Structural Studies of

the Integrator Complex — pre-UsnRNA 3'-end Processing Machinery

Yixuan Wu

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

© 2018
Yixuan Wu
All rights reserved

ABSTRACT

Structural Studies of the Integrator Complex -- pre-UsnRNA 3'-end Processing Machinery

Yixuan Wu

The Integrator complex (INT) is a metazoan-specific group of proteins associated with RNA polymerase II (Pol II) that has important functions in the 3'-end processing of noncodng RNAs, including uridine-rich small nuclear RNA (UsnRNA) and enhancer RNA (eRNA). Recently, INT has also been reported to be involved in Pol II transcriptional regulation of protein-encoding genes. INT contains at least 14 subunits, but the function of each subunit is difficult to predicted, because most subunits lack identifiable domains and display little similarity with other proteins. The endonuclease activity of INT is carried out by its subunit 11 (IntS11), which belongs to the metallo-β-lactamase superfamily and is a paralog of CPSF-73, the endonuclease for pre-mRNA 3'-end processing. IntS11 forms a stable complex with INT subunit 9 (IntS9) through their C-terminal domains (CTDs). This dissertation describes the crystal structure of the IntS9-IntS11 CTD complex at 2.1-Å resolution and summaries the structure-based biochemical and functional studies. The complex is composed of a continuous nine-stranded β-sheet with four strands from IntS9 CTD and five from IntS11 CTD. Highly conserved residues are located in the interface between the two CTDs. The structural observations on the complex are confirmed by yeast two-hybrid assays and coimmunoprecipitation experiments. Functional studies demonstrate that the Int9-IntS11 interaction is crucial for proper INT function in snRNA 3'-end processing.

The dissertation also presents the structural studies of a newly found mammalian mRNA deNADding enzyme, Nudt12. We determined the crystal structure of mouse Nudt12 in complex

with the deNADding product AMP and three Mg^{2+} ions at 1.6-Å resolution. The structure provides exquisite insights into the molecular basis of the deNADding activity within the NAD pyrophosphate. Previous studies have reported that NAD-capped mRNAs in mammalian cells are hydrolyzed by the DXO deNADding enzyme. Together with biochemical and functional studies, we demonstrate that Nudt12 is a second mammalian deNADding enzyme structurally and mechanistically distinct from DXO and targets different RNAs.

TABLE OF CONTENTS

LIST of TABLES AND FIGURES

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Liang Tong for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my PhD study.

Besides my advisor, I would like to thank Prof. Lawrence Chasin and Prof. Alexander Tzagoloff for being my committee and for their insightful comments and encouragement. I would like to thank Prof. Qing Fan and Prof. Marko Jovanovic for serving on my dissertation committee.

My sincere thanks also goes to my undergraduate advisors Prof. Yigong Shi and Prof. Nieng Yan, who invited me into the world of structural biology. Without their precious support I would not be able to come to Columbia University and pursue a PhD in structural biology.

I would like to thank my fellow labmates for creating a productive and friendly lab environment. I want to thank Dazhi Tan, Shukun Luo, Vivien Wang, Timothy Tran, Kehui Xiang, Ashley Jurado, and Jia Wei for helping me accommodate to the lab and for giving me a lot of suggestions on my research.

I would like to thank my friends. Without them, my graduate life would be less interesting. I want to thank Yinglu Zhang, Yadong Sun and Yi Wang for the lovely lunch time and game time in the past few years. I want to thank Chen Chen for being my classmate for 10 years, from college to graduate school. We watched a lot of movies, shows, exhibitions together. You enriched my life in New York. I want to thank Yaqiong Chen who has been my roommate since college. Thank you for taking care of Kris when I was on vacations. I want to thank my

best friend Dong Li for coming together with me to Columbia University and for sharing my happiness and sadness all the time.

Last but not the least, I would like to thank my parents. Although you are far away back in China, you always support me and love me without conditions. Thank you for encouraging me in all my pursuits. I always knew that you believed in me and wanted the best for me.

DEDICATION

This PhD research Thesis is dedicated to all members of my family, especially to my father, Yan Wu; my mother, Hua Guan.

PREFACE

This dissertation summarizes the projects I have researched over the course of my graduate study. The bulk of my work was studying the Integrator complex, which is involved in non-coding RNA 3'-end processing and RNA polymerase II transcriptional regulation. I also describe my work on structural determination of a mammalian deNADding enzyme, Nudt12.

CHAPTER ONE:

Integrator complex – Pre-UsnRNA 3'-end Processing Machinery

Section 1.1: Overview of Transcription of UsnRNA

In humans, protein-coding genes only occupy a small fraction of the genome, yet it is thought that more than 40% of the genome can be transcribed into RNAs (Cheng et al., 2005). The untranslated RNAs, also known as non-coding RNAs (ncRNAs), are highly abundant in cells and some are functionally important. The Uridine-rich small nuclear RNAs (UsnRNAs, snRNAs) comprise a small group of functional ncRNAs in the nucleoplasm. Except for U6 and U6atac small nuclear RNAs that are transcribed by RNA polymerase I or RNA polymerase II, the remaining snRNAs are all transcribed by RNA polymerase II (Pol II) (Jawdekar & Henry, 2008; Matera & Wang, 2014). UsnRNAs usually associate with proteins to form small nuclear ribonucleoprotein particles (snRNPs) (Matera, Terns, & Terns, 2007; Matera & Wang, 2014). U1, U2, U4/U6, U5 snRNAs are required for removal of introns in protein-coding mRNAs (Lee & Rio, 2015; Matera & Wang, 2014). U7 snRNA is required for 3'-end formation of replication-dependent histone mRNA (Romeo & Schumperli, 2016) and U3 snRNA is required for processing of rRNA (Henras, Plisson-Chastang, O'Donohue, Chakraborty, & Gleizes, 2015).

Unlike mRNA, Pol II transcribed snRNA has a relatively simple gene structure (Figure 1). It has a TATA-less promoter which comprises an enhancer-like DSE (distal sequence element) and an essential snRNA gene specific PSE (proximal sequence element) (Chen & Wagner, 2010). The transcripts of snRNA have no introns and are non-polyadenylated. The 3'-end formation of pre-snRNA is mediated by snRNA-specific 3'-box to generate the mature snRNA (Egloff, O'Reilly, & Murphy, 2008).

To initiate snRNA transcription, transcription factors Oct1, Sp1, NF1 and Staf bind to sites in the DSE (Jawdekar & Henry, 2008). Oct-1 stabilizes binding of PTF (PSE-binding transcription factor)/SNAPc to the PSE through direct interaction (Murphy, 1997; Murphy, Yoon, Gerster, & Roeder, 1992). PTF helps to recruit the TBP (TATA-binding protein) and TAFs (TBP-associated factors) to the Pol II-transcribed snRNA genes to form the pre-

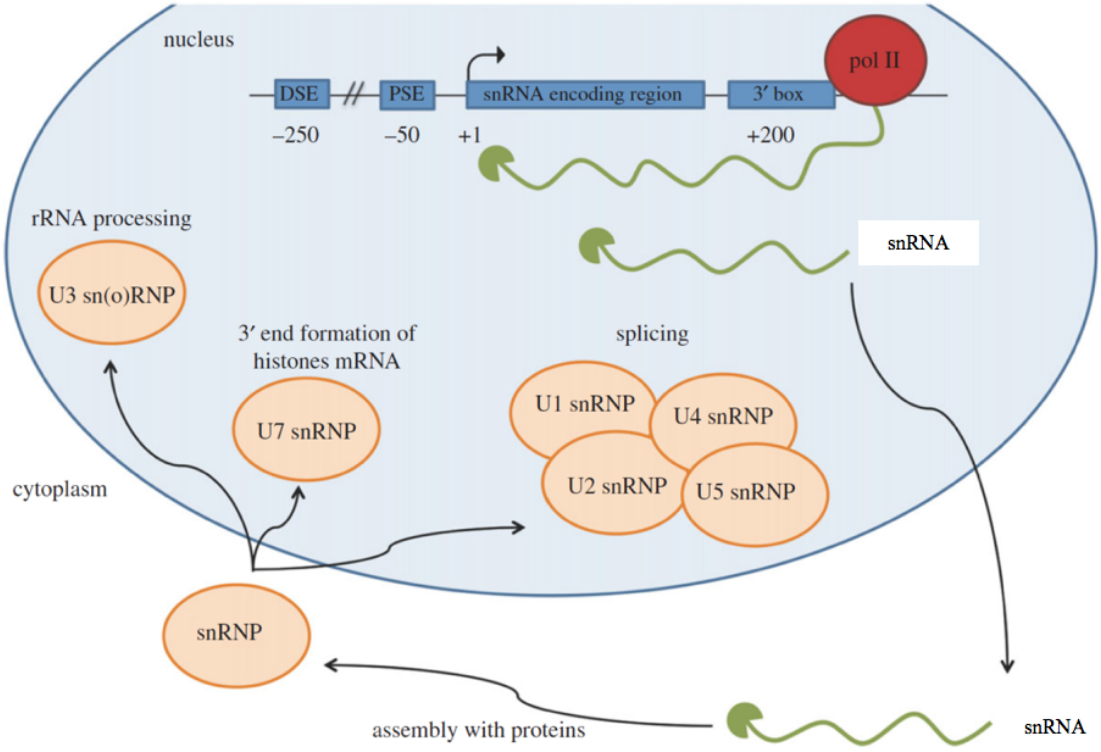


Figure 1 Expression of human snRNA genes. The arrow on the gene represents the start site of transcription and the numbers below the line indicate the position of the elements with respect to the transcription start site. The snRNA transcript is represented in green with the cap in the 5'-end. The pre-snRNA is matured by co-transcriptional capping and 3'-end trimming. The cleaved snRNA is exported into cytoplasm and assembles with the snRNP proteins. The functions of various snRNPs after reimport into the nucleus are noted. Modified from (Guiro & Murphy, 2017).

initiation complex (Zaborowska, Taylor, Roeder, & Murphy, 2012). The pre-initiation complex recruits Pol II to the snRNA gene promoter. The phosphorylation of the carboxyl terminal

domain (CTD) of Rpb1, the largest subunit of Pol II, plays important roles in snRNA expression. The CTD of Rpb1 comprises 52 repeats of the consensus heptapeptide $Y_1S_2P_3T_4S_5P_6S_7$. With different phosphorylation status of the heptad repeats, CTD regulates the recruitment of different transcription factors and snRNA processing factors (Figure 2) (Guiro & Murphy, 2017). Soon after initiation of transcription, the subunit of TFIIH, CDK7 (cyclin-dependent kinase 7) phosphorylates Ser5 and Ser7 (Akhtar et al., 2009). Ser7P mediate recruitment of RPAP2, which dephosphorylates Ser5P and recruits an incomplete Integrator complex including IntS1, IntS4, IntS5, IntS6 and IntS7 (Egloff, Zaborowska, Laitem, Kiss, & Murphy, 2012). The catalytic subunit IntS11 is recruited to activate 3'-box recognition and RNA 3'-end cleavage after Ser2 is phosphorylated by CDK9 kinase subunit of P-TEFb (positive-transcription elongation factor b) (Egloff et al., 2010). After 3'-end cleavage, the snRNA is exported to the cytoplasm through the activity of the snRNA-specific export factor PHAX (phosphorylated adapter for RNA export) (Ohno, Segref, Bachi, Wilm, & Mattaj, 2000) and then undergoes the snRNP biogenesis pathway (Matera et al., 2007).

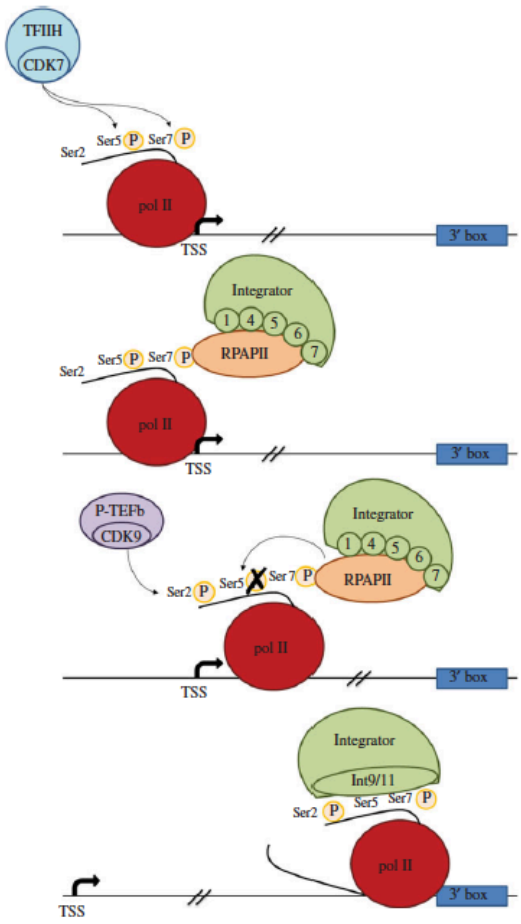


Figure 2 Pol II CTD phosphorylation events in snRNA gene transcription. From (Guiro & Murphy, 2017).

Section 1.2: The Integrator Complex in snRNA 3'-end Processing

The Integrator complex (INT) was discovered due to its binding affinity to the CTD of Rpb1 (Baillat et al., 2005). There were twelve subunits identified in INT from the original affinity purification (Baillat et al., 2005). The twelve INT subunits were named from IntS1 to IntS12 based on their molecular weights, where IntS1 is the largest subunit with a molecular mass of 244 kDa. and IntS12 is the smallest at 48 kDa (in humans) (Figure 3) (Baillat et al., 2005; Baillat & Wagner, 2015). Two additional INT subunits, C12orf11 (Asunder) and C15orf44 (also known as VWA9 or CG4785), were identified by subsequent proteomic analyses and were

termed IntS13 and IntS14 respectively (Figure 3) (Chen et al., 2012).

