests of baculovirus A and B described in (**c**) and (**d**). (**f**) LDS-PAGE analysis of expression test of a baculovirus expressing INTS10, INTS13 and INTS14. The prefix 't' before a protein name indicated it is tagged with 6xHis-MBP affinity tag.

On the other hand, INTS12 associated weakly to this complex (INTS10/13/14) and INTS5/8 heterodimers did not show any substantial interaction with the INTS10/13/14 heterotrimeric subcomplex (Figure 4.3 c, d, and e). The INTS13/14 heterodimer as well as INTS10 were expressed and could be purified to homogeneity (Figure S5a and b, Figure S5 e and f). The interaction between INTS10 and INTS13/14 heterodimer was additionally confirmed via amylose affinity pulldown using purified 6xHis-MBP-INTS10 and INTS13/14 (Figure S7b lane

8) and the three proteins can be purified together when co-expressed in insect cells (Figure 4.3f).

4.1.4 Identification of interaction partners of INTS1

The Integrator subunit INTS1 is a 250 kDa DUF (domain of unknown function) containing protein (Figure 4.1 INTS1). It is predicted to be rich in alpha helical secondary structures indicative of protein-protein interactions. In *Drosophila*, INTS12 has been shown to interact with INTS1 (Jiandong Chen et al., 2013). Co-expression of the full length INTS1 and INTS12 showed that the two subunits have direct protein-protein interaction (Figure 4.4a). However due to poor expression of these two subunits and their instability during purification, INTS1/12 heterodimer could not be purified. Also, co-expression analysis showed that, INTS1 has some interaction with INTS3/6 as well as INTS13/14 heterodimers (Figure 4.4b and c).

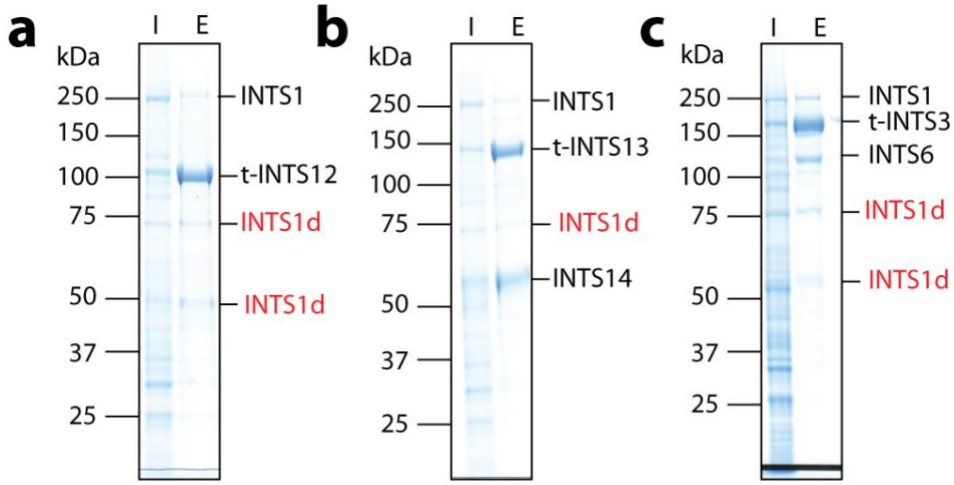


Figure 4.4. Identification of interacting partners of INTS1. Input (I) and Elution fractions (E) of amylose affinity pulldown of protein expression tests were analyzed on LDS-PAGE. Protein bands in the elution fraction were identified by mass spectrometry. INTS1 degradation products are indicated as INTS1d and colored red. The prefix letter 't' represents 6xHis-MBP affinity tag (a) LDS-PAGE analysis of co-expression of INTS1 and INTS12. (b) LDS-PAGE analysis of co-expression of INTS1, INTS13 and INTS14. (c) LDS-PAGE analysis of co-expression of INTS1, INTS3 and INTS6

The heterodimers, INTS2/7 and INTS5/8 as well as INTS4/9/11 heterotrimer did not show interaction with INTS1 in the co-expression and amylose affinity pulldown assays (results not shown).

In summary, in sections 4.1.3 and 4.1.4, I have identified previously unknown interaction between INTS5 and INTS8 subunits as well as between INTS13 and INTS14 subunits of INT.

Interaction between INTS10 and the identified INTS13/14 heterodimer was also established. I could also recapitulate the interaction between INTS1 and INTS12 subunits reported in *Drosophila* here in the human INT. In additional, I showed that INTS1 interacts with the INTS3/6 as well as the INTS13/14 heterodimers.

4.1.5 Co-infection of 3 baculoviruses and partial purification of INT

A combination of previously known interaction partners and the results of the systematic co-expression assays informed the creation of three expression vectors with each vector containing interacting subunits of INT whenever possible (Section 3.3.5.1).

Viruses were made from each of these expression vectors (Figure 4.5a) and Hi5 cells were infected at 1:6000 v/v for each virus. The cells were harvested after 72 hrs and a test purification was done (Figure 4.5b). There was severe degradation of several subunits, notably INTS1, INTS3, INTS6, and DDX26B suggestive of suboptimal expression and purification conditions. Protein purified by ion exchange chromatography was fractionated into two peaks on analytical gel filtration showing the complex is not homogenous. Mass spectrometry analysis identified peptides for all the subunits of INT in the two elution peaks of gel filtration (Figure 4.5c). The first peak at 1 ml may represent the intact full-length INT while the second peak may be a heterogenous mixture of subcomplexes of INT. Negative stain analysis of the two elution peaks revealed heterogenous population of particles with varying sizes for both peaks. This may represent different views of INT and/or INT subcomplexes. The sample may not be homogenous as there were three subunits with the same affinity tag and hence different populations of complexes may have been enriched. Particles in the first peak (Figure 4.5d) have broader size distribution with the majority of the particles having diameter in the range of 250 - 450 Å which is the approximate expected size of the full-length INT (approximately 1.5 MDa). In contrast, particles in the second peak (Figure 4.5e) have relatively smaller particle diameters with a narrower particle dimeter distribution representing heterogenous population of subcomplexes of INT.

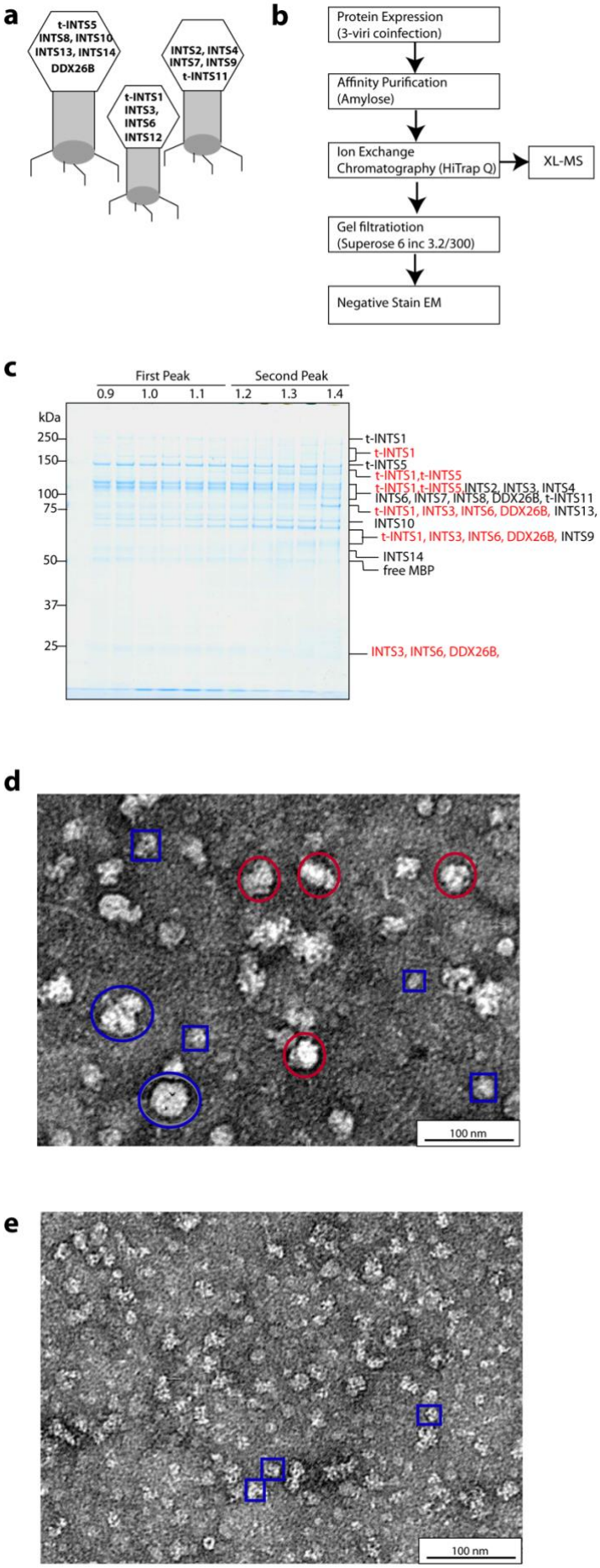


Figure 4.5. Co-infection and partial purification of INT. (**a**) Cartoon representation of the three baculoviruses expressing subcomplexes of INT. The subunits co-expressed are indicated in each cartoon. The prefix letter 't' implies the specific subunit has 6xHis-MBP affinity tag. (**b**) A flowchart for the experimental design showing the expression and purification strategy used. (**c**) LDS-PAGE analysis of fractions from gel filtration on Superose 6 Increase 3.2/300 column and mass spectrometric identification of protein. Proteins running at their expected molecular weight are indicated in black while degradation products are indicated in red. The retention volumes of the fractions analyzed are indicated above the gel. (**d**) and (**e**) are representative negative stain micrographs of the first and second peak respectively from the gel filtration. Particles of the expected size (approximately 450 Å or 45 nm) are indicated with red circles. Blue circles represent oligomeric particles or negative stain artifacts and blue squares are representative particles smaller than the expected size for the full-length INT and might be subcomplexes. Micrographs were acquired at 37,000x magnification (3.3 Å/pixel). A scale bar is provided.

4.1.6 XL-MS on partially purified INT identifies new interaction partners

To have an appreciation of how the various subcomplexes may be interacting in the partially purified INT, a sample of the HiTrap Q elution peak corresponding to the full INT was crosslinked with 1 mM BS3 for 30 min on ice and subjected to XL-MS. There were very few inter subunit crosslinks observed. High confidence crosslinks were observed between known interacting subunits for examples INTS3 and INTS6 as well as INTS13 and INTS14 serving as internal positive controls.

Interestingly, INTS3 had high number of high-ranking crosslinks to DDX26B. Specifically, K842 of DDX26B had high scoring crosslinks with K557 of INTS3 (Figure 4.6a and b). This was outstanding because of the very few inter subunit crosslinks observed in general. This prompted the idea that DDX26B might have strong direct protein-protein interaction with INTS3/6 heterodimer. To test this INTS3, INTS6 and DDX26B were cloned together for co-expression. Indeed, all three proteins were present in good stoichiometry in the elution of amylose affinity pulldown using affinity tag on INTS6 (Figure 4.6c). This confirms for the first time DDX26B has direct protein-protein interaction with a subunit (INTS3) of the Integrator complex. Affinity tagging of INTS3 resulted in overrepresentation of this subunit (Figure S6) whereas affinity tag on INTS6 produces a stoichiometric complex. The proteins can be co-purified in pulldown assays but turn to aggregate (oligomerize) when analyzed by gel filtration suggesting that there is an interaction partner they might be missing. I came to this conclusion after several expression and purification protocols were tested with no success at preventing the aggregation or oligomerization.

4.1.7 INTS3/6-DDX26B heterotrimer interacts with INTS5/8 heterodimer

To identify which subunits/subcomplexes have direct physical interaction with the identified INTS3/6-DDX26B heterotrimer, a co-infection and amylose affinity purification assay was designed. This entailed co-infecting Hi5 insect cells with a V_1 baculovirus expressing this trimer (virus 0) and V_1 baculoviruses expressing tagged subunits/subcomplexes of INT as shown in Figure 4.7a and b. The cultures were monitored and harvested after 48 hrs.

Amylose affinity pulldown followed by western blot analysis (using anti-INTS3 antibody) of input and elution fractions showed that all the subunit/subcomplexes have some physical interaction with the INTS3/6-DDX26B heterotrimer showing that this trimer might represent a central core of INT (Figure 4.7b). Pulldown from control cells expressing only INTS3/6-

DDX26B without affinity tag had no signal for INTS3 in the elution fraction showing that INTS3 does not bind nonspecifically to the beads.

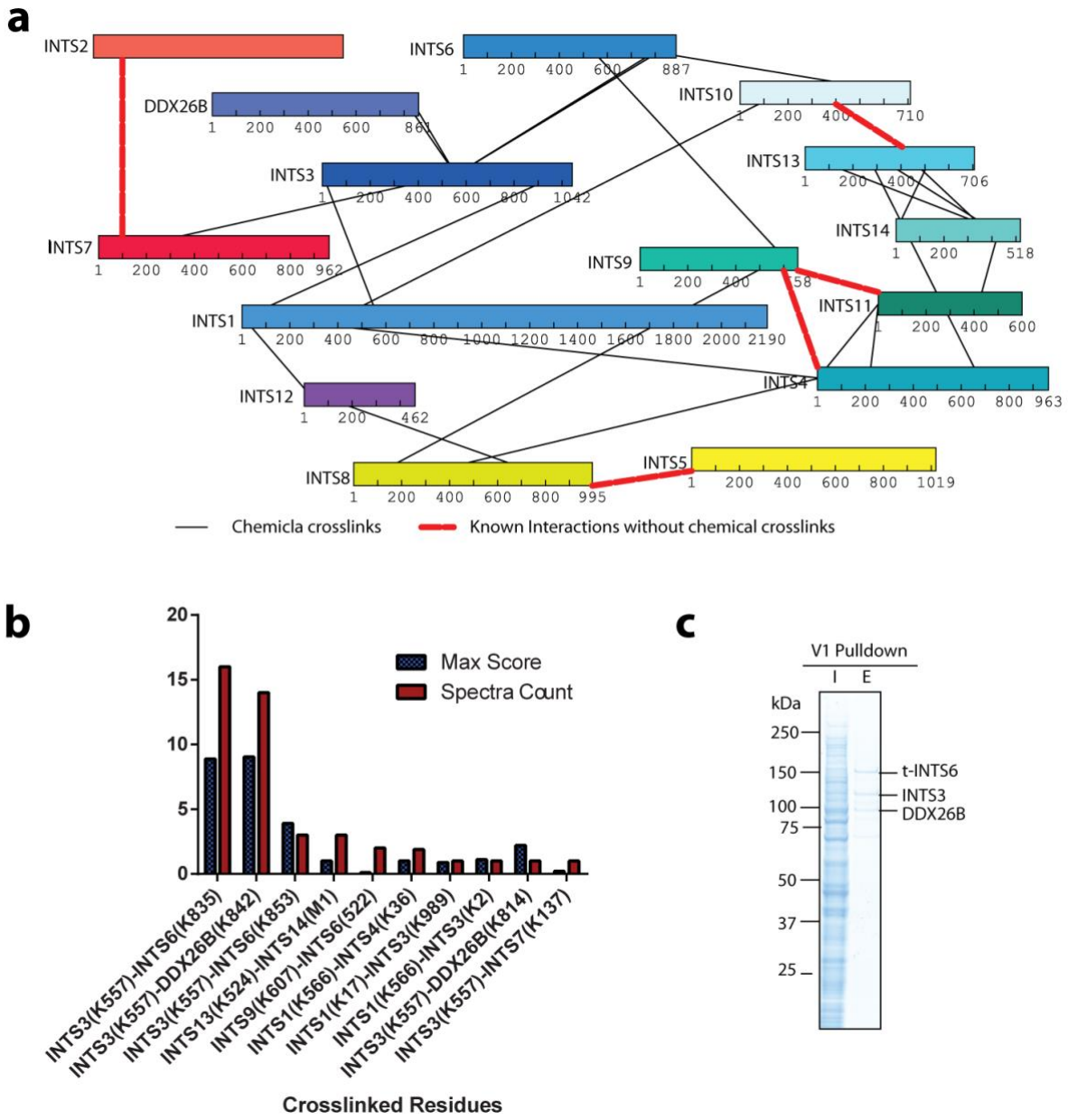
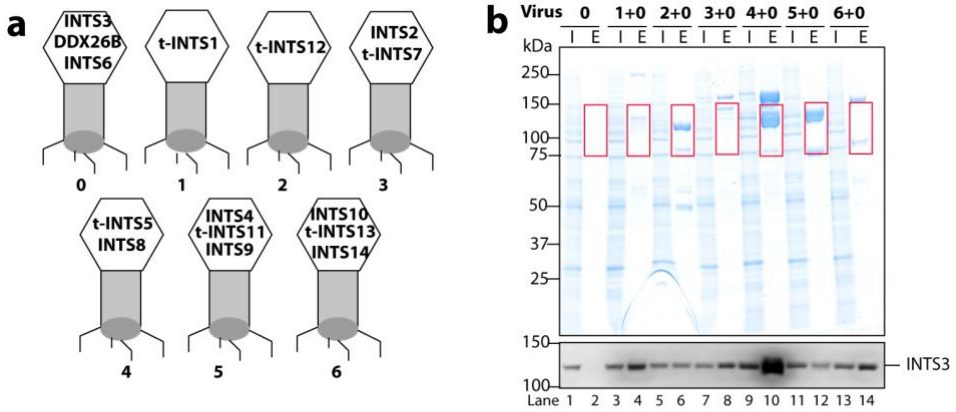


Figure 4.6. Crosslinking mass spectrometry identifies interaction between INTS3 and DDX26B. (a) Cartoon representing the crosslinking network of the INT obtained from BS3 mediated XL-MS of the partially purified INT. INT subunits are depicted as rectangles and BS3 mediated chemical crosslinks as black line segments. Red dashed lines represent known interacting subunits without chemical crosslinks. (b) A bar chart plot of the scores (top ten) of the crosslinks as well as their spectra counts. Specific crosslinked Lys residues are indicated beneath the bar chart. A file containing all crosslinks can be provided upon request. (c) An LDS-PAGE analysis of amylose affinity pulldown of SF21 cells expressing INTS3/6-DDX26B heterotrimer created based on the XL-MS results. Protein identities were confirmed by mass spectrometry



c

Mass Spectrometric Exclusive Unique Spectrum count							
Protein	**0**	**1+0**	**2+0**	**3+0**	**4+0**	**5+0**	**6+0**
INTS3	20	73	49	57	105	42	50
INTS6	16	76	49	58	87	39	38
DDX26B	13	51	31	27	50	32	42

Figure 4.7. Identification of interaction partner(s) of INTS3/6-DDX26B heterotrimer. (**a**) Cartoon representation of baculoviruses (1-6) expressing subunits/subcomplexes of INT. The prefix letter 't' stands for 6xHis-MBP affinity tag. The subunits of INT expressed by each virus are indicated. (**b**) An LDS-PAGE analysis of input (I) and elution fractions of amylose affinity pulldown (E) from Hi5 insect cells expressing various combinations of baculoviruses in (a). Numbers on top of the gel represents which baculoviruses were co-infected. Red rectangles indicate the fragment of the LDS-PAGE gel analyzed by mass spectrometry. An anti INTS3 western blot of the same fractions in the LDS-PAGE is appended to the LDS-PAGE gel. (**c**) Exclusive unique spectra count for INTS3, INTS6 and DDX26B from regions of the pulldown indicated with the red rectangles in (b).

Of interest, INTS5/8 heterodimer showed the strongest signal (enrichment) of INTS3 in the elution fraction. This suggests that, of all constructs tested, the INTS5/8 heterodimer is the most stably associated with the INTS3/6-DDX26B heterotrimer. Mass spectrometric analysis of the fragments of the LDS-PAGE gel corresponding to the sizes of INTS3, INTS6 and DDX26B (between 75 kDa and 150 kDa) showed with red rectangles in Figure 4.7b confirmed the western blot results. The unique spectra count for INTS3 and INTS6 were about 1.5 - 2 folds higher in the pulldown with INTS5/8 heterodimer compared to the other subunits/subcomplexes while DDX26B was enriched to lesser amount (Figure 4.7c). Pulldown with INTS1 showed the second highest spectra counts for INTS3, INTS6 and DDX26B confirming the earlier observation that INTS1 interacts with INTS3/6 heterodimer (Figure 4.4c). The interaction between INTS3/6-DDX26B and INTS5/8 was reproduced by *in vitro* pulldown using partially purified proteins (Figure S7d lane 8). Based on these results, I created a pentameric co-expression vector of INTS3, INTS5, INTS6, INTS8, and DDX26B.

4.1.8 Co-expression and purification of the INTS3/5/6/8-DDX26B heteropentamer

The INTS3/6-DDX26B heterotrimer and INTS5/8 heterodimer forms oligomers and run at the void volume of gel filtration columns during purification (Figure S3 and S6). This suggests the subcomplexes may be misfolded or some protein-protein interaction surfaces are exposed (hydrophobic). Therefore, it was expedient to evaluate the oligomerization behavior of the resultant pentameric complex under purification conditions given that the constituent subcomplexes forms oligomers. To these ends, I made a baculovirus harboring the expression cassettes for INT subunits in the INTS3/5/6/8-DDX26B heteropentamer (Figure 4.8a). Preliminary pulldown experiments from SF21 cells infected with the V_1 baculovirus confirmed expression and interaction of the constituent subunits (not shown).

After pioneering test runs, the final purification strategy for the heteropentamer included an amylose affinity purification followed by anion exchange chromatography on a 5 ml HiTrap Q and gel filtration chromatography on Sephacryl S-300. Mass spectrometric analysis of the elution peak revealed all subunits of the heteropentamer are present in the peak fraction (Figure 4.8b). Based on this result it is clear that the heteropentameric complex is stable under this purification condition. The elution peak from gel filtration runs at 40 ml which coincides with the theoretical void volume of the 120 ml gel filtration column (Figure 4.8c). The ratio of UV absorption at 260 nm to 280 nm was 0.9. These observations suggested the complex might be oligomerizing and the 260/280 ratio suggests DNA contamination. To ascertain the oligomerization state of the purified complex, negative stain transmission electron microscopy was conducted on a uranyl acetate stained sample of the elution peak. The negative stain micrograph revealed majority of the particles are in the expected size range of 200-300 Å for a complex of an estimated molecular weight of 450 – 500 kDa assuming single copies of each protein is in the complex. This shows that, despite eluting close to the void volume of the gel filtration column, the heteropentamer is not aggregated. There are however few particles that are bigger than the expected dimensions of the complex of interest which might be oligomers. These bigger particles may also represent clustering of particles induced by the harsh uranyl acetate staining condition.

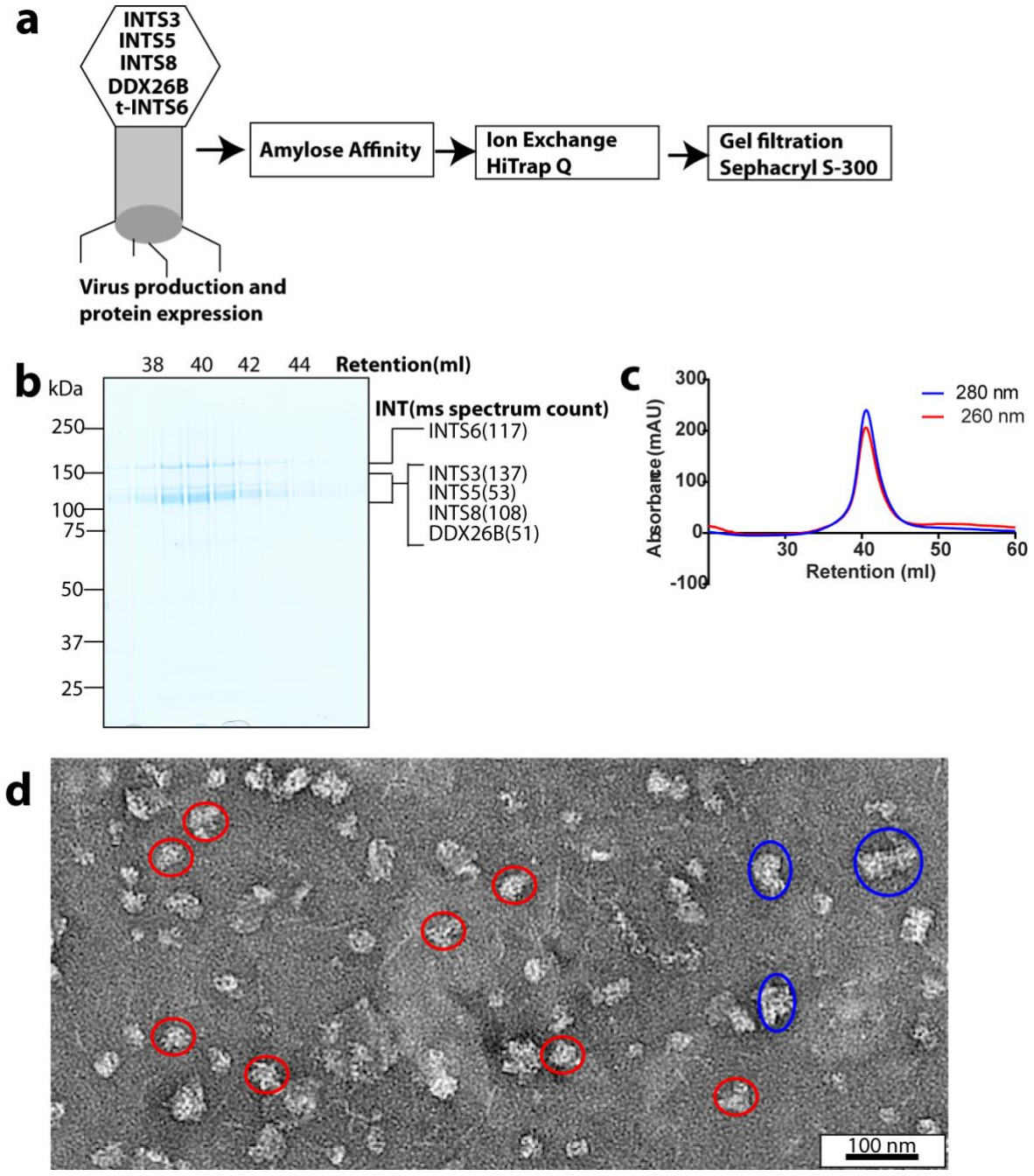


Figure 4.8. Expression, Purification and characterization of INTS3/5/6/8-DDX26B heteropentameric subcomplex. (**a**) A cartoon representing the expression and purification strategy. A baculovirus expressing all the subunits of the heteropentamer with a 6xHis-MBP (t) tag on INTS6 is shown. (**b**) An LDS-PAGE analysis of fractions from gel filtration chromatography on Sephacryl S-300 and mass spectrometric identification of proteins there-in. Mass spectrometric unique spectra counts for each protein is indicated in the brackets. (**c**) A chromatogram of the gel filtration run shown in (b). (**d**) A representative negative stain micrograph of the peak fraction of the gel filtration chromatography of the heteropentamer. Selected particles of the approximate expected diameter are indicated with red circles and oligomers or staining-induced clusters are indicated with blue circles/ovals. A scale bar is provided.

4.1.9 The INTS3/5/6/8-DDX26B heteropentamer interacts with INTS2/7 heterodimer

Partially purified INTS2/7 heterodimer showed some interaction with INTS5/8 heterodimer albeit not stoichiometric (Figure S7d, Lane11). Co-infection assay in Figure 4.7 demonstrates a weak interaction between INTS2/7 heterodimer and INTS3/6-DDX26B. In contrast, the INTS2/7 heterodimer did not interact with the INTS10/13/14 heterotrimer and the cleavage module (INTS4/9/11) (Figure S7d Lane 9 and lane 10 respectively). These results predicted an interaction between INTS2/7 heterodimer and the INTS3/5/6/8-DDX26B heteropentamer. To test this, I made a baculovirus harboring the expression cassettes of INTS2, INTS3, INTS5, INTS6, INTS7, INTS8, and DDX26B with affinity tag on either INTS5, or INTS6 or INTS7 (Figure 4.9a left panel). The elution fractions from amylose affinity pulldown from SF21 cells expressing the V_1 viruses were analyzed on LDS-PAGE and blotted with anti INTS2 and INTS5 antibodies. INTS2 co-purifies with INTS5 in all three constructs according to the western blot results (Figure 4.9a right panel). Mass spectrometric identification of proteins in each fraction confirms the western blot result showing that INTS2/7 heterodimer interacts with the heteropentameric subcomplex and could be co-purified by one affinity tag. This heteroheptameric subcomplex of INTS2, INTS3, INTS5, INTS6, INTS7, INTS8 and DDX26B is hereafter referred to as core Integrator (core-INT).

4.1.10 Purification of INTS2/3/5/6/7/8-DDX26B (Core-INT)

Several expression and purification conditions were tested to identify the most optimal conditions for purifying core-INT. It was important to tag a subunit that would purify a stoichiometric complex. 6xHis-MBP tag at the N-termini of INTS5, INTS6 and INTS7 were tested. Tagging of INTS5 resulted in overrepresentation of INTS5/8 heterodimer while tagging INTS6 and INTS7 produced relatively better results when compared to tagged INTS5 (Figure 4.9a left panel). INTS6 however degrades and it was necessary to affinity tag this subunit to reduce its degradation. I first tagged the N-terminus of INTS6; however, the affinity tag could not be removed by the TEV protease. Secondary structure prediction showed formation of a beta sheet at the very N-terminus of INTS6, which may incorporate the TEV cleavage site. The affinity tag at the C-terminus of INTS6 was then attempted. This did not solve the degradation of INTS6 and N-terminus affinity tag on INTS6 was chosen as the tag of choice for co-purification of this subcomplex of the INT.

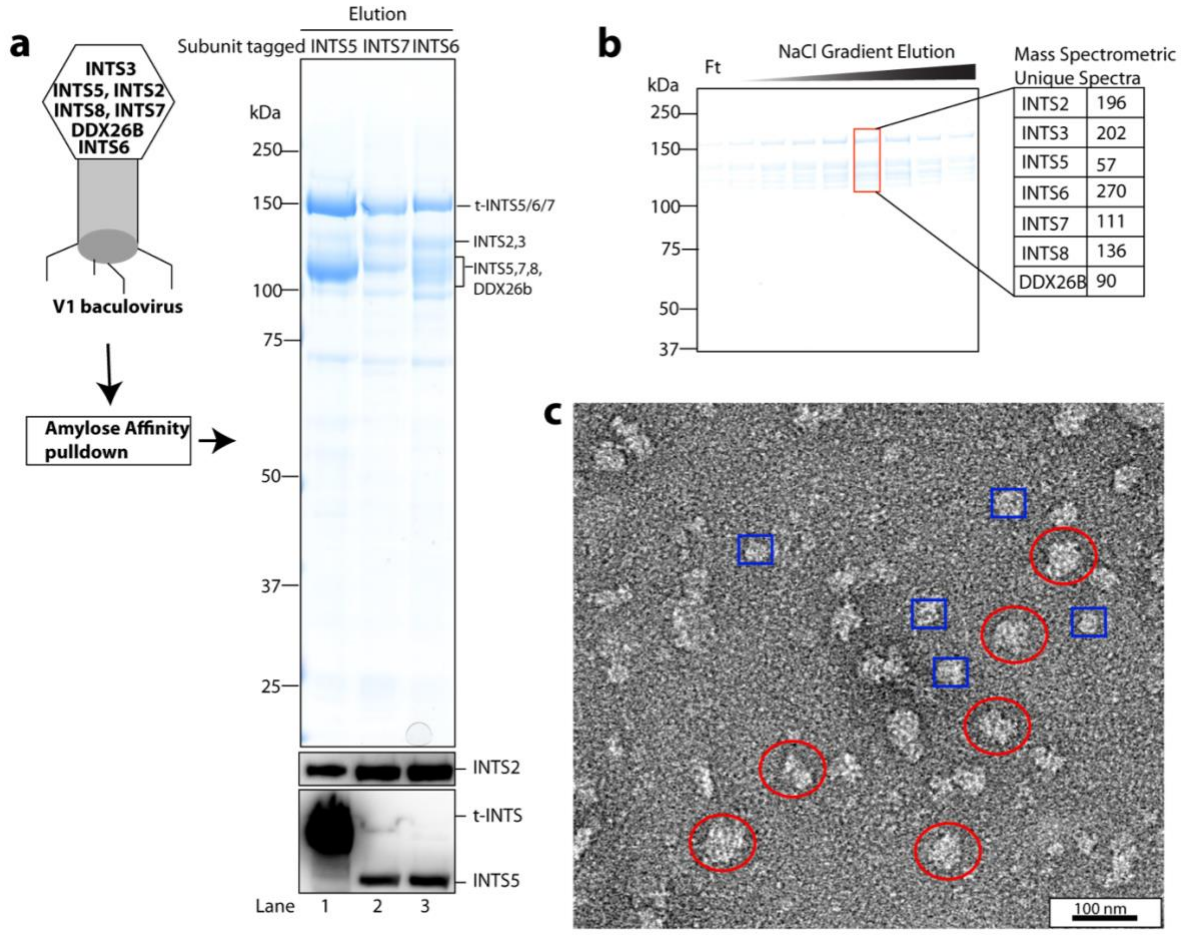


Figure 4.9. Identification and purification of heptameric core-INT. (a) A cartoon of a baculovirus expressing heptameric core-INT comprising INTS2, INTS3, INTS5, INTS6, INTS7, INTS8, DDX26B (left panel). An LDS-PAGE analysis of the elution fractions of amylose affinity pulldown from SF21 cells infected with the V_1 baculoviruses of core-INT with affinity tag on INTS5 (Lane 1) or INTS7 (Lane 2), or INTS6 (lane 3). Proteins were identified by mass spectrometry. Western blot analysis of the same fractions with anti INTS2 and anti INTS5 antibodies is appended to the bottom of the LDS-PAGE. (**b**) An LDS-PAGE analysis of the peak elution fractions of a HiTrap Q anion exchange purification of the heptameric core-INT. Mass spectrometric unique spectra count of the core-INT subunits in the peak fraction is affixed to the LDS-PAGE. (**c**) A representative uranyl acetate negative stain electron micrograph of the peak fraction from (b). Particles meeting the expected size are shown with red circles/oval. A few smaller particles are shown with blue squares. A scale bare is provided.