INT is present only in metazoans. The total molecular weight of 14 subunits is over one million Dalton, with eight out of fourteen subunits possessing molecular weights greater than 100 kDa. Most INT subunits have no identifiable paralogs within the human genome and lack recognizable domains that would suggest a role in RNA processing. The most common predicted motifs within INT subunits are alpha-helical repeats such as HEAT, ARM and TPR, or VWA domain (Chen & Wagner, 2010), which indicate potential protein-protein interaction surfaces. There are two exceptions, IntS11 and IntS9, which do share homology with CPSF-73 and CPSF-100 respectively (Baillat et al., 2005). CPSF-73 and CPSF-100 are involved in the cleavage of pre-mRNAs and replication-dependent histone pre-mRNAs and belong to a class of zinc-dependent nucleases called the β-CASP (CPSF, Artemis, SMN1/PSO2) family (Callebaut, Moshous, Mornon, & de Villartay, 2002b; Mandel et al., 2006) (Figure 4).

INT has been implicated in the 3'-end formation of snRNA since its first description (Baillat et al., 2005). In addition to its association with Pol II CTD, INT was shown to be present at the promoter, body and 3'-end of the snRNA by chromatin immunoprecipitation (ChIP) experiments (Baillat et al., 2005). Most importantly, RNAi-mediated knock-down of various INT subunits leads to the accumulation of misprocessed pre-UsnRNA (Baillat et al., 2005). A possible mechanism for snRNA 3'-end processing is that once INT is recruited to the promoter of snRNA and associates with Pol II CTD, the 3’-box is recognized by certain INT subunits so that the INT catalytic subunits can cleave the nascent pre-snRNA (Figure 5) (Guiro & Murphy, 2017). IntS11 is the presumable catalytic subunit in INT responsible for pre-snRNA cleavage since their high sequence similarity to CPSF-73. Overexpression of an IntS11 catalytic mutant interfered with snRNA 3'-end processing (Baillat et al., 2005), indicating the importance of its

endonuclease activity in snRNA 3'-end processing. ChIP analysis of U2 snRNA shows that IntS11 is mainly associated with the 3'-box, whereas IntS5 is found associated with both the snRNA gene promoter region and the 3'-box (Egloff et al., 2012). Therefore, the INT complex may assemble in a stepwise manner, with some subunits playing a role early in transcription initiation and elongation, while others are later recruited for snRNA 3'-end processing.

Besides the function in UsnRNA 3'-end cleavage, INT was also shown to play a role in UsnRNA gene transcription termination. Knockdown of IntS11 or IntS9 results in an increased level of Pol II occupancy downstream of the 3'-end cleavage site, demonstrating that snRNA 3'-end processing is linked to transcription termination (O'Reilly et al., 2014). However, the mechanism by which Pol II is released remains to be determined.

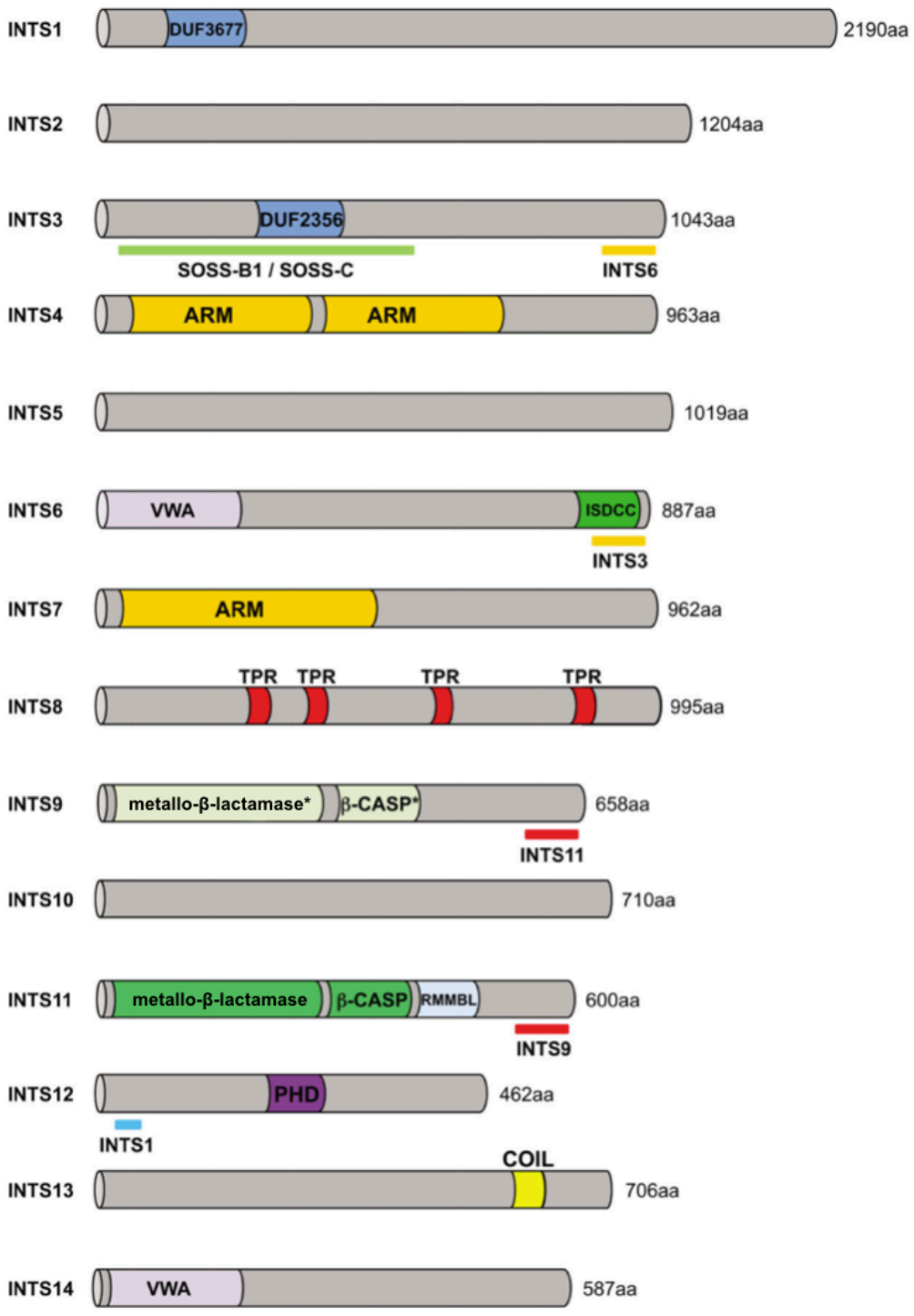


Figure 3 Integrator subunit domain schematic. Predicted protein domains of 14 Integrator subunits are illustrated. The length indicated are from human orthologues (in amino acids, aa). Modified from (Baillat & Wagner, 2015)

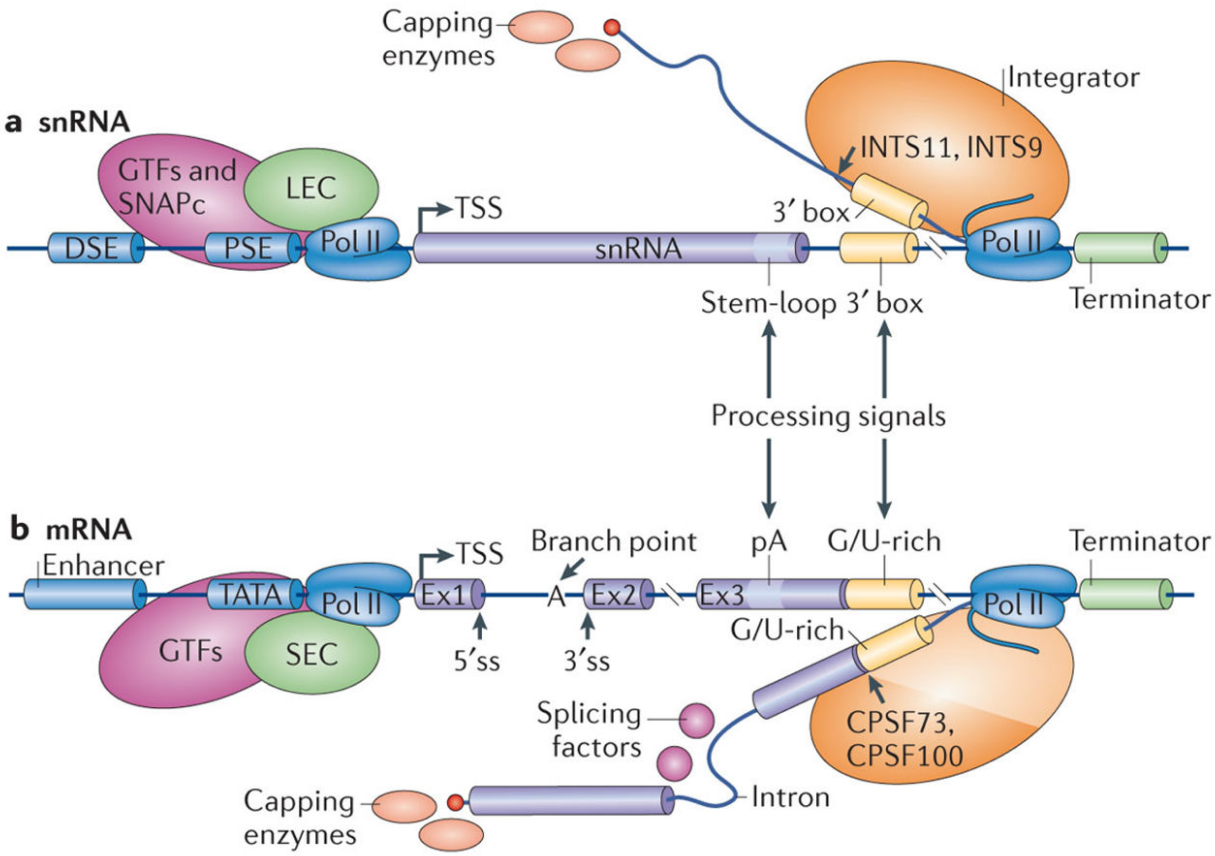


Figure 4. Comparison of transcription and processing of (a) snRNA and (b) mRNA. The INT is required for recognition of snRNA downstream processing signals, including 3'-box. Two of its subunits, IntS11 and IntS9, share sequence similarity to the mRNA 3'-end processing factors CPSF-73 and CPSF-100. For both snRNAs and mRNAs, 5'-end capping and 3'-end cleavage are thought to occur co-transcriptionally. From (Matera & Wang, 2014).

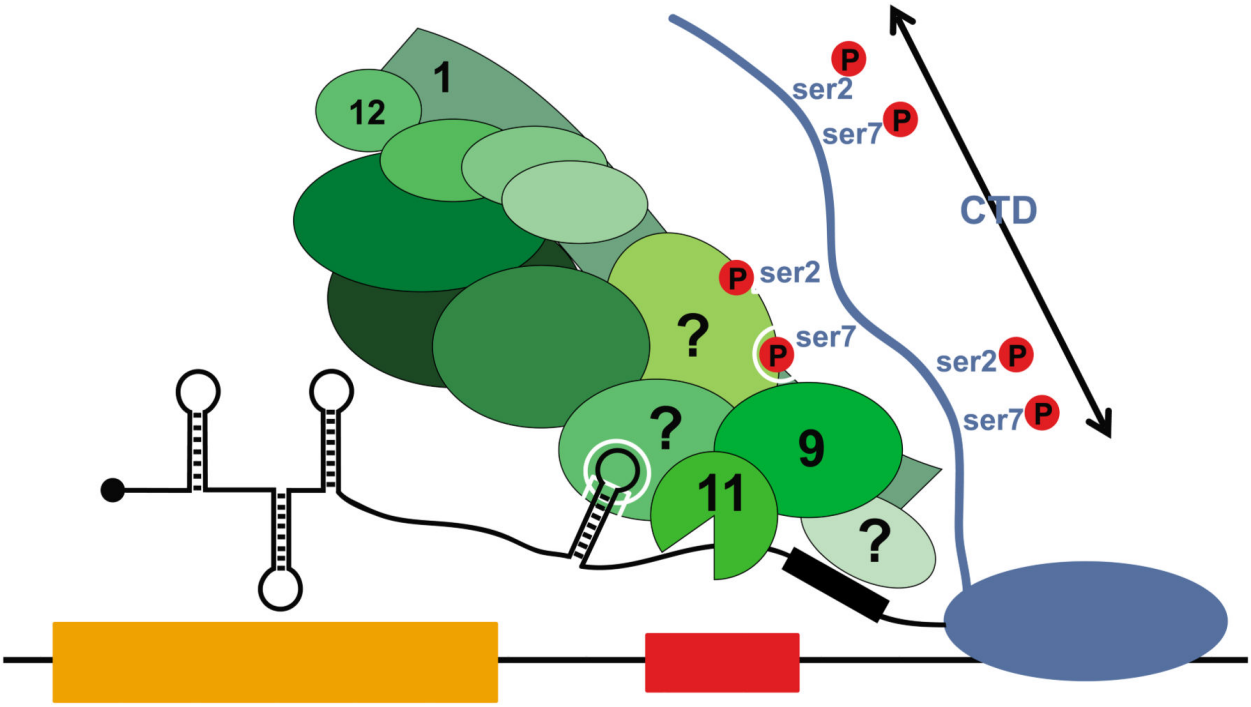


Figure 5 Model of INT function in UsnRNA processing (Guiro & Murphy, 2017). INT is recruited early in the UsnRNA transcription cycle and associates with Pol II CTD through recognition of the ser7P/ser2P dyad. Upon transcription, the UsnRNA terminal stem loop and 3'-box are recognized by unknown factors. The heterodimeric cleavage factor IntS9/IntS11 then carries out UsnRNA cleavage.

Section 1.3 Role of Integrator Complex in Pol II Transcriptional Pause-Release

In addition to its function in UsnRNA biosynthesis, INT has been recently shown to play a critical role in the activation of protein-coding genes, especially in the Pol II pause-release and elongation (Gardini et al., 2014; Skaar et al., 2015; Stadelmayer et al., 2014). Pol II pausing is an important step in controlling gene expression and regulating biological processes. It occurs late during the initiation stage of transcription, beginning with the recruitment of Pol II and basal transcription factors to the promoter to form the pre-initiation complex (PIC) (Luse, 2013). Pol II escapes the promoter after transcribing 9-10 nt nascent RNA, and serine 5 residues of Rpb1 CTD heptad repeats are phosphorylated. Pol II may pause when the transcript is 20-50 nt long, and the

productive elongation is activated only after pause release (Zhou, Li, & Price, 2012). Pol II pausing occurs prevalently both at highly transcribed genes and at those genes that are transcribed at very low levels and are poised for activation (e.g. heat shock genes) (Gaertner & Zeitlinger, 2014). The paused PoI II is stabilized by promoter-associated transcription factors (TFs) together with negative elongation factor (NELF) and DRB-sensitivity-inducing factor (DSIF) (Yamaguchi, Shibata, & Handa, 2013). Release of paused Pol II requires the kinase activity of the P-TEFb complex, which is composed of the cyclin T1 and cyclin-dependent kinase 9 (CDK9) (Yamaguchi et al., 2013; Zhou et al., 2012). Once P-TEFb is recruited to the promoter by specific TFs and cofactors, it phosphorylates the repressive DSIF/NELF complex, causing NELF to dissociate from Pol II and transforming DSIF into a positive elongation factor (Yamaguchi et al., 2013; Zhou et al., 2012). P-TEFb also phosphorylates the CTD of Pol II at Ser2 to create a platform for binding RNA processing and chromatin modifying factors that facilitate productive RNA synthesis (Peterlin & Price, 2006). The detailed mechanism of P-TEFb recruitment to transcriptionally active genes is not clear. Recent biochemical studies demonstrated a large multi-subunit complex termed Super Elongation Complex (SEC) which contains an active P-TEFb as well as other elongation factors (Lin et al., 2010; Luo, Lin, & Shilatifard, 2012). The discovery of SEC suggests the involvement of additional factors in Pol II pause-release.

Surprisingly, INT has been recently found to play an important role in the regulation of transcription initiation as well as Pol II pause-release at immediate early genes (IEGs) following transcriptional activation by epidermal growth factor (EGF) in human cells (Gardini et al., 2014). IEGs are known to be regulated through Pol II pause-release. Under starvation conditions, low levels of INT were detected using antibodies against IntS1, IntS11, and IntS9 at IEG

transcription start sites (TSSs); however, after EGF stimulation, INT occupancy markedly increased at the TSS and within the body of the gene. Depletion of IntS1 and IntS11 through short hairpin RNA (shRNA) abrogated responsiveness of IEGs to EGF stimulation. Moreover, knockdown of Integrator subunits caused failure of Pol II to escape pausing and progress into productive elongation. Further ChIP-seq analysis showed that depletion of INT also abolished the stimulus-dependent recruitment of two SEC components, ELL2 and AFF4, to IEGs. Collectively, this study indicates that INT is also recruited to protein-coding genes and plays a critical role in transcription initiation and Pol II pause-release by association with the SEC complex (Figure 6) (Baillat & Wagner, 2015).

Another study revealed the function of INT in regulating Pol II pause-release by interacting with NELF (Stadelmayer et al., 2014). Both biochemical and functional assays showed that INT subunits specifically controlled NELF-mediated Pol II pause-release at coding genes. Surprisingly, IntS3 and IntS11 had the opposite effect on NELF-mediated Pol II pausing. Genes containing NELF and IntS3 correlates with low Pol II density at the TSS and increased Pol II in gene bodies. In complete agreement, IntS3 knockdown reduces Pol II occupancy over gene bodies. By contrast, IntS11 depletion increases Pol II occupancy and RNA-seq read density over gene bodies but not over termination sites, resulting in defects in Pol II processivity and RNA processing. In this study, INT-mediated regulation of coding genes is restricted to NELF-target genes. Moreover, the data suggest that the catalytic activity of IntS11 may play a role in regulating Pol II pause-release and Pol II processivity. Interestingly, in INT target coding genes, a 3'-box sequence was found close to the termination sites. The presence of the 3'-box was associated with a decrease in RNA levels over termination sites in IntS11-depleted cells and suggested a functional contribution of the 3'-box in IntS11-mediated regulation of RNA

processing. Therefore, INT may play an important role in coupling the Pol II pause-release to mRNA processing.

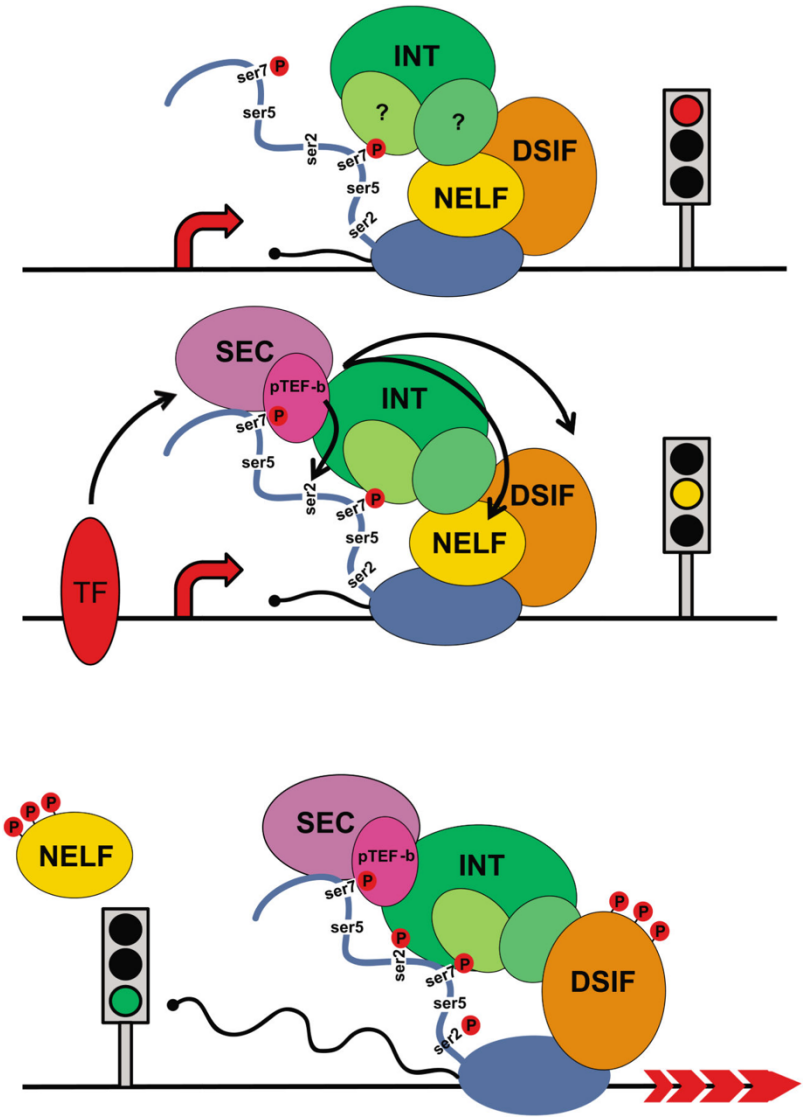


Figure 6 INT role in Pol II pause-release (Baillat & Wagner, 2015). Top, under starvation conditions, Pol II starts the transcription and pauses 40-60 nt downstream of the TTS. INT together with the NELF and DSIF is associated with Pol II CTD through phosphorylation state recognition. Middle, upon stimulation, INT is further enriched at the pause site and recruits P-TEFb and SEC. Bottom, P-TEFb phosphorylates the Pol II CTD Ser2, DSIF and NELF. NELF dissociates from the complex, and DSIF becomes a positive regulator of elongation. Pol II enters the stage of productive elongation.

Section 1.4 Integrator mediates enhancer RNA (eRNA) biogenesis

Recently, INT was shown to mediate the biogenesis of enhancer RNA (eRNA) as well (Lai, Gardini, Zhang, & Shiekhattar, 2015). Enhancers are distal regulatory elements which are involved in tissue- and temporal-specific regulation of gene expression in metazoans. Enhancers regulate transcription of particular genes allowing cell-type and cell-state specificities of gene expression (Pennacchio, Bickmore, Dean, Nobrega, & Bejerano, 2013). Enhancer RNAs (eRNAs) are transcripts derived from enhancers and are suggested to play a role in transcriptional regulation both in *cis* and in *trans* (Pennacchio et al., 2013).

Lai et al. revealed that INT was recruited to enhancers and super-enhancers in a signal-dependent manner using starving HeLa cells (Lai et al., 2015). After starving for 48 h, HeLa cells were stimulated with EGF to induce IEGs. INT occupancy over enhancers were measured using antibodies against IntS1, IntS9 and IntS11 before and after EGF stimulation. Before EGF induction, enhancers were occupied by a detectable amount of INT. After addition of EGF, INT was further recruited to enhancers. Depletion of IntS11 and IntS1 using shRNAs reduced the production of eRNA after EGF stimulation and a significant decrement in level of EGF-responsive protein-coding gene transcripts in proximity of these EGF-induced enhancers was also observed. Stimulus-induced enhancer-promoter chromatin looping were also abrogated upon INT depletion, thus inhibiting enhancer and promoter communication. After INT depletion, eRNAs were found to remain bound to Pol II and their primary transcripts accumulated. This observation suggests a block in 3'-end cleavage of primary eRNA transcripts and a defect in termination. Indeed, IntS11 catalytic activity was demonstrated to be required to regulate eRNA induction since the catalytic mutant of IntS11 failed to rescue the EGF-induced eRNA levels

following endogenous IntS11 depletion. These results confirmed the important role of INT cleavage activity in eRNA transcription termination.

Interestingly, eRNA has been reported to facilitate the transition of paused Pol II into productive elongation by acting as a decoy for NELF complex upon induction of IEGs (Schaukowitch et al., 2014). This model could link the diverse functions of INT during transcription. When at promoter, INT regulates pause-release factors, leading to modulation of productive transcriptional elongation. When at enhancers, INT governs eRNA maturation and enhancer-promoter communications.

Section 1.5 The Pathophysiology of Integrator

As mentioned above, INT is a multifunctional complex that regulates Pol II-mediated transcription at different gene classes. The improper function of INT is expected to alter diverse cellular pathways and biological processes. Indeed, several INT subunits have been found to play important roles in DNA damage response, adipose tissue differentiation, embryonic development, ciliogenesis, lung homeostasis, the Cajal body (CB), and herpesvirus microRNA 3'-end formation (Hata & Nakayama, 2007; Jodoin et al., 2013; Kapp, Abrams, Marlow, & Mullins, 2013; Kheirallah, de Moor, Faiz, Sayers, & Hall, 2017; Otani et al., 2013; Takata, Nishijima, Maeshima, & Shibahara, 2012; Tao, Cai, & Sampath, 2009; Zhang, Ma, & Yu, 2013). Dysfunction of INT may also lead to various pathologies in humans. Genetic and molecular studies have detected the presence of mutations or altered expression levels in genes coding for INT subunits in several diseases, including malignancies (Rienzo & Casamassimi, 2016). Table 1 summaries some identified pathophysiological roles of different INT subunits.

Table 1 Human INT subunits and their possible pathophysiological roles (Rienzo & Casamassimi, 2016).

Name	Recognizable domains	Pathophysiological role
IntS1	DUF3677	Development at early blastocyst stage Hematopoietic development
IntS2	—	Gastric cancer peritoneal carcinomatosis
IntS3	DUF2356	DNA damage response and maintenance of genome stability Overexpressed in hepatocellular carcinoma
IntS4	ARM, HEAT	Development
IntS5	—	Hematopoietic development Postmenopausal osteoporosis
IntS6	VWA, ISDCC	Adipogenesis, Overgrowth, DNA damage response Pathogenesis of hepatocellular carcinoma (with its pseudogene INTS6P1) Childhood B-precursor acute lymphoblastic leukemia
IntS7	ARM	DNA damage response Development
IntS8	TRP	Putative biomarker for gastric cancer Mutated in peripheral T cell lymphoma
IntS9	metallo-β-lactamase, β-CASP	Possible role in malignancies
IntS10	—	Nicotine dependence (GWAS) Childhood B-precursor acute lymphoblastic leukemia
IntS11	metallo-β-lactamase, β-CASP	Adipogenesis Hematopoietic development Possible role in malignancies
IntS12	PHD	Lung function and pulmonary diseases
IntS13	COIL	Spermatogenesis and oogenesis Cell division
IntS14	VWA	Elevated expression in SV40-immortalized cells, cancer cells, and NSCLC tissues

Section 1.6 Integrator Cleavage Module

Although INT was originally purified as a single entity, ChIP experiments indicate that INT appears to act as a modular complex on the genome. Different INT subunits give distinct occupancy patterns on UsnRNA genes in human or on the HSP70 gene in fly. On human U2

snRNA, IntS5 shows predominant occupancy from the promoter to the 3'-box, whereas IntS11 is mainly associated with the 3'-box (Egloff et al., 2012). For *Drosophila* HSP70 gene, IntS12 is present at the promoter and peaks at the transcriptional pausing site while IntS9 is shifted toward the 3'-end of the gene with a marked peak in the gene body (Gardini et al., 2014). These observations would indicate a modular complex of INT that assembles in a stepwise manner with different functions during the transcription cycle. A module containing IntS5 or IntS12 could be recruited early with a role in transcription initiation and pausing, while IntS9 and IntS11 are recruited later to form a module that functions in elongation and 3'-end processing.

While much progress has been made since the discovery of INT, many issues still remain unsolved. Little is known about the putative existence of modules and sub-complexes within the complex. Also, the interactions between the identified subunits, as well as between INT subunits and other protein factors involved in transcription regulation need to be further investigated. Moreover, it is very difficult to predict the function and the mechanism of the majority of INT subunits, because most subunits share little sequence homology with other proteins and have few recognizable domains, and the structural information on these subunits is very limited.