For optimizing the purification, I varied pH, salt (type and concentration), cell lysis strategy (sonication/French press) and additives. Lysis by French press did not make any significant change to the stability of the complex when compared to sonication and therefore brief sonication was chosen (see methods). I obtain similar results when KCl or NaCl was used as the main salt. However, the complex completely degraded when MES pH 6.5 was used for

purification whereas purification in HEPES pH 7.4 and Tris pH 8 yielded similarly better results. For the purposes of downstream crosslinking experiments, HEPES pH 7.4 was chosen over Tris pH 8 as a buffering agent. It was also observed that longer expression (longer than 72 hrs) resulted in low yield. Hi5 cells expressing this complex were therefore harvested 48 – 60 hours after infection (best case was 48 hours). Also, baculovirus expressing this complex turn to decay unusually fast (about a month) compared to baculoviruses expressing other constructs which are infective after a year of storage at 4 ^{o}C.

The final purification protocol included an amylose affinity step followed by anion exchange chromatography which removes over stoichiometric INTS6 and its major degradation product (Figure 4.9b). The elution from ion exchange on HiTrap Q column was evaluated for homogeneity and oligomerization using negative stain electron microscopy. The micrograph in Figure 4.9c shows that the peak fraction from HiTrap Q elution had a heterogenous population of complexes with little or no aggregates. This heterogeneity may come from INTS3/6 heterodimer and/or INTS3/6-DDX26B heterotrimer and/or INTS3/5/6/8-DDX26B heteropentamer which has the same behavior on ion exchange as core-INT. Gel filtration chromatography was not a favored choice to separate these different populations of complexes as it could not separate the complex (core-INT) from aggregates of INTS3/6 heterodimer and INTS3/6-DDX26B heterotrimer and from INTS3/5/6/8-DDX26B heteropentamer because they have similar sizes and run close to the void volume of the gel filtration columns tested. The complex was therefore further purified via a shallow 15-30 % sucrose gradient. Smaller subcomplexes including INTS3/6 heterodimer were separated in early fractions (lower percentage sucrose) of 1 - 5. The heptameric complex had peak fractions from 7 to 11 which tailed into higher fraction (Figure 4.10a). Mass spectrometric analysis confirmed the presence of all subunits showing this is a stable complex.

To evaluate the homogeneity of the purified complex, negative stain electron microscopy was done on the 11th fraction from the sucrose density gradient (Figure 4.10c). The negative stain micrograph had homogenous single particles with diameter ~250 - 350 Å. The particles did not differ drastically from the heteropentameric complex (INTS3/5/6/8-DDX26B) (Figure 4.8c). The sucrose density gradient successfully separated the heptameric complex from smaller subcomplexes that were present in the peak fraction from anion exchange chromatography on HiTrap Q column.

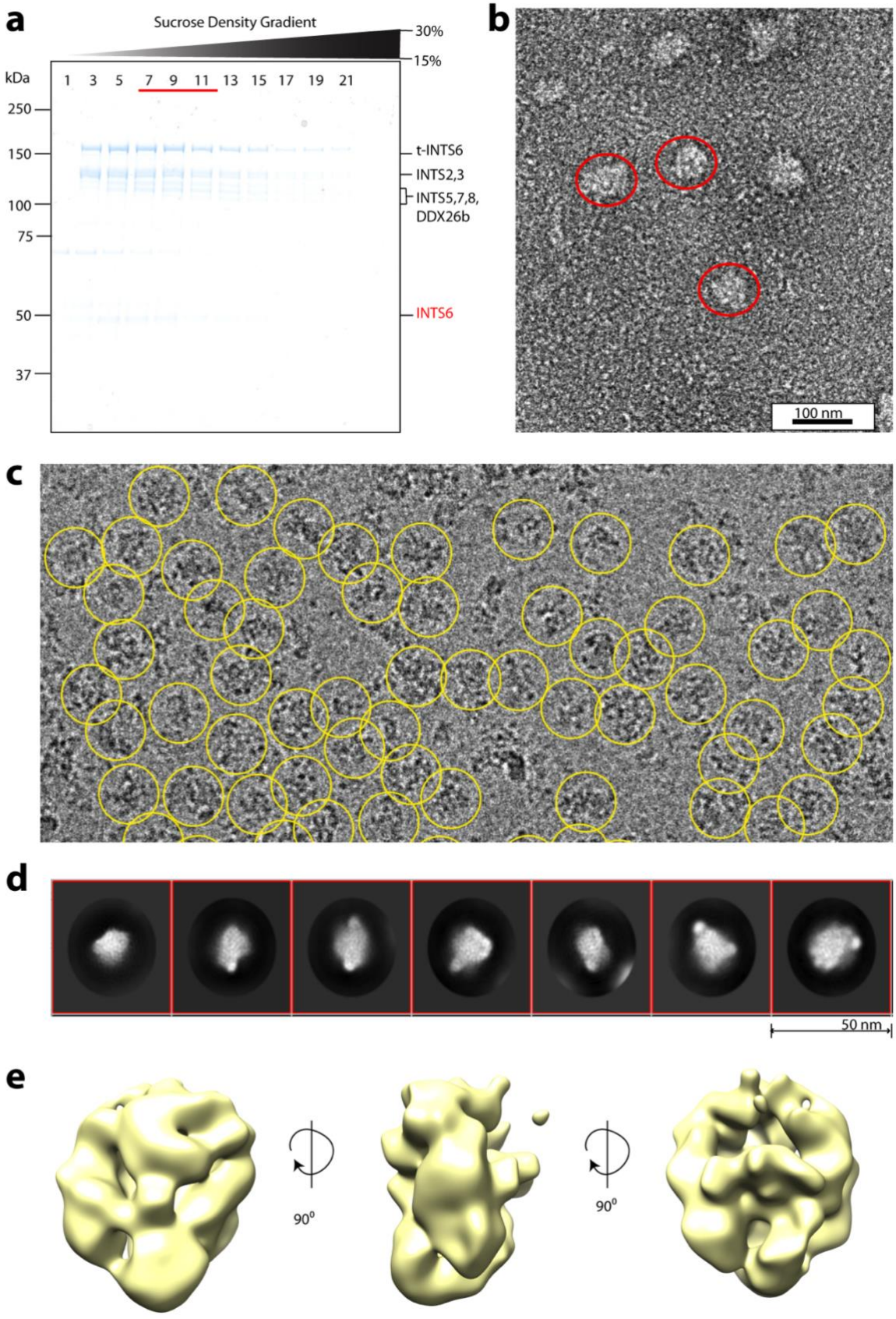


Figure 4.10. Purification and characterization of core-INT. (**a**) LDS-PAGE analysis of a 15-30% sucrose density gradient of core-INT. Odd numbered fractions were analyzed. Proteins in peak fraction (11) were identified by mass spectrometry. (**b**) A section of a uranyl acetate negative stain electron microscopy

micrograph of the peak fraction of the sucrose density gradient. Particles of the core-INT are indicated with Red circles/ovals. Micrograph was acquired at 37,000x magnification and a pixel size of 3.3 Å/pixel A scale bar is provided. (**c**) A section of cryo electron microscopy micrograph at 120,000 magnification and a pixel size of 1.23 Å/pixel. Particles were automatically picked in WARP (Tegunov & Cramer, 2019) (yellow circles). (**d**) A diagram of selected 2D class averages of core-INT processed in cryoSPARC2 (Punjani et al., 2017). A particle box size of 500 Å was used. (**e**) Different views of a low-resolution electron microscopy map of core-INT. Initial model was generated from few selected classes in cryoSPARC2. The initial model was used for iterative classification and refinement of the data which converged in the map presented with an estimated resolution of 20 Å.

Given how homogenous the sample from sucrose gradient was under negative stain electron microscopy conditions, its behavior in ice was evaluated using cryo-EM. To this end, fractions 9 - 11 from sucrose density gradient were fixed by crosslinking with 0.1% (v/v) glutarylaldehyde (see methods). Samples were dialyzed overnight to remove the crosslinker and sucrose before cryo-grids were prepared (see method). Initial screening of the cryo-grids in a 200 kV Glacios transmission electron microscope revealed homogenous single particles (Figure 4.10c). The particles from cryo-EM screening looked similar in shape and dimension to those observed in the negative stain micrographs. The diameter of the particles in ice was between 250 - 350 Å which may represent different views of a complex of 600 - 700 kDa in size. The quality of the particles on the cryo grids were suitable for data collection and single particle analysis. An overnight data set was collected on the 200 kV Glacios transmission electron microscope at a magnification of 120,000x and a pixel size of 1.23 Å/pixels which yielded ~250,000 particles picked and preprocessed in Warp (Tegunov & Cramer, 2019). 2D class averages were calculated using cryoSPARC (Punjani et al., 2017) (Figure 4.10d). The calculated 2D class averages lacked high resolution features suggesting a poor alignment of the particles. An initial 3D model was calculated from a selected group of the 2D class averages which was used for further iterative classification and refinement of the data set yielding a doughnut shaped 3D volume for the core-INT at an estimated resolution of 20 Å. I could not obtain high resolution features of the core-INT from this data set due to misalignment of the particles. This may be due to conformational heterogeneity/flexibility in the complex. It is also possible core-INT maybe be denatured by freezing during cryo-grid preparation (Nogales et al., 2016).

4.1.11. XL-MS reveal inter-subunit interactions within core-INT

The INTS3/6-DDX26B heterotrimer was identified as a subcomplex of INT based on high confidence chemical crosslinks between INTS3/6 heterodimer and DDX26B (Figure 4.6a). There was no crosslink between INTS3/6-DDX26B heterotrimer and subunits of INTS5/8 heterodimer as well as between INTS2/7 heterodimer and INTS5/8 heterodimer although pulldown experiments show interaction between these modules (Figure S7). Using co-infection assays, I showed INTS3/6-DDX26B heterotrimer interacts with INTS5/8 heterodimer resulting in a heteropentameric subcomplex (Figure 4.7 and 4.8) which in turn interacts with INTS2/7 heterodimer (Figure 4.9) to form the core-INT. To have a glimpse into how the subunits are interacting within the core-INT, we employed XL-MS using BS3 as well as the zero length crosslinker, 1-Ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC). Core-INT purified from ion exchange chromatography was crosslinked with 3 mM BS3 or 5 - 100 mM EDC at room temperature for 30 min. For EDC mediated XL-MS, I did a titration of the crosslinker from 0 – 100 mM and run the crosslinked core-INT on LDS-PAGE. Concentrations that showed sufficient crosslinked products (5-100 mM) were excised from the gel for mass spectrometric analysis.

Previously observed BS3 chemical crosslinks between INTS3 and DDX26B as well as between INTS3 and INTS6 were reproduced (compare Figure 4.6a and Figure 4.11a). Additionally, there were many high confidence crosslinks between C-terminal half of INTS6 to the N-terminal half of INTS3. Also, DDX26B had widespread crosslinks to INTS6 which were not observed in Figure 4.6a. INTS8 of the INTS5/8 heterodimer had BS3 crosslinks to all subunits of the core INT (especially to INTS6) but INTS7 and INTS5. This suggests that, INTS8 makes several contacts with subunits of the core INT and brings with it INTS5. There was no inter-subunit BS3 crosslink between INTS5 and the other subunits of core INT. This might be due to lack of Lys residues in the interacting interface of INTS5 or INTS5 is buried in the inner core of the complex inaccessible to the crosslinker. INTS2 of the INTS2/7 heterodimer had extensive crosslinks to INTS6. Particularly, the C-terminal half had many crosslinks to the C-terminal half of INTS6 in the same region that has crosslinks to INTS3 creating an interaction hub between the three proteins. INTS2 additionally has BS3 crosslinks to its interaction partner, INTS7 unlike in Figure 4.6a where there were no crosslinks observed between subunits of this heterodimer.

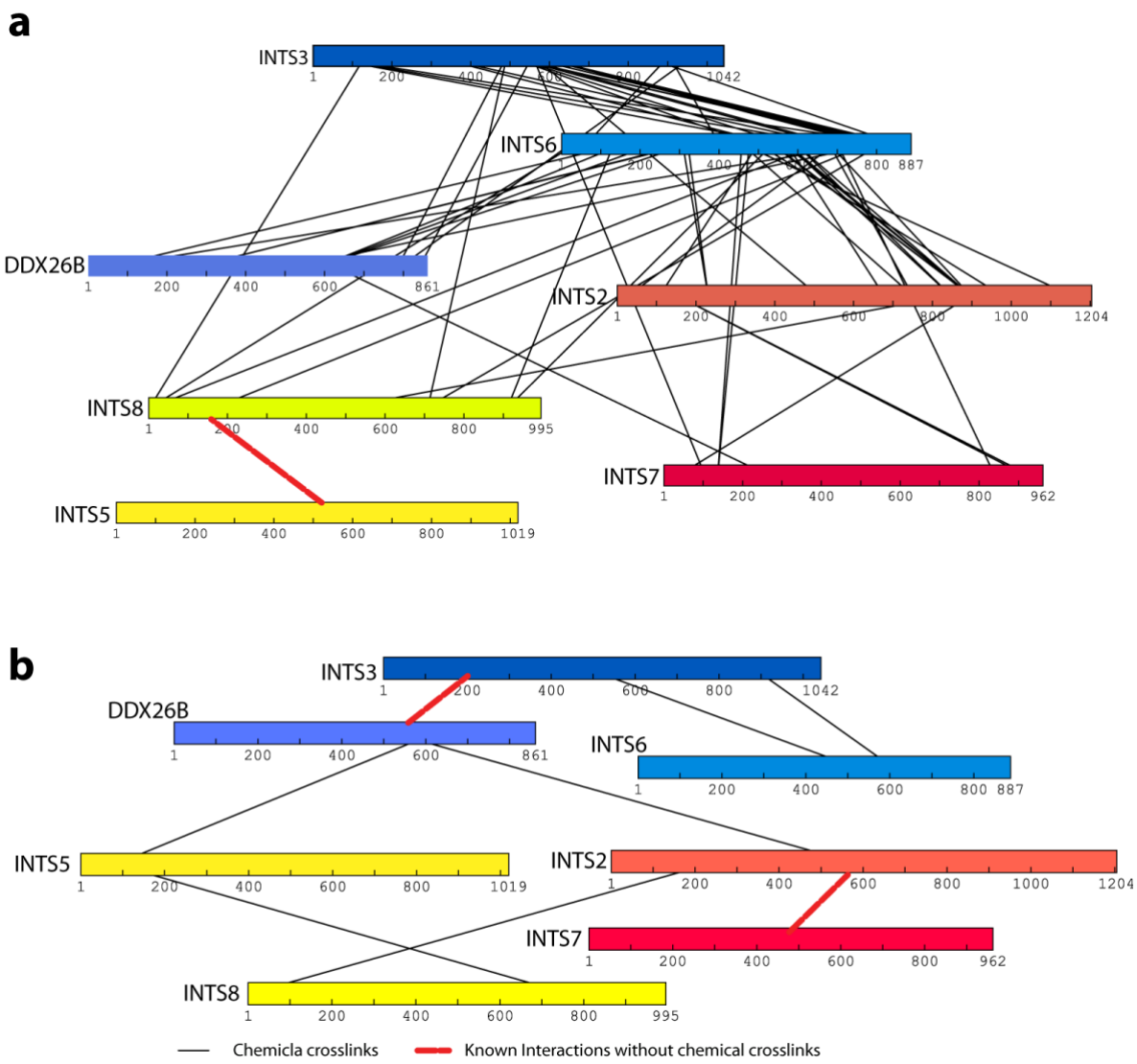


Figure 4.11. Crosslinking mass spectrometry reveals inter-subunit connections in core-INT. (**a**) A cartoon representing BS3 chemical crosslinking network of core-INT. Subunits are indicated as rectangles and chemical crosslinking between Lys residues are indicated with black line segments. Red dashed lines between INTS5 and INTS8 indicates they are known interacting subunits. (**b**) A cartoon of EDC mediated chemical crosslinking network within core-INT. Coloring scheme and legend is the same as in (**a**). A file with all observed crosslinks can be provided upon request.

There were also BS3 chemical crosslinks between INTS2 and INTS8 which may buttress the observed interaction between INTS2/7 and INTS5/8 heterodimers (Figure S7 lane 11). BS3 crosslinking depends on the availability of Lys residue within a crosslinking distance (< 30 Å) (Leitner et al., 2010, 2016). This constrain means that only protein-protein interaction interfaces with Lys residues can be captured with BS3 mediated XL-MS. EDC, a so called zero length crosslinker, provides a complementary option to buttress BS3 XL-MS results because it crosslinks amino groups of Lys side chains (and alpha-amino groups of N-terminal residues) to carboxyl groups of Asp and Glu side chains and carboxyl group of C-terminal residues. The

caveat to EDC XL-MS is that, due to its short length, crosslinking groups must be in close proximity. This distance restrain leads to very low number of crosslinked peptides when EDC is used for XL-MS experiment. Nonetheless, EDC crosslinks are very valuable because of this distance restrain (Leitner et al., 2010). EDC mediated XL-MS was done on the core-INT. As expected, there were very few crosslinked peptides observed compared to BS3 mediated XL-MS (Figure 4.11b) despite excess EDC in the crosslinking reaction. EDC mediated chemical crosslinks were observed between the N-terminal region of INT5 and the C-terminal regions of DDX26B and INTS8. Interaction between these regions of these subunits were not captured by BS3 mediated XL-MS exemplifying how EDC mediated XL-MS can complement BS3 mediated XL-MS.

4.1.12. Purification of INTS4/9/11 and INTS10/13/14 heterotrimers

Interaction between INTS4 and INTS9/11 heterodimer led to a better expression and solubility of INTS9/11 heterodimer in affinity pulldown experiments (Figure 4.2a). To arrive at the construct that showed good affinity purification, I tested different tagging options. 6xHis-MBP affinity tag on INTS4 resulted in poor expression of INTS4 and hence the complex could not be purified. On the contrary, having a 6xHis tag on INTS4 allowed good expression of INTS4 and purification of the INTS4/9/11 trimer (Albrecht et al., 2018). This suggests that the bulky MBP tag might be interfering with the assembly of INTS4/9/11 trimer. In purification of INTS4 alone, 6xHis-MBP affinity tag did not have any adverse effect (Figure S5c and d). Additionally, 6xHis-MBP affinity tagging of INTS11 was also tested. This tagging option proved to be efficient yielding high purity and stoichiometric INTS4/9/11(Figure 4.2a). The final purification strategy used included a Ni affinity purification followed by amylose affinity. The affinity tag was then removed using 6xHis tagged TEV protease followed by Ni affinity purification to remove the TEV protease and the affinity tag. The trimeric complex was finally purified by gel filtration. The stoichiometric trimer eluted with a peak fraction at 16.5 ml on a Superose 6 Increase 10/300 gel filtration column (Figure 4.12a). Albrecht and colleagues had a similar result when they purified the INTS4/9/11 heterotrimer using a Tris based buffer system and characterized it as the cleavage module (CM) of the Integrator complex (Albrecht et al., 2018).

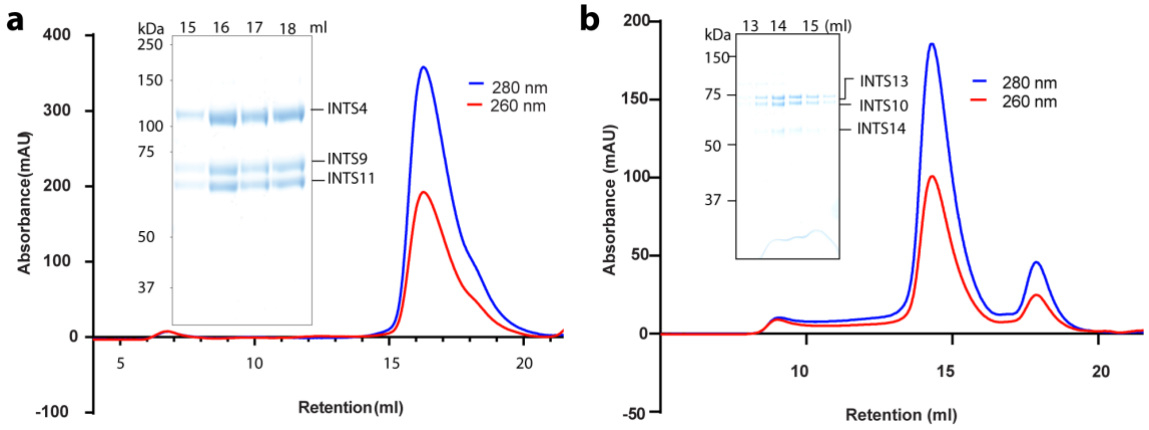


Figure 4.12. Purification of INTS4/9/11 and INTS10/13/14 trimeric subcomplexes of INT. A gel filtration chromatogram displaying UV absorbance of collected fractions at 260 and 280 nm. Fractions analyzed are indicated on top of the LDS-PAGE gel of the purification of INTS4/9/11 (**a**) and INTS10/13/14 (**b**).

The protein Asunder (Asu) and Von Willebrand associated protein 9 (VWA9) have been identified as additional subunits of INT in an RNAi screen in *Drosophila* and named INTS13 and INTS14 respectively (Jiandong Chen et al., 2012b). Chen and coworkers also showed interaction between *Drosophila* IntS10 and IntS14. By systematic co-expression, I discovered in this study that INTS13 and INTS14 forms a heterodimer (Figure 4.3b) which is stable under gel filtration conditions (Figure S5a and b). Furthermore, I demonstrated that the INTS13/14 heterodimer interacts with INTS10 using co-expression and pulldown assays (Figure 4.3d, e and f). To test the stability of this complex, a purification strategy was established which involves affinity purification via HisTrap and amylose affinity chromatography and a final gel filtration step to show the complex is homogeneous (Figure 4.12b). To get full length proteins it was necessary to affinity tag INTS13 as it turn to have N-terminus degradation. Using 6xHis affinity tag works as good as using the MBP affinity tag on INTS13 for the purification of the trimeric subcomplex.

4.1.13 The Cleavage module interacts with INTS10/13/14 heterotrimer.

Amylose affinity pulldown using purified cleavage module (CM) and INTS13/14 heterodimer shows these subcomplexes interact (Figure S7a lane 6). It was therefore expected that the cleavage module will interact with the INTS10/13/14 heterotrimeric subcomplex. To test this, I immobilized purified 6xHis-MBP tag INTS10 on amylose resin and incubated it with the purified INTS13/14 heterodimer and the CM. Results in Figure S7c Lane 7 shows the formation of a hexameric complex of INTS4, INTS9, INTS10, INTS11, INTS13 and INTS14. To further corroborate this pulldown result, I formed the hexameric complex using purified CM and

INTS10/13/14 heterotrimer. The reconstituted hexamer was purified from the free heterotrimers via analytical gel filtration using Superose 6 Increase 3.2/300 column (Figure 4.13a). There was a clear shift in the retention volume of the CM showing the formation of a bigger complex. Based on these results, the INTS10/13/14 heterotrimer is named the cleavage module interacting module (CMIM). To test whether this interaction between the two modules is stable enough for co-purification, a baculovirus expressing the six subunits with an affinity tag on INTS13 was created. It was important to tag INTS13 because it suffers a small N-terminus degradation which abolishes the interaction between the two modules (results not shown). Amylose affinity pulldown from SF21 insect cells infected with a baculovirus harboring the expression construct for the hexameric subcomplex confirms the formation of the hexameric complex (Figure 4.13b). However, attempts to co-purify the hexameric complex routinely led to the purification of just the CM or the CMIM when INTS11 or INTS13 was affinity tagged respectively. This suggests the interaction between the two modules is not strong enough to withstand the purification conditions tested (results not shown). Furthermore, attempts to visualize the hexameric complex under cryo conditions failed as the complex fell apart into the constituent subcomplexes. This was the case with or without fixation with various crosslinkers. This observation also shows that the interaction between the two modules is rather weak confirming results from the attempted co-purification.

The hexameric complex was further characterized using BS3 mediated XL-MS to understand how the subunits within the two trimers are interacting and how the two trimers are interacting to form the hexamer (Figure 4.13c). As expected, high confidence crosslinks were observed between the known interacting C-terminal regions of INTS9 and INTS11 (Wu et al., 2017) showing the validity of the crosslinking experiment. There were also a lot of high confidence crosslinks between the region spanning the predicted HEAT repeat of INTS4 and the C-terminal region of INTS11. This agrees well with the proposed binding of INTS4 via its HEAT repeats to the C-terminal interacting domains of INTS9 and INTS11 (Albrecht et al., 2018). A predicted N-terminal loop of INTS4 (Residue 1-20) also had a lot of crosslinks to one specific Lys residue (Lys159) of INTS9 suggesting INTS4 makes addition contact with INTS9 forming a triangular interaction network. The crosslink of INTS9(Lys159) to INTS4 was also observed when INT was identified as a substrate of PP2A (Herzog et al., 2012b).

The CMIM on the other hand, forms a linear interaction network with no crosslinks between INTS10 and INTS14. *Drosophila* IntS10 and 14 have been shown to interact (Jiandong Chen et al., 2012b). There is only 21% and 30% sequence identity between *Drosophila* and human INTS10 and INTS14 respectively from amino acid sequence alignment. It is therefore likely

the interaction between the two subunits is not conserved in the human INT. It is also possible the interaction interface between the two subunits does not have Lys residues.

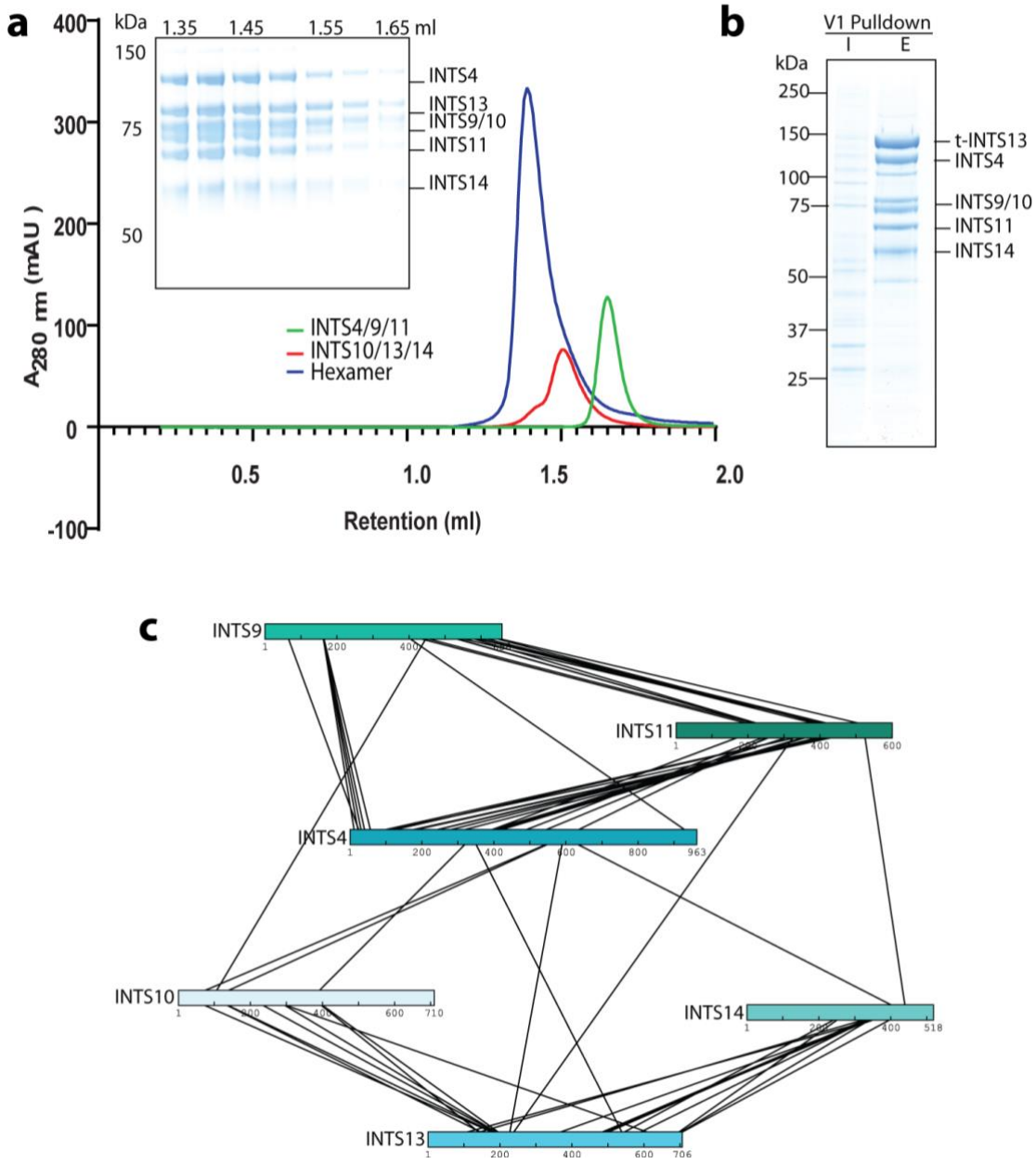


Figure 4.13. Protein-protein interaction between the cleavage module and INTS10/13/14 heterotrimer. (**a**) Assembly of a hexameric complex of the cleavage module and INTS10/13/14 using analytical gel filtration. Chromatograms showing the elution profiles of cleavage module (green), the INTS10/13/14 heterotrimer (red) and the resultant hexameric complex (blue) are shown. An LDS-PAGE gel of fractions from the hexameric complex is shown with the retention volumes of the fractions analyzed. (**b**) LDS-PAGE analysis of amylose affinity pulldown from SF21 insect cells expressing the hexameric complex. Input (I) and elution (E) fractions were analyzed. Proteins bands in the elution fraction were identified by mass spectrometry. The prefix letter 't' stands for 6xHis-MBP affinity tag. (**c**) A cartoon representing the BS3 mediated chemical crosslinking network

within the hexameric complex. Colored rectangles represent different subunits of the complex with amino acid residue numbers provided. Black line segments represent observed chemical crosslinks. A file with all observed crosslinks with their spectra counts and max scores can be provided upon request.