As mentioned in Section 1.2, IntS11 and IntS9 are paralogues of the cleavage and polyadenylation specificity factors CPSF-73 and CPSF-100. CPSF-73 carries out the endonuleolytic cleavage for the processing of either pre-mRNA or replication-dependent histone pre-mRNA (Dominski, Yang, & Marzluff, 2005; Mandel et al., 2006). It is known that CPSF-73 and CPSF-100 together with a scaffold protein called Symplekin form a complex named Core Cleavage Complex (CCC) (Michalski & Steiniger, 2015; Sullivan, Steiniger, & Marzluff, 2009) or mCF (mammalian cleavage factor) (Shi & Manley, 2015). These three proteins then interact with other factors in the processing machinery to catalyze cleavage of the pre-mRNA. Depletion

of Symplekin from *Drosophila* S2 cells leads to extreme misprocessing of histone pre-mRNA, which is similar to depletion of CPSF-73 and CPSF-100 (Sullivan et al., 2009; Wagner et al., 2007).

In the context of INT, no candidate 'Symplekin-like' subunit could be readily identified, because there are no other INT subunits sharing homology with members of the cleavage and polyadenylation machinery except for IntS9 and IntS11. A recent study using modified yeast two-hybrid screen determined that IntS4 binds to IntS9-IntS11 heterodimer, indicating a possible 'Symplekin-like' role of IntS4 (T. R. Albrecht et al., 2018). Although there is no significant homology detected between Symplekin and IntS4, there are some similarities in the two proteins. IntS4 has a similar size to Symplekin. Both proteins possess HEAT repeats. The structure prediction of IntS4 identified an array of N-terminal HEAT repeats over the first 800 amino acids. The remaining 200 amino acids at the C-terminal end are predicted to be rich in β-strands. IntS4 was found to bind to IntS9 and IntS11 only when they form the IntS9-IntS11 heterodimer through their CTDs. Similar to depletion of IntS11 or IntS9, depletion of IntS4 also leads to severe UsnRNA misprocssing. Consistently, siRNA knockdown of these subunits in HeLa cells causes complete disassembly of Cajal bodies and histone locus bodies (T. R. Albrecht et al., 2018; Takata et al., 2012). Cajal bodies and histone locus bodies are two major nuclear bodies that are primarily nucleated around snRNA and clustered replication-dependent histone genes and are involved in snRNA biogenesis and histone mRNA production (T. R. Albrecht et al., 2018). Altogether, these data propose a model that IntS4, IntS9, and IntS11 constitute a minimal Integrator cleavage module.

Section 1.7 Conclusion

Since its discovery, INT has been added to the Pol II-mediated transcription machinery as one of the major components. INT plays roles in transcriptional initiation, stimulus-dependent Pol II pause-release, and termination of various gene classes. The multiple functions could be explained by the existence of different INT sub-complexes or modules, which could be sequentially recruited during different transcriptional stages; alternatively, subunits within INT could adopt major conformational changes when involved in different processes of the transcription cycle. However, future evidences are needed to prove these hypotheses.

Here, we performed structural studies of several INT subunits that are indicated to be essential for INT functions in various processes. **Chapter 2** presents the crystal structure of human IntS9-IntS11 CTD complex, which uncovers an extensive molecular interface that could allow recognition by other subunits of INT. Based on structural observations, biochemical and functional studies were conducted, demonstrating the IntS9-IntS11 CTD interaction is functional important.

In **Chapter 3**, I will present my work on structure determination of Nudt12, a novel mammalian deNADding enzyme. The redox cofactor nicotinamide adenine dinucleotide (NAD) was recently reported to be covalently linked to the 5'-end of eukaryotic mRNAs (Jiao et al., 2017). The mammalian non-canonical decapping enzyme, DXO, possesses deNADding activity by removing the entire NAD moiety from the 5'-end of NAD-capped RNA in cells (Jiao et al., 2017). Our crystal structure of mouse Nudt12 in complex with the deNADding product AMP and three Mg^{2+} interprets the molecular basis of Nudt12 deNADding activity. The structural observations as well as results from biochemical and functional studies demonstrate that Nudt12 is a second mammalian deNADding enzyme structurally and mechanistically distinct from DXO.

CHAPTER TWO：

Structural and Functional Studies of Human IntS9-IntS11 CTD Complex

Section 2.1: Introduction

The Integrator complex (INT) is characterized as the machinery responsible for 3'-end processing of noncoding RNAs, including the uridine rich small nuclear RNA (UsnRNA) and enhancer RNA (eRNA), but its molecular mechanism of action is poorly understood. So far, a large amount of efforts has been made to understand the 3'-end processing of the other two classes of Pol II transcripts, poly(A)+ mRNA and the replication-dependent histone mRNA. Although the cis elements of these two kinds of RNAs are quite different and their respective 3'-end formation complexes comprise different protein factors, both poly(A)+ pre-mRNA and histone pre-mRNA are cleaved by a cleavage core consist of CPSF-73, CPSF-100 and Symplekin (Jurado, Tan, Jiao, Kiledjian, & Tong, 2014; Millevoi & Vagner, 2010; Romeo & Schumperli, 2016).

CPSF-73 is an RNA endonuclease and belongs to the metallo-β-lactamase superfamily (Mandel et al., 2006). The general architecture of metallo-β-lactamase proteins consists of an αβ/βα sandwich with two zinc ion binding sites located at the top of the β sandwich. The residues that coordinate Zn^{2+} ions are conserved among active metallo-β-lactamase proteins (Aravind, 1999). The active site of metallo-β-lactamase family proteins are formed by 5 signature motifs (motif 1 to 5) (Aravind, 1999). A unique feature for CPSF-73 and other nucleic acid endonucleases within metallo-β-lactamase family is that motif 5 is replaced with three additional motifs (motif A to C) and a motif called the β-CASP (CPSF, Artemis, SNM1/PSO2) is inserted within the metallo-β-lactamase domain (Callebaut, Moshous, Mornon, & de Villartay, 2002a).

The crystal structure of CPSF-73 catalytic segments including the metallo-β-lactamase domain and β-CASP motif demonstrated its endonuclease activity (Mandel et al., 2006). The inserted β-CASP motif forms a separate domain that is located above the metallo-β-lactamase domain, where both N- and C- terminal regions of the protein are involved in forming the metallo-β-lactamase domain (Figure 7) (Mandel et al., 2006). As shown in the structure, the active site is positioned in the interface between β-CASP motif and the metallo-β-lactamase domain which seems to be inaccessible to its substrates. Indeed, *in vitro* assay using standard conditions for 3'-end processing reactions showed weak endonuclease activity of CPSF-73, which was consistent with the observations from the structure (Mandel et al., 2006). This has led to the hypothesis that β-CASP endonuclease may require conformational changes to activate catalysis.

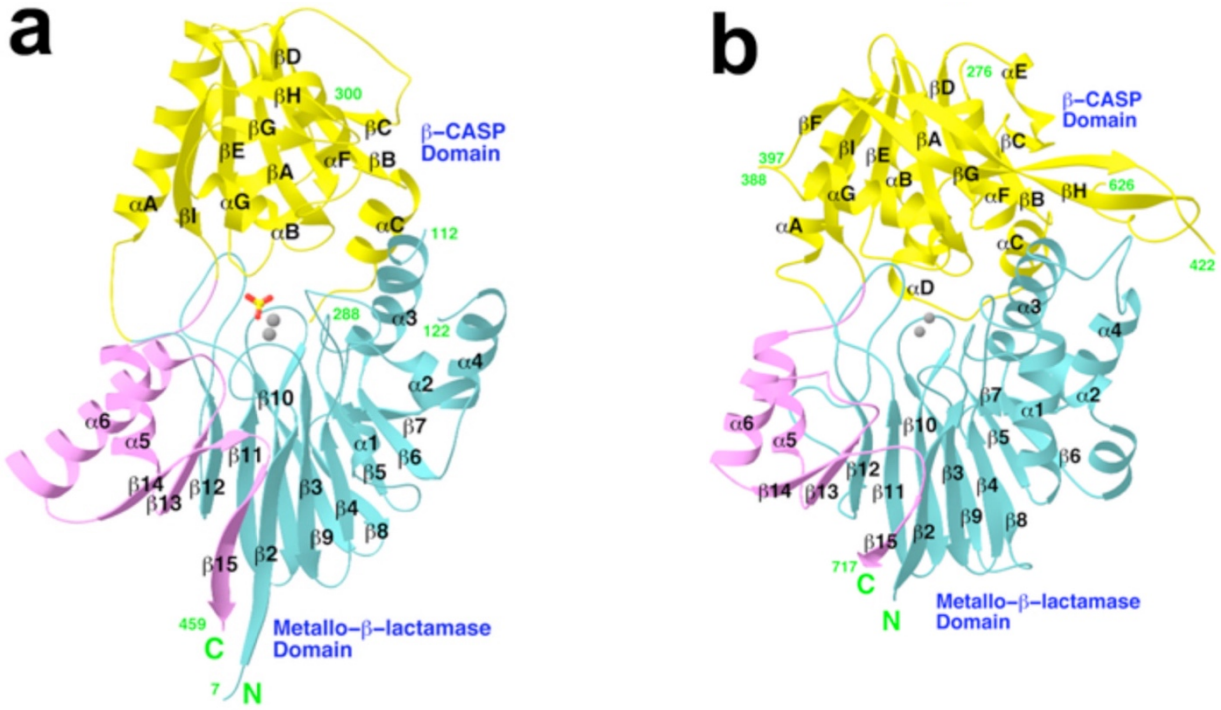


Figure 7 Structures of human CPSF-73 and yeast CPSF-100 (Ydh1p). a, Schematic representation of the structure of human CPSF-73. The β-strands and α-helices are labeled, and the two zinc atoms in the active site are shown as gray spheres. The sulfate ion is shown as a stick model. b, Schematic representation of the structure of yeast CPSF-100. The zinc atoms in the CPSF-73 structure are shown for reference. From (Mandel et al., 2006).

CPSF-100 is another cleavage and polyadenylation specificity factor in the pre-mRNA 3'-end cleavage core. CPSF-100 is thought to interact with CPSF-73 to form a heterodimer that is recruited to pre-mRNA substrates to elicit processing (Dominski, Yang, Purdy, Wagner, & Marzluff, 2005). CPSF-100 also belongs to the β-CASP family. However, CPSF-100 contains residue substitutions at the active site so that it seems to have no endonuclease activity. The overall structure of yeast CPSF-100 is similar to that of human CPSF-73 (Mandel et al., 2006). The function of the inactive CPSF-100 is unknown. One hypothesis is that it helps activate CPSF-73 by inducing conformational changes through heterodimerization.

As mentioned in Chapter 1, IntS11 and IntS9 are paralogs of CPSF-73 and CPSF-100 in INT. IntS11 and IntS9 are both members of β-CASP family (Baillat et al., 2005). IntS11 shares high homology with CPSF-73 within the N-terminal metallo-β-lactamase domain and the β-CASP domain (40% identity). The sequence conservation between IntS9 and CPSF-100 is weaker, and IntS9 has two insertions in the metallo-β-lactamase domain (Wu, Albrecht, Baillat, Wagner, & Tong, 2017). Similar to CPSF-100, IntS9 also has residue changes in the active site and is predicted to be inactive. The existence of inactive form of cleavage factor in both machineries indicates their functional importance for 3'-end cleavage.

Beside the metallo-β-lactamase and β-CASP domains, IntS9, IntS11, CPSF-73, and CPSF-100 all contain a C-terminal domain (CTD) (Figure 8). The sequence conservation of CTDs is much poorer. Previous studies showed that the CTDs of IntS9 and IntS11 are required and sufficient to mediate their heterodimerization (T. R. Albrecht & Wagner, 2012). The interaction between IntS9 and IntS11 is critical for snRNA 3'-end processing (T. R. Albrecht & Wagner, 2012). In order to understand the molecular basis of this association, we have determined the crystal structure of the IntS9-IntS11 CTD complex (Wu et al., 2017).

Additionally, we designed mutations based on the structure that disrupt the IntS9-IntS11 interaction. We demonstrated that disruption of IntS9-IntS11 interaction also abolishes U7 snRNA 3'-end processing.

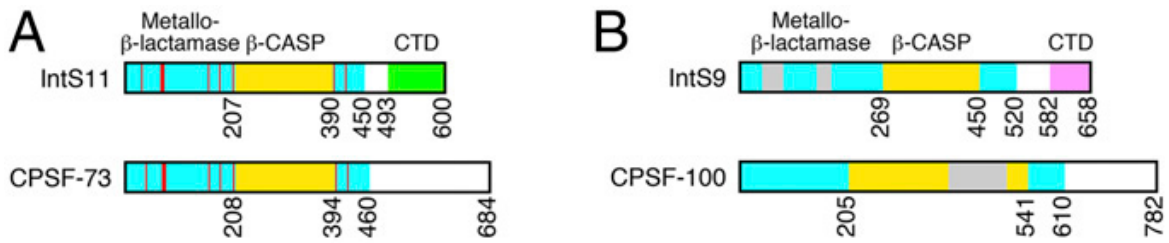


Figure 8 Domain organizations of (A) IntS11, CPSF-73, (B) IntS9 and CPSF-100. The metallo-β-lactamase domain is in cyan. The β-CASP motif is in yellow. Red lines indicate the conserved residues in the active site. From (Wu et al., 2017).

Section 2.2 Experimental Procedures

Protein Expression and Purification

The C-terminal domain of human IntS11 (residues 491–600) was subcloned into the pET28a vector (Novagen), which introduced an N-terminal His-tag. The C-terminal domain of human IntS9 (residues 582–658) was subcloned into the pCDFDuet vector (Novagen) without any affinity tag. The two proteins were coexpressed in *Escherichia coli* BL21Star (DE3) cells at 23 °C for 16–20 h. The cells were lysed by sonication in a buffer containing 20 mM Tris (pH 8.5), 200 mM NaCl, and 5% (vol/vol) glycerol. The IntS9–IntS11 heterodimer was purified by Ni-NTA (Qiagen) chromatography. The eluted protein was treated overnight with thrombin at 4 °C to remove the His-tag and was further purified by gel filtration chromatography (Sephacryl

S-300; GE Healthcare). The purified protein was concentrated to 30 mg/mL in a solution containing 20 mM Tris (pH 8.5), 200 mM NaCl, and 10 mM DTT before being flash-frozen in liquid nitrogen and stored at −80 °C.