The N-terminal region of INTS13 had several crosslinks to the N-terminal half of INTS10 whereas the C-terminal half had crosslinks to INTS14. This way, INTS13 connects INTS10 and INTS14 in this heterotrimer in the human INT. There were very few inter-subcomplex crosslinks compared to the crosslinks within the two subcomplexes in the hexamer. Whilst this lack of inter-subcomplex crosslinks can be explained in terms of lack of Lys residues at interaction interfaces, it is likely due to the transient, weak and conformationally flexible nature of the interaction between the two modules. This later explanation is supported by the instability of the hexameric complex during purification and under cryo conditions.

4.1.14 Reconstitution of 10-subunit and 13-subunit subcomplexes of INT

The core-INT (section 4.1.9 and 4.1.10), the cleavage module (CM) and the cleavage module interacting module (CMIM) (section 4.1.12) together constitute 13 of the 15 subunits of the INT. The purification of these modules has been established (described in the aforementioned sections). The interaction between these modules was tested to ensure they are functional subcomplexes of the INT. Firstly, since the interaction between the CM and the CMIM was not stable under purification condition, I tested the stability of the interaction between the core-INT and the CMIM under purification conditions. The CMIM had shown a better interaction with INTS3/6-DDX26B trimeric component of the core-INT compared to the CM (Figure 4.7 and Figure S7d lane 6) whereas INTS5/8 and INTS2/7 heterodimers did not have any interactions with either the CM or the CMIM. To test interaction between the core-INT and the CMIM, I co-infected Hi5 insect cells with baculoviruses expressing the two subcomplexes with affinity tag on INTS6 as in the core-INT (Figure 4.14a). The purification of the 10-subunit complex of core-INT and CMIM was done using the purification strategy described for core-INT. Analysis of the elution fractions from ion exchange on LDS-PAGE followed by mass spectrometric identification of the proteins showed INTS10, INTS13 and INTS14 (subunits of the CMIM) co-purify with the core-INT via a single affinity tag on core-INT subunit, INTS6 (Figure 4.14b). The CMIM does not bind to HiTrap Q anion exchange column when purified on its own and must have high affinity binding to the core-INT to form a stable complex under ion exchange conditions up to 500 mM NaCl. This shows that, the CMIM interacts stably with the core-INT having survived a full purification protocol including ion exchange chromatography contrary

to its interaction with the CM described in section 4.1.13. However, the yield of this 10-subunits subcomplexes was extremely low (< 100 µg from 4-5 L of insect cells).

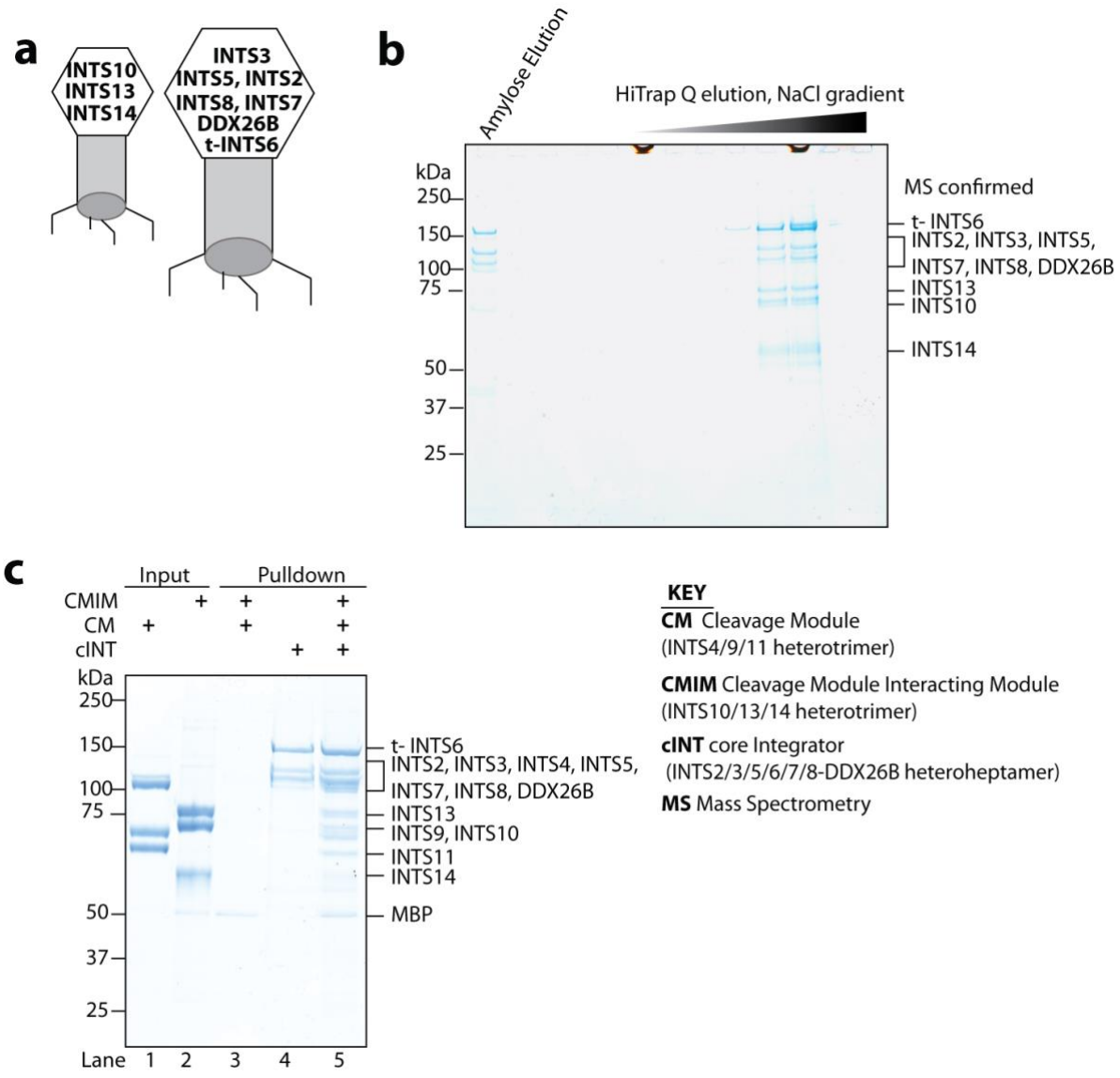


Figure 4.14. Interaction between the core-INT, cleavage module and the cleavage module interacting module. (**a**) A cartoon representing baculoviruses expressing core-INT and cleavage module interacting module. The prefix 't' on INTS6 represent 6xHis-MBP affinity tag. (**b**) An LDS-PAGE analysis of fractions from a HiTrap Q ion exchange purification of the 10-subunits subcomplex of INT from co-infection of the two baculoviruses in (a). The elution from amylose affinity purification was directly applied to the HiTrap and the flow through from the HiTrap (Amylose elution) was collected and analyzed alongside the peak elution fractions from the HiTrap Q. Proteins in the elution were identified by mass spectrometry. (**c**) Amylose affinity pulldown. Affinity tagged core-INT was immobilized on amylose beads and then incubated with CM and CMIM or none. As a background binding control of CM and CMIM were also incubated with beads without core-INT. Elution fractions from all three set-ups were analyzed. A key explaining acronyms is provided.

Since the yield of the core-INT is slightly better, it was better to reconstitute the complex from purified subcomplexes. I tested the formation of a 13-subunit complex of these three modules of INT from purified components using amylose affinity pulldown. To this end, purified core-INT was immobilized on amylose beads using the affinity tag on INTS6. Bound core-INT was then incubated with CM and CMIM. A background binding control was also done where only CM and CMIM were incubated with the beads without the affinity tagged core-INT. LDS-PAGE analysis of the maltose elution from the background control shows the CM and CMIM do not bind nonspecifically to the beads (Figure 4.14c lane 3) whiles core-INT bound as expected (Figure 4.14c lane 4). The two modules bound to the core-INT and co-eluted showing the formation of a 13 subunits subcomplex of the INT (Figure 4.14c lane 5). The complex was also reproduced using analytical gel filtration (not shown).

4.1.15 Reconstitution of the full Integrator complex from purified components

The reconstituted 13-subunit subcomplex of the INT described in the preceding section (4.1.14) was missing the 250 kDa INTS1 and the 49 kDa PHD zinc finger containing INTS12 subunits of INT. Although these subunits are known to interact in *Drosophila* (Jiandong Chen et al., 2013) and now in human (Figure 4.4a), this heterodimer could not be purified because of the poor expression and degradation of the subunits during purification. INTS1 showed interactions with other subcomplexes such as INTS13/14 heterodimer (Figure 4.4b) and INTS3/6 (Figure 4.4c). Purification of neither complexes yielded full length INTS1. Attempted purifications of the full length INTS1 lead to low yield and aggregated protein (Figure S1). Purification of INTS12 led to a severe degradation of the full-length protein to ~25 kDa N-terminal fragment. Protein disorder prediction of INTS12 showed a small structured domain of residues 1-50 shown to bind INTS1 in *Drosophila* (Jiandong Chen et al., 2013) and the PHD domain (159 - 215). The rest of the protein is predicted to be disordered (Figure 4.1 INTS12). The N-terminal region INTS12(1-194) was cloned and a purification was attempted (Figure S2). This truncated domain of INTS12 was also aggregated suggesting it lacks some interaction partner. I used amylose affinity pulldown to test whether these partially purified INTS1 and INTS12 interact with the 13-subunits INT subcomplex described in section 4.14.

To achieve this, affinity tagged core-INT was arrested on amylose beads and then incubated with the CM, CMIM, INTS1 and INTS12. A background binding of the untagged subcomplexes/subunits was tested by incubating them with amylose beads in the absence of the affinity tagged core-INT. An LDS-PAGE analysis of elution fractions showed in Figure 4.15a

lane 8 indicates the untagged subcomplexes/subunits do not interact nonspecifically with the beads. The CM and the CMIM co-purified with the core-INT in the elution of the test (Figure 4.15a lane 9,10 and 11) as observed previously in Figure 4.14c lane5.

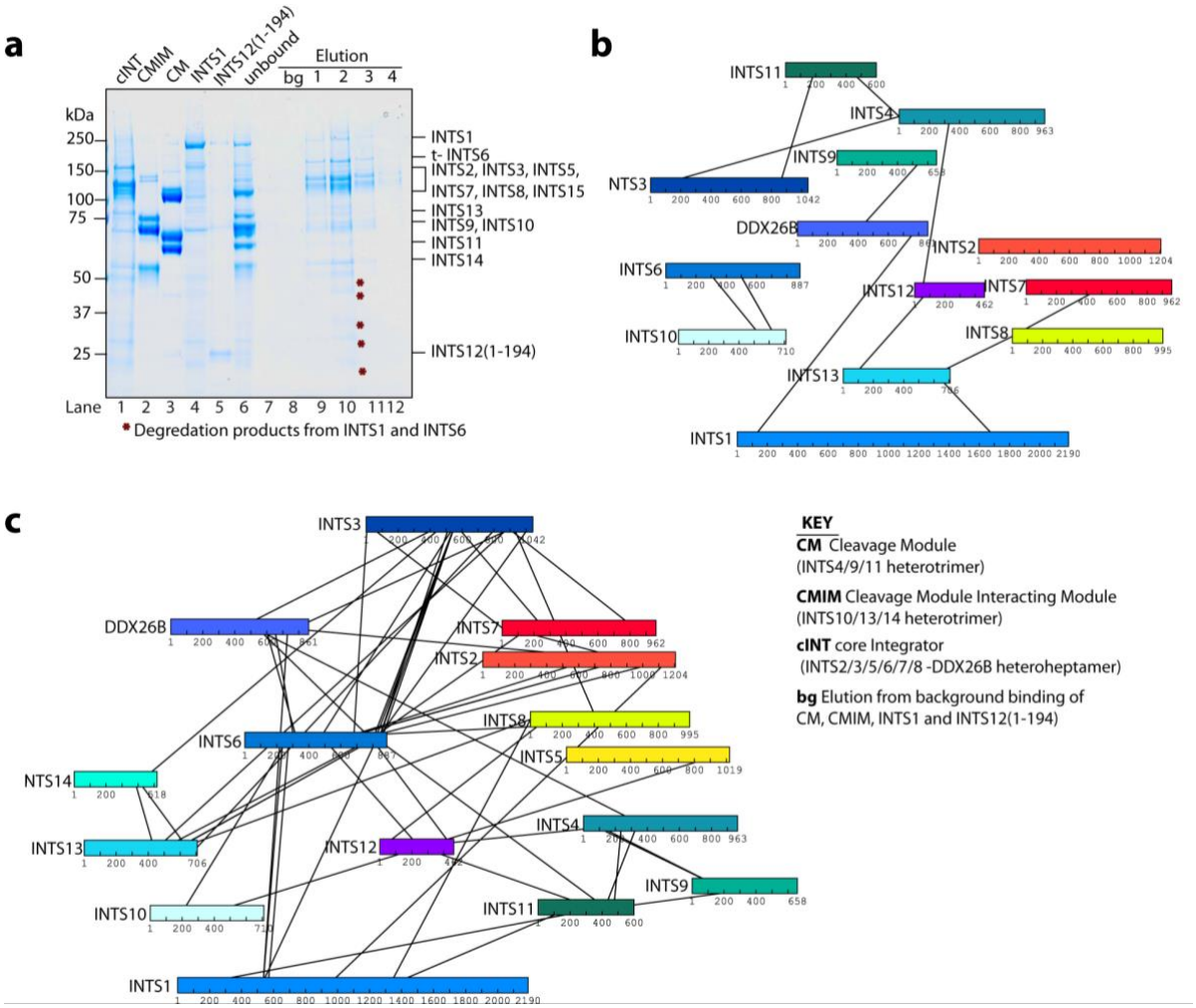


Figure 4.15. Reconstitution of full INT and its crosslinking network. (**a**) Amylose affinity pulldown. Affinity tagged core-INT was immobilized on amylose beads and then incubated with CM and CMIM, INTS1 and INTS12. As a background binding control, INTS1, INTS12, CM and CMIM were also incubated with beads without core-INT. Elution fractions from the background test (bg) as well four sequential elutions from the test binding were analyzed on LDS-PAGE. (**b**) A cartoon representing EDC mediated chemical crosslinking network of the reconstituted INT. Subunits are indicated as rectangles and chemical crosslinks between Lys residues are indicated with black line segments. (**c**) A cartoon of BS3 mediated chemical crosslinking network within INT. Coloring scheme and legend is the same as in (b). A key explaining acronyms is provided. A file detailing all crosslinks and their scores can be provided on request.

Interestingly, partially purified INTS1 and INTS12 also co-purified with the 13-subunits indicating the formation of the full INT (Figure 4.15a lane 9,10 and 11). This indicates that, the

partially purified INTS1 and INTS12 are in a state that they can be bound by their cognate interaction partners/interfaces within the INT.

To further characterize how the different subunits and subcomplexes of INT are interacting, we used BS3 and EDC mediated XL-MS (Figure 4.15b and c). In contrast to XL-MS results presented in section 4.1.6, here a complex assembled and purified via a single affinity tag (on INTS6 subunit of core-INT) was used ensuring a better purity and homogeneity. The elution fractions from the amylose affinity pulldown was crosslinked with 3 mM BS3 or a gradient from 5 – 100 mM EDC as described for the core-INT (section 4.1.11). For EDC crosslinking, samples were loaded on LDS-PAGE and bands corresponding to crosslinked complexes were excised for mass spectrometric analysis. There were very few crosslinks peptides observed for the EDC crosslinked samples and INTS5 and INTS14 were not detected at all (Figure 4.15b). This might be due to the stringent distance restrains imposed by the EDC crosslinker. The in-gel digestion method used is usually associated with significant sample loses (Leitner et al., 2010) and may contribute to the low number of crosslinked peptides observed. The EDC mediated crosslinks between subunits of the core-INT observed in Figure 4.11b could not be reproduced. There were low confident crosslinks observed between subunits of different subcomplexes. The BS3 mediated XL-MS experiment on the other hand produces many high confidence crosslinks (Figure 4.15c). The crosslinking pattern within the core-INT subunits in this experiment was mostly consistent with the results of crosslinking the core-INT alone (Figure 4.11a) suggesting the core-INT did not undergo drastic conformational changes upon binding the other subcomplexes and subunits. There were high confidence crosslinks between subunits of the CM but not as many crosslinks were observed compared to when the complex of the CM and CMIM was crosslinked (Figure 4.13c). Notably, crosslinks between the C-terminal interacting domains of INTS9 and INTS11 were not observed. This might be an experimental artefact or the CM might be in a different conformation (potentially extended/relaxed) when in context of the full INT. The crosslinks observed between the CM and CMIM were also not observed in the context of the full INT. It is possible that the two modules may no longer be in close proximity in the context of the full INT. Crosslinks between the C-terminal regions of INTS13 and 14 were preserved between this experiment and when the complex of CM and CMIM was crosslinked whereas crosslinks between the N-terminal regions of INTS13 and INTS10 were no longer present. Interestingly, there were many high confidence crosslinks between the C-terminal region of INTS13 to the C-terminal regions of INTS3 and INTS6 of the core-INT. This may represent some of the interaction interfaces between the two modules accounting for the stability of the interaction between them seen in

Figure 4.14. There were crosslinks observed between different subunits to both INTS1 and INTS12 suggesting how these two subunits might be incorporated into the Integrator complex. To get a better overview of how these two subunits interact with the rest of the subunits of the INT, their purification must be improved.

4.1.16 Identification of a soluble domain in INTS1 and its interaction with INTS12

The largest subunit of INT, INTS1 was very challenging to express and purify (Figure S1). INTS1 does not have any known folded domain except the predicted N-terminal DUF3677 domain. Its amino acid sequence has a lot of predicted disordered domains (Figure 4.1 INTS1). I could not co-purify the full length INTS1 with its identified interaction partner (INTS12) or subcomplexes (INTS13/14 and INTS3/6 heterodimers) (Figure 4.4) as INTS1 was poorly expressed and degrades. This shows that, these interaction partners are not able to stabilize the full length INTS1. To test which part of INTS1 is expressible and soluble on its own, the INTS1 protein was divided into an N-terminal half (residues 1-1010) and a C-terminal half (residues 1010-2190). I expressed these variants as well as the DUF3677-containing N-terminal domain (residues 1-294) and a small predicted domain in the C-terminal region (residues 2045 – 2190) in the absence of solubility tags. A Ni affinity pulldown from Hi5 insect cells expressing 6xHis tagged versions of these variants showed that only the DUF3677-containg N-terminal domain (residues 1-294) was clearly overexpressed (Figure 4.16a lane 1) while the other variants were at background levels. Attempts to purify this soluble domain showed it was co-purifying with a chaperon and forming oligomers (Figure S1e and f). This result implies that INTS1(1-294) is lacking some interaction partner. To identify which subcomplex of INT binds this domain of INTS1, a co-infection assay between a baculovirus expressing 6xHis tagged INTS1(1-294) and baculoviruses expressing core-INT, CM or the CMIM was done. Each of these modules had an affinity tagged subunit for subsequent amylose affinity pulldown. Additionally, I made a baculovirus expressing a 6xHis tagged INTS1(1-294) and INTS12(1-194) to test if these two domains interact (Figure 4.16b).

A western blot analysis of input and elution samples of amylose affinity pulldown from Hi5 cells expressing 6xHis tagged INTS1(1-294) shows the protein is expressed and does not bind amylose beads nonspecifically (Figure 4.16c lane 1 and 2). The same analysis was done on Hi5 cells co-infected with combinations of baculoviruses shown is Figure 4.16b and c. There was signal for 6xHis tagged INTS1(1-294) in the elution fractions indicating this domain of INTS1 have some interaction with all the subcomplexes of INT. Given that the same number of cells was used in the pulldowns and all parameters were kept constant for all experiments (except

the expression levels), it appears that the CMIM (A+D) has a better interaction with this domain of INTS1 compared to the core-INT and the CM. Given the core-INT is poorly expressed compared to the CM, it might have a better interaction with INTS1(1-294) compared to the CM which is well expressed.

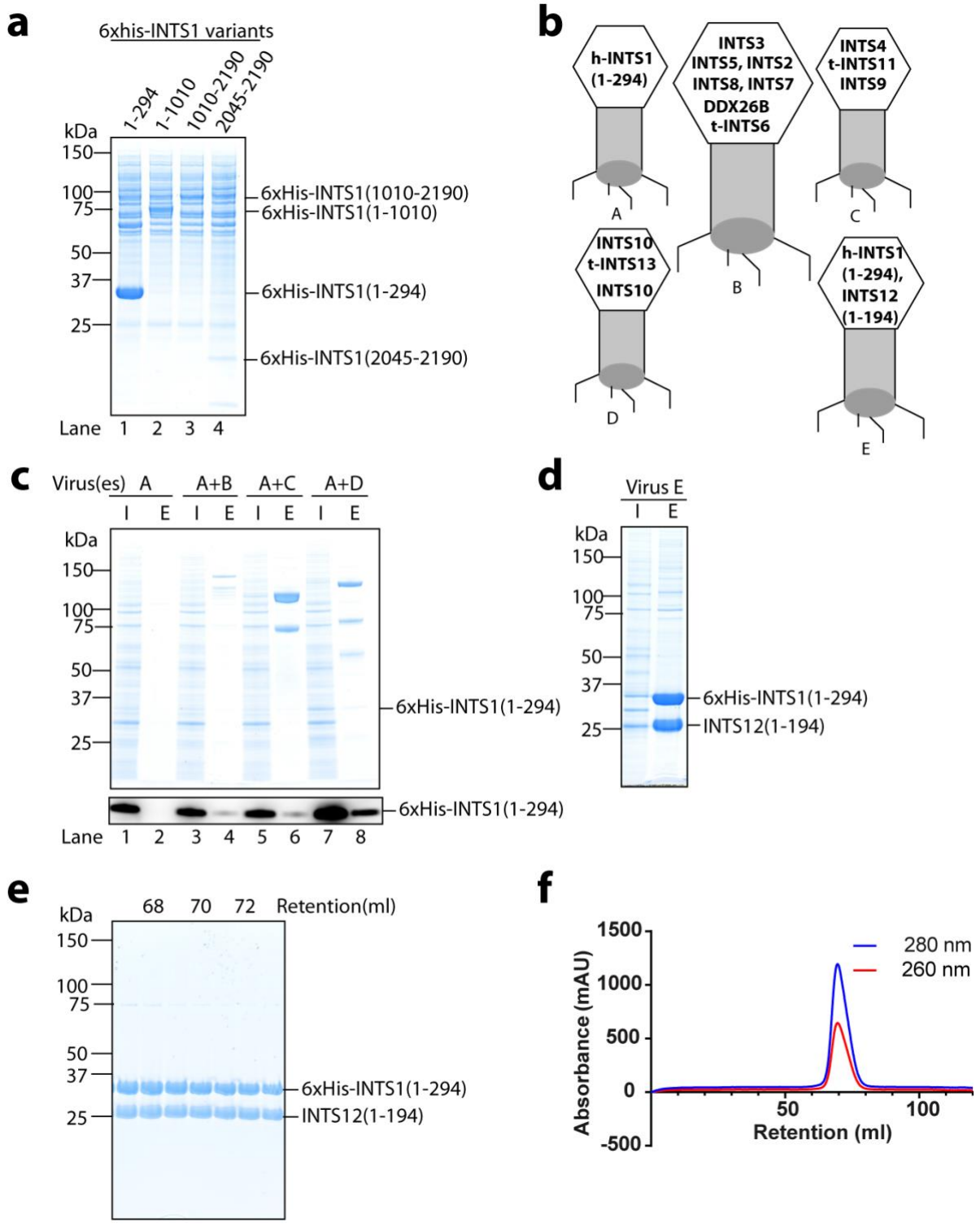


Figure 4.16. Identification of a soluble domain in INTS1 and its interactions with INT subcomplexes INTS12. (**a**) expression and solubility test of INTS1 domains. An LDS-PAGE analysis of elution samples of Ni affinity pulldown from insect cells expressing variants of INTS1. Residues encompassing each domain is indicated on top of the gel. (**b**) A cartoon representing baculoviruses used in co-infection assay. The INT subunits in each baculovirus are indicated and the subunit with affinity tag is labeled with 't'. (**c**) Amylose affinity pulldown from co-infection assay. An LDS-PAGE analysis of input (I) and elution (E) samples of pulldown from Hi5 cells infected with various combinations of baculoviruses in (b). The virus(es) used in each infection is/are indicated on top of the gel. A western blot analysis for 6xHis-INTS1(1-294) from the same samples is appended to the bottom of the LDS-PADE gel. (**d**) Ni affinity pulldown of 6xHis-INTS1(1-294) and INTS12(1-194). An LDS-PAGE analysis of input (I) and elution (E) samples of a Ni affinity pulldown from Hi5 cells expressing 6xHis-INTS1(1-294) and INTS12(1-194). (**e**) An LDS-PAGE analysis of fractions from gel filtration chromatographic purification of INTS1(1-294)/INTS12(1-194) heterodimer showing they form a homogenous complex. The retention volume of the fractions analyzed are indicated on top of the gel. (**f**) A gel filtration chromatogram of INTS1(1-294)/INTS12(1-194) purification showing the two domains interact and elute as a single homogenous peak.

The N-terminal domain of INTS1 is likely involved in the interaction observed between INTS13/14 heterodimer and INTS1 (Figure 4.4b).

An N-terminal microdomain in *Drosophila* INTS12 (residue 1-45) has been shown to interact with INTS1 (Jiandong Chen et al., 2013). A co-expression experiment demonstrated that human INTS1 and INTS12 also have interaction (Figure 4.4a). It is not known which part of INTS1 interacts with INTS12. To test whether the soluble and stable N-terminal domain of INTS1 (residue 1-294) interacts with the stable N-terminal domain of INTS12 described in Figure S2, I made a baculovirus expressing the two domains with a 6xHis tag on INTS1(1-294). A Ni affinity pulldown from Hi5 cells infected with this baculovirus shows the two domains interact and are co-purified via the affinity tag on INTS1(1-294) (Figure 4.16d). Both of the INTS1 and INTS12 stable and soluble domains forms oligomers when purified independently (Figure S1e, f and S2d, e) suggesting they lack interaction partner(s). Co-purification of these interacting domains however prevented this oligomerization (Figure 4.16e and f) showing that the residues that might be responsible for the oligomerization are involved in the interaction between these domains.

In conclusion, the inter-subunit interaction network within the Integrator complex has been delineated using co-infection in insect cells coupled to affinity purification and crosslinking mass spectrometry. INT can be divided into four subcomplexes namely the core-INT (comprising INTS2, INTS3, INTS5, INTS6, INTS7, INTS8 and DDX26B), the cleavage module (made up of INTS4, INTS9 and INTS11), the cleavage module interacting module

(composed of INTS10, INTS13 and INTS14) and INTS1/12 heterodimer (or their interacting N-terminal domains). The purification of these modules has been established and importantly the full INT can be reconstituted from the separately purified subcomplexes. I also provided a crosslinking network of the subunits (using BS3 mediated XL-MS) providing a proxy of how the subunits are connected within the Integrator complex in the absence of high-resolution structures.

4.2 Interaction between INT and RNA Pol II elongation complex

The role of the Integrator complex in RNA Polymerase II (Pol II) transcription has been widely studied and reviewed in (Baillat & Wagner, 2015; Rienzo & Casamassimi, 2016a). These studies suggest that the INT (or certain subunits) must have direct interactions with key players of Pol II transcription. However, these studies lack direct evidence of interaction between INT and Pol II transcription complexes using purified components. Here, I traced interaction between the Integrator complex and INTS3 and the pause elongation complex of Pol II-DSIF-NELF (PEC) (Vos, Farnung, Urlaub, et al., 2018) using the reconstituted INT and subunit INTS3.

4.2.1 INTS3 interacts with NELF and the Pause Elongation Complex (PEC)

The INTS3 subunit of INT was shown to be a part of the SOSS complex where is was called SOSS-A (Huang et al., 2009; Li et al., 2009; Ren et al., 2014; Skaar et al., 2009). This subunit was also shown to co-fractionate with the negative elongation complex (NELF) when NELF associated protein complexes from co-immunoprecipitation was separated on glycerol gradient (Stadelmayer et al., 2014). This suggests INTS3 may have a functional association with the NELF complex independent of INT. To test whether this interaction can be recapitulated in a purified system, I created a baculovirus co-expressing the four-subunits NELF complex with a 6xHis tag on NELF -C and MBP tagged INTS3. The NELF subunits, A and E have long stretches of disordered regions named tentacles (Vos, Farnung, Urlaub, et al., 2018). To test whether these tentacles are important for interaction with INTS3, the NELF tentacle mutants were also co-expressed with INTS3. Amylose affinity pulldown from Hi5 insect cells co-expressing MBP tag INTS3 and the NELF variants showed that INTS3 interacts with wild-type NELF as well as the tentacle mutants (Figure 4.17a). These results confirm the observed *in vivo* interaction between the NELF complex and INTS3 (Stadelmayer et al., 2014) and further shows that the disordered regions in NELF -A and -E are dispensable for this interaction (Figure 4.17a). To assess which regions of INTS3 are important for interaction with NELF, various

truncation mutants of INTS3 with MBP tags were created (Figure 4.17b). These truncations were co-expressed with wild-type NELF complex and amylose affinity pulldown followed by western blot for 6xHis tagged NELF -C was used to evaluate interactions. The results in Figure 4.17c shows that all the variants of INTS3 interact with the NELF complex. This implies multiple regions of the INTS3 protein contact various regions/subunits of the NELF complex. Attempts to purify full length INTS3 normally results in aggregation (not shown). Purification of INTS3-NELF complex shows the complex is stable under the purification conditions used (Figure 4.17d). Interestingly, the excess INTS3 by-product of the co-expression was not aggregated proposing co-expression with the NELF complex may help the folding and solubility of INTS3 (Figure 4.17e).