Protein Crystallization

Crystals of the IntS9–IntS11 CTD complex were obtained at 20 °C using the sitting-drop vapor-diffusion method. The reservoir solution contained 0.1 M Bis-Tris (pH 6.5) and 21–24% (wt/vol) PEG 3350. The protein concentration was 10 mg/mL. Crystals took 2 wk to grow to full size. A heavy-atom derivative was prepared by soaking native crystals in the mother liquor with 1 mM HgCl for 3 h. All crystals were cryo-protected by the reservoir solution supplemented with 5% (vol/vol) ethylene glycol and were flash-frozen in liquid nitrogen for data collection at 100 K.

Data Collection and Processing

X-ray diffraction data of native and heavy-atom–derivative crystals were collected at a wavelength of 0.979 Å on an ADSC Q315R CCD at the 5.0.1 beamline of Advanced Light Source (ALS). The diffraction images were processed with the HKL program (Otwinowski & Minor, 1997). The crystals belonged to space group $P2_1$ with cell dimensions of a = 63.0 Å, b = 67.8 Å, c = 98.6 Å, and β = 100.6°. There are four copies of the IntS9–IntS11 complex in the crystallographic asymmetric unit. A native dataset was collected to 2.1-Å resolution, and the derivative dataset was collected to 2.3-Å resolution.

Structure Determination and Refinement

Four Hg atoms were located and used for phasing by the AutoSol routine in PHENIX (Adams et al., 2002b), using the single isomorphous replacement (SIR) method. Most of the protein residues were automatically built by the AutoBuild routine in PHENIX, and further manual building was carried out with the program Coot (Emsley & Cowtan, 2004b). The structure was refined using PHENIX. The crystallographic information is summarized in Table 2.

Table 2 Summary of crystallographic information

Structure	Ints9-Ints11 CTD Complex
Data Collection	
Space Group	$P2_1$
Cell Dimensions	
a, b, c (Å)	63.0, 67.8, 98.6
α, β, γ (°)	90, 100.6, 90
Resolution (Å)	25-2.1 (2.2-2.1)
R_{merge} (%)	7.8 (42.4)
I/σI	20.5 (3.5)
Completeness	99.6 (96.8)
Redundancy	6.8 (5.4)
Refinement	
Resolution (Å)	25-2.1
No. Reflections	47760
R_{work}/R_{free} (%)	16.7 (22.0)
No. Atoms	
Protein	5636
Ligand/Ion	0
Water	419
B-Factors	
Protein	38.3
Ligand/Ion	-
Water	39.2
R.M.S. Deviations	
Bond Lengths (Å)	0.012
Bond Angles (°)	1.3

The Numbers In Parentheses Are For The Highest Resolution Shell.

Section 2.3 Results and Discussion

Section 2.3.1 Constructs Design and Soluble Protein Expression

Previous studies have shown the binding site within IntS11 for IntS9 is approximately from residue 418-597 (*Drosophila)*, and the interacting region within IntS9 is between residue 565 and 658 (T. R. Albrecht & Wagner, 2012). The initial constructs for protein expression were made using *Drosophila* genes. For *Drosophila* IntS11 (dIntS11) CTD, constructs were designed with constant C-terminus while varying the N-terminus (Figure 9). For *Drosophila* IntS9 (dIntS9), constructs were designed with variant N- and C-termini (Figure 9). To obtain IntS9-IntS11 CTD complex, Int9 and IntS11 CTD genes were subcloned into pET28a/pET26b (with His-tag) or pCDFDuet (without His-tag). Different combination of dIntS11 and dIntS9 CTD constructs were achieved by co-transforming pET and pCDFDuet vectors into *E. coli* expression system.

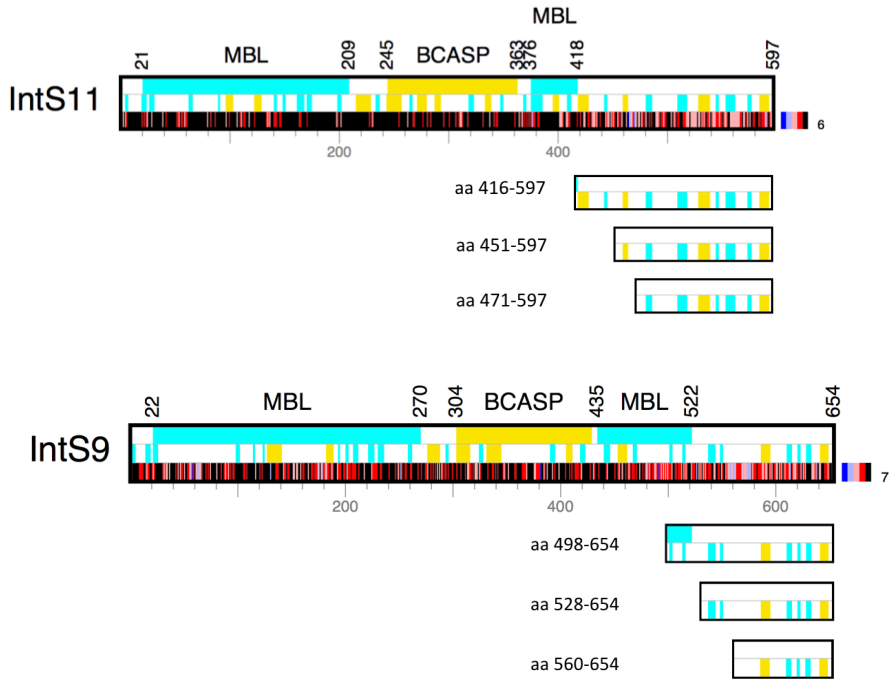
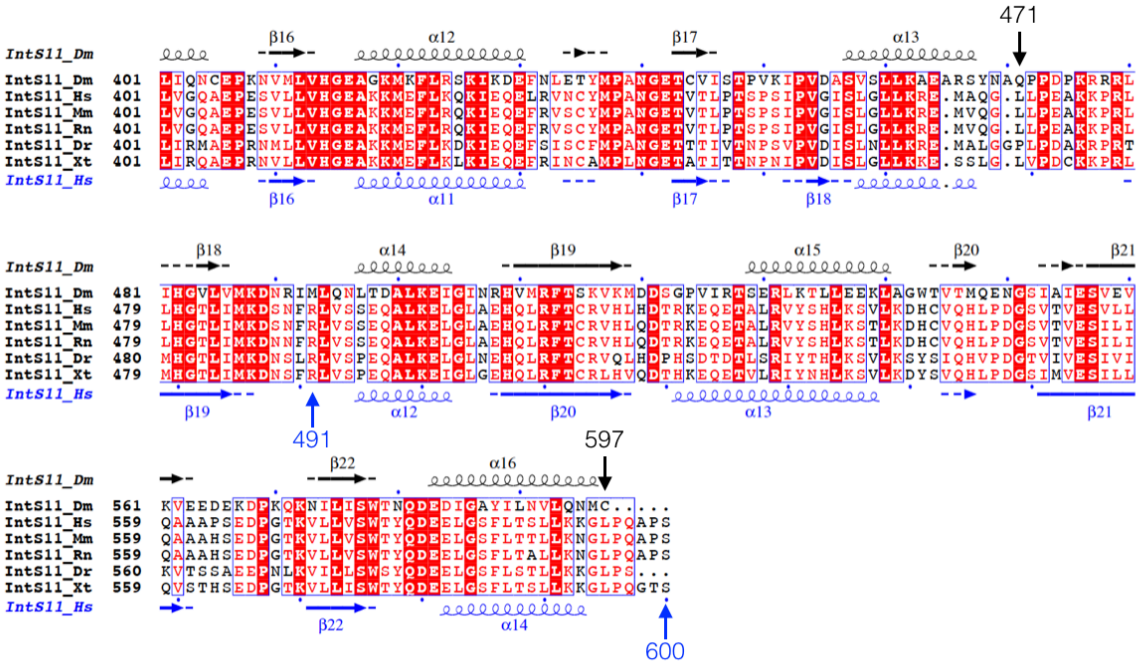


Figure 9 Constructs for dIntS11 and dIntS9 CTD. The domain organization, secondary structure prediction, and conservation information of IntS11 and IntS9 are shown in bars with bold border. The top

row indicates domain organization, where metallo-β-lactamase domain (MBL) is in cyan and β-CASP motif (BCASP) is in yellow. The middle row shows the secondary structure prediction. α-helix is represented in yellow and β-strand is in cyan. The bottom row shows the sequence conservation (IntS11: 6 orthologs; IntS9: 7 orthologs). The blue means the least conserved and the black means absolutely conserved. The constructs of each *Drosophila* protein are shown as six short bars.

After some expression trials, we found that dIntS9 CTD required His-tag to achieve optimal yield. The minimal region within dIntS11 for binding dIntS9 CTD is residue 471-597, and the smallest dIntS9 CTD construct that could pull down dIntS11 CTD contains residue 560 to 654 (Figure 10).

The constructs for human IntS11 (hIntS11) and human IntS9 (hIntS9) were designed based on the information from *Drosophila* protein expression. Residue 491-600 of hIntS11 and residue 582-658 of hIntS9 were two smallest constructs that we found to form heterodimer (Figure 10).



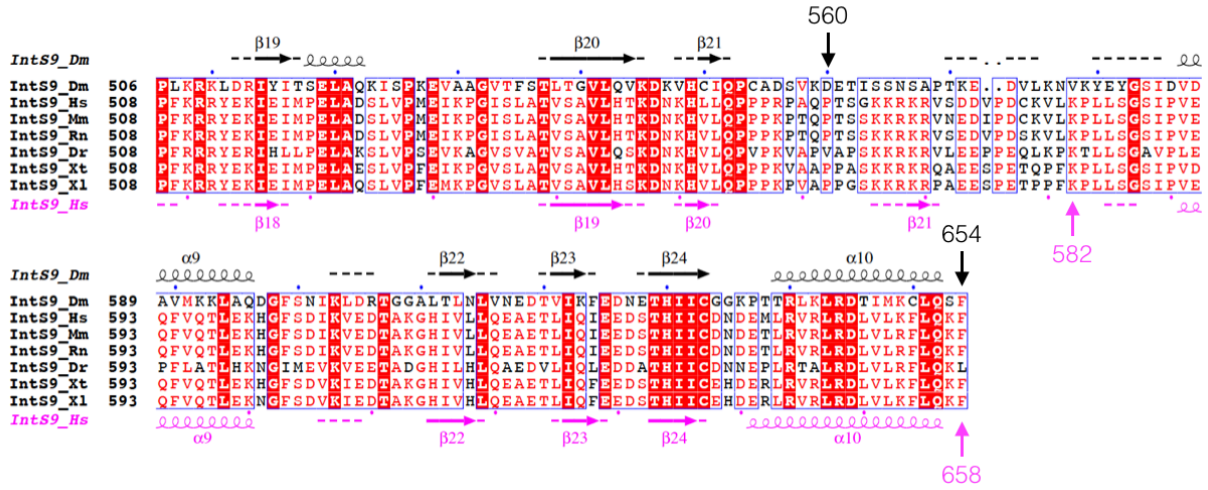


Figure 10 Sequence alignment of IntS11 CTD and IntS9 CTD. The predicted secondary structure elements in *Drosophila* IntS11 and IntS9 are shown in black in above the sequences, and the predicted secondary structure elements in human IntS11 and IntS9 are shown in blue and pink respectively below the sequences. The vertical arrows above or below the sequences indicate the minimal constructs for CTD interaction. Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Rn, *Rattus norvegicus*; Dr, *Danio rerio*; Xt, *Xenopus tropicalis*; Xl, *Xenopus laevlis*. Modified from an output from ESPript (Gouet, Courcelle, Stuart, & Metoz, 1999).

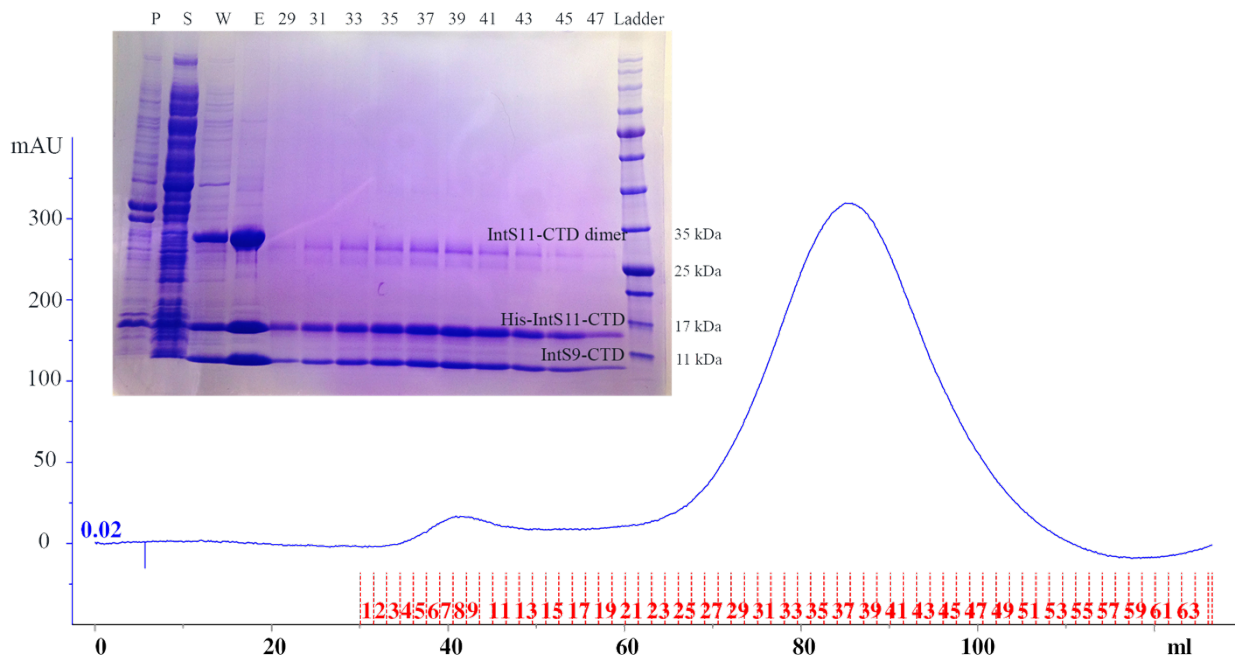


Figure 11 Size exclusion chromatography of human IntS9-IntS11 CTD complex. The eluted fractions were monitored for protein content using A280. The peak fractions were analyzed by SDS-PAGE and visualized by Coomassie blue staining (gel inset).