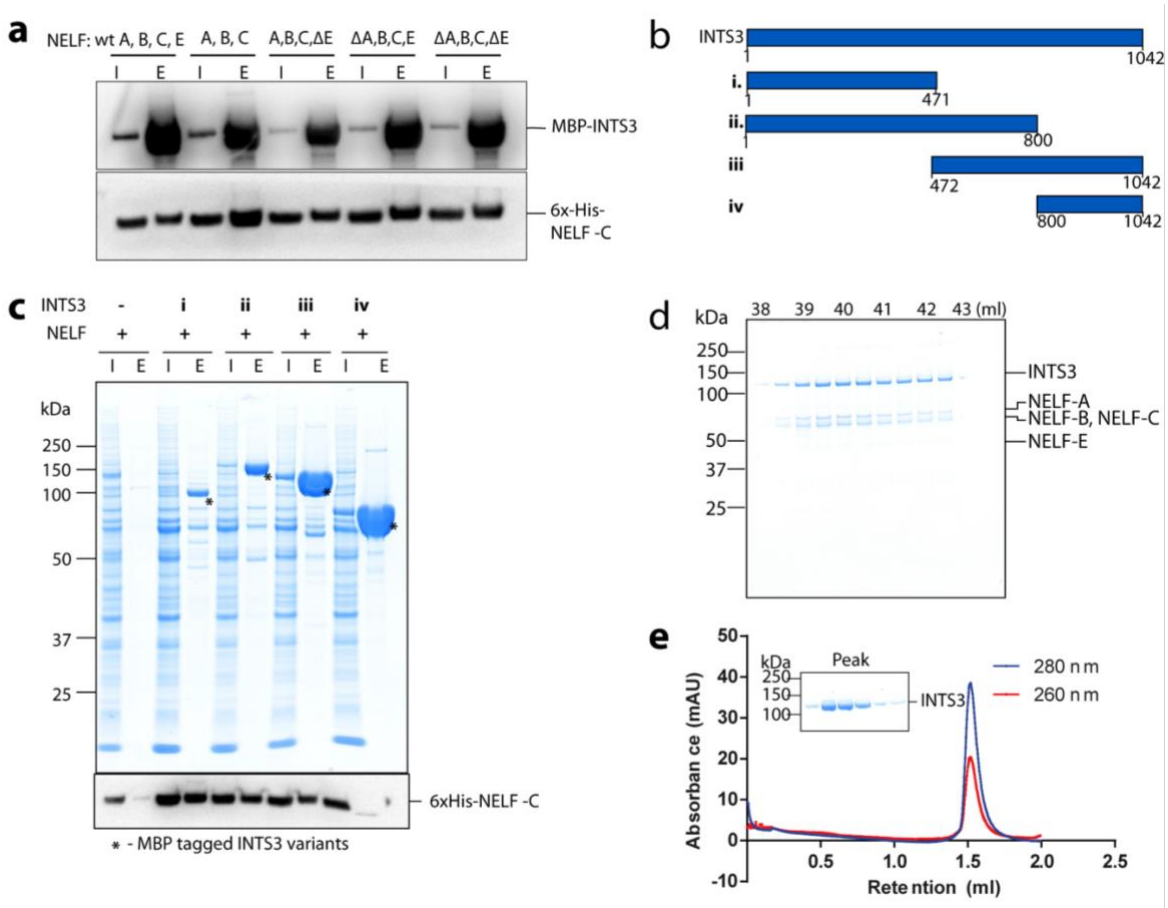


Figure 4.17. Interaction of INTS3 and NELF complex. (**a**) INTS3 interacts with wt-NELF and NELF tentacle mutants. Western blot analysis of input (I) and elution fractions (E) of amylose affinity pulldown from Hi5 cells co-expressing INTS3 and the various variants of the NELF complex. A, B, C, and E represents the NELF subunits, ΔA and ΔE represent NELF A(1-188) and E(1-138) tentacle mutants respectively. (**b**) Cartoon representing INTS3 truncation mutants with the residue numbers indicated. (**c**) Interactions between INTS3 variants and wt-NELF. LDS-PAGE and western blot analysis of input (I) and elution (E) fractions of amylose affinity pulldown from Hi5 cells co-expressing the NELF complex and the various truncations of INTS3 indicated on top of the gel. In variant iv, the western blot signal is masked by the high concentration of that

variant which co-migrates with NELF -C at the same apparent molecular weight. (**d**) Purification of INTS3-NELF complex. An LDS-PAGE analysis of peak fractions of a gel filtration chromatographic purification of INTS3-NELF complex. The retention volume of the fractions analyzed are indicated on top of the gel. (**e**) Purification of INTS3. A gel filtration chromatogram of the excess INTS3 by-product of INTS3-NELF complex purification showing the elution peak. A slice of an LDS-PAGE analysis of the peak fractions is appended.

The negative elongation complex (NELF) is so named for its role in repressing transcription elongation (Narita et al., 2003; Yamaguchi et al., 1999). A structure of repressed (paused) transcription elongation complex of Pol II, NELF and DSIF (PEC) was recently solved (Vos, Farnung, Urlaub, et al., 2018). I have recapitulated an interaction between INTS3 and NELF (Figure 4.17). The follow-up question was, how does the interaction between NELF and INTS3 affect NELF binding to the elongation complex? To address this question, the PEC was formed in the presence of INTS3. Results presented in Figure 4.18a and b show that a PEC can be formed in the presence of INTS3. The presence of INTS3 in the PEC did not have any discernible impact on *in vitro* transcription elongation and pausing (not shown).

We used XL-MS to further characterize how INTS3 is interacting with the PEC. To achieve this, the peak fractions of INTS3-PEC (Figure 4.18a and b) was crosslinked with 2 mM BS3 and the crosslinked product was analyzed by mass spectrometry. The previously observed crosslinking pattern within the PEC (Vos, Farnung, Urlaub, et al., 2018) was reproduced showing the complex was intact and did not undergo any dramatic conformational changes upon binding INTS3 (Figure 4.18c). As expected, INTS3 mostly crosslinked to the subunits of the NELF complex with few crosslinks to RPB2, RPB5 and RBP7 subunits of Pol II. Crosslinks to the Pol II subunits have low scores except the one to RPB5. There were no crosslinks between INTS3 and DSIF subunits, SPT4 and SPT5. The crosslinks between INTS3 and the NELF subunits span the entire length of INTS3 protein showing that, INTS3 makes multiple contacts with NELF confirming the results of the INTS3 truncations in Figure 4.17c. The RBP5 and RBP7 subunits of Pol II are located close to the NELF complex in the structure of the PEC (PDB 6GML) allowing INTS3 to contact these subunits of Pol II and hence the crosslinks to these subunits. The residue of RBP2 subunit that crosslinked to INTS3 on the contrary is located on the side of Pol II opposite to the side bound by NELF. The crosslink between INTS3 and RPB2 therefore may not represent a specific interaction between these proteins.

I attempted to determine a cryo-EM structure of the INTS3-PEC complex but obtained only the structures of PEC, Pol II-DSIF and Pol II (results not shown). This suggests INTS3 dissociated from NELF and for that matter PEC during cryo-EM sample preparation. Only 10% of the particles was the PEC showing that NELF has a low occupancy on the Pol II-DSIF complex.

This was also observed previously by Vos and colleagues (Vos et al., 2018) suggesting that additional factors might be needed to stabilize the paused complex of Pol II-DSIF-NELF.

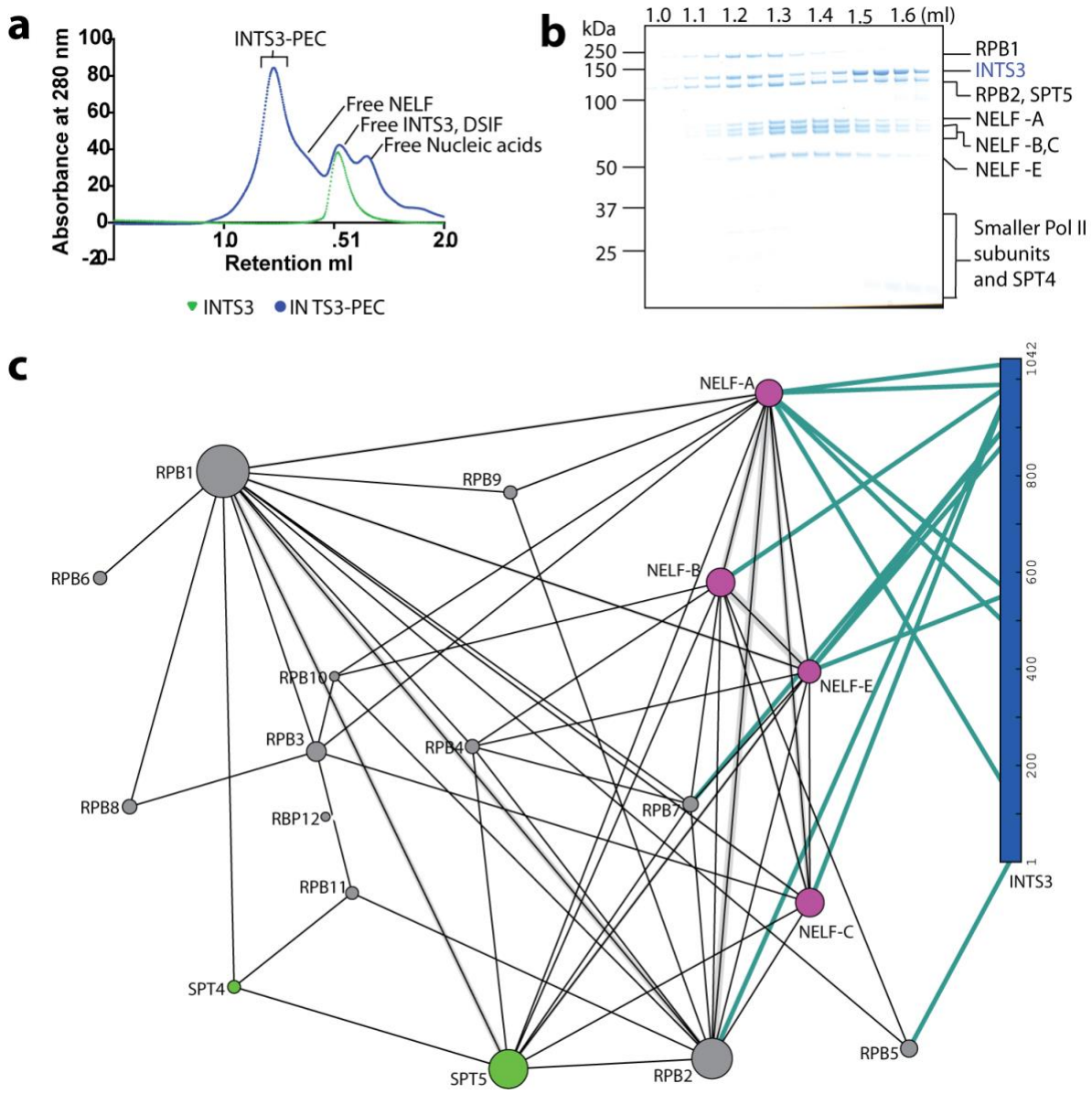


Figure 4.18. Interaction of INTS3 and the pause elongation complex (PEC). (a) Formation of an INTS3 – PEC. A gel filtration chromatogram (absorbance at 280 nm) showing formation of the INTS3-PEC (blue). The elution profile of INTS3 alone is shown in green and peaks of excess factors and nucleic acids are labeled. (**b**) Formation of INTS3-PEC. An LDS-PAGE analysis of the fractions from the gel filtration chromatographic separation of the INTS3-PEC complex. The retention volumes of the fractions analyzed are indicated on top of the gel. (**c**) A cartoon representing the crosslinking network between INTS3 (blue rectangle) and the PEC. Grey circles represent subunits of Pol II, light green circles (DSIF subunits SPT5 and SPT4) and magenta (NELF subunits). Crosslinks within the PEC are shown by black line segments and crosslinks between INTS3 and PEC are shown by the green line segments

4.2.2 The reconstituted INT interacts with the PEC

INT associates with Pol II to perform its roles in transcription elongation and termination. The reported roles of INT in transcription often involve DSIF and NELF which are a part of the PEC (Baillat & Wagner, 2015; Rienzo & Casamassimi, 2016a). The PEC therefore represents an attractive starting point to study the involvement of INT in Pol II transcription and regulation *in vitro*. During these initial studies, I did not explore the roles of CTD phosphorylations on the interaction between INT and PEC. Nonetheless endogenously purified Pol II may have the needed CTD phosphorylation for interaction with INT. To test whether INT binds to the PEC, PEC was formed as described previously (Vos, Farnung, Urlaub, et al., 2018) but in the presence of INT. The subcomplexes of INT were pre-incubated together to allow the assembly of INT. The preformed INT was added to Pol II before NELF and DSIF were added. I then purified the complex on a 15-30% sucrose density gradient. As controls, PEC formed from similar amount of Pol II, NELF and DSIF as well as the constituent subcomplexes of INT were also run on the gradient. The INTS1 and INTS12 interacting domains, the CM, and CMIM were in the top 8 fractions of the gradient (Figure 4.19a-c) and core-INT run as shown in Figure 4.10a. The PEC has a peak from fraction 8 to 10 which tails to fraction 13 (Figure 4.19e). In the presence of INT, Pol II, DSIF and NELF migrated to fraction 14, 15 and 16 showing the formation of a bigger complex of INT and PEC hereafter referred to as INT-PEC (Figure 4.19f and Table S1). The CM and CMIM alone do not bind the PEC when tested. Mass spectrometric analysis of fraction 16 identified unique peptides for all the expected proteins in the INT-PEC (Table S1). Based on this result, it can be concluded that the reconstituted INT interacts with the PEC.

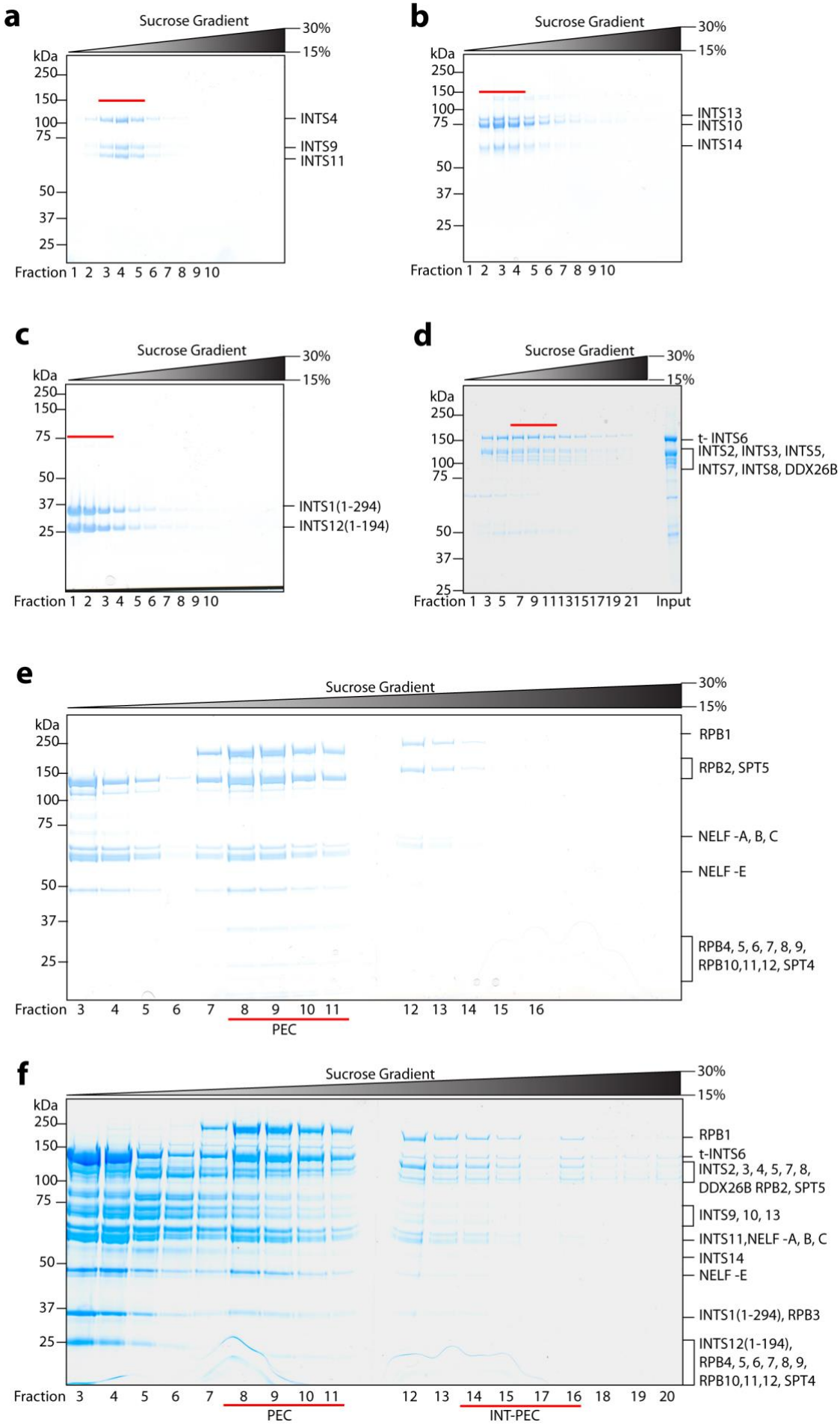

a
Sucrose Gradient
30%
15%
kDa
250
150
100
75
50
37
25
INTS4
INTS9
INTS11
Fraction 1 2 3 4 5 6 7 8 9 10
b
Sucrose Gradient
30%
15%
kDa
250
150
100
75
50
37
25
INTS13
INTS10
INTS14
Fraction 1 2 3 4 5 6 7 8 9 10
c
Sucrose Gradient
30%
15%
kDa
250
150
75
50
37
25
INTS1(1-294)
INTS12(1-194)
Fraction 1 2 3 4 5 6 7 8 9 10
d
Sucrose Gradient
30%
15%
kDa
250
150
100
75
50
37
25
t- INTS6
INTS2, INTS3, INTS5,
INTS7, INTS8, DDX26B
Fraction 1 3 5 7 9 11 13 15 17 19 21 Input
e
Sucrose Gradient
30%
15%
kDa
250
150
100
75
50
37
25
RPB1
RPB2, SPT5
NELF -A, B, C
NELF -E
RPB4, 5, 6, 7, 8, 9,
RPB10,11,12, SPT4
Fraction 3 4 5 6 7 8 9 10 11 12 13 14 15 16
PEC
f
Sucrose Gradient
30%
15%
kDa
250
150
100
75
50
37
25
RPB1
t-INTS6
INTS2, 3, 4, 5, 7, 8,
DDX26B RPB2, SPT5
INTS9, 10, 13
INTS11,NELF -A, B, C
INTS14
NELF -E
INTS1(1-294), RPB3
INTS12(1-194),
RPB4, 5, 6, 7, 8, 9,
RPB10,11,12, SPT4
Fraction 3 4 5 6 7 8 9 10 11 12 13 14 15 17 16 18 19 20
PEC
INT-PEC

Figure 4.19 Interaction between INT and PEC. An LDS-PAGE analysis of fractions from 15-30% sucrose density gradient of CM (**a**), CMIM (**b**), INTS1/12 interacting domains (**c**), core-INT same as Figure 4.10a (**d**). PEC (**e**) and INT-PEC (**f**). The fractions analyzed are indicated at the bottom of each gel. The red line segment shows the peak fractions. Note that in (**b**) INTS13 partly degrades and runs at the same molecular weight as INTS10. And in (**f**) fraction 17 was mistakenly loaded before 16.

4.2.3 XL-MS identifies subunits involved in the interaction between INT and PEC

To uncover which subunits of INT contact PEC, we used BS3 mediated XL-MS.

To this end, I formed the PEC and purified it on analytical gel filtration to remove excess NELF and DSIF. Preformed INT was then added to PEC and incubated for 30 min at 30 ºC. The complex was directly crosslinked without further purification due to limiting amounts of core-INT. Initial attempts to purify the INT-PEC prior to XL-MS analysis resulted in sample lost and very few crosslinks.

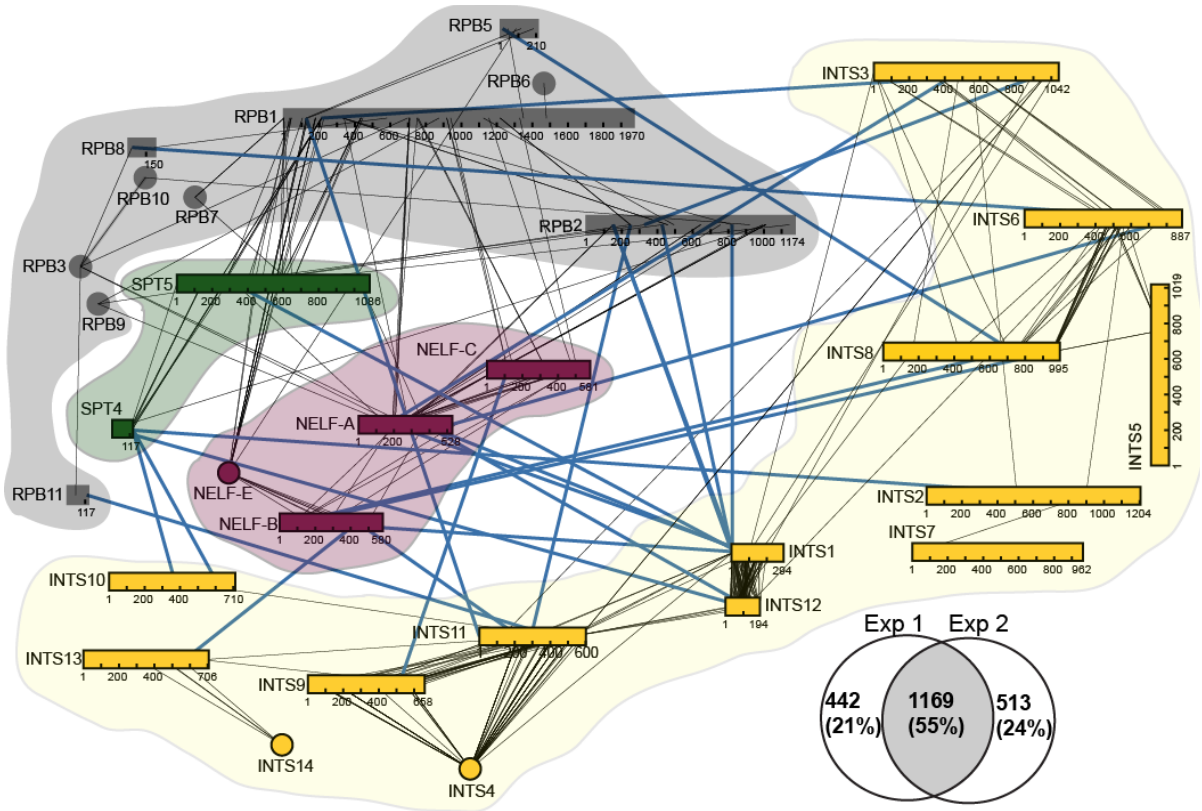


Figure 4.20. Overview of BS3 crosslinking network of INT-PEC. The Venn diagram shows the overlap between two independent experiments. Only crosslinks in the intersection were used for the final visualization. Pol II subunits are colored gray and highlighted in a gray background, DSIF subunits are colored green and highlighted by a green transparent background, NELF subunits are colored hot pink and highlighted with a transparent pink background and INT subunits are colored

orange and grouped by a yellow background. For clarity, only inter crosslinks are shown. Crosslinks between INT and PEC are shown by blue lines while other crosslinks are shown by gray lines. Crosslinks within INT, NELF, Pol II were also not shown for clarity. Crosslinks between Pol II and NELF -A and NELF -C are shown because INT crosslink to the same regions of those NELF subunits.

Crosslinking experiments were repeated two times and crosslinks that were reproduced in both experiments were used as final results. We observed 1169 crosslinks that were reproduced in the two independent experiments amounting to 55% of all crosslinks (Figure 4 Venn diagram). Crosslink network between INT subunits is largely reminiscent to our results when INT alone was crosslinked shown in Figure 4.15 except crosslinks between INTS5 and INTS8 as well as crosslinks from INTS11 to INTS1 and INTS12 suggestive of some conformational changes in INT upon binding PEC. Crosslinks within PEC were mostly consistent with results from Vos and colleagues (Vos, Farnung, Urlaub, et al., 2018) with some deviations. Notably, we did observe reproducible crosslinks between the C-termini of NELF-A and NELF-E to the subunits of DSIF.

Results showed in Figure 4.20 reveal crosslinks between several subunits of INT to PEC. INTS1, INTS3, INTS6, INTS8, INTS10, INTS11 and INTS12 have the most crosslinks to PEC, suggesting that these subunits play a major role in the interaction between INT and PEC. INTS2, INTS9, and INTS13 had fewer crosslinks to PEC. INTS4, INTS5, INTS7, and INTS14 had no crosslinks to PEC.

INTS1, INTS3, INTS6, INTS8, INTS11 and INTS12 subunits of INT have crosslinks to RPB1, RPB2, RPB5 RPB8 and RPB11 subunits of Pol II. INTS1, INTS3, INTS6, INTS8, INTS9, INTS11 and INTS12 have crosslinks to subunits of the NELF complex. INTS1, INTS2, INTS10, and INTS11 have crosslinks to DSIF subunits especially SPT4. For a visual appreciation of the location of INT crosslinks in real space, I mapped them on the model of PEC (PDB 6GML) and shown as orange spheres (Figure 4.21). INT crosslinks are located on the front view of Pol II between NELF and around the downstream DNA and extends to SPT4 and the upstream DNA and exiting RNA (Figure 4.21a). There were no crosslinks on the back view of Pol II 180^{o} away from the front view (Figure 4.21b). It appears INT traces the path of NELF and DSIF on Pol II as crosslinks between INT and Pol II are mostly on regions of Pol II that have crosslinks to these factors. This highlights how INT, NELF and DSIF may collaborates in regulating Pol II during pause – release on protein coding genes and during termination of snRNA genes. INTS1(1-294) has the most and the highest scoring crosslinks. Notably, INTS1(K17) to RPB2(K151) had a lot of hits and the highest score. The N-terminus of INTS1 have additional crosslinks to RPB2(K821) and SPT5(K322) which are located 24 Å

and 32 Å from RPB2(K151) respectively in PEC structure (PDB 6GML) showing the N-terminal of INTS1 is located in this region and may contact the upstream DNA. All these residues are in unstructured loops around the RNA exit tunnel of Pol II and can interact with the predicted unstructured N-terminus of INTS1. The N-terminus of INTS1 have many crosslinks to INTS11 and may localize INTS11 and the CM close to the exiting RNA for processing. INTS1 have high score crosslinks to the NELF complex as well. INTS1(K100) has crosslinks to NELF -A(K190) and (K371) which are located in the flexible 'tentacle' of this protein and to NELF -B(K146). INTS12(1-294) crosslinks to the terminal residue of SPT4 (A2) which is located close to the region where most of INTS1 crosslinks are located. INTS8 may form an important bridge between NELF and Pol II having crosslinks to NELF -B (K85 and K118) and RPB5 (K12). Similar to INTS8, INTS3 have crosslinks to NELF -A(K161), RPB1(K213) and RPB2(K327) suggesting how this subunit might bridge the interaction between NELF and Pol II.

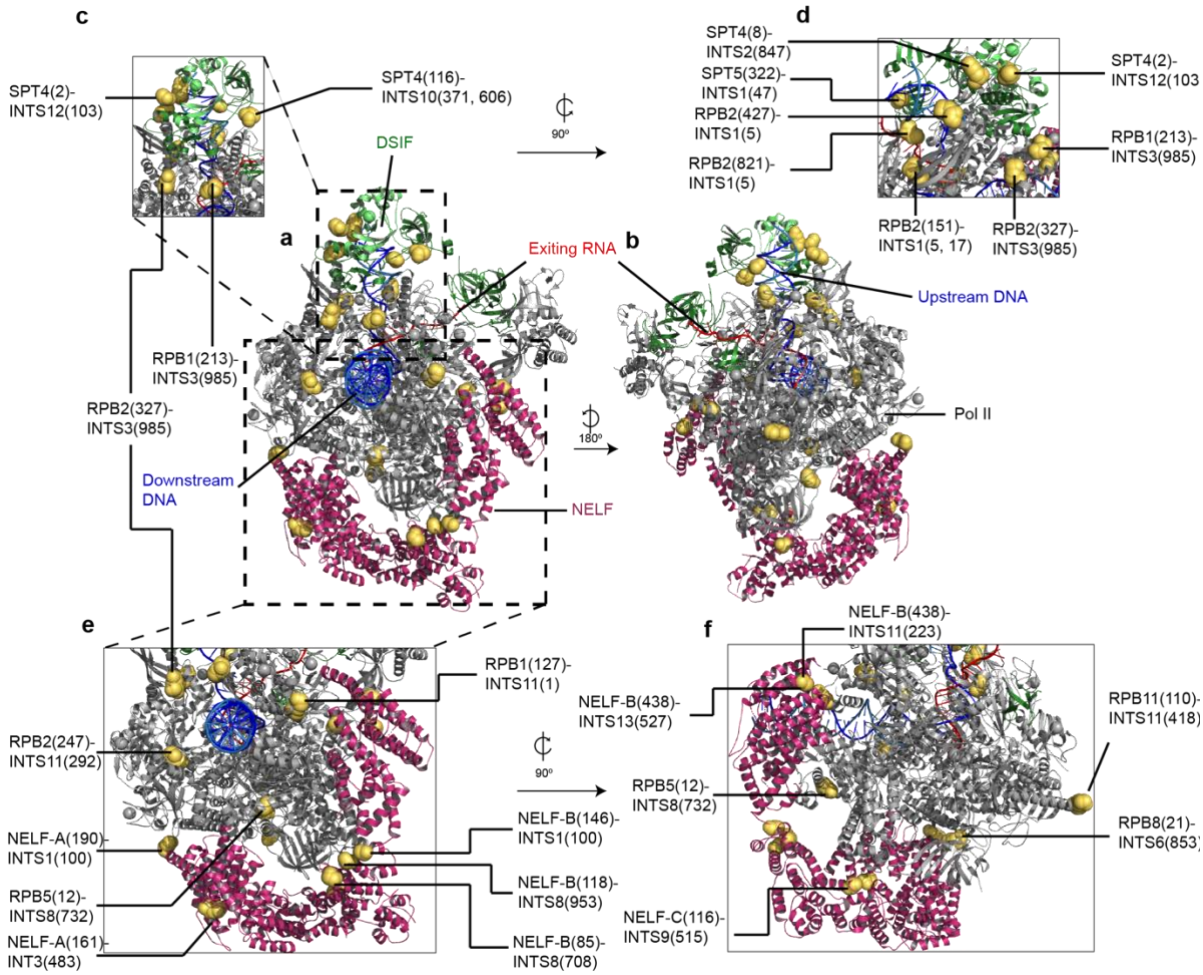


Figure 4.21 Mapping of crosslinks between INT and PEC on the PEC atomic model (PDB 6GML). Crosslinking residues of the PEC are shown as orange spheres and labeled together with the residue of INT they crosslinked to. (**a**) A 'front' view of PEC with the downstream DNA pointing out of the page. Coloring scheme is the same as in Figure 4.20. Most of INT crosslinks are located on this view of PEC. Residues that crosslink to INT are shown as orange spheres. (**b**) A 'back' view of PEC 180º around the vertical axis relative to the front view in (**a**). The upstream DNA in indicated. (**c-f**) Zoomed-in of regions with crosslinks. Crosslinked residues are labeled together with the INT residue they crosslinked to. (**c**) and (**d**) are related by a 90º anticlockwise rotation and (**e**) and (**f**) by a 90º clockwise rotation

In conclusion, I recapitulated interaction between NELF and INTS3 using purified proteins. I also demonstrated for the first time, *in vitro* interaction between INT and paused transcription elongation complex of Pol II, DSIF and NELF. There are multiple INT subunits involved in this interaction, specifically INTS1, INTS3, INTS8 and INTS11 had high confidence crosslinks to both NELF complex and Pol II compared to other subunits of INT.

5 Discussion and Conclusions

5.1 Recombinant production of INT and its inter-subunit interaction network

Pioneers in biochemical reconstitution of transcriptional complexes relied on chromatographic fractionation of cell/tissue homogenates, which are laborious and often results in minute amounts of protein (Roeder & Rutter, 1969; Sekimizu et al., 1976). *In vitro* biochemical studies in the transcription field were revolutionized by advancements in recombinant DNA technology. This technology allows for the expression and purification of proteins in good quality and quantity that will otherwise not be available from endogenous sources (Lis, 2019). With the advent of cloning, *Escherichia coli* and *Saccharomyces cerevisiae* proved useful resources for recombinant protein production because they are easy to manipulate genetically and can be cultured in large amounts at a relatively low cost (Nevalainen et al., 2005; Peti & Page, 2007; Puig et al., 2001; Vieira Gomes et al., 2018). Overexpression in *E. coli* and *S. cerevisiae* may not be ideal for proteins from higher eukaryotes because the necessary machinery for protein folding and installing post translational modification may be lacking (Vieira Gomes et al., 2018). In such cases, insect and mammalian cells are important alternatives although they are relatively difficult to manipulate genetically and are more expensive to culture. Nonetheless, the baculovirus - insect cell recombinant protein expression system have gained high grounds in the expression of large molecular complexes which led to high resolution cryo-EM structures of TFIIH (Schilbach et al., 2017), the anaphase promoting complex (Zhang et al., 2013) just to name a few. This is due to the availability of well-established technologies for efficient cloning of genes and consistent production of baculoviruses and cell lines for protein expression (Berger et al., 2004; Gradia et al., 2017; Luckow et al., 1993; Trowitzsch et al., 2010).

In this study, I used the baculovirus - insect cell recombinant protein expression system to reconstitute INT. It was not possible to clone all 15 subunits into a single baculovirus because of the large size of the genes. The largest virus I constructed was 34,000 bp, coding for 8 subunits of INT. This challenge can be circumvented by co-infection with multiple baculoviruses. I therefore employed a bottom-up approach where I identified interacting subunits and then combined them into bigger subcomplexes of INT for co-expression and purification (Section 4.1). Initial expression tests in insect cells showed that all subunits of INT can be expressed individually. This prompted the idea to purify all the subunits independently

and reconstitute INT *in vitro*. However, purifications often resulted in oligomerization of most of the subunits, which is suggestive of suboptimal purification conditions or misfolding due to a lack of an interaction partner. After testing several buffer conditions without successfully solving the problem of oligomerization, I opted to identify interacting subunits by co-infection and pulldown assays. Using this strategy, I discovered previously unknown interactions between subunits of INT (Section 4.1.3).

Combining the results of the systematic co-infection assays and the information about known interacting subunits allowed the creation of 3 baculoviruses for co-infection. Purification of INT from Hi5 cells co-infected with the three baculoviruses and XL-MS experiment led to the identification of novel interaction between INTS3/6 heterodimer and DDX26B (Section 4.1.5 and 4.1.6). This trimer formed the basis for the identification of a heteropentameric subcomplex (Section 4.1.7) and then the heteroheptameric core-INT (Section 4.1.9). This way the 15 subunits of INT were grouped into 4 subcomplexes named core-INT (7 subunits), CM (3 subunits), CMIM (3 subunits) and INTS1/12 (2 subunits) (Figure 5.1).

The construct expressing core-INT is 34,000 bp in size. It was difficult to clone additional subunits into it and when done successfully, there was little or no protein expressed. I therefore established purification protocols for the CM, CMIM and identified soluble and interaction domains in INTS1 and INTS12 for *in vitro* reconstitution of INT. Using amylose affinity pulldown and analytical gel filtration, I discovered an interaction between the CM and CMIM. This interaction was however not stable under purification conditions tested implying that the interaction might be weak and transient. The CMIM interacts with the core-INT forming a complex which is stable in 500 mM NaCl during purification. To further probe the interaction between the subcomplexes of INT, XL-MS was conducted on the reconstituted INT. INT was formed by pulldown using MBP tag on INTS6. From the XL-MS results, the CMIM and INTS1/12 heterodimer have a lot of crosslinks to the core-INT whereas the CM has fewer crosslinks (Section 4.1.15). It is therefore possible that the CM is a dissociable module of INT while the other subunits form the functional core of INT (Figure 5.1).

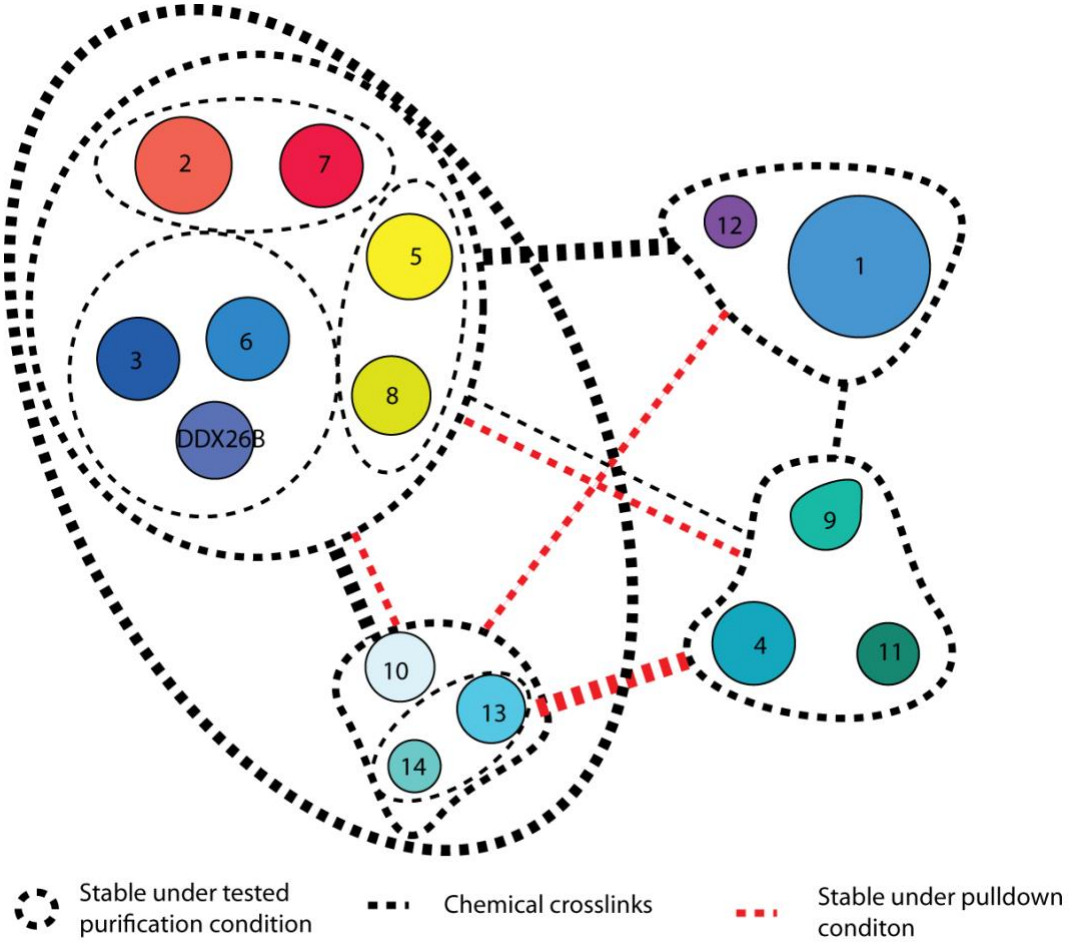


Figure 5.1. A model of subunit-interactions within INT. INT subunits are represented by circles/ovals scaled to the number of amino acids in the specific subunit. Subcomplexes that are stable under a tested purification condition are encircles with dashed circles. Black dashed line indicated interaction between subcomplexes inferred from XL-MS. Lines are weighted according to the number of crosslinks observed between the subcomplexes. Red dashed lines represent interactions shown by pulldown assay.

5.2 Modularity of INT

Large macromolecular complexes are often divided into submodules. Both the Mediator complex and TFIIH have dissociable kinase modules (Kornberg, 2005; Schilbach et al., 2017). Since the discovery of INT, it has been speculated that INT might also exhibit modular architecture given its large size (Baillat et al., 2005; Baillat & Wagner, 2015; Rienzo & Casamassimi, 2016b). The question remains whether the core-INT, CM, CMIM and INTS1/12 heterodimer subcomplexes reported here represent functional modules of INT *in vivo*. The CM which shares substantial similarities with the cleavage module of the mammalian CPF was described by another independent group and characterized as the cleavage module of INT

(Albrecht et al., 2018). Chromatin immunoprecipitation coupled to sequencing (CHIP-seq) analysis of INT subunits shows that INTS5 of the core-INT binds the promoter region while INTS11 subunit of the CM binds preferentially to the 3' region of snRNA genes (Egloff et al., 2012). Furthermore, *Drosophila* INTS12 has a CHIP-seq peak at the promoter region of Hsp70 whereas INTS9 subunit of the CM peaks at the 3' end (Gardini et al., 2014). These results show at least INTS9/11 dimer are recruited to genes independent of the rest of the INT insinuating that they might be a part of dissociable module of INT.

INTS13 subunit of the CMIM was identified to play a role in enhancer activation together with NAB2 at poised enhancers (Barbieri et al., 2018). Mass spectrometric analysis of anti-INTS13 immuno-precipitated material separated by gel filtration shows enrichment of INTS10, INTS14 and INTS13 in a lower molecular weight fraction, distinct from the full INT (Barbieri et al., 2018). This suggests that the subunits within the CMIM interact *in vivo* and the CMIM may represent functional module of INT. *In vivo* interaction of *Drosophila* INTS1 and INTS12 has been reported (Jiandong Chen et al., 2013). However, it is not known whether this is a functional module of INT. INTS5, INTS6 and INTS7 subunits of core-INT but not INTS11 co-purified with RPAP2 (Gardini et al., 2014). It can therefore be speculated that these subunits are in one module of INT *in vivo*.

5.3 The CM of INT is similar to the CM of mammalian CPF

The INTS9/11 heterodimer has clear similarities with CPSF73 and its regulatory subunit CPSF100 of the CPF. All these proteins have an N-terminal metallo beta-lactamase and a beta-CASP domains and a less characterized C-terminal domain. INTS11 and CPSF73 are endoribonucleases. INTS9 and CPSF100 have the same folds as INTS11 and CPSF73 but are not active endoribonucleases (Albrecht & Wagner, 2012; Baillat et al., 2005; Mandel et al., 2006). INTS4 interacts with INTS9/11 heterodimer to form the CM of INT. INTS4 has N-terminal HEAT repeats similar to Symplekin (SYMPK), which is the third protein in the CM of the mammalian CPF (Albrecht et al., 2018; Y. Zhang et al., 2019) (Figure 5.2).

I purified the CM (Section 4.1.12) and attempted to determine its structure by x-ray crystallography and cryo-EM. I could not identify crystallization conditions for the full-length CM and after limited proteolysis. I then switched to cryo-EM where I could obtain a low-resolution 3D map of the CM at ~22 Å (Figure S8). The CM appears to be trilobal in shape similar to the CM of the mammalian CPF which was solved to a medium resolution by cryo-EM (Zhang et al., 2019). The two complexes likely have similar 3D structure due to the similarity in the amino acid sequence of their subunits. It also emerges that they have a lot of conformational flexibility. This flexibility may account for poor alignment of their particles and

hence the low resolution (Nogales et al., 2016). Addition of interaction partners may restrict this flexibility and allow for a better resolution. A recent study of the histone pre-mRNA 3' formation machinery which includes CM of the CPF was a clear demonstration of this notion. Symplekin, CPSF100 and CPSF73 made extensive interactions with both proteins and nucleic acids to adopt a rigid conformation. In this conformation, a high resolution structure was determine which led to an atomic model for the histone pre-mRNA 3' formation machinery including the CM of the CPF (Sun et al., 2020). In similar direction, I attempted cryo-EM studies on a complex of the CM and the CMIM. However, the complex did not survive tested cryo conditions and fell apart or got denatured.

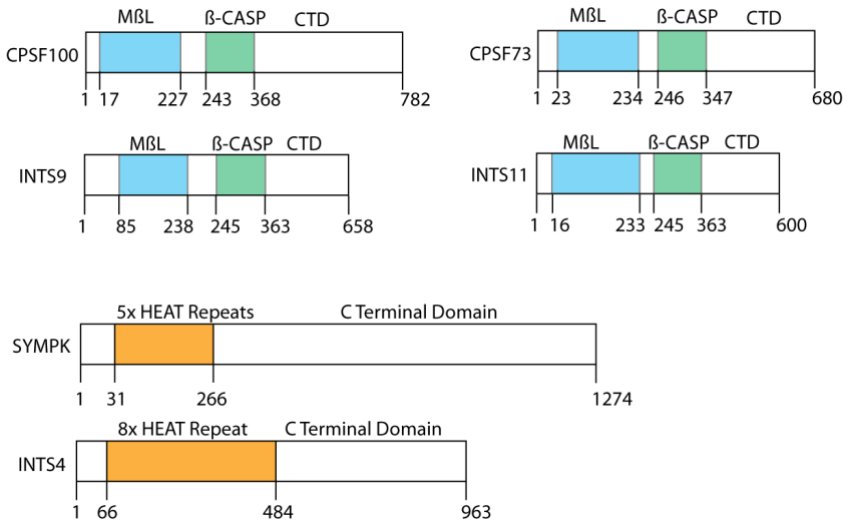


Figure 5.2. Homology between CMs of INT and CPF. A cartoon representing subunits of the CMs. Similar subunits are drawn together and scaled to their respective number of amino acids. Conserved domains namely metallo beta-lactamase (MβL), β-CASP, and HEAT repeats are indicated. Less conserved C-terminal domains (CTD) are also shown.

5.4 INT interacts with PEC

The reconstituted INT interacts with PEC in agreement with the widely reported role of INT in Pol II pause-release (Baillat & Wagner, 2015; Elrod et al., 2019; Gardini et al., 2014; Rienzo & Casamassimi, 2016b; Stadelmayer et al., 2014). Result of XL-MS shows that INTS1, INTS3, INTS6, INTS8, INTS10, INTS11 and INTS12 play major roles in the interaction between INT and PEC (Section 4.2.3). These subunits have multiple high confidence crosslinks to subunits of NELF and Pol II. It emerges that INT interacts with both NELF and DSIF in agreement with data showing that INTS4, INTS6, INTS9, INTS11 and INTS12 have binary interaction with NELF -A and INTS6 interacts with SPT5 (DSIF subunit) in pulldown assays (Yamamoto et al.,

2014). There might be other crucial interactions that were not captured by XL-MS because Lys residues are not found at the respective protein interfaces - a limitation of BS3 mediated XL-MS. Also, interaction between INT and the exiting RNA as well as DNA may be important for stabilizing the complex.

Phosphorylation of Pol II CTD on Ser7 and Ser2 of the heptad repeat has been shown to be important for INT recruitment (Baillat et al., 2005; Egloff et al., 2007, 2010). However, INT does not have any predictable CTD/phosphorylated CTD interaction domain. INT recruitment via a phosphorylated CTD is thought to be mediated by other proteins such as RPAP2 which binds CTD phosphorylated on Ser7 and interacts with INT (Egloff et al., 2012). In this study, Pol II was not phosphorylated *in vitro* prior to the formation of INT-PEC complex. However, endogenously purified Pol II may already carry various phosphorylations on the CTD. We detected high-confidence crosslinks between INTS1, INTS4, INTS6, INTS11, INTS12, INTS13 and INTS14 to various subunits of Pol. This implies INT has intimate interactions with Pol II apart from a possible interaction with the CTD. This suggests that, phosphorylated CTD might be crucial for only the initial specific recruitment of INT to the transcriptional complex. After this phosho-CTD-dependent specific recruitment, INT then makes intricate interactions with other subunits of Pol II and NELF to elicit its function in transcription regulation. Some of these interactions are with loops buried deep inside the Pol II active center suggesting that INT might influence RNA synthesis upon binding the PEC. Future work shall focus on the impact of INT binding on RNA synthesis and Pol II pause-release *in vitro*.

5.5 INTS3 interacts with NELF and PEC, a potential role of SOSS complex in Pol II transcription

An *in vivo* interaction between INTS3 and NELF which is functionally distinct from INT was reported suggesting INTS3 may have other roles in transcription outside INT (Stadelmayer et al., 2014). Using purified proteins, I demonstrated that INTS3 physically interacts with multiple subunits of NELF. The NELF -A and -E tentacles (Vos, Farnung, Urlaub, et al., 2018) are not important for this interaction. Furthermore, I showed INTS3 interacts with the PEC mostly via interactions with NELF (Section 4.2.1). The functional relevance of the interaction between INTS3 and NELF is not known (Stadelmayer et al., 2014). INTS3 belongs also to the SOSS complex involved in single stranded DNA (ssDNA) binding, DNA damage response and genome stability (Huang et al., 2009; Ren et al., 2014; Skaar et al., 2009). It is possible that the SOSS complex interacts with paused Pol II elongation complex via interactions between INTS3 and NELF. Long pauses of Pol II may lead to exposed ssDNA and R-loop formation (Proudfoot, 2016). The SOSS complex may thus be recruited by NELF the major pausing factor (Kwak &

Lis, 2013; Vos, Farnung, Urlaub, et al., 2018; Yamaguchi et al., 1999). This would protect the ssDNA and prevent DNA single strand and double strand breaks. INTS3 and SOSS-B1/NABP2 subunit of the SOSS complex co-immunoprecipitated with NELF -B, SPT5 and Pol II from Hela cells showing that the SOSS complex associates with Pol II in cells (Skaar et al., 2015). Further experiments are needed to understand the role of the SOSS complex in Pol II transcription and to distinguish the role of INTS3 in SOSS complex from its roles in INT.

6. Future Perspectives

This study is paramount to *in vitro* studies that shall reveal structural and biochemical details of the roles of INT in Pol II transcription. Some of these *in vitro* results may inform *in vivo* experiments that can help understand the roles of INT in the congested nuclear environment of the cell. A few of these experiments are suggested below.

Structural analysis of INT

I have described a protocol for the reconstitution of the 15 subunit INT and used negative stain and cryo-EM to evaluate the quality of the different INT subcomplexes. Cryo-EM on core-INT and CM produced low resolution maps revealing the overall shapes of these subcomplexes. These maps cannot be interpreted further because of the limited resolution. Pulldown and XL-MS experiments also revealed how various subunits within INT are interacting at the protein level. Future structural studies are needed to provide molecular details on how the subunits are interacting within INT. A high-resolution cryo-EM structure of the full INT can provide this molecular insight. Cryo-EM requires relatively low amount of protein (Nogales et al., 2016) and can be done despite the low yield of core-INT per purification. The full INT may be reconstituted from the subcomplexes via amylose affinity pulldown, analytical gel filtration or sucrose density gradient. The preliminary cryo-EM studies on the core-INT and the CM shows these subcomplexes might be flexible or denatured under cryo conditions resulting in low resolution. The full INT may behave differently as flexible domains may be bound by interaction partners potentially restricting their conformational freedom. The problem of protein denaturation during cryo sample preparation is often caused by protein adsorption to the air-water interface. This problem can be solved by using cryo-grids coated with materials like carbon support or graphene oxide which adsorbs the protein particles (D'Imprima et al., 2019). Moreover, different detergents can be used to saturate the air-water interface thereby protecting the proteins from denaturation (Chen et al., 2019).

Structural studies on INT subcomplexes can be expanded by using x-ray crystallography. More suitable targets for such approach would be individual domains of INT subunits, which could be identified bioinformatically, and smaller subcomplexes. For example, the soluble interacting domains of INTS1/12 (Section 4.1.16) could be suitable for crystallization attempts. The production of smaller subcomplexes comprised of different subunits can be guided by the XL-MS data provided in this study. This can provide molecular insight into how various subunits

are interacting within their respective subcomplexes and within INT. Structures of very flexible domains may be tackled by nuclear magnetic resonance (NMR) spectroscopy.

Biochemical and structural analysis of INT in the context of Pol II transcription

This study provided initial results on the interaction between INT and PEC *in vitro* which is prerequisite to gaining a mechanistic understanding of the roles of INT in Pol II pause - release. The next step will be to evaluate the influence of this interaction on Pol II transcription using RNA extension *in vitro* transcription and pausing assays described by Vos and coworkers (Vos, Farnung, Boehning, et al., 2018; Vos, Farnung, Urlaub, et al., 2018). By using the full INT or the purified subcomplexes, it will be possible to pin down which specific subunit/subcomplexes are important for specific roles of INT in Pol II pause - release and transcription in general. *In vitro* transcription termination assays can also be used to dissect termination roles of INT. This could be done by *in vitro* transcription through the 3' box of an snRNA DNA template. The contribution of specific subcomplexes in such termination assays may be deciphered by excluding them from the assay. Pol II CTD kinases such as CDK7 and CDK9 and phosphatases like RPAP2 and PP2A have been shown to be important for the *in vivo* functions of INT (Egloff et al., 2007, 2010, 2012; Solis-Mezarino & Herzog, 2017). With the *in vitro* reconstituted transcription system, the complex interplay of various CTD phosphorylation states can be studied in a more systematic way. This could be done by testing the impact of phosphorylation and dephosphorylation by different combinations of the kinases and phosphatases on the recruitments of INT and on *in vitro* transcription by Pol II in the presence of INT. These functional studies may help to identify functionally relevant targets for structural studies.

The INT-PEC complex described here (Sections 4.2.2 and 4.2.3) could be a starting complex for understanding mechanistically, how INT associates with Pol II elongation complexes using cryo-EM. This may help to clarify how INT subunits are interacting with Pol II, DSIF and NELF subunits during promoter proximal pausing/early transcription elongation. Such a structure, complemented by functional studies described above, can provide molecular details on the contribution of INT in Pol II transcriptional pause and release. Furthermore, key players which have been shown to interact with both INT and Pol II, such as RPAP2 (Egloff et al., 2012), may be needed to stabilize the complex of INT and Pol II transcription elongation complexes for cryo-EM. In vitro phosphorylation of the CTD may also be important for the formation of a stable complex for cryo-EM analysis, as demonstrated before for the activated transcription complex (Vos, Farnung, Boehning, et al., 2018). In case this holistic approach proves to be too challenging, one could also identify the subcomplex of INT which stably

associate with Pol II elongation complexes for structural analysis. Indeed, to fully understand how INT is involved in Pol II transcription, structures of several intermediate complexes will be necessary. For example, structures of Pol II-INT, Pol II-DSIF-INT and Pol II-NELF-INT may provide important snapshots of conformational changes that occur when INT binds an elongating Pol II. Initial biochemical analysis will be needed to ascertain the stability of such complexes prior to cryo-EM studies.

In vivo functional studies of INT

Attempts to understands the roles of INT *in vivo* have thus far relied on knockdown of INT subunits by RNA interference (RNAi). The RNA output is then measured by quantitative reverse transcriptase PCR (qRT-PCR) of specific target genes or RNA sequencing genome-wide or RNA sequencing of specific genes (Baillat et al., 2005; Skaar et al., 2015; Yamamoto et al., 2014). RNAi may have off-target effects and often includes prolonged treatment of cells with a virus harboring the shRNA which can perturb the cell leading to artificial effect. More transient methods for depleting proteins such as degrons (Nishimura et al., 2009) may be employed to test the importance of specific subunits of INT on Pol II transcriptional output genome-wide. Recent advances in CRISPR-cas technology (Baillat et al., 2016; Pickar-Oliver & Gersbach, 2019; Ran et al., 2013) can provide a boost in tagging specific subunits of INT for degradation. Coupling transient depletion of INT to additional stimuli such as heat shock and stimulation by specific factors like EGF will be instrumental in understanding signal dependent roles of INT in transcription regulation. Also, metabolic labelling coupled to transient transcript sequencing (TT-seq) (Schwalb et al., 2016) may tell us how perturbations of INT affect the synthesis rate of Pol II transcription. Additionally, other genome-wide techniques such as PRO-seq, GRO-seq, mNET-seq in combination with TT-seq can provide further characterization of the contribution of INT in Pol II transcription especially in promoter proximal pause and release (Kwak et al., 2013; Lis, 2019; Nojima et al., 2015). This approach was exemplified by Gressel and colleagues on CDK9 (Gressel et al., 2017).

Only few subunits of INT have been used so far to identify the genomic targets of INT in CHIP-seq experiment due to the lack of high quality antibodies for all subunits (Baillat & Wagner, 2015). Using CRISPR-cas technique, one may tag each subunit of INT with affinity tags such as FLAG or HA which will allow efficient CHIP-seq experiments to identify genomic targets of all subunits of INT. This will clarify the key question of whether all subunits of INT are simultaneously present at various locations on a gene. It will further help to identify stable

subcomplexes of INT *in vivo* which may represent functional modules of the complex by co-purification.

Metabolically labeled newly synthesized RNA can be crosslinked to proteins directly binding then by exposure to UV light. When coupled to immunoprecipitation of specific subunits of INT using protocols such as PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) (Danan et al., 2016; Spitzer et al., 2014) one can identify on a genome-wide scale, RNAs bound and processed by INT with nucleotide resolution.

7 Supplementary methods and results

7.1 Supplementary methods

7.1.1 Expression and Purification of INTS4

Expression

The ORF for full length INTS4 was cloned into the MacroLab 438-C vector which contains an N-terminal 6xHis-MBP affinity tag. V_0 and V_1 viruses were produced for this construct. Protein expression and harvesting of cultures was done according to the protocol described for the cleavage module (INTS4/9/11) in section 3.4.1. The Hi5 cells expressing INTS4 was resuspended in 35 ml of lysis buffer (20 mM Tris 7.5, 500 mM NaCl, 1 mM EDTA, 1 mM DTT, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine). The harvested cells were flash frozen in liquid nitrogen and store at -80 ^{o}C until purification.

Purification

Frozen pellets were thawed in a water bath at 25 ^{o}C and lysed by sonication with 30% amplitude for 2 min with 0.6 s pulse on and 0.4 s pulse off. The lysate was cleared by centrifugation at 87,207xg for 1 hr and filtered with a 0.8 µm syringe filter. The cleared lysate was applied to a self-packed amylose column with a total bed volume of 15 ml pre-equilibrated in wash buffer (20 mM Tris pH 7.5, 500 mM NaCl, 1 mM EDTA, 1 mM DTT). The amylose column was washed with 100 ml of wash buffer before eluting with amylose elution buffer containing 20 mM Tris pH 7.5, 500 mM NaCl, 1 mM EDTA, 1 mM DTT, 100 mM maltose. The fractions containing the protein of interest were pooled and the volume split in two. One part was concentrated and applied to a Superose 6 Increase 10/300 gel filtration column equilibrated in 20 mM Tris pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM EDTA, 1 mM DTT. The peak fractions were collected and concentrated. The protein was aliquoted (5 µl), flash frozen in liquid nitrogen and stored at -80 ^{o}C. The other half was treated overnight with catalytic amounts of TEV protease to remove the affinity tag. Subsequently, the sample was applied to an equilibrated 5 mL HisTrap column in a reverse nickel affinity step to remove the 6xHis-tagged TEV protease, the affinity tag and proteins with undigested affinity tag. The unbound protein was collected, concentrated using a 50 kDa cut-off Amicon ultra centrifugation filter and applied to a Superose 6 Increase 10/300 gel filtration column equilibrated in gel filtration buffer. The elution peak was concentrated, the protein was aliquoted (5 µl), flash frozen in liquid nitrogen and stored at -80 °C.

7.1.2 Expression and purification of INTS10

Expression

The ORF for full length INTS10 was cloned into the MacroLab 438-C vector which contains an N-terminal 6xHis-MBP affinity tag. V_0 and V_1 viruses were produced for this construct. Protein expression and harvesting of cultures was done according to the protocol described for the cleavage module (INTS4/9/11) in section 3.4.1. The harvested Hi5 cells expressing 6xHis-MBP-INTS10 was resuspended in 35 ml of lysis buffer (20 mM HEPES pH 7.4, 300 mM NaCl, 30 mM imidazole, 10% glycerol, 1 mM DTT, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine) per litre of culture. The harvested cells were flash frozen in liquid nitrogen and stored at -80 ^{o}C until purification.

Purification

Lysis of cells and clearance of the lysate was done as described for INTS4 in section 7.1.1 above. The cleared lysate was applied to a pre-equilibrated 5 ml HisTrap column at a flow rate of 2 ml/min and the column was washed with 100 ml of wash buffer (20 mM HEPES pH 7.4, 300 mM NaCl, 10% glycerol, 1 mM DTT). A self-packed amylose column with a total bed volume of 15 ml pre-equilibrated in wash buffer was connected in tandem to the 5 ml HisTrap column. The bound protein was eluted from the HisTrap column onto the amylose column using Ni elution buffer containing 20 mM HEPES pH 7.4, 300 mM NaCl, 10% glycerol, 1 mM DTT, 250 mM imidazole. The HisTrap column was detached and the amylose column was washed with 100 ml of wash buffer before eluting with amylose elution buffer containing 20 mM HEPES pH 7.4, 300 mM NaCl, 10% glycerol, 1 mM DTT, 100 mM maltose. The fractions containing the protein of interest were pooled. INTS10 precipitated when the affinity tag (which is also a solubility tag) was removed by TEV protease digest. The protein from amylose elution was therefore collected, concentrated using a 30 kDa cut-off Amicon ultra centrifugal filter and applied to a Superose 6 Increase 10/300 gel filtration column equilibrated in gel filtration buffer containing 20 mM HEPES pH 7.4, 500 mM NaCl, 10% glycerol, 1 mM DTT. The elution peak was concentrated, aliquoted, flash frozen and stored in 5 µl aliquots at -80 ^{o}C.

7.1.3 Expression and purification of INTS13 – INTS14 heterodimer

Expression and purification of this heterodimer was done as described for INTS10. Briefly, the ORF of INTS13 was cloned into the MacroLab vector 438-C and that of INTS14 was cloned into a 438-A vector. The two vectors were combined by LIC and V_0 and V_1 baculoviruses were produced. Protein expression was done in Hi5 insect cells as described for INTS10. Protein purification was done using the buffer conditions and strategy described for INTS10. Both affinity-tagged and untagged versions of this heterodimer were produced.

7.1.4 Expression and partial purification of INTS1 and INTS1(1-294)

Expression

The ORF for full length INTS1 was cloned into the MacroLab 438-C and 438-B vectors which contain an N-terminus 6xHis-MBP and 6xHis tags respectively. Truncated versions of INTS1 consisting of residues 1-294, 1-1010, 1010-2190 and 2045-2190 were created from the INTS1-438 -B construct by round-the-horn PCR and cloning. V_0 and V_1 viruses were produced for these constructs and protein expression from each variant of INTS1 was tested by amylose affinity pulldown for the full length INTS1 and Ni affinity pulldown for the truncation variants. Protein expression of the full length or the INTS1(1 - 294) truncation variant were induced by infecting 2x 600 ml of Hi5 insect cells at $1x10^6$ cells/ml with 300 µl of the respective V1 viruses in a 3 l flask. Cell, density, viability and size were monitored in 24 hrs intervals and cells were diluted to maintain a density of $1x10^6$ cells/ml. Cultures were harvested when viability dropped below 85% (usually within 72 hrs) by centrifugation at 328xg for 30 min. The supernatant was discarded and the cell pellet was resuspended in 35 ml of lysis buffer (20 mM HEPES pH 7.4, 300 mM NaCl, 30 mM imidazole, 10% glycerol, 1 mM DTT, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine) per litre of culture. The harvested cells were flash frozen in liquid nitrogen and store at -80 ^oC until purification.

Purification

Purification of full length INTS1 was very problematic as INTS1 degrades during purification and forms oligomers/aggregates when analysed by gel filtration chromatography. Different buffer systems and purification strategies were tried to no avail. The final protocol that was adopted for partial purification was as follows. Cell pellets were thawed in a water bath at room temperature and lysed by sonication with 30% amplitude for 2 min with 0.4 s pulse on and 0.6 s pulse off. The lysate was cleared by centrifugation at 87,207xg for 1 hr and filtered with a 0.8 µm syringe filter. The cleared lysate was applied to a pre-equilibrated 5 ml HisTrap column at a flow rate of 2 ml/min and the column was washed with 100 ml of high salt wash buffer (20 mM HEPES pH 7.4, 1 M NaCl, 10% glycerol, 30 mM imidazole, 1 mM DTT) followed by 50 ml of low salt wash buffer (20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 30 mM

imidazole, 1 mM DTT). A self-packed amylose column with a total bed volume of 15 ml pre-equilibrated in wash buffer was connected in tandem to the 5 ml HisTrap column and the 6xHis-MBP tagged INTS1 was eluted from the 5ml HisTrap column onto the amylose column using Ni elution buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 1 mM DTT, 250 mM imidazole. The HisTrap column was detached and the amylose column was washed with 100 ml of wash buffer before eluting with amylose elution buffer containing 20 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 1 mM DTT, 100 mM maltose.

For the truncated variant (INTS1(1-294)) which had only a 6xHis tag, the amylose step was skipped and the protein was eluted with 0-100% gradient of Ni elution buffer containing 500 mM instead of 250 mM imidazole. The samples from amylose/nickel elution peaks were applied to a 5 ml HiTrap Q column equilibrated in low salt wash buffer. The column was washed with 100 ml of low salt wash buffer and eluted with a 0-100% gradient of high salt wash buffer. Elution fractions containing proteins of interest were collected, concentrated using a 50 and 10 kDa cut-off Amicon ultra centrifugal filter for full length INTS1 and INTS1(1-294) respectively and stored is 5 µl aliquots at -80 ^{o}C. This was used for pulldown assays.

Several other buffer systems with variations in PH, salt concentration and additives were attempted for the purification of INTS1.

7.1.5 Expression and partial purification of INTS12 and INTS12(1-194)

Expression and purification of INTS12 and its truncation variant was as described for INTS1 and INTS1(1 - 294) respectively. INTS12 was cloned into a MacroLab 438-C and 438-B and the truncation was generated in INTS12-438-B construct using round-the-horn PCR. Similar to INTS1, the purification of full-length INTS12 and its truncated version was very difficult. Using several different purification strategies the protein could be partially purified, although not to high purity.