The human IntS9-IntS11 CTD complex was purified using Ni-NTA followed by Sephacryl S-300 gel filtration chromatography (Figure 11). Based on SDS-PAGE, the bottom two bands represent His-IntS11-CTD (13.6kDa) and IntS9-CTD (8.9kDa) respectively (Figure 11). There was a third band at around 27kDa enriched in protein eluate from Ni-NTA which we initially thought to be an impurity protein. But this band coeluted with the CTD complex and got much fainter during gel filtration. Later we confirmed this band to be the IntS11 CTD dimer band and was formed through an intermolecular disulfide bond. In the elution buffer, where no DTT existed, IntS11 CTD formed dimers through disulfide bonds. In gel filtration buffer, there

was 10mM DTT. The disulfide bonds were reduced so that the dimer band disappeared. Although the SDS-PAGE sample buffer contains the reducing agent β-mercaptoethanol (β-ME) which should have reduced the disulfide bond, the sample buffer I used was made years ago and the β-ME had decomposed.

Section 2.3.2 Crystal Screening and Optimization

The purified human IntS9-IntS11 CTD complex was screened for crystallization conditions with several commercial crystallization kits (Cudney, Patel, Weisgraber, & Newhouse, 1994; Jancarik, Scott, Milligan, Koshland, & Kim, 1991) by sitting drop vapor diffusion at 20 C°. Rectangular plate shaped crystals appeared in a condition containing 0.1M Tris (pH 7.0), 20% (w/v) PEG 1000 (Figure 12a). The crystal diffracted to 2.6-Å resolution with a space group $P2_1$. But the crystal was unable to be reproduced, so we could not solve the structure because of phase problem.

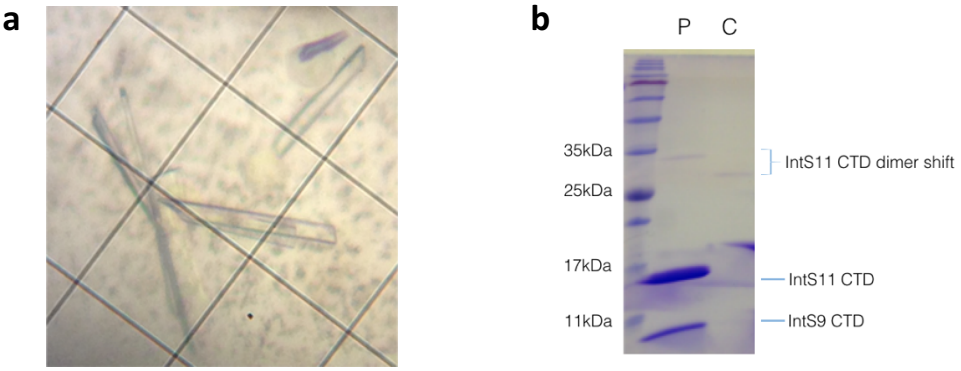


Figure 12 Initial crystals for human IntS9-IntS11 CTD complex. (a) Rectangular plate shaped crystals in the drop. (b) Lane P, purified protein solution sample for crystallization; Lane C, crystal sample, crystals were dissolved in loading buffer and loaded on SDS-PAGE. The blue brace shows the shift of IntS11 dimer.

In order to figure out why the protein could not crystallize, initial crystals were analyzed by SDS-PAGE (Figure 12b). By comparing the crystal sample with the original protein sample, we found that the IntS11 CTD dimer band shifted downward by 2~3kDa. We suspected that the protein might be digested by a contaminating protease in the contaminate during crystallization, and the degradation products were easier to crystallize. Based on this information, we performed limited proteolysis using three different proteases, trypsin, chymotrypsin and subtilisn (Figure 13).

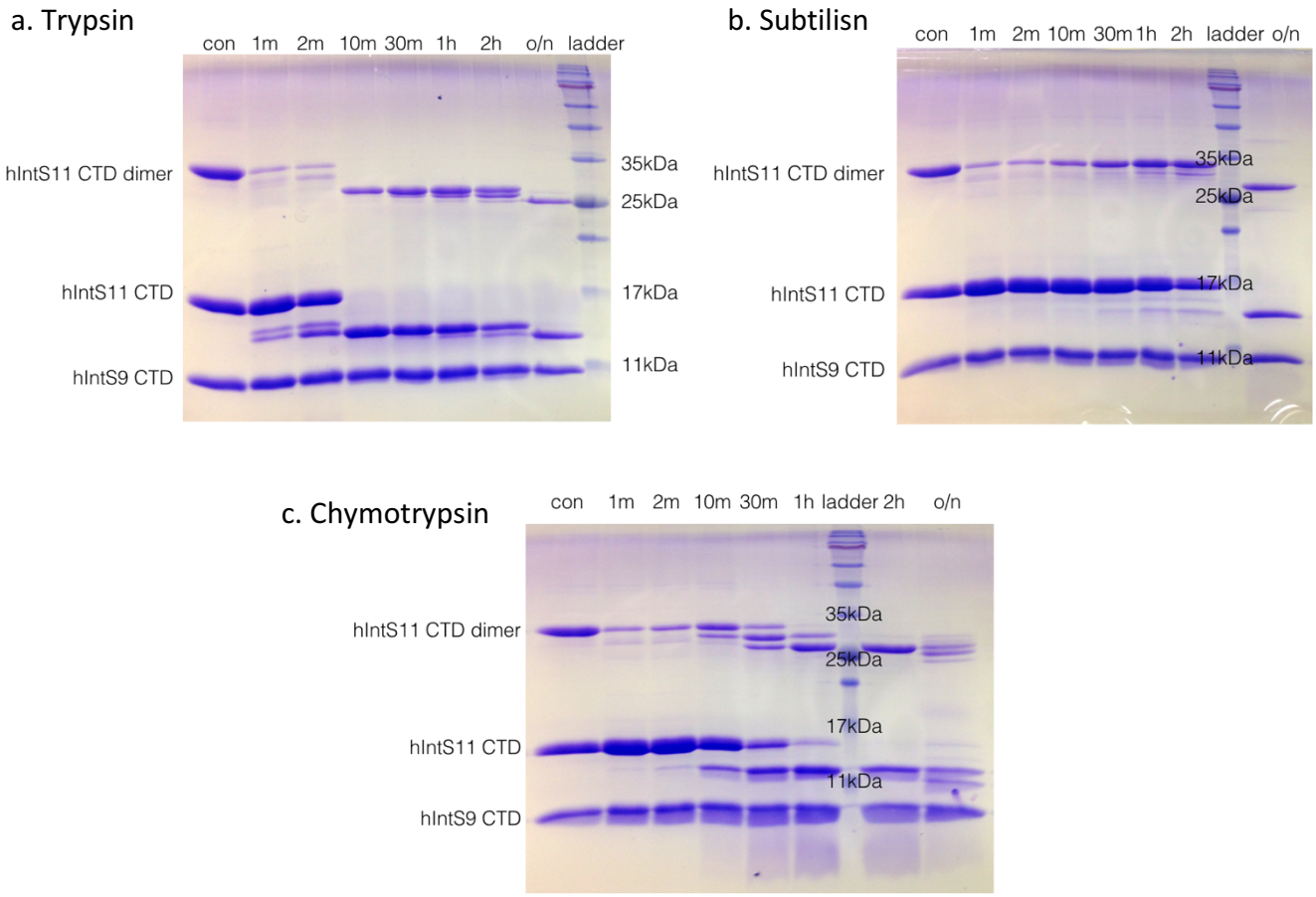


Figure 13 Limited proteolysis of human IntS9-IntS11 CTD complex. IntS9-IntS11 CTD complex protein was diluted to 1 mg/ml. Protease was added to the reaction with a 1:1000 protease:protein ratio. Samples

from each reaction were taken at different time and were boiled and frozen in -20 ℃ before loading on SDS-PAGE. (a) Trypsin at room temperature. (b) Subtilisn at 4 ℃. (c) Chymotrypsin at room temperature.

The limited proteolysis results showed that IntS9 CTD (residue 582-658) bands were very stable in all three reactions (Figure 13). The N-His-IntS11 CTD (residue 491-600) construct was eventually digested into a smaller protein with the size similar to what we observed in crystal samples. The digestion products for three different proteases shifted to around the same position on SDS-PAGE, which indicated a compact region of the protein. Before we design any new truncations, we thought what was removed from the original construct by protease might be the flexible N-His tag. Proteins without His-tag were then purified and used for crystallization (Figure 14). Fortunately, crystals were observed in several conditions this time (Figure 15).

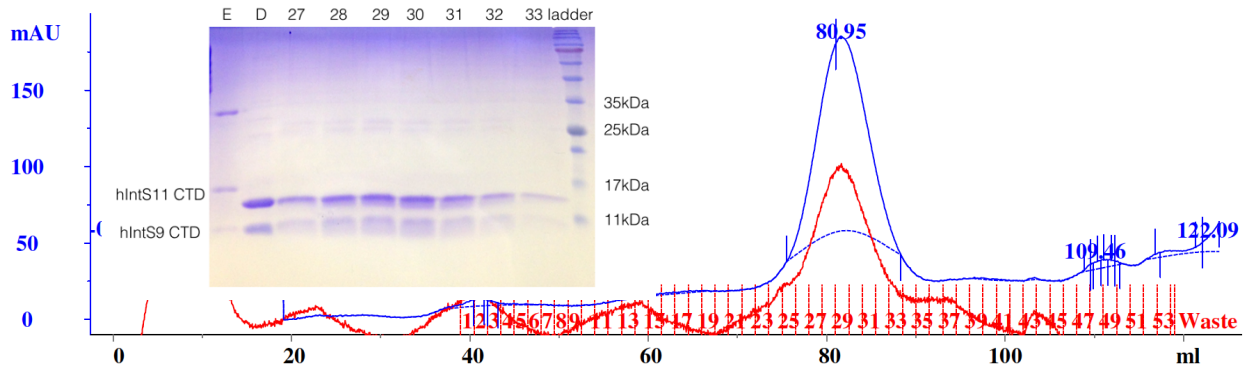


Figure 14 Gel filtration of human IntS9-IntS11 CTD complex after overnight thrombin digestion at 4 ℃.

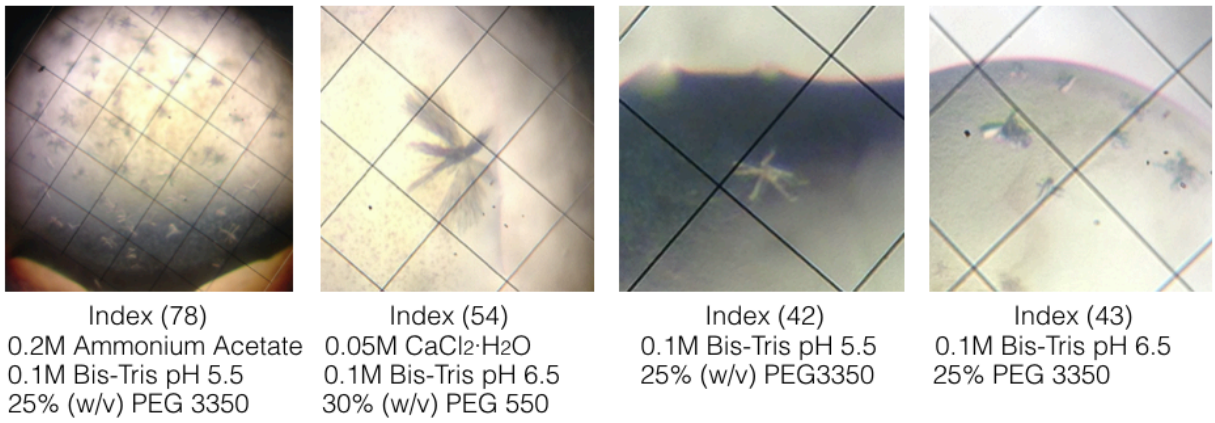


Figure 15 Crystals of thrombin treated human IntS9-IntS11 CTD complex from screening.

The common contents from these conditions are Bis-Tris pH 5.5-6.5 and PEG 3350. Grid screen was set up to reproduce and optimize the crystals. Single crystals appeared within one week and grew into full-size within two weeks. Most crystals were hollow (Figure 16). The optimized the crystals were able to diffract to 2.1-Å and were very reproducible.

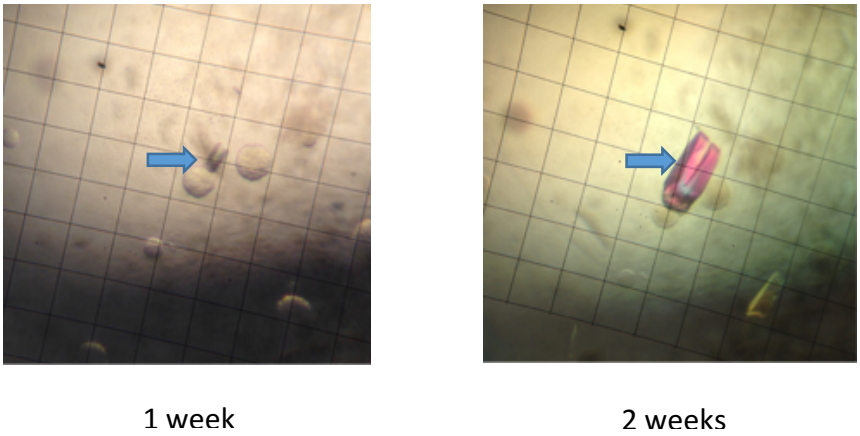


Figure 16 Optimized native human IntS9-IntS11 CTD complex crystal.

Section 2.3.3 Structure Determination

Selenomethionine Substituted Proteins

In order to determine the phase to solve the structure, we tried to express and purify selenomethionine substituted proteins for single or multiple wavelength anomalous diffraction (SAD or MAD) (Hendrickson, Horton, & Lemaster, 1990). The crystallizable constructs were transformed into methionine auxotroph B834 (DE3) cells (Novagen) and grown in LeMaster media supplemented with selenomethionine (Hendrickson et al., 1990). However, the yield of selenomethioninyl IntS9 CTD dropped dramatically. After purification, there was excess IntS11 CTD in the protein solution. After crystallization screening, we did not get any hits from this sample.

Heavy Atom Soaking

Since we were unable to get selenomethionine crystals, we tried to determine the structure by heavy atom soaking. Native crystals were soaked in the mother liquor with four heavy atom compounds for 3 hours or overnight (Table 3). Three of them did not disrupt the diffraction quality of the crystals and seven data sets were collected with these crystals.

After processing the data sets, the Hg-derivative dataset showed large differences to the native dataset by high derivative R factor and four Hg atoms were located and used for phasing using single isomorphous replacement (SIR) by the AutoSol routine in PHENIX(Adams et al., 2002b).

Table 3 Summary of heavy atom derivatives

Soak	Soaking Time	Resolution (Å)
Native	None	2.1
1mM K_2PtCl_4	3 hours	No diffraction
1mM $KAu(CN)_2$	3 hours/overnight	2.5/2.7
1mM $HgCl_2$	3 hours/overnight	2.3/2.2
1mM Trimethyl Pb(IV) Acetate	3 hours/overnight	2/2.1

Section 2.3.4 Structure of the IntS9 and IntS11 CTDs

The structure of IntS9-IntS11 CTD complex has been determined at 2.1-Å resolution (Figure 17). The refined atomic model has good agreement with the X-ray diffraction data and the expected geometric parameters. 98.1% of the residues are in the favored region of the Ramachandran plot, and no residues are in the disallowed region.

The structure of IntS9 CTD (residues 582-658) contains a four-stranded, antiparallel β-sheet (β2-β5). One face of the β-sheet is covered by two helices (α1-α2). A two-stranded antiparallel β-sheet (β1, β6) formed by residues near the beginning and end of the domain likely provides further stability to this domain.

The structure of IntS11 CTD (residues 493-596) contains a five-stranded, antiparallel β-sheet (β1-β5). Two helices cover one of its faces (α2-α3). A short helix (α1) precedes the first β-sheet is partly stabilized by interactions with IntS9.

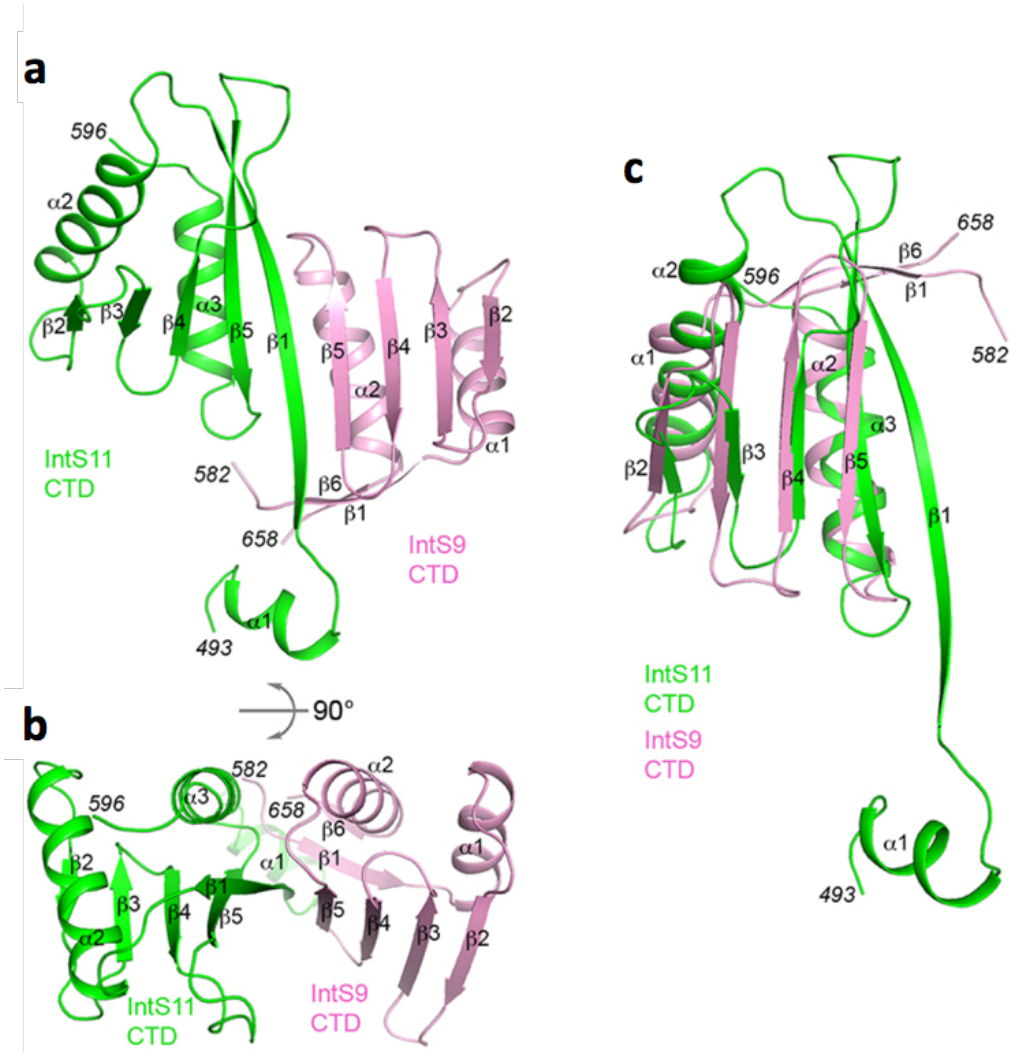


Figure 17 Crystal structure of the human IntS9-IntS11 CTD complex. **(a)** Structure of the human IntS9-IntS11 CTD complex. The IntS9 CTD is in pink, and the IntS11 CTD in green. **(b)** Structure of the human IntS9-IntS11 CTD complex, viewed after 90° rotation around the horizontal axis. **(c)** Overlay of the structure of IntS9 CTD (pink) with that of IntS11 CTD (green). The structure figures were produced with PyMOL (www.pymol.org). From (Wu et al., 2017)

The last four strands (β2-β5) of the IntS11 CTD has a similar fold to that for the β-sheet in IntS9 CTD. The two CTD structures can be superposed with an rmsd of 2.4-Å for 64 equivalent Cα atoms, although the sequence identity between the two proteins in this region is

only 16%. The two helices covering the β-sheet are located at similar positions in the two structures as well. A unique feature of IntS11 CTD is strand β1. It is the longest strand in the structure and is located in the center of the interface with IntS9.

Close structural homologs for IntS9 CTD include the CTD of an atypical Sm-like archaeal protein (Mura, Phillips, Kozhukhovsky, & Eisenberg, 2003) and the platform subdomain of the AP-2 complex β subunit (Schmid et al., 2006), based on a DaliLite search (Figure 18) (Holm, Kaariainen, Rosenstrom, & Schenkel, 2008). Close structural homologs for IntS11 CTD include the kinase associated-1domain (KA domain) at the C terminus of yeast septin-associated kinases and human MARK/PAR1 kinases (Moravcevic et al., 2010), the C-terminal domain of the catalytic subunit of AMP-activated protein kinase (AMPK, SNF1 in yeast) that mediates heterotrimer formation (Amodeo, Rudolph, & Tong, 2007; Townley & Shapiro, 2007; Xiao et al., 2007), and the N-terminal domain of BamC, part of the β-barrel assembly machinery (Figure 18) (R. Albrecht & Zeth, 2011). These structural homologs do not offer much insight into the functions of the two CTDs.

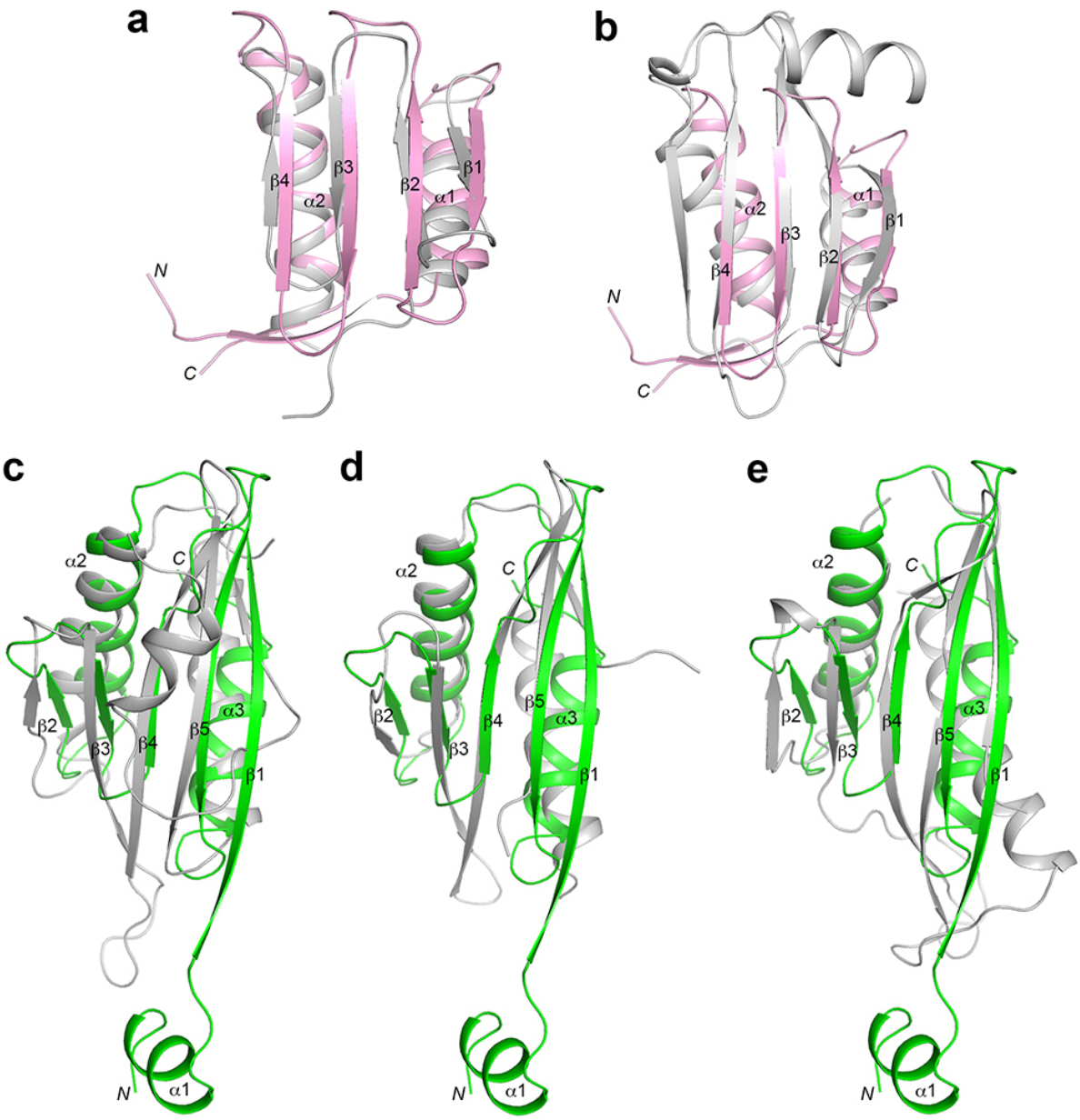


Figure 18 Structural homologs of IntS9 and IntS11 CTDs. (a and b) Overlay of the IntS9 CTD (pink) with the C-terminal domain of an atypical Sm-like archaea protein [Protein Data Bank (PDB) ID code 1M5Q; Z-score 7.3; 17% identity] (gray) (a) and the platform subdomain of the AP-2 complex β subunit (PDB ID code 2IV9, Z-score 6.8; 4% identity) (b). (c–e) Overlay of the IntS11 CTD (green) with the KA1 domain of yeast Kcc4 (PDB ID code 3OSM; Z-score 7.9, 8% identity) (c), the C-terminal domain of the catalytic subunit of AMPK (PDB ID code 4EAK, Z-score 6.9, 16% identity) (d), and the N-terminal domain of BamC (PDB ID code 2YH6, Z-score 6.3, 10% identity) (e). From (Wu et al., 2017).