7.1.6 Expression and partial purification of INTS5/8 heterodimer

Expression

The ORF of INTS5 was cloned into a MacroLab 438-C vector and the ORF of INTS8 was cloned into a 438-A vector. The two vectors were combined by LIC. V_0 and V_1 viruses for the construct expressing 6xHis-MBP-INTS5 and INTS8 were made. Expression was induced in Hi5 cells by infecting 600 ml of cells at $1x10^6$ cells/ml with V_1 virus at 1:2000 v/v virus to culture ratio. Cultures were monitored for 72 hours and cells were harvested by spinning at 328xg for 30 min. Pellets were re-suspended in lysis buffer containing 50 mM HEPES pH 7.4,

500 mM NaCl, 10% glycerol, 30 mM imidazole, 1 mM DTT, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine (50ml per litre of culture) and flash froze in liquid nitrogen before storage at -80 ºC.

Purification.

The purification of this heterodimer was challenging due to the formation of aggregates when analysed by gel filtration. The following protocol was used for partial purification.

Pellets were quickly thawed in a water bath at room temperature and transferred onto ice. Lysis was done by sonication (30% amplitude for 5 min, 0.6 s pulse on and 0.4 s pulse off). After sonication, the lysate was cleared by centrifugation at 87,207xg for 30 min and the supernatant was transferred into an ultracentrifuge tube and further spun for 1 hr at 45000 rpm using a Ti45 rotor in an ultracentrifuge (Beckman Coulter). Subsequently, supernatant was then filtered through a 0.8 µm filter and applied to a 5ml HisTrap column using a peristaltic pump. The column was transferred to an Aekta system and washed with 100 ml of the lysis buffer without protease inhibitors. Bound protein was eluted with a gradient (0-100%) of nickel elution buffer containing 50 mM HEPES pH 7.4, 500 mM NaCl, 10% glycerol, 500 mM imidazole, 1 mM DTT. The peak fractions of the elution were pooled and 2 mg of 6xHis tagged TEV protease was added and dialyzed against a dialysis buffer (50 mM HEPES pH 7.4, 150 mM NaCl, 10% glycerol, 30 mM imidazole, 1 mM DTT) overnight at 4 ºC to remove the affinity tag and decrease the imidazole concentration. A 5 ml HisTrap, 5ml HiTrap Q and 5ml HiTrap Sp were connected in tandem in the following order: HisTrap followed by HiTrap Q followed by HiTrap Sp. The columns were equilibrated in dialysis buffer and the TEV-digested sampled was applied using the peristaltic pump and the flow through and wash fractions were collected. After sample application, the columns were separated and the TEV protease, affinity tags and proteins with undigested affinity tag bound to the HisTrap column were eluted with 25 ml of the nickel elution buffer. The ion exchange columns were separately eluted with 25 ml of ion exchange elution buffer (50 mM HEPES pH 7.4, 500 mM NaCl, 10% glycerol, 30 mM imidazole, 1 mM DTT). Samples (4 µl) were taken from all fraction and analysed on LDS-PAGE. Protein of interest bound the anion exchange column and this fraction was take and concentrated to 2 ml for gel filtration. A Superdex 200 Increase 10/300 and a Superpose 6 Increase 10/300 gel filtration columns were equilibrated into gel filtration buffer containing 50 mM HEPES pH 7.4, 200 mM NaCl, 10% glycerol, 1 mM DTT and 1 ml aliquot of the protein was applied to each. The protein of interest eluted close to the void volume of each column suggesting this heterodimer forms oligomers. The peak fractions from both gel filtration runs were pooled and applied to a 1 ml Mono Q anion exchange column pre-equilibrated in gel filtration buffer. After

washing with 10 ml of the gel filtration buffer, the column was eluted with a gradient (0-100%) of buffer QE (50 mM HEPES pH 7.4, 1000 mM NaCl, 10% glycerol, 1 mM DTT). The peak fractions were pooled and dialyzed against GF buffer overnight. The sample was concentrated, aliquoted, flash frozen and stored at -80 ºC. This was used for pulldown assays with other subcomplexes. For the purification of affinity tagged variant of this heterodimer, the second anion exchange on Mono Q was omitted.

7.1.7 Expression and partial purification of INTS2/7 heterodimers

Expression

Expression was done as described for INTS5/8 heterodimer from a baculovirus harbouring the ORF for INTS2 and 6xHis-MBP-INTS7. The cultures were harvested and cell pellet was resuspended in a lysis buffer containing 25 mM Tris–HCl pH 8, 0.2 M NaCl, 10 mM BME, 10 uM ZnSO4, 10% glycerol, 30mM imidazole, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine (50ml per litre of culture) and stored at -80 ºC.

Purification

The purification of INTS2/7 heterodimer was also very problematic as it forms oligomers/aggregates when analysed by gel filtration chromatography. Different buffer systems and purification strategies were tried to no avail. The final protocol that was adapted for partial purification was as follows. Cell pellets were thawed in a water bath at room temperature and lysed by sonication with 30% amplitude for 2 min with 0.4 s pulse on and 0.6 s pulse off. The lysate was cleared by centrifugation at 87,207xg for 1 hr and filtered with a 0.8 µm syringe filter. The cleared lysate was applied to a 5 ml HisTrap column (pre-equilibrated in wash buffer) at a flow rate of 2 ml/min and the column was washed with 100 ml of wash buffer (25 mM Tris–HCl pH 8, 0.2 M NaCl, 10 mM BME, 10 µM $ZnSO_4$, 10% glycerol, 30 mM imidazole). A self-packed amylose column with a total bed volume of 15 ml pre-equilibrated in wash buffer was connected in tandem to the 5 ml HisTrap column and the bound protein was eluted from the HisTrap column onto the amylose column using Ni elution buffer containing 25 mM Tris–HCl pH 8, 0.2 M NaCl, 10 mM BME, 10 uM $ZnSO_4$, 10% glycerol, 250 mM imidazole. The HisTrap column was detached and the amylose column was washed with 100 ml of wash buffer before eluting with amylose elution buffer containing 25 mM Tris–HCl pH 8, 0.2 M NaCl, 10 mM BME, 10 uM ZnSO4, 10% glycerol, 30 mM imidazole, 100 mM maltose. Fractions containing the INTS2/7 heterodimeric complex were pooled and treated overnight with 6xHis-TEV protease and lambda phosphatase (home made in our laboratory) in the presence of 1 mM $MnCl_2$ to remove the affinity tag and potential post-translational phosphorylations from the insect cells. The protein was applied again to an equilibrated 5 ml HisTrap in a reverse nickel

affinity step to remove the TEV protease, the affinity tag and proteins with undigested affinity tag. The unbound protein was collected and applied to a 5 ml HiTrap Q column equilibrated in low salt wash buffer. The column was wash with 100 ml of wash buffer and eluted with a gradient (0-100%) of high slat wash buffer (25 mM Tris–HCl pH 8, 1 M NaCl, 10 mM BME, 10 µM ZnSO4, 10% glycerol, 30 mM imidazole. Elution fractions containing proteins of interest were collected, concentrated using a 100 kDa cut-off Amicon ultra centrifugal filter and stored is 5 µl aliquots at -80 ^{o}C. This was used for amylose pulldown experiments with other subcomplexes without further gel filtration chromatography as protein elutes in the void.

7.1.8 Expression and partial purification of INTS(3/6)-DDX26B heterotrimer.

Expression

The ORF for INTS3 was cloned into the MacroLab 438-C vector and the ORFs of INTS6 and DDX26B were each cloned into the 438-A vector using LIC. The three vectors were combined into a multiprotein expression vector using LIC (see methods) and V_0 and V_1 baculoviruses were produced in Sf9 and SF21 insect cells respectively. Protein expression of this construct was as described for the heteropentameric subcomplex (INTS3/5/6/8-DDX26B) in section 3.4.3. The cells were harvested by centrifugation at 328xg for 30 min and the pellets were re-suspended in lysis buffer containing 20 mM HEPES pH 7.4, 300 mM NaCl, 10% glycerol, 10 mM BME, 0.284 µg/ml leupeptin, 1.37 µg/ml pepstatin A, 0.17 mg/ml PMSF, 0.33 mg/ml benzamidine (50 ml per litre of culture) and flash froze in liquid nitrogen before storage at -80 ^{o}C.

Purification

Lysis of cells and clearance of lysate was described (Section 3.4.3). The cleared lysate was applied to a pre-equilibrated self-packed amylose column with a total bed volume of 15 ml at a flow rate of 1 ml/min and the column was washed with 100 ml of wash buffer (20 mM HEPES pH 7.4, 300 mM NaCl, 10% glycerol, 10 mM BME). The bound protein was eluted with amylose elution buffer containing 20 mM HEPES pH 7.4, 300 mM NaCl, 10% glycerol, 10 mM BME, 100 mM maltose. The fractions containing the trimeric complex were pooled and diluted with zero-salt buffer to bring the NaCl concentration to ~200 mM. The diluted protein was applied to a 5 ml HiTrap Q column equilibrated in low salt ion exchange buffer containing (20 mM HEPES pH 7.4, 200 mM NaCl, 10% glycerol, 10 mM BME). The column was washed with the low salt buffer and bound protein was eluted with a gradient (0-100%) of high salt buffer (20 mM HEPES pH 7.4, 1 M NaCl, 10% glycerol, 10 mM BME) over 100 ml. The fractions containing the trimeric complex of 6xHis-MBP-INTS3, INTS6 and DDX26B were pooled and concentrated to 1 ml using a 100K cut-off Amicon Ultra Centrifugal filter (MERCK

Millipore). The protein was applied to a Superose 6 Increase 10/300 gel filtration column equilibrated in the low salt ion exchange buffer and fractionated into 0.5 ml fractions. The fractions containing the pure monomeric trimer were pooled, concentrated and stored in 5 µl aliquots at -80 ^{o}C.

7.1.9 *In vitro* pulldowns with purified subunits and subcomplexes

To test the interaction between purified components of INT, amylose affinity pulldown was used. The experimental set-up was similar to the one described in section 3.4.5.

7.1.10 Cryo-EM analysis of the cleavage module of INT (INTS4/9/11).

Cryo-EM data for this sample was collected under cryogenic conditions using an FEI Titan Krios G2 transmission electron microscope operated in EFTEM mode at 300 kV, energy filter slit set to 20 eV and working with a K3 direct detector (Gatan). Data collection was done automatically with the serialEM (Mastronarde, 2005) software at 81,000x magnification (1.05 Ä/pixel). Three images were collected per foil hole with an electron dose rate of 23.15 e-/px/s and an exposure time of 1.99 s resulting in a total dose of 40 e/Å^2. The images were fractionated over 40 frames and a defocus range of 1.5 - 3 µm was used. The movie frames were aligned, motion and contrast transfer function corrected in WARP (Tegunov & Cramer, 2019). Particles were automatically picked in WARP. Calculation of 2D class averages, generation of initial model and 3D volume calculations were done in RELION 2.2 (Kimanius et al., 2016; Scheres, 2012).

7.2 Supplementary Results

Purification of INT subunits and initial subcomplexes

The co-expression assays described in the results section (section 4.1) were crucial for the identification of physically interacting subunits/subcomplexes of INT. The design of some of these co-expression assays were informed by in vitro pulldown assays using purified/partially purified subunits/subcomplexes of INT. The results of co-expression with few exceptions were supported by in vitro pulldown assay using purified components. The results of the purifications (attempted) of the subunits/subcomplexes of INT used mainly for pulldown assays are described below.

7.2.1 Expression and partial purification of INTS1 and INTS1(1-294)

The largest subunit of INT, INTS1 was one of the most difficult subunits of INT to work with because it is poorly expressed, degrades, and oligomerizes during purifications. When expressed separately as a single subunit, the degradation could be reduced and some intact full length INTS1 makes it to gel filtration stage of purification. During gel filtration chromatography, the protein elutes in the void volume suggesting oligomerization/aggregation (Figure S1). Several additives including but not limited to detergents, different salts, different pH buffers, and chelating agents such as EDTA were added to the purification buffer to remove/reduce the aggregation but to no avail. Figure S1c and d are representative LDS-PAGE and gel filtration chromatogram of a typical purification of INTS1.

From domain and disorder predictions, INTS1 has a small domain of unknown function (DUF) at the N-terminus and the rest of the sequence consist mostly of disordered loops with high protein binding propensities interspersed by helical secondary structures (Figure 4.1 INTS1 and Figure S1a and b). The expression and solubility tests of the N- and C-terminal halves as well as the C-terminal region (amino acids 2045 to 2190) in the absence of solubility tags showed that they are poorly expressed as no clear enrichment over background proteins was observed (Figure 4.16a lane 2, 3 and 4). Conversely, the DUF3677 containing N-terminal region (residues 1-294) showed clear overexpression and solubility when expressed alone without any solubility tag (Figure 4.16a lane 1). Despite the clear expression and solubility of this domain, it also formed oligomers when purified separately as shown by gel filtration chromatography (Figure S1e and f). This observation suggests that this stable and soluble domain of INTS1 is lacking some interaction partner to prevent its oligomerization.

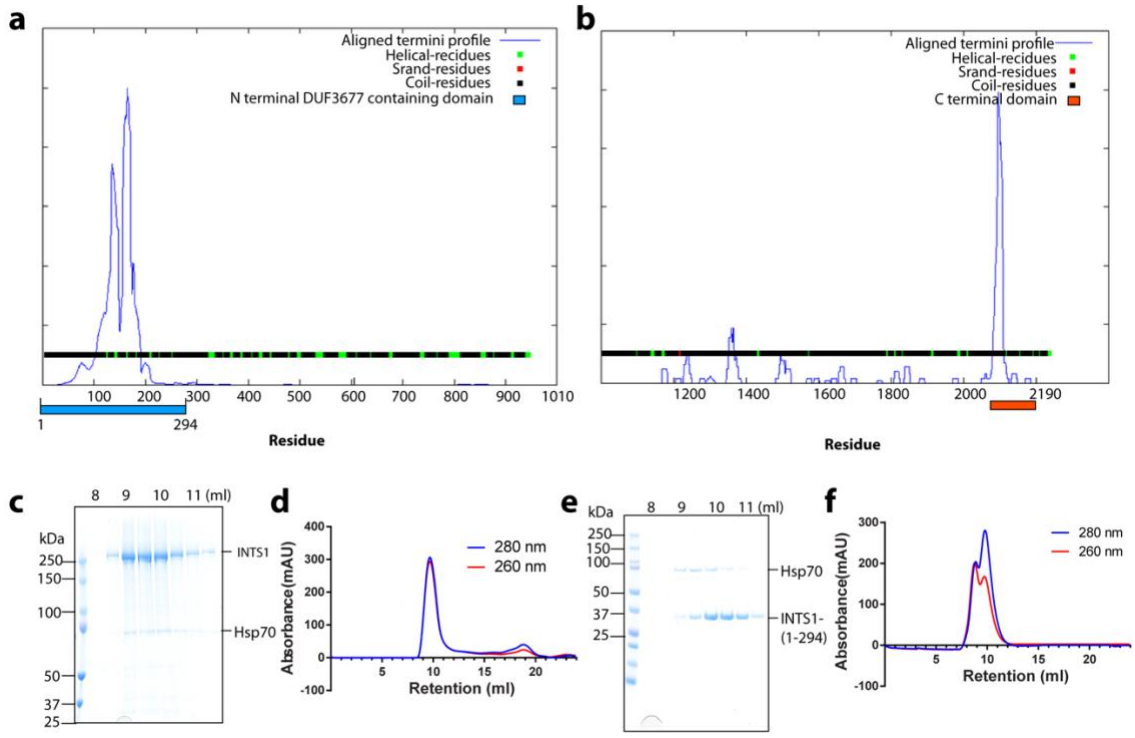


Figure S1. Partial purification of INTS1 and INTS1(1-294) variant. Domain and secondary structure prediction output of INTS1(1-1010) (**a**) and INTS1(1010-2190) (**b**) using the online tool DomPred. A key explaining the features is provided in the top right corner of each plot. It was necessary to split the amino acid sequence into two because the software could not handle the full sequence. An LDS-PAGE (**c**) and gel filtration chromatogram (**d**) of a typical purification of full length INTS1 showing that the protein elutes at the void volume of the column. An LDS-PAGE analysis (**e**) and gel filtration chromatogram (**f**) of a typical purification of the INTS1(1-294) variant of INTS1 showing it eluted at the void volume of the column together with the Hsp70 chaperone. Protein identities were confirmed by mass spectrometry.

7.2.2 Expression and partial purification of INTS12 and INTS12(1-194)

The PHD domain-containing INTS12 is predicted to be the most disordered subunit of INT (Figure 4.1 INTS12 and Figure S2c) and consequently one of the most challenging subunits of INT to express and purify. Similar to the purification of INTS1, several attempts were made to purify this subunit. The protein degrades from its C-terminus to a small N-terminal part which oligomerizes and co-purify with Hsp70 from the insect cell as shown by gel filtration chromatography (Figure S2a and b). Therefore, a truncation variant encompassing the N-terminal micro domain and the PHD domain (residues 1 – 294) which lacked most of the C-terminal low complexity amino acid sequence was created (Figure S2c). This variant of INTS12 was stable but also co-purifies with the chaperon and oligomerizes as validated by gel filtration chromatography (Figure S2d and e) suggesting it is lacking one or more key interaction partner(s) to prevent this oligomerization.

7.2.3 Expression and partial purification of INTS5/8 heterodimer

The INTS5/8 heterodimer was identified during the systematic co-expression of subunits of INTS (Section 4.1.3, Figure 4.3a). Unlike INTS1 and INTS12, the subunits of this heterodimer express well and are stable under the purification conditions tested. However, similar to INTS1 and INTS12 but to a lesser extent, the purified heterodimer forms oligomers and eluted partly at the void volume of all gel filtration columns tested (Figure S3c). This suggests that the heterodimer might have some exposed hydrophobic surfaces that are potentially involved in protein-protein interactions which accounts for the various oligomeric populations observed on gel filtration. A primary sequence analysis of INTS5 showed that there are no known domains predicted in the sequence and that there are widespread unstructured regions albeit with high protein binding propensities (Figure 4.1 INTS5). A sequence analysis of INTS8 on the other hand showed the presence of 4 TPR motives known for protein-protein interaction (Zeytuni & Zarivach, 2012b) as well as disordered regions that could potentially act as protein binding sites (Figure 4.1 INTS8). The cornucopia of protein binding regions in these two proteins may imply that the heterodimer turns to self-associate in the absence of cognate interaction partners and hence aggregates/oligomerizes.

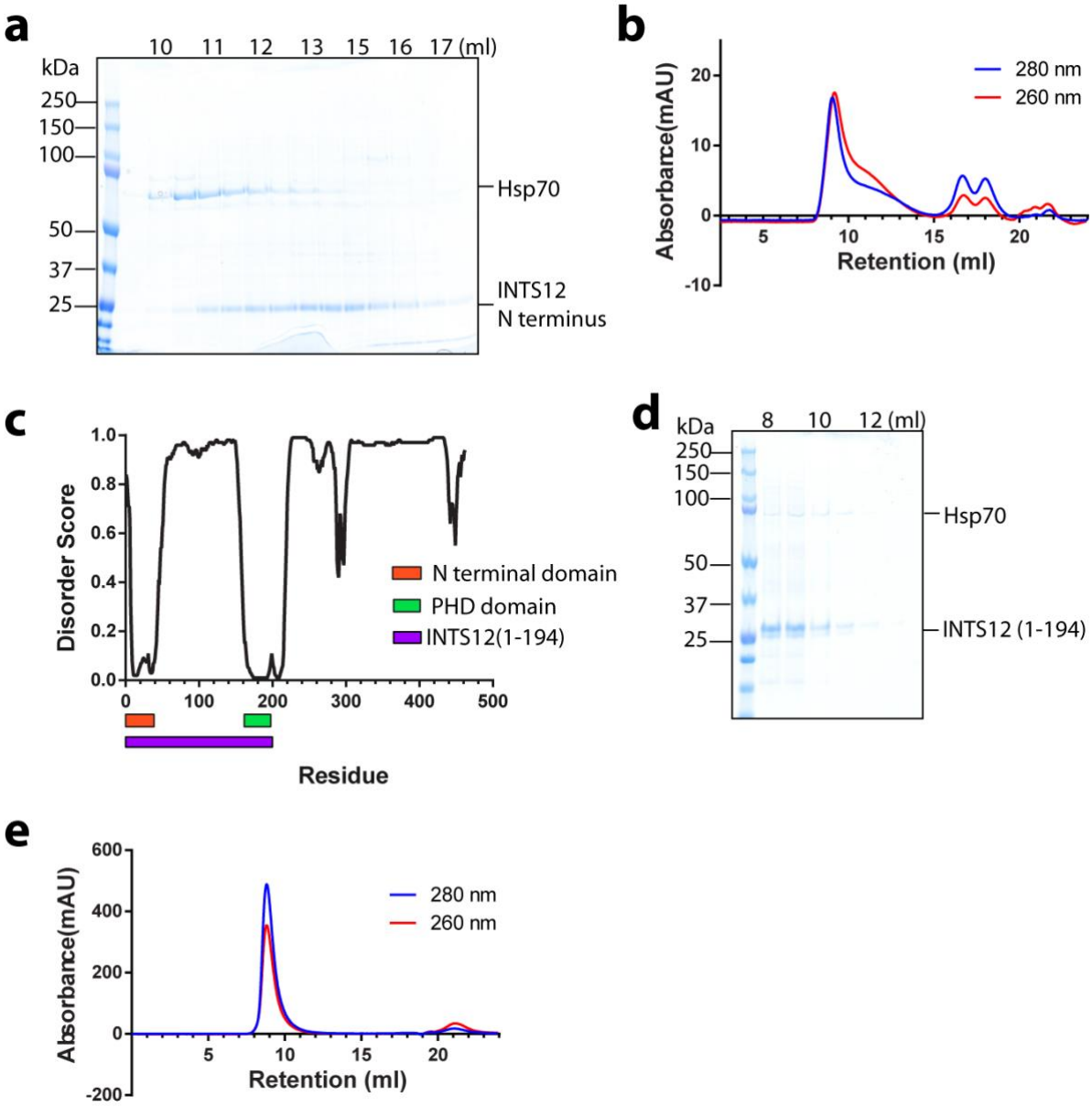


Figure S2. Partial purification of INTS12 and INTS12(1-194) variant. An LDS-PAGE analysis of fractions from gel filtration chromatography (**a**) and a corresponding chromatogram (**b**) of a representative attempted purification of INTS12. Proteins were identified by mass spectrometry. (**c**). Amino acid sequence analysis of INTS12. Predicted and stable domains are indicated with rectangles. An LDS-PAGE (**d**) and gel filtration chromatographic (**e**) analysis of a typical attempted purification of INTS12(1-294) variant. Both the full-length and truncated INTS12 elutes at the void volume of the gel filtration column. For each chromatogram, UV absorption at 280 and 260 nm are shown.

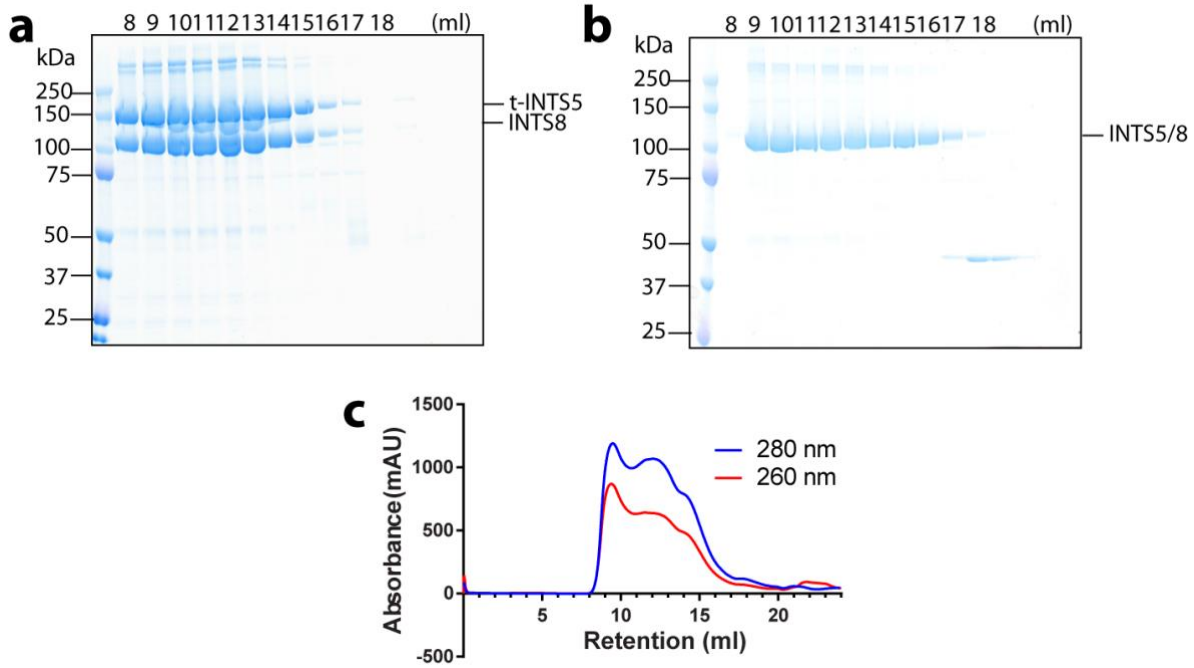


Figure S3. Partial purification of INTS5/8 heterodimer. An LDS-PAGE analysis of fractions of a gel filtration chromatographic purification of affinity tagged INTS5/8 heterodimer **(a)** and untagged INTS5/8 heterodimer **(b).** The fractions analyzed are indicated on top of each gel. 't' indicates 6xHis-MBP affinity tag. The identities of the proteins were confirmed by mass spectrometry. Note in **(b)** INTS5 (108 kDa) without affinity tag runs at the same molecular weight as INTS8 (113 kDa). **(c).** A representative gel filtration chromatogram of INTS5/8 heterodimer showing the UV absorbance at 280 nm and 260 nm.

7.2.4 Expression and partial purification of INTS2/7 heterodimers

The idea of protein-protein interaction between INTS2 and INTS7 was prompted by an observed chemical crosslink between this two subunits in a study characterizing the substrates of PP2A (Solis-Mezarino & Herzog, 2017). As expected, co-expression of the two subunits showed that they form a stable heterodimer (Figure 4.2b). Similar to the INTS5/8 heterodimer, the INTS2/7 heterodimer oligomerizes under all purification conditions tested (Figure S4). The amino acid sequence analysis predictions of INTS2 and INTS5 are similar. They both indicate structured regions interspersed with potentially protein binding disordered regions with no known domains (Figure 4.1 INTS2). The N-terminal region of INTS7 (residues 1 - 500) is predicted to form an armadillo-like repeat while the C-terminal region is predicted to have disordered loops with high protein binding probabilities. The INTS2/7 heterodimer is predicted to have a lot of protein-protein interactions similar to the INTS5/8 heterodimer. Therefore, it may self-associate in the absence of associated interaction partners and form oligomers to satisfy exposed hydrophobic interaction surfaces in the aqueous environment.

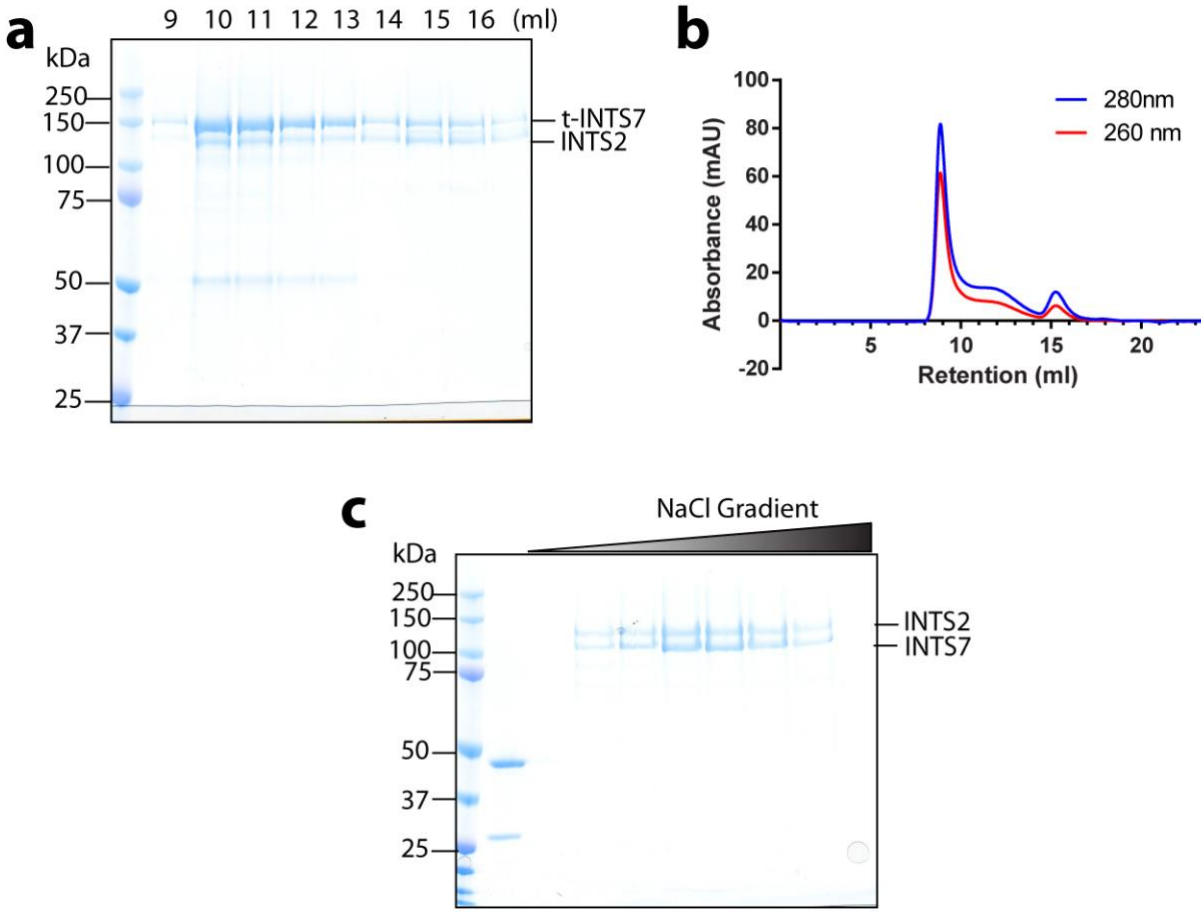


Figure S4. Partial purification of INTS2/7 heterodimer. (**a**) An LDS-PAGE analysis of fractions of a gel filtration chromatographic purification of the affinity tagged INTS2/7 heterodimer. The fractions analyzed are indicated on top of the gel. The proteins were confirmed by mass spectrometry. 't' indicates 6xHis-MBP affinity tag (**b**) A typical gel filtration chromatogram of the purification of the affinity tagged or untagged versions of INTS2/7 heterodimer showing that the dimer elutes in the void volume. The UV absorption at 260 and 280 nm is shown. (**c**) An LDS-PAGE analysis of the purification of the INTS2/7 heterodimer without affinity tag after Mono Q anion exchange chromatography.

7.2.5 Purification of INTS4, INTS10 and INTS13/14 heterodimer

The interaction between INTS13 and INTS14 subunits of the INT was also discovered during the systematic co-expression assays (Figure 4.3b). This subcomplex expressed well and can be purified to homogeneity (Figure S5a and b). Unlike INTS2/7 and INTS5/8 heterodimers, it did not oligomerize (Figure S5a). The void peak in Figure S5b contained affinity tagged variant of INTS13 which was resistant to TEV protease digest. This fraction was probably aggregated or misfolded and was separated from the monomeric INTS13/14 heterodimer.

Also, the single subunits INTS4 and INTS10 were purified for *in vitro* pulldown assays. The purifications of these two subunits were the most straightforward as they expressed well, were soluble and did not oligomerize on gel filtration. Both proteins have a void peak suggestive of

some oligomerization but represented only a small fraction of the total yield and was discarded (Figure S5c and d and Figure S5e and f).

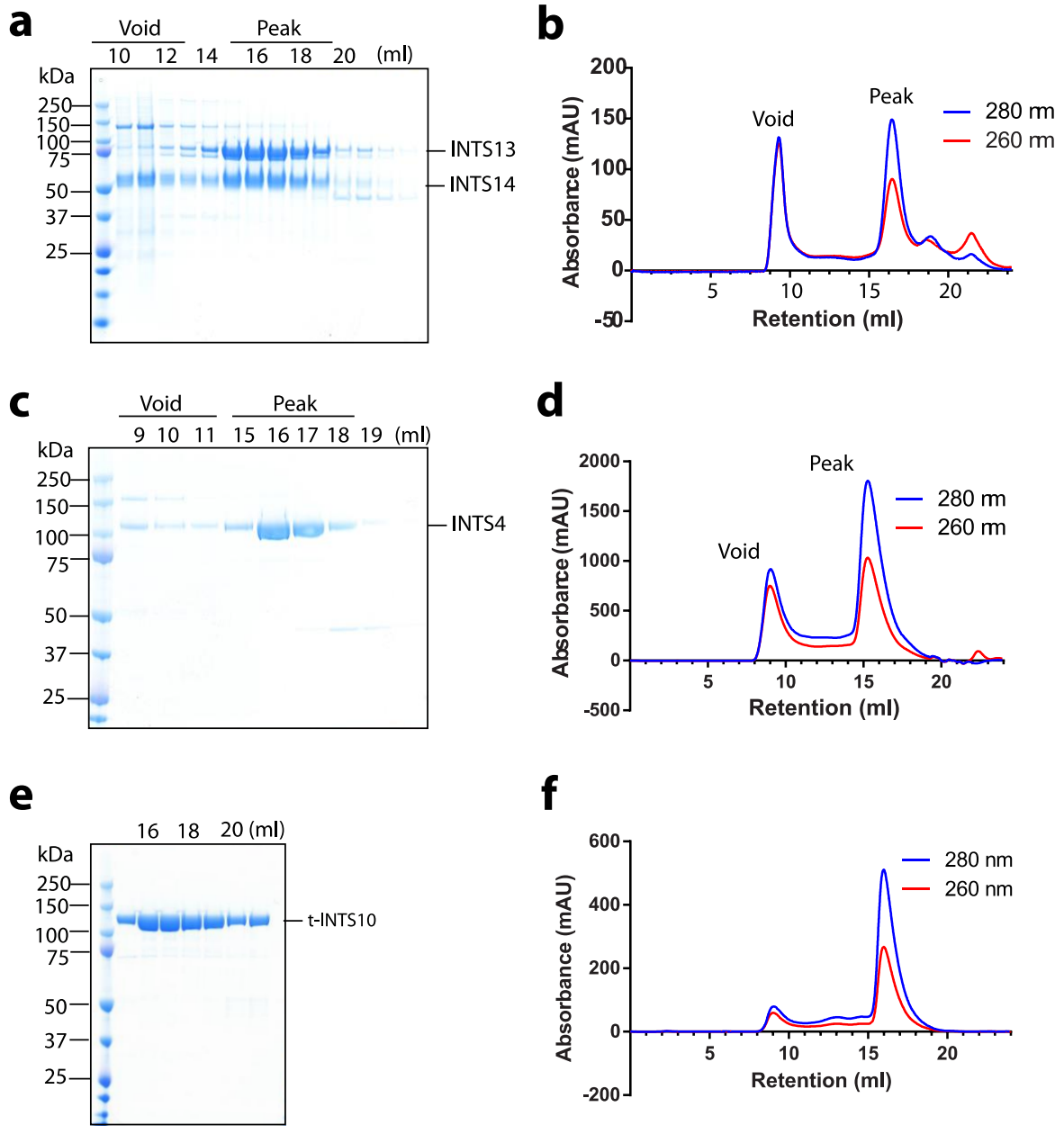


Figure S5. Purification of INTS4, INTS10 and INTS13/14 heterodimer. An LDS-PAGE of the gel filtration chromatography analysis of INTS4, INTS10 and INTS13/14 heterodimer. The retention volume of the fractions analyzed are indicated on top of each gel. Protein identities were confirmed by mass spectrometry. The UV absorbance at 280 and 260 nm is shown (**a**) An LDS-PAGE analysis of gel filtration fractions of INTS13/14 heterodimer. (**b**) A gel filtration chromatogram for the purification of INTS13/14 heterodimer. (**c**) An LDS-PAGE analysis of gel filtration fractions of INTS4. (**d**) A gel filtration chromatogram for the purification of

INTS4. (**e**) An LDS-PAGE analysis of the gel filtration fractions of INTS10. (**d**) A gel filtration chromatogram for the purification of INT10.

7.2.6 Expression and partial purification of INTS3/6-DDX26B

This trimer was discovered by co-expression of the full INT, partial purification and crosslinking mass spectrometry (Figure 4.6). It was purified in order to perform *in vitro* pulldown assays against other purified subcomplexes/subunits. This subcomplex also showed some aggregation/oligomerization together with a fraction of monomeric protein (Figure S6). The tagged INTS3 was over-stoichiometric and the excess could not be separated from the complex (Figure S6a). The void peak contained some degradation products of INTS6 and the insect cell's Hsp70 chaperone suggestive of misfolding or exposed hydrophobic surfaces. The fractions containing monomeric complex (peak) devoid of chaperon and INTS6 degradation were used for pulldown assays.

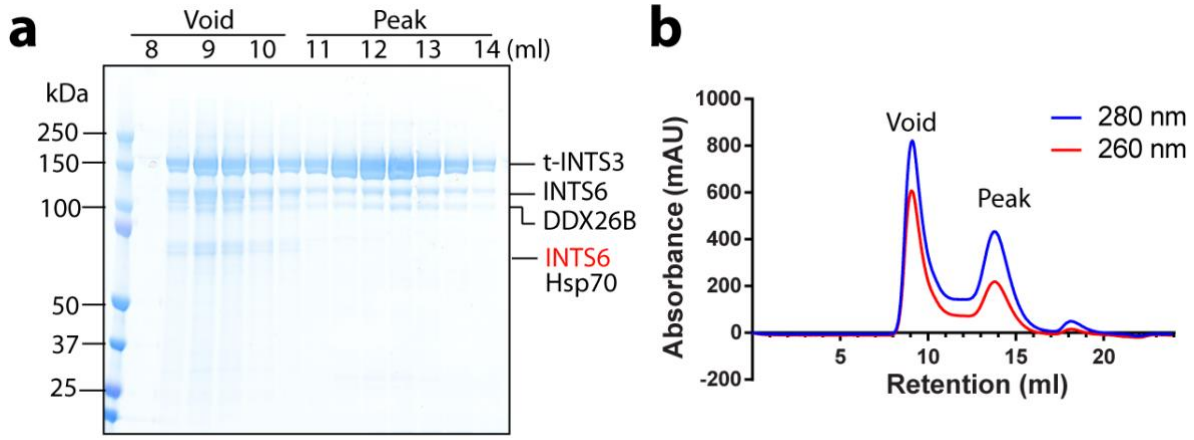


Figure S6. Purification of INTS3/6-DDX26B heterotrimer. (**a**) An LDS-PAGE analysis of elution fractions of gel filtration chromatography of the heterotrimer. The fractions analyzed are indicated on top of the gel. Identification of proteins was done by mass spectrometry. (**b**) A gel filtration chromatogram of the heterotrimer. The void peak and the peak of interest are indicated. The UV absorption at 280 nm and 260 nm are plotted.

7.2.7 Identification of inter subunit/subcomplex interaction via amylose affinity pulldown

As part of the preliminary experiments to identify interaction subunits/subcomplexes and to complement results from co-infection assay, pulldown experiments were done using affinity tagged subunit or subcomplex against untagged subunits or subcomplexes.

In this way, the interaction between the cleavage module (INTS4/9/11) and INTS13/14 heterodimer was discovered (Figure S7a lane 6) which provided the hint for the hexameric subcomplex described in section 4.1.13. Interaction between INTS10 and INTS13/14 from co-expression (Figure 4.3) was also supported (Figure S7b lane 8) and finally the hexameric subcomplex was first formed via affinity pulldown (Figure S7c lane 7). There was a weak interaction between the INTS3/6-DDX26B heterotrimer and the INTS10/13/14 heterotrimer (Figure S7c lane 6). The INTS3/6-DDX26B heterotrimer also interacts with the INTS5/8 heterodimer (Figure S7c lane 8). The INTS5/8 heterotrimer also showed some interaction with INTS2/7 heterodimer (Figure S7c lane 11). Interactions between the INTS3/6-DDX26B heterotrimer and INTS5/8 heterodimer supported the heteropentameric subcomplex in section 4.1.8 and interaction between INTS2/7 heterodimer and INTS5/8 heterodimer informed the experiments leading to the discovery of core-INT (Section 4.1.9). Furthermore, the subunits/subcomplexes which did not interact in this assay helped design and streamline the co-expression assays.

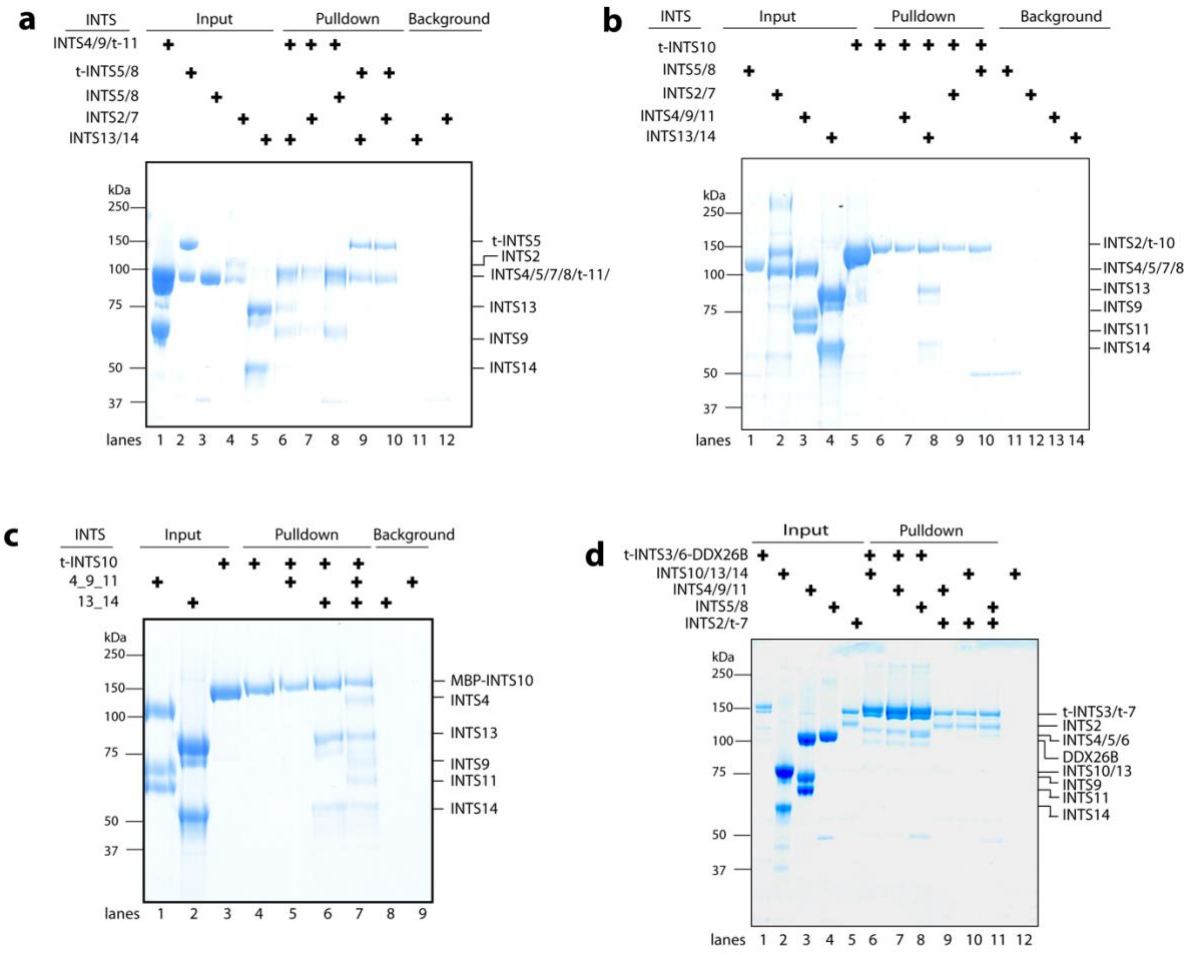


Figure S7. Interaction between purified subunits/subcomplexes of INT. LDS-PAGE analyses of input samples and elution fractions from pulldown or background binding of untagged subunits/subcomplexes. The plus sign (+) indicates that the specific subunit/subcomplex in that row was added in that particular pulldown experiment. 't' indicates the 6xHis-MBP affinity tag. **(a)** Amylose affinity pulldown between affinity tagged INTS4/9/11 and INTS5/8 against untagged INTS2/7, INTS5/8 and INTS13/14. **(b)** Amylose affinity pulldown of affinity tagged INTS10 against INTS2/7, INTS5/8, INTS13/14 heterodimers and INTS4/9/11 heterotrimer. **(c)** Formation of hexameric complex of 6xHis-MBP-INTS10, INTS13/14 heterodimer and INTS4/9/11 heterotrimer. **(d)** Amylose affinity pulldown between affinity tagged INTS3/6-DDX26B and INTS2/7 against untagged INTS5/8 heterodimer, INTS4/9/11 and INTS10/13/14 heterotrimers.

7.2.8 Structural characterization of the cleavage module (INTS4/9/11)

The cleavage module was one of the first subcomplexes of INT to be successfully purified. Crystallization of the subcomplex proved unsuccessful (results not shown). Cryo-EM has been successfully used to determine the structure of complexes of equivalent size of 200 kDa (Fan et al., 2019; Kumar et al., 2019). Prior to cryo-EM experiments on the cleavage module, negative stain EM was used to access the quality of the purified cleavage module (Figure S8a to c). The negative stain micrograph revealed uniformly distribute particles of approximately 10 nm in diameter showing that the cleavage module is not aggregated. 2D class averages were calculated in RELION and revealed different views of the particles. An *ab initio* 3D model was generated from a subset of the data in RELION and used for iterative 3D classification and refinement resulting in four different 3D classes (Figure S8c). The 3D classes calculated form the negative stain data revealed a trilobed structure for the cleavage module.

The cleavage module was further analysed by cryo-EM (Figure S8d to f). Particles were uniformly distributed on cryo grids indicating the sample is applicable for cryo-EM data collection (Figure S8d). A cryo-EM data set was collected and processed (see methods). The calculated 2D class averages did not show any high-resolution features suggesting the particles could not be aligned (Figure S8e). The 3D classes obtained from the 2D class averages had an estimated resolution of 20 Å and also lacked high-resolution information (Figure S8f). The low-resolution structure revealed a trilobed shape for the cleavage module similar to the shape obtained from negative stain EM. Also, the overall shape of the cleavage module suggests that its lobes are very mobile and have a huge conformational freedom. This may be the reason for the misalignment observed in the 2D and 3D class averages. Such conformational flexibility may also account for the inability of the cleavage module to crystallize.

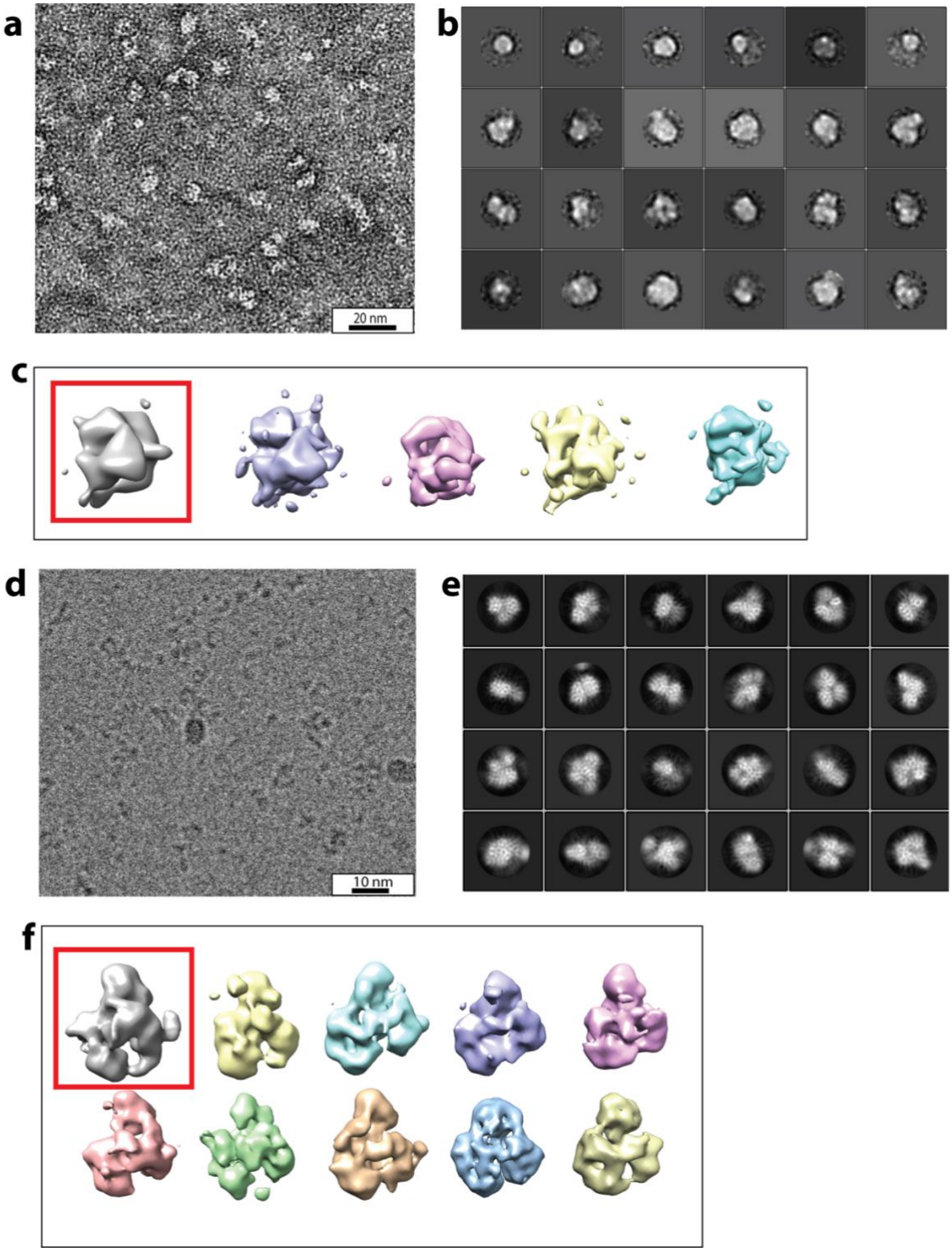


Figure S8. Structural characterization of the cleavage module. (**a**) A representative negative stain micrograph of the cleavage module. A scale bar is provided. (**b**) An image of selected 2D class averages of the cleavage module calculated from the negative stain data set. (**c**) Representative 3D classes of the cleavage module obtained from the negative stain data set. The ab initio model is indicated with the red box**.** (**d**), (**e**) and (**f**) are parallels of (**a**), (**b**) and (**c**) respectively representing cryo-EM analysis of the cleavage module.

Table S1. Mass spectrometric identification of protein in INT-PEC. Band in fraction 16 of the sucrose density gradient of INT-PEC was used for the analysis.

INT	Exclusive Unique Spectra
INTS1(1-294)	21
INTS2	99
INTS3	126
INTS4	110
INTS5	34
INTS6	153
INTS7	61
INTS8	75
INTS9	46
INTS10	49
INTS11	96
INTS12(1-194)	9
INTS13	48
INTS14	31
DDX26B	61
Pol II	
RPB1	98
RPB2	183
RPB3	17
RPB4	11
RPB6	6
RPB7	18
RPB8	12
RPB9	4
RPB10	10
NELF	
NELF -A	50

NELT -B	73
NELF -C	79
NELF -E	28
DSIF	
SPT5	142
SPT4	3

References

Albrecht, T. R., Shevtsov, S. P., Wu, Y., Mascibroda, L. G., Peart, N. J., Huang, K.-L., Sawyer, I. A., Tong, L., Dundr, M., & Wagner, E. J. (2018). Integrator subunit 4 is a 'Symplekin-like' scaffold that associates with INTS9/11 to form the Integrator cleavage module. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gky100

Albrecht, T. R., & Wagner, E. J. (2012). snRNA 3' end formation requires heterodimeric association of integrator subunits. *Molecular and Cellular Biology*, *32*(6), 1112–1123. https://doi.org/10.1128/MCB.06511-11

Allison, L. A., Moyle, M., Shales, M., & James Ingles, C. (1985). Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. *Cell*, *42*(2), 599–610. https://doi.org/10.1016/0092-8674(85)90117-5

Andel, F., Ladurner, a G., Inouye, C., Tjian, R., & Nogales, E. (1999). Three-dimensional structure of the human TFIID-IIA-IIB complex. *Science (New York, N.Y.)*, *286*(5447), 2153–2156.

Baejen, C., Andreani, J., Torkler, P., Battaglia, S., Schwalb, B., Lidschreiber, M., Maier, K. C., Boltendahl, A., Rus, P., Esslinger, S., Söding, J., & Cramer, P. (2017). Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes. *Molecular Cell*, *66*(1), 38-49.e6. https://doi.org/10.1016/j.molcel.2017.02.009

Baillat, D., Hakimi, M. A., Näär, A. M., Shilatifard, A., Cooch, N., & Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell*, *123*(2), 265–276. https://doi.org/10.1016/j.cell.2005.08.019

Baillat, D., Russell, W. K., & Wagner, E. J. (2016). CRISPR-Cas9 mediated genetic engineering for the purification of the endogenous integrator complex from mammalian cells. *Protein Expression and Purification*, *128*, 101–108. https://doi.org/10.1016/j.pep.2016.08.011

Baillat, D., & Wagner, E. J. (2015). Integrator: Surprisingly diverse functions in gene expression. In *Trends in Biochemical Sciences* (Vol. 40, Issue 5, pp. 257–264). https://doi.org/10.1016/j.tibs.2015.03.005

Barbieri, E., Trizzino, M., Welsh, S. A., Owens, T. A., Calabretta, B., Carroll, M., Sarma, K., & Gardini, A. (2018). Targeted Enhancer Activation by a Subunit of the Integrator Complex. *Molecular Cell*, *71*(1), 103-116.e7. https://doi.org/10.1016/j.molcel.2018.05.031

Berger, I., Fitzgerald, D. J., & Richmond, T. J. (2004). Baculovirus expression system for heterologous multiprotein complexes. *Nature Biotechnology*, *22*(12), 1583–1587. https://doi.org/10.1038/nbt1036

Bernecky, C., Herzog, F., Baumeister, W., Plitzko, J. M., & Cramer, P. (2016). Structure of transcribing mammalian RNA polymerase II. *Nature*, 1–14. https://doi.org/10.1038/nature16482

Bernecky, C., Plitzko, J. M., & Cramer, P. (2017). Structure of a transcribing RNA polymerase II–DSIF complex reveals a multidentate DNA–RNA clamp. *Nature Structural & Molecular Biology*, *24*(10), 809–815. https://doi.org/10.1038/nsmb.3465

Bieniossek, C., Papai, G., Schaffitzel, C., Garzoni, F., Chaillet, M., Scheer, E., Papadopoulos, P., Tora, L., Schultz, P., & Berger, I. (2013). The architecture of human general transcription factor TFIID core complex. *Nature*, *493*(7434), 699–702. https://doi.org/10.1038/nature11791

Bitinaite, J., & Nichols, N. M. (2009). DNA Cloning and Engineering by Uracil Excision. *Current Protocols in Molecular Biology*, *86*(1). https://doi.org/10.1002/0471142727.mb0321s86

Borukhov, S., Lee, J., & Laptenko, O. (2005). Bacterial transcription elongation factors: New insights into molecular mechanism of action. *Molecular Microbiology*, *55*(5), 1315–1324. https://doi.org/10.1111/j.1365-2958.2004.04481.x

Buratowski, S., Hahn, S., Guarente, L., & Sharp, P. A. (1989). Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell*, *56*(4), 549–561. https://doi.org/10.1016/0092-8674(89)90578-3

Burgess, R. R., Travers, A. A., Dunn, J. J., & Bautz, E. K. F. (1969). Factor stimulating transcription by RNA polymerase. *Nature*, *221*(5175), 43–46. https://doi.org/10.1038/221043a0

Casañal, A., Kumar, A., Hill, C. H., Easter, A. D., Emsley, P., Degliesposti, G., Gordiyenko, Y., Santhanam, B., Wolf, J., Wiederhold, K., Dornan, G. L., Skehel, M., Robinson, C. V, & Passmore, L. A. (2017). Architecture of eukaryotic mRNA 3′-end processing machinery. *Science*, *358*(6366), 1056–1059. https://doi.org/10.1126/science.aao6535

Chen, James, Noble, A. J., Kang, J. Y., & Darst, S. A. (2019). Eliminating effects of particle adsorption to the air/water interface in single-particle cryo-electron microscopy: Bacterial RNA polymerase and CHAPSO. *Journal of Structural Biology: X*, *1*(October 2018), 100005. https://doi.org/10.1016/j.yjsbx.2019.100005

Chen, Jiandong, Ezzeddine, N., Waltenspiel, B., Albrecht, T. R., Warren, W. D., Marzluff, W.

F., & Wagner, E. J. (2012a). An RNAi screen identifies additional members of the Drosophila Integrator complex and a requirement for cyclin C/Cdk8 in snRNA 3'-end formation. *RNA (New York, N.Y.)*, *18*(12), 2148–2156. https://doi.org/10.1261/rna.035725.112

Chen, Jiandong, Ezzeddine, N., Waltenspiel, B., Albrecht, T. R., Warren, W. D., Marzluff, W. F., & Wagner, E. J. (2012b). An RNAi screen identifies additional members of the Drosophila Integrator complex and a requirement for cyclin C/Cdk8 in snRNA 3'-end formation. *RNA*, *18*(12), 2148–2156. https://doi.org/10.1261/rna.035725.112

Chen, Jiandong, Waltenspiel, B., Warren, W. D., & Wagner, E. J. (2013). Functional analysis of the integrator subunit 12 identifies a microdomain that mediates activation of the drosophila integrator complex. *Journal of Biological Chemistry*, *288*(7), 4867–4877. https://doi.org/10.1074/jbc.M112.425892

Chen, Z. L., Meng, J. M., Cao, Y., Yin, J. L., Fang, R. Q., Fan, S. B., Liu, C., Zeng, W. F., Ding, Y. H., Tan, D., Wu, L., Zhou, W. J., Chi, H., Sun, R. X., Dong, M. Q., & He, S. M. (2019). A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-11337-z

Cheung, A. C. M., & Cramer, P. (2011). Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature*, *471*(7337), 249–253. https://doi.org/10.1038/nature09785

Chodosh, L. A., Firell, A., Samuels, M., & Sharp, P. A. (1989). *Inhibits Transcription Elongation by RNA Polymerase II in Vitro *. 264*(4), 2250–2257.

Cohen, S., Puget, N., Lin, Y. L., Clouaire, T., Aguirrebengoa, M., Rocher, V., Pasero, P., Canitrot, Y., & Legube, G. (2018). Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-02894-w

Combe, C. W., Fischer, L., & Rappsilber, J. (2015). xiNET: Cross-link network maps with residue resolution. *Molecular and Cellular Proteomics*, *14*(4), 1137–1147. https://doi.org/10.1074/mcp.O114.042259

Conaway, R. C., & Conaway, J. W. (2019). The hunt for RNA polymerase II elongation factors: a historical perspective. *Nature Structural and Molecular Biology*, *26*(9), 771–776. https://doi.org/10.1038/s41594-019-0283-1

Core, L., & Adelman, K. (2019). Promoter-proximal pausing of RNA polymerase II: A nexus of gene regulation. In *Genes and Development*. https://doi.org/10.1101/gad.325142.119

Cramer, P. (2019). Organization and regulation of gene transcription. *Nature*, *573*(7772), 45–

54. https://doi.org/10.1038/s41586-019-1517-4

Cramer, P., Bushnel, D. A., & Kornberg, R. D. (2001). Structural Basis of Transcription: RNA Polymerase II at 2.8 Angstrom Resolution. *Science*, *292*(5523), 1863–1876. https://doi.org/10.1126/science.1059493

Cramer, P., Bushnell, D. A., Fu, J., Gnatt, A. L., Maier-Davis, B., Thompson, N. E., Burgess, R. R., Edwards, A. M., David, P. R., & Kornberg, R. D. (2000). Architecture of RNA polymerase II and implications for the transcription mechanism. *Science*, *288*(5466), 640–649. https://doi.org/10.1126/science.288.5466.640

Creamer, T. J., Darby, M. M., Jamonnak, N., Schaughency, P., Hao, H., Wheelan, S. J., & Corden, J. L. (2011). Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genetics*, *7*(10). https://doi.org/10.1371/journal.pgen.1002329

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, *227*, 561–563. https://doi.org/10.1007/978-1-4020-6754-9_2672

D'Imprima, E., Floris, D., Joppe, M., Sánchez, R., Grininger, M., & Kühlbrandt, W. (2019). Protein denaturation at the air-water interface and how to prevent it. *ELife*, *8*, 1–18. https://doi.org/10.7554/eLife.42747

Dahmus, M. E. (1996). Reversible phosphorylation of the C-terminal domain of RNA polymerase II. *Journal of Biological Chemistry*, *271*(32), 19009–19012. https://doi.org/10.1074/jbc.271.32.19009

Danan, C., Manickavel, S., & Hafner, M. (2016). PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites HHS Public Access. *Methods Mol Biol*, *1358*, 153–173. https://doi.org/10.1007/978-1-4939-3067-8_10

Dienemann, C., Schwalb, B., Schilbach, S., & Cramer, P. (2019). Promoter Distortion and Opening in the RNA Polymerase II Cleft. *Molecular Cell*, *73*(1), 97-106.e4. https://doi.org/10.1016/j.molcel.2018.10.014

Dyson, H. J., & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. In *Nature Reviews Molecular Cell Biology* (Vol. 6, Issue 3, pp. 197–208). https://doi.org/10.1038/nrm1589

Egloff, S., O'Reilly, D., Chapman, R. D., Taylor, A., Tanzhaus, K., Pitts, L., Eick, D., & Murphy, S. (2007). Serine-7 of the RNA polymerase II CTD is specifically required for snRNA gene expression. *Science*, *318*(5857), 1777–1779. https://doi.org/10.1126/science.1145989

Egloff, S., Szczepaniak, S. A., Dienstbier, M., Taylor, A., Knight, S., & Murphy, S. (2010).