Section 2.3.5 Overall Structure of the IntS9-IntS11 CTD complex

The complex of IntS9-IntS11 CTDs is formed by placing the β-sheets of the two domains side by side so that a parallel β-sheet is formed by strand β5 of IntS9 and strand β1 of IntS11. This juxtaposition creates a nine-stranded β-sheet in the IntS9-IntS11 CTD heterodimer. Most strands in this β-sheet are in antiparallel with only the two strands at the subunit interface being in parallel (Figure 19). The four flanking helices cover the same face of the β-sheet, whereas the

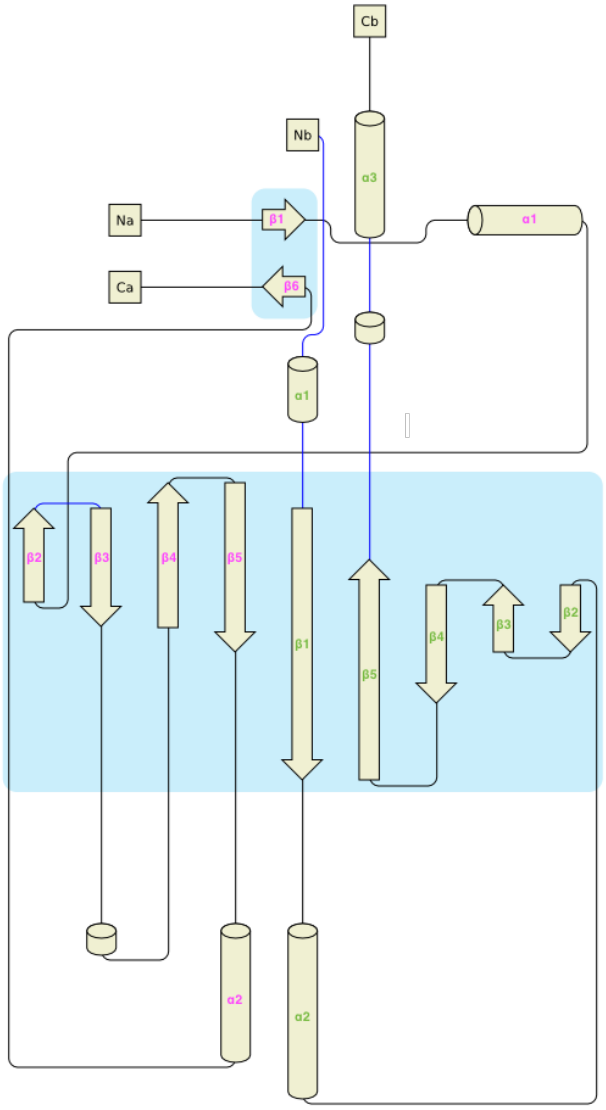


Figure 19 Topology diagram of the IntS9-IntS11 CTD complex. The secondary structure of IntS9 CTD is labeled in pink and the one of IntS11 CTD is labeled in green. Modified output from Pro-origami (Stivala, Wybrow, Wirth, Whisstock, & Stuckey, 2011).

other face of the β-sheet is exposed to the solvent (Figure 17b). Between the two β-strands at the center of the interface, there are seven hydrogen bonds formed (Figure 20a). In addition, many

side chains are also involved in the formation of the heterodimer, and ~1,200 Å^2 of the surface area of each subunit is buried in this interface. The neighboring side chains of the two strands on the exposed face are in contact with each other. The N-terminal helix (α1) of IntS11 contacts the N-terminal segment of IntS9, likely stabilizing both proteins in this region of the interface.

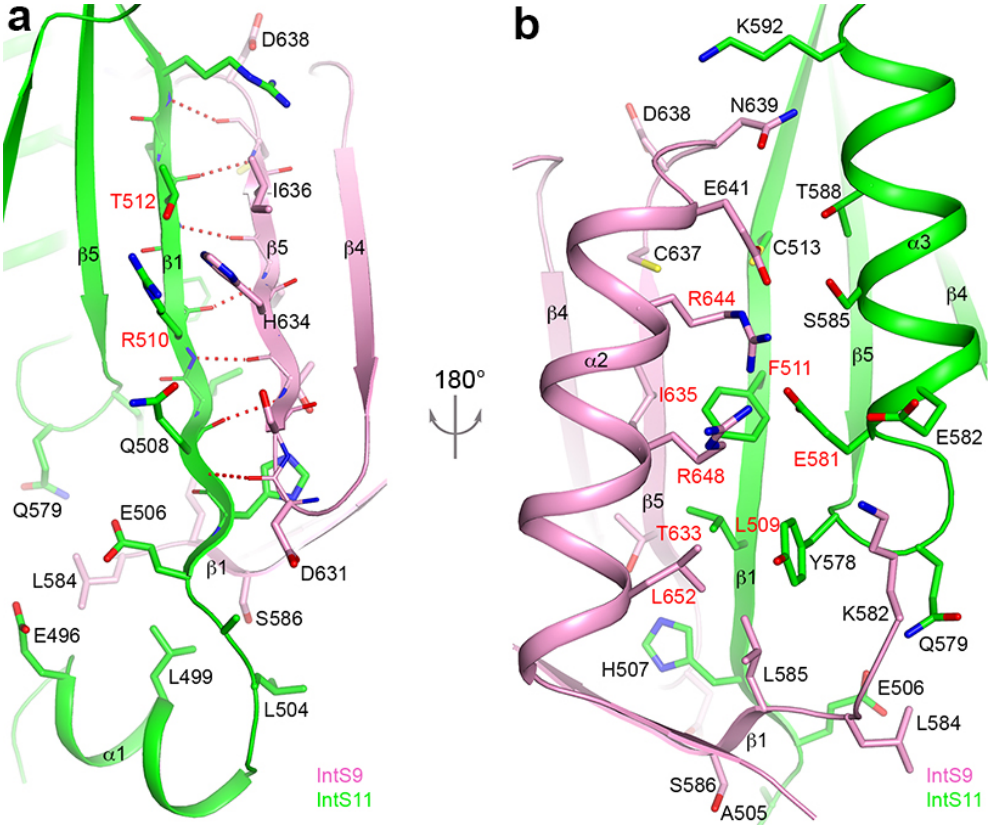


Figure 20 Detailed interactions at the interface of the IntS9-IntS11 complex. (a) Hydrogen bonds between IntS11 β1 strand and IntS9 β5 strand are indicated by the dashed lines in red. The neighboring side chains of the two strands are in contact. Residues in IntS11 α1 helix interact with residues in IntS9. From (Wu et al., 2017).

On the other face of the β sheet, helix α2 of IntS9 and helix α3 of IntS11 are positioned next to each other. This allows favorable interactions among some of their side chains as well as side chains of the two β strands in the center of the interface (Figure 20b). Residues from the two β-strands are buried inside the interface and are mostly hydrophobic. Residues from the two helices are exposed to the solvent and are mostly hydrophilic or charged. Most of the residues at this interface are highly conserved among IntS11 and IntS9 homologs, especially near the center of the interface (Figure 21).

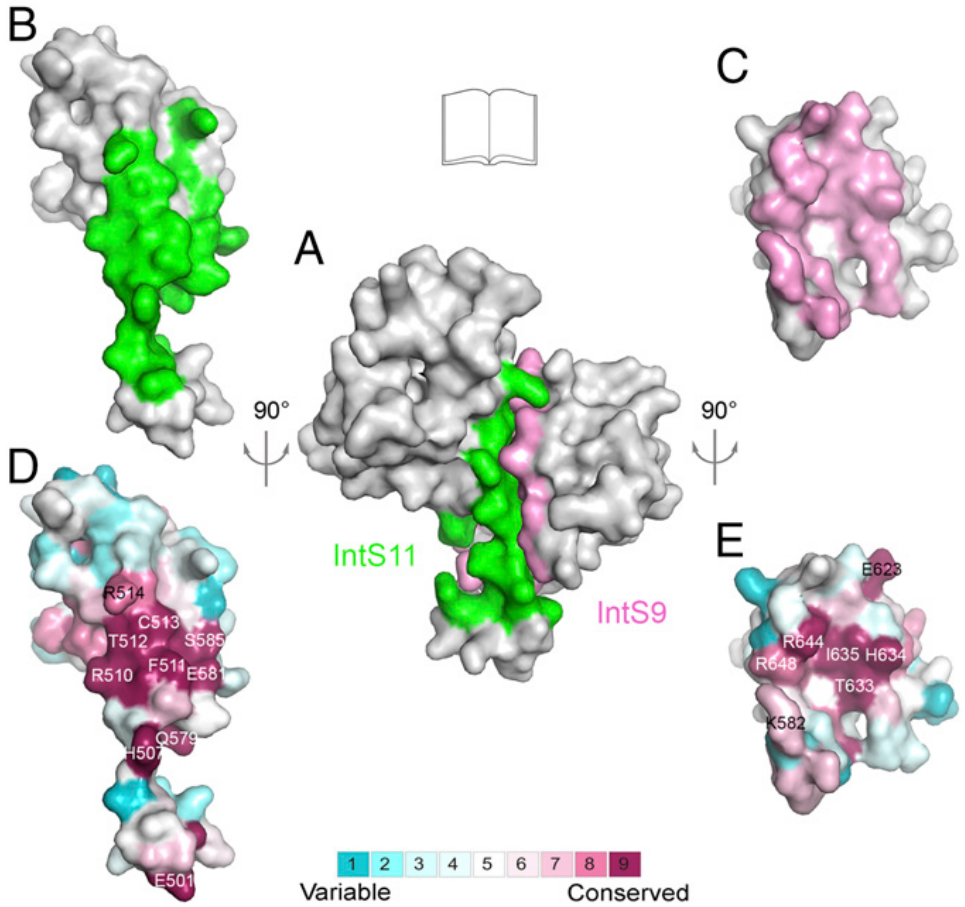


Figure 21 Molecular surface of the IntS9-IntS11 CTD complex. (A) Molecular surface of the IntS9-IntS11 CTD complex. Residues in the interface of IntS9 are colored in pink, and those in IntS11 are colored in green. The other residues are in gray. (B) An "open-book" view of the IntS9-IntS11 interface showing the surface area of IntS11 in contact with IntS9 after 90° rotation around the vertical axis. (C) An "open-book" view of the IntS9-IntS11 interface showing the surface area of IntS9 in contact with IntS11 after 90° rotation around the vertical axis. (D) Molecular surface of IntS11 colored by sequence

conservation produced by ConSurf (Armon, Graur, & Ben-Tal, 2001). Highly conserved residues are labeled. The color scheme runs from dark red (highly conserved) to cyan (poorly conserved) (color bar at bottom). (E) Molecular surface of IntS9 colored by sequence conservation. From (Wu et al., 2017).

There are four copies of the IntS9-IntS11 CTD complex in the crystallographic asymmetric unit. The overall structures of the two subunits in the four complexes are similar, with rmsds of ~0.5Å for equivalent Cα atoms between any pair of them. The overall structures of the four complexes are similar as well, especially for the β-sheet and the four flanking helices (Figure 22). Several loop regions show large differences in the conformations, suggesting that these regions are somewhat flexible. We observed a disulfide bond formed by the two Cys542 residues from two IntS11 subunits in neighboring complexes, which covalently links two complexes. The other cysteine residues in the structure are in the fully reduced state. Cys542 is located just before strand β2 in IntS11, there are some conformational differences observed in this strand among the four complexes. But since the disulfide bond creates only a small region of contact between two complexes, it is unlikely to affect the overall structure of the complex. This disulfide bond is likely the same disulfide bond we observed in Figure 11. We ran the protein solution samples and crystal samples on the non-reducing (w/o β-ME) SDS-PAGE (Figure 23). A minor amount of IntS11 CTD formed dimers in protein solution, while in the crystal, IntS11 CTD proteins were dimers. The dimer band disappeared by adding DTT in the loading samples. It seems like the hIntS11 CTD proteins coincidently form dimers through the disulfide bond in solution without strong reducing agent. During crystallization, the dimerization of IntS9-IntS11 CTD complex through the disulfide bond could possibly facilitate crystal packing.

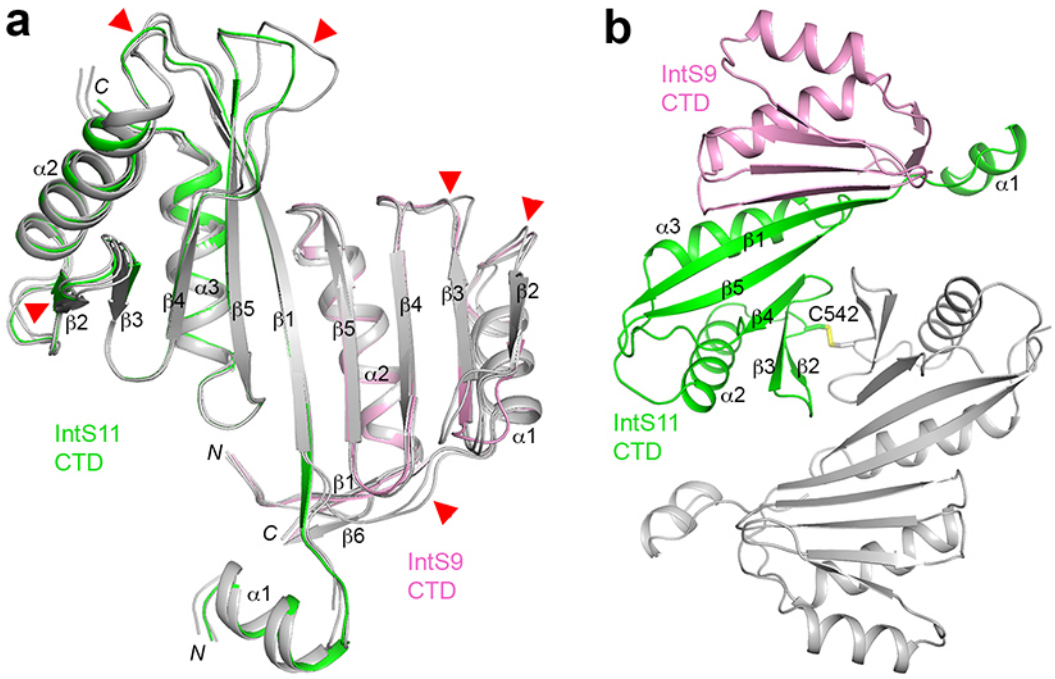


Figure 22 Structural comparisons of the four copies of IntS9-IntS11 CTD complexes. **(a)**. Overlay of the structures of the four IntS9-IntS11 CTD complexes. One complex is shown in color, and the other three in gray. Regions of large conformational differences