The integrator complex recognizes a new double mark on the RNA polymerase II carboxyl-terminal domain. *Journal of Biological Chemistry*, *285*(27), 20564–20569. https://doi.org/10.1074/jbc.M110.132530

Egloff, S., Zaborowska, J., Laitem, C., Kiss, T., & Murphy, S. (2012). Ser7 phosphorylation of the CTD recruits the RPAP2 ser5 phosphatase to snRNA genes. *Molecular Cell*, *45*(1), 111–122. https://doi.org/10.1016/j.molcel.2011.11.006

Ehara, H., Yokoyama, T., Shigematsu, H., Yokoyama, S., Shirouzu, M., & Sekine, S.-I. (2017). Structure of the complete elongation complex of RNA polymerase II with basal factors. *Science (New York, N.Y.)*, *357*(6354), 921–924. https://doi.org/10.1126/science.aan8552

Eick, D., & Geyer, M. (2013). The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code. *Chemical Reviews*, *113*(11), 8456–8490. https://doi.org/10.1021/cr400071f

Elrod, N. D., Henriques, T., Huang, K. L., Tatomer, D. C., Wilusz, J. E., Wagner, E. J., & Adelman, K. (2019). The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. *Molecular Cell*, *76*(5), 738-752.e7. https://doi.org/10.1016/j.molcel.2019.10.034

Engel, C., Sainsbury, S., Cheung, A. C., Kostrewa, D., & Cramer, P. (2013). RNA polymerase i structure and transcription regulation. *Nature*, *502*(7473), 650–655. https://doi.org/10.1038/nature12712

Fan, X., Wang, J., Zhang, X., Yang, Z., Zhang, J. C., Zhao, L., Peng, H. L., Lei, J., & Wang, H. W. (2019). Single particle cryo-EM reconstruction of 52 kDa streptavidin at 3.2 Angstrom resolution. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-10368-w

Fong, N., Brannan, K., Erickson, B., Kim, H., Cortazar, M. A., Sheridan, R. M., Nguyen, T., Karp, S., & Bentley, D. L. (2015). Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Molecular Cell*, *60*(2), 256–267. https://doi.org/10.1016/j.molcel.2015.09.026

Fuda, N. J., Ardehali, M. B., & Lis, J. T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, *461*(7261), 186–192. https://doi.org/10.1038/nature08449

Gardini, A., Baillat, D., Cesaroni, M., Hu, D., Marinis, J. M., Wagner, E. J., Lazar, M. A., Shilatifard, A., & Shiekhattar, R. (2014). Integrator regulates transcriptional initiation and pause release following activation. *Molecular Cell*, *56*(1), 128–139. https://doi.org/10.1016/j.molcel.2014.08.004

Gilmour, D. S., & Lis, J. T. (1986). RNA Polymerase II Interacts with the Promoter Region of the Noninduced hsp7O Gene in Drosophila melanogaster Cells. *Brenner's Encyclopedia of Genetics*, *6*(11), 3984–3989. https://doi.org/10.1016/B978-0-12-374984-0.01350-4

Gómez-Orte, E., Sáenz-Narciso, B., Zheleva, A., Ezcurra, B., de Toro, M., López, R., Gastaca, I., Nilsen, H., Sacristán, M. P., Schnabel, R., & Cabello, J. (2019). Disruption of the Caenorhabditis elegans Integrator complex triggers a non-conventional transcriptional mechanism beyond snRNA genes. *PLoS Genetics*, *15*(2), e1007981. https://doi.org/10.1371/journal.pgen.1007981

Goodfellow, S. J., & Zomerdijk, J. C. B. M. (2013). Basic Mechanisms in RNA Polymerase I Transcription of the Ribosomal RNA Genes. In T. K. Kundu (Ed.), *Epigenetics: Development and Disease*, (Vol. 61, pp. 211–236). Springer Netherlands. https://doi.org/10.1007/978-94-007-4525-4_10

Goodrich, J. A., & Tjian, R. (1994). Transcription factors IIE and IIH and ATP hydrolysis direct promoter clearance by RNA polymerase II. *Cell*, *77*(1), 145–156. https://doi.org/10.1016/0092-8674(94)90242-9

Gradia, S. D., Ishida, J. P., Tsai, M. S., Jeans, C., Tainer, J. A., & Fuss, J. O. (2017). MacroBac: New Technologies for Robust and Efficient Large-Scale Production of Recombinant Multiprotein Complexes. In *Methods in Enzymology*. https://doi.org/10.1016/bs.mie.2017.03.008

Gressel, S., Schwalb, B., Decker, T. M., Qin, W., Leonhardt, H., Eick, D., & Cramer, P. (2017). CDK9-dependent RNA polymerase II pausing controls transcription initiation. *ELife*, *6*, 1–24. https://doi.org/10.7554/eLife.29736

Guiro, J., & Murphy, S. (2017). Regulation of expression of human RNA polymerase II-Transcribed snRNA genes. *Open Biology*, *7*(6), 3–11. https://doi.org/10.1098/rsob.170073

Hantsche, M., & Cramer, P. (2016). The Structural Basis of Transcription: 10 Years After the Nobel Prize in Chemistry. *Angewandte Chemie - International Edition*, *55*(52), 15972–15981. https://doi.org/10.1002/anie.201608066

Hardesty, B., Tsalkova, T., & Kramer, G. (1999). Co-translational folding. *Current Opinion in Structural Biology*, *9*(1), 111–114. https://doi.org/10.1016/S0959-440X(99)80014-1

He, Y., Fang, J., Taatjes, D. J., & Nogales, E. (2013). Structural visualization of key steps in human transcription initiation. *Nature*, *495*(7442), 481–486. https://doi.org/10.1038/nature11991

Herzog, F., Kahraman, A., Boehringer, D., Mak, R., Bracher, A., Walzthoeni, T., Leitner, A., Beck, M., Hartl, F.-U., Ban, N., Malmstrom, L., & Aebersold, R. (2012a). Structural

Probing of a Protein Phosphatase 2A Network by Chemical Cross-Linking and Mass Spectrometry. *Science*, *337*(6100), 1348–1352. https://doi.org/10.1126/science.1221483

Herzog, F., Kahraman, A., Boehringer, D., Mak, R., Bracher, A., Walzthoeni, T., Leitner, A., Beck, M., Hartl, F.-U., Ban, N., Malmstrom, L., & Aebersold, R. (2012b). Structural Probing of a Protein Phosphatase 2A Network by Chemical Cross-Linking and Mass Spectrometry. *Science*, *337*(6100), 1348–1352. https://doi.org/10.1126/science.1221483

Huang, J., Gong, Z., Ghosal, G., & Chen, J. (2009). SOSS Complexes Participate in the Maintenance of Genomic Stability. *Molecular Cell*, *35*(3), 384–393. https://doi.org/10.1016/j.molcel.2009.06.011

Hurwitz, J. (2005). The discovery of RNA polymerase. *Journal of Biological Chemistry*, *280*(52), 42477–42485. https://doi.org/10.1074/jbc.X500006200

Hurwitz, J., Bresler, A., & Diringer, R. (1960). *Biochemical and Biophysical Research Communication*. *3*(1), 15–18.

Izban, M. G., & Luse, D. S. (1992). The RNA polymerase II ternary complex cleaves the nascent transcript in a 3′ → 5′ direction in the presence of elongation factor SII. *Genes and Development*, *6*(7), 1342–1356. https://doi.org/10.1101/gad.6.7.1342

Jasiak, A. J., Armache, K.-J., Martens, B., Jansen, R.-P., & Cramer, P. (2006). Structural Biology of RNA Polymerase III: Subcomplex C17/25 X-Ray Structure and 11 Subunit Enzyme Model. *Molecular Cell*, *23*(1), 71–81. https://doi.org/10.1016/j.molcel.2006.05.013

Jawdekar, G. W., & Henry, R. W. (2008). Transcriptional regulation of human small nuclear RNA genes. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, *1779*(5), 295–305. https://doi.org/10.1016/j.bbagrm.2008.04.001

Jeruzalmi, D., & Steitz, T. A. (1998). Structure of T7 RNA polymerase complexed to the transcriptional inhibitor T7 lysozyme. *EMBO Journal*, *17*(14), 4101–4113. https://doi.org/10.1093/emboj/17.14.4101

Kim, J. L., Nikolov, D. B., & Burley, S. K. (1993). Co-crystal structure of TBP recognizing the minor groove of a TATA element. *Nature*, *365*(6446), 520–527. https://doi.org/10.1038/365520a0

Kim, M., Krogan, N. J., Vasiljeva, L., Rando, O. J., Nedea, E., Greenblatt, J. F., & Buratowski, S. (2004). The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature*, *432*(7016), 517–522. https://doi.org/10.1038/nature03041

Kimanius, D., Forsberg, B. O., Scheres, S. H. W., & Lindahl, E. (2016). Accelerated cryo-EM structure determination with parallelisation using GPUS in RELION-2. *ELife*, *5*, 1–21.

https://doi.org/10.7554/eLife.18722

Kornberg, R. D. (2005). Mediator and the mechanism of transcriptional activation. *Trends in Biochemical Sciences*, *30*(5), 235–239. https://doi.org/10.1016/j.tibs.2005.03.011

Kuhn, C. D., Geiger, S. R., Baumli, S., Gartmann, M., Gerber, J., Jennebach, S., Mielke, T., Tschochner, H., Beckmann, R., & Cramer, P. (2007). Functional Architecture of RNA Polymerase I. *Cell*, *131*(7), 1260–1272. https://doi.org/10.1016/j.cell.2007.10.051

Kumar, A., Clerici, M., Muckenfuss, L. M., Passmore, L. A., & Jinek, M. (2019). Mechanistic insights into mRNA 3′-end processing. In *Current Opinion in Structural Biology* (Vol. 59, pp. 143–150). Elsevier Ltd. https://doi.org/10.1016/j.sbi.2019.08.001

Kwak, H., Fuda, N. J., Core, L. J., & Lis, J. T. (2013). Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science*, *339*(6122), 950–953. https://doi.org/10.1126/science.1229386

Kwak, H., & Lis, J. T. (2013). Control of Transcriptional Elongation. *Annual Review of Genetics*, *47*(1), 483–508. https://doi.org/10.1146/annurev-genet-110711-155440

Lai, F., Gardini, A., Zhang, A., & Shiekhattar, R. (2015). Integrator mediates the biogenesis of enhancer RNAs. *Nature*, *525*(7569), 399–403. https://doi.org/10.1038/nature14906

Larochelle, M., Robert, M.-A., Hébert, J.-N., Liu, X., Matteau, D., Rodrigue, S., Tian, B., Jacques, P.-É., & Bachand, F. (2018). Common mechanism of transcription termination at coding and noncoding RNA genes in fission yeast. *Nature Communications*, *9*(1), 4364. https://doi.org/10.1038/s41467-018-06546-x

Leitner, A., Faini, M., Stengel, F., & Aebersold, R. (2016). Crosslinking and Mass Spectrometry: An Integrated Technology to Understand the Structure and Function of Molecular Machines. *Trends in Biochemical Sciences*, *41*(1), 20–32. https://doi.org/10.1016/j.tibs.2015.10.008

Leitner, A., Walzthoeni, T., Kahraman, A., Herzog, F., Rinner, O., Beck, M., & Aebersold, R. (2010). Probing native protein structures by chemical cross-linking, mass spectrometry, and bioinformatics. *Molecular and Cellular Proteomics*, *9*(8), 1634–1649. https://doi.org/10.1074/mcp.R000001-MCP201

Li, Y., Bolderson, E., Kumar, R., Muniandy, P. A., Xue, Y., Richard, D. J., Seidman, M., Pandita, T. K., Khanna, K. K., & Wang, W. (2009). hSSB1 and hSSB2 form similar multiprotein complexes that participate in DNA damage response. *Journal of Biological Chemistry*, *284*(35), 23525–23531. https://doi.org/10.1074/jbc.C109.039586

Licatalosi, D. D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J. B., & Bentley, D. L. (2002). Functional interaction of yeast pre-mRNA 3′ end processing factors with RNA

polymerase II. *Molecular Cell*. https://doi.org/10.1016/S1097-2765(02)00518-X

Lis, J. T. (2019). A 50 year history of technologies that drove discovery in eukaryotic transcription regulation. *Nature Structural & Molecular Biology*, *26*(9), 777–782. https://doi.org/10.1038/s41594-019-0288-9

Liu, H., & Naismith, J. H. (2008). An efficient one-step site-directed deletion, insertion, single and multiple-site plasmid mutagenesis protocol. *BMC Biotechnology*, *8*(1), 91. https://doi.org/10.1186/1472-6750-8-91

Lobo, S. M., & Hernandez, N. (1989). A 7 bp mutation converts a human RNA polymerase II snRNA promoter into an RNA polymerase III promoter. *Cell*, *58*(1), 55–67. https://doi.org/10.1016/0092-8674(89)90402-9

Lorch, Y., & Kornberg, R. D. (2017). Chromatin-remodeling for transcription. *Quarterly Reviews of Biophysics*, *50*, 1–15. https://doi.org/10.1017/S003358351700004X

Lorch, Y., LaPointe, J. W., & Kornberg, R. D. (1987). Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*, *49*(2), 203–210. https://doi.org/10.1016/0092-8674(87)90561-7

Luckow, V. A., Lee, S. C., Barry, G. F., & Olins, P. O. (1993). Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in Escherichia coli. *Journal of Virology*, *67*(8), 4566–4579. https://doi.org/10.1128/JVI.67.8.4566-4579.1993

Malovannaya, A., Lanz, R. B., Jung, S. Y., Bulynko, Y., Le, N. T., Chan, D. W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., Kim, B. J., Li, C., Chen, R., Li, W., Wang, Y., O'Malley, B. W., & Qin, J. (2011). Analysis of the human endogenous coregulator complexome. *Cell*, *145*(5), 787–799. https://doi.org/10.1016/j.cell.2011.05.006

Mandel, C. R., Kaneko, S., Zhang, H., Gebauer, D., Vethantham, V., Manley, J. L., & Tong, L. (2006). Polyadenylation factor CPSF-73 is the pre-mRNA 3'-end-processing endonuclease. *Nature*, *444*(7121), 953–956. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3866582&tool=pmcentrez&rendertype=abstract

Marshall, N. F., & Price, D. H. (1995). Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *Journal of Biological Chemistry*, *270*(21), 12335–12338. https://doi.org/10.1074/jbc.270.21.12335

Marzluff, W. F., Wagner, E. J., & Duronio, R. J. (2008). Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nature Reviews Genetics*, *9*(11), 843–854. https://doi.org/10.1038/nrg2438

Mastronarde, D. N. (2005). Automated electron microscope tomography using robust prediction of specimen movements. *Journal of Structural Biology*. https://doi.org/10.1016/j.jsb.2005.07.007

Mattaj, I. W., Dathan, N. A., Parry, H. D., Carbon, P., & Krol, A. (1988). Changing the RNA polymerase specificity of U snRNA gene promoters. *Cell*, *55*(3), 435–442. https://doi.org/10.1016/0092-8674(88)90029-3

Mayer, A., Heidemann, M., Lidschreiber, M., Schreieck, A., Sun, M., Hintermair, C., Kremmer, E., Eick, D., & Cramer, P. (2012). CTD Tyrosine phosphorylation impairs termination factor recruitment to RNA polymerase II. *Science*, *336*(6089), 1723–1725. https://doi.org/10.1126/science.1219651

Mitchell, P., & Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, *245*(4916), 371–378. https://doi.org/10.1126/science.2667136

Moreland, R. J., Tirode, F., Yan, Q., Conaway, J. W., Egly, J. M., & Conaway, R. C. (1999). A role for the TFIIH XPB DNA helicase in promoter escape by RNA polymerase II. *Journal of Biological Chemistry*, *274*(32), 22127–22130. https://doi.org/10.1074/jbc.274.32.22127

Mühlbacher, W., Sainsbury, S., Hemann, M., Hantsche, M., Neyer, S., Herzog, F., & Cramer, P. (2014). Conserved architecture of the core RNA polymerase II initiation complex. *Nature Communications*, *5*, 4310. https://doi.org/10.1038/ncomms5310

Narita, T., Yamaguchi, Y., Yano, K., Sugimoto, S., Chanarat, S., Wada, T., Kim, D. -k., Hasegawa, J., Omori, M., Inukai, N., Endoh, M., Yamada, T., & Handa, H. (2003). Human Transcription Elongation Factor NELF: Identification of Novel Subunits and Reconstitution of the Functionally Active Complex. *Molecular and Cellular Biology*, *23*(6), 1863–1873. https://doi.org/10.1128/mcb.23.6.1863-1873.2003

Neuman de Vegvar, H. E., Lund, E., & Dahlberg, J. E. (1986). 3′ end formation of U1 snRNA precursors is coupled to transcription from snRNA promoters. *Cell*, *47*(2), 259–266. https://doi.org/10.1016/0092-8674(86)90448-4

Nevalainen, K. M. H., Te'o, V. S. J., & Bergquist, P. L. (2005). Heterologous protein expression in filamentous fungi. *Trends in Biotechnology*, *23*(9), 468–474. https://doi.org/10.1016/j.tibtech.2005.06.002

Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T., & Kanemaki, M. (2009). An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nature Methods*, *6*(12), 917–922. https://doi.org/10.1038/nmeth.1401

Nogales, E., Louder, R. K., & He, Y. (2016). Cryo-EM in the study of challenging systems: the human transcription pre-initiation complex. *Current Opinion in Structural Biology*, *40*, 120–127. https://doi.org/10.1016/j.sbi.2016.09.009

Nojima, T., Gomes, T., Grosso, A. R. F., Kimura, H., Dye, M. J., Dhir, S., Carmo-Fonseca, M., & Proudfoot, N. J. (2015). Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, *161*(3), 526–540. https://doi.org/10.1016/J.CELL.2015.03.027

Perales, R., & Bentley, D. (2009). "Cotranscriptionality": The Transcription Elongation Complex as a Nexus for Nuclear Transactions. *Molecular Cell*, *36*(2), 178–191. https://doi.org/10.1016/j.molcel.2009.09.018

Peti, W., & Page, R. (2007). Strategies to maximize heterologous protein expression in Escherichia coli with minimal cost. *Protein Expression and Purification*, *51*(1), 1–10. https://doi.org/10.1016/j.pep.2006.06.024

Pickar-Oliver, A., & Gersbach, C. A. (2019). The next generation of CRISPR–Cas technologies and applications. In *Nature Reviews Molecular Cell Biology* (Vol. 20, Issue 8, pp. 490–507). Nature Publishing Group. https://doi.org/10.1038/s41580-019-0131-5

Porrua, O., Boudvillain, M., & Libri, D. (2016). Transcription Termination: Variations on Common Themes. *Trends in Genetics*, *32*(8), 508–522. https://doi.org/10.1016/j.tig.2016.05.007

Porrua, O., & Libri, D. (2013). A bacterial-like mechanism for transcription termination by the Sen1p helicase in budding yeast. *Nature Structural and Molecular Biology*, *20*(7), 884–891. https://doi.org/10.1038/nsmb.2592

Proudfoot, N. J. (2016). Transcriptional translation in mammals: Stopping de RNA polymerase II juggernaut. *Science*, *352*(6291), 1–22. https://doi.org/10.1126/science

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., & Séraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods (San Diego, Calif.)*, *24*(3), 218–229. https://doi.org/10.1006/meth.2001.1183

Punjani, A., Rubinstein, J. L., Fleet, D. J., & Brubaker, M. A. (2017). cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, *14*(3), 290–296. https://doi.org/10.1038/nmeth.4169

Quan, J., & Tian, J. (2009). Circular Polymerase Extension Cloning of Complex Gene Libraries and Pathways. *PLoS ONE*, *4*(7), e6441. https://doi.org/10.1371/journal.pone.0006441

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A., & Zhang, F. (2013). Genome

engineering using the CRISPR-Cas9 system. *Nat. Protocols*, *8*(11), 2281–2308. https://doi.org/10.1038/nprot.2013.143\rhttp://www.nature.com/nprot/journal/v8/n11/abs/nprot.2013.143.html#supplementary-information

Reines, D., Ghanouni, P., Li, Q., Mote, J., & Jr. (1992). The RNA Polymerase II Elongation Complex Factor-Dependent Transcription elongation involves nascent RNA cleavage. *The Journal of Biological Chemistry*, *267*(22), 15516–15522.

Ren, W., Chen, H., Sun, Q., Tang, X., Lim, S. C., Huang, J., & Song, H. (2014). Structural Basis of SOSS1 Complex Assembly and Recognition of ssDNA. *Cell Reports*, *6*(6), 982–991. https://doi.org/10.1016/j.celrep.2014.02.020

Rienzo, M., & Casamassimi, A. (2016a). Integrator complex and transcription regulation: Recent findings and pathophysiology. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1859*(10), 1269–1280. https://doi.org/10.1016/j.bbagrm.2016.07.008

Rienzo, M., & Casamassimi, A. (2016b). Integrator complex and transcription regulation: Recent findings and pathophysiology. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1859*(10), 1269–1280. https://doi.org/10.1016/j.bbagrm.2016.07.008

Riva, M., Schultz, P., & Carles, C. (1998). The RNA cleavage activity of RNA polymerase III is mediated by an essential TFIIS-like subunit and is. *Genes & Development*, 3857–3871.

Roeder, R. G., & Rutter, W. J. (1969). Multiple Forms of DNA-dependent RNA Polymerase in Eukaryotic Organisms. *Nature*, *224*(5216), 234–237. https://doi.org/10.1038/224234a0

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews. Molecular Cell Biology*, *16*(3), 129–143. https://doi.org/10.1038/nrm3952

Sainsbury, S., Niesser, J., & Cramer, P. (2013). Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. *Nature*, *493*(7432), 437–440. https://doi.org/10.1038/nature11715

Scheres, S. H. W. (2012). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. *Journal of Structural Biology*, *180*(3), 519–530. https://doi.org/10.1016/j.jsb.2012.09.006

Schilbach, S., Hantsche, M., Tegunov, D., Dienemann, C., Wigge, C., Urlaub, H., & Cramer, P. (2017). Structures of transcription pre-initiation complex with TFIIH and Mediator. *Nature*, *551*(7679), 204–209. https://doi.org/10.1038/nature24282

Schwalb, B., Michel, M., Zacher, B., Hauf, K. F., Demel, C., Tresch, A., Gagneur, J., & Cramer,

P. (2016). TT-seq maps the human transient transcriptome. *Science*, *352*(6290), 1225–1228. https://doi.org/10.1126/science.aad9841

Sekimizu, K., Kobayashi, N., Mizuno, D., & Natori, S. (1976). Purification of a Factor from Ehrlich Ascites Tumor Cells Specifically Stimulating RNA Polymerase II. *Biochemistry*, *15*(23), 5064–5070. https://doi.org/10.1021/bi00668a018

Shah, N., Maqbool, M. A., Yahia, Y., El Aabidine, A. Z., Esnault, C., Forné, I., Decker, T.-M., Martin, D., Schüller, R., Krebs, S., Blum, H., Imhof, A., Eick, D., & Andrau, J.-C. (2018). Tyrosine-1 of RNA Polymerase II CTD Controls Global Termination of Gene Transcription in Mammals. *Molecular Cell*, *69*(1), 48-61.e6. https://doi.org/10.1016/j.molcel.2017.12.009

Skaar, J. R., Ferris, A. L., Wu, X., Saraf, A., Khanna, K. K., Florens, L., Washburn, M. P., Hughes, S. H., & Pagano, M. (2015). The Integrator complex controls the termination of transcription at diverse classes of gene targets. *Cell Research*, *25*(3), 288–305. https://doi.org/10.1038/cr.2015.19

Skaar, J. R., Richard, D. J., Saraf, A., Toschi, A., Bolderson, E., Florens, L., Washburn, M. P., Khanna, K. K., & Pagano, M. (2009). INTS3 controls the hSSB1-mediated DNA damage response. *Journal of Cell Biology*, *187*(1), 25–32. https://doi.org/10.1083/jcb.200907026

Skourti-Stathaki, K., Proudfoot, N. J., & Gromak, N. (2011). Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. *Molecular Cell*, *42*(6), 794–805. https://doi.org/10.1016/j.molcel.2011.04.026

Solis-Mezarino, V., & Herzog, F. (2017). compleXView: a server for the interpretation of protein abundance and connectivity information to identify protein complexes. *Nucleic Acids Research*, *45*(W1), W276–W284. https://doi.org/10.1093/nar/gkx411

Spitzer, J., Hafner, M., Landthaler, M., Ascano, M., Farazi, T., Wardle, G., Nusbaum, J., Khorshid, M., Burger, L., Zavolan, M., & Tuschl, T. (2014). PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): A Step-By-Step Protocol to the Transcriptome-Wide Identification of Binding Sites of RNA-Binding Proteins. In *Methods in Enzymology* (Vol. 539, pp. 113–161). Academic Press Inc. https://doi.org/10.1016/B978-0-12-420120-0.00008-6

Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., Raffel, R., Sobhian, B., Severac, D., Rialle, S., Parrinello, H., Cuvier, O., & Benkirane, M. (2014). Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nat Commun*, *5*, 5531. https://doi.org/10.1038/ncomms6531

Stevens, A. (1960). Incorporation of the adenine ribonucleotide into RNA by cell fractions from E. coli B. *Biochemical and Biophysical Research Communications*, *3*(1), 92–96. https://doi.org/10.1016/0006-291X(60)90110-8

Sun, Y., Zhang, Y., Aik, W. S., Yang, X.-C., Marzluff, W. F., Walz, T., Dominski, Z., & Tong, L. (2020). Structure of an active human histone pre-mRNA 3′-end processing machinery. *Science*, *367*(6478), 700–703. https://doi.org/10.1126/science.aaz7758

Swanson, M. S., & Winston, F. (1992). SPT4, SPT5 and SPT6 interactions: Effects on transcription and viability in Saccharomyces cerevisiae. *Genetics*, *132*(2), 325–336.

Tatomer, D. C., Elrod, N. D., Liang, D., Xiao, M., Jiang, J. Z., Jonathan, M., Huang, K., Wagner, E. J., Cherry, S., & Wilusz, J. E. (2019). The Integrator complex cleaves nascent mRNAs to attenuate transcription. *BioRxiv Molecular Biology*, 1–14. https://doi.org/10.1101/748319

Tegunov, D., & Cramer, P. (2019). Real-time cryo-electron microscopy data preprocessing with Warp. *Nature Methods*, *16*(11), 1146–1152. https://doi.org/10.1038/s41592-019-0580-y

Tewari, R., Bailes, E., Bunting, K. A., & Coates, J. C. (2010). Armadillo-repeat protein functions: Questions for little creatures. In *Trends in Cell Biology* (Vol. 20, Issue 8, pp. 470–481). https://doi.org/10.1016/j.tcb.2010.05.003

Thummel, C. S., Burtis, K. C., & Hogness, D. S. (1990). Spatial and temporal patterns of E74 transcription during Drosophila development. *Cell*, *61*(1), 101–111. https://doi.org/10.1016/0092-8674(90)90218-4

Trowitzsch, S., Bieniossek, C., Nie, Y., Garzoni, F., & Berger, I. (2010). New baculovirus expression tools for recombinant protein complex production. *Journal of Structural Biology*, *172*(1), 45–54. https://doi.org/10.1016/j.jsb.2010.02.010

Vannini, A., & Cramer, P. (2012). Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Molecular Cell*, *45*(4), 439–446. https://doi.org/10.1016/j.molcel.2012.01.023

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., & Meinhart, A. (2008). The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nature Structural and Molecular Biology*, *15*(8), 795–804. https://doi.org/10.1038/nsmb.1468

Vieira Gomes, A., Souza Carmo, T., Silva Carvalho, L., Mendonça Bahia, F., & Parachin, N. (2018). Comparison of Yeasts as Hosts for Recombinant Protein Production. *Microorganisms*, *6*(2), 38. https://doi.org/10.3390/microorganisms6020038

Vos, S. M., Farnung, L., Boehning, M., Wigge, C., Linden, A., Urlaub, H., & Cramer, P. (2018).

Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. *Nature*, *560*(7720), 607–612. https://doi.org/10.1038/s41586-018-0440-4

Vos, S. M., Farnung, L., Urlaub, H., & Cramer, P. (2018). Structure of paused transcription complex Pol II–DSIF–NELF. *Nature*, *560*(7720), 601–606. https://doi.org/10.1038/s41586-018-0442-2

Wada, T., Takagi, T., Yamaguchi, Y., Ferdous, A., Imai, T., Hirose, S., Sugimoto, S., Yano, K., Hartzog, G. A., Winston, F., Buratowski, S., & Handa, H. (1998). DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes and Development*, *12*(3), 343–356. https://doi.org/10.1101/gad.12.3.343

Waddell, C. S., & Craig, N. L. (1989). Tn7 transposition: recognition of the attTn7 target sequence. *Proceedings of the National Academy of Sciences*, *86*(11), 3958–3962. https://doi.org/10.1073/pnas.86.11.3958

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of Molecular Biology*, *337*(3), 635–645. https://doi.org/10.1016/j.jmb.2004.02.002

Weinmann, R., Raskas, H. J., & Roeder, R. G. (1974). Role of DNA dependent RNA polymerases II and III in transcription of the adenovirus genome late in productive infection. *Proceedings of the National Academy of Sciences of the United States of America*, *71*(9), 3426–3430. https://doi.org/10.1073/pnas.71.9.3426

Weinmann, Roberto, & Roeder, R. G. (1974). Role of DNA-Dependent RNA Polymerase III in the Transcription of the tRNA and 5S RNA Genes (mouse myeloma cells/isolated nuclei/a-amanitin). *Proceedings of the National Academy of Sciences of the United States of America*, *71*(5), 1790–1794. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC388326/pdf/pnas00058-0212.pdf

Wu, Y., Albrecht, T. R., Baillat, D., Wagner, E. J., & Tong, L. (2017). Molecular basis for the interaction between Integrator subunits IntS9 and IntS11 and its functional importance. *Proceedings of the National Academy of Sciences*, *114*(17), 4394–4399. https://doi.org/10.1073/pnas.1616605114

Xie, M., Zhang, W., Shu, M.-D., Xu, A., Lenis, D. A., DiMaio, D., & Steitz, J. A. (2015). The host Integrator complex acts in transcription-independent maturation of herpesvirus microRNA 3’ ends. *Genes & Development*, *29*(14), 1552–1564. https://doi.org/10.1101/gad.266973.115

Yamaguchi, Y., Takagi, T., Wada, T., Yano, K., Furuya, A., Sugimoto, S., Hasegawa, J., &

Handa, H. (1999). NELF, a Multisubunit Complex Containing RD, Cooperates with DSIF to Repress RNA Polymerase II Elongation. *Cell*, *97*(1), 41–51. https://doi.org/10.1016/S0092-8674(00)80713-8

Yamamoto, J., Hagiwara, Y., Chiba, K., Isobe, T., Narita, T., Handa, H., & Yamaguchi, Y. (2014). DSIF and NELF interact with Integrator to specify the correct post-transcriptional fate of snRNA genes. *Nature Communications*, *5*(4263), 1–10. https://doi.org/10.1038/ncomms5263

Yang, B., Wu, Y. J., Zhu, M., Fan, S. B., Lin, J., Zhang, K., Li, S., Chi, H., Li, Y. X., Chen, H. F., Luo, S. K., Ding, Y. H., Wang, L. H., Hao, Z., Xiu, L. Y., Chen, S., Ye, K., He, S. M., & Dong, M. Q. (2012). Identification of cross-linked peptides from complex samples. *Nature Methods*, *9*(9), 904–906. https://doi.org/10.1038/nmeth.2099

Yoshimura, S. H., & Hirano, T. (2016). HEAT repeats - versatile arrays of amphiphilic helices working in crowded environments? In *Journal of Cell Science* (Vol. 129, Issue 21, pp. 3963–3970). https://doi.org/10.1242/jcs.185710

Zaborowska, J., Egloff, S., & Murphy, S. (2016). The pol II CTD: new twists in the tail. *Nature Structural & Molecular Biology*, *23*(9), 771–777. https://doi.org/10.1038/nsmb.3285

Zaborowska, J., Taylor, A., Roeder, R. G., & Murphy, S. (2012). A novel TBP-TAF complex on RNA polymerase II-transcribed snRNA genes. *Transcription*, *3*(2), 92–104. https://doi.org/10.4161/trns.19783

Zeytuni, N., & Zarivach, R. (2012a). Structural and functional discussion of the tetra-trico-peptide repeat, a protein interaction module. In *Structure*. https://doi.org/10.1016/j.str.2012.01.006

Zeytuni, N., & Zarivach, R. (2012b). Structural and Functional Discussion of the Tetra-Trico-Peptide Repeat, a Protein Interaction Module. *Structure*, *20*(3), 397–405. https://doi.org/10.1016/j.str.2012.01.006

Zhang, H., Rigo, F., & Martinson, H. G. (2015). Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Molecular Cell*, *59*(3), 437–448. https://doi.org/10.1016/j.molcel.2015.06.008

Zhang, Y., Sun, Y., Shi, Y., Walz, T., & Tong, L. (2019). Structural Insights into the Human Pre-mRNA 3′-End Processing Machinery. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2019.11.005

Zhang, Z., Yang, J., Kong, E. H., Chao, W. C. H., Morris, E. P., da Fonseca, P. C. a, & Barford, D. (2013). Recombinant expression, reconstitution and structure of human anaphase-

promoting complex (APC/C). *The Biochemical Journal*, *449*(2), 365–371. https://doi.org/10.1042/BJ20121374

Zhou, M., & Law, J. A. (2015). RNA Pol IV and V in gene silencing: Rebel polymerases evolving away from Pol II's rules. *Current Opinion in Plant Biology*, *27*(5), 154–164. https://doi.org/10.1016/j.pbi.2015.07.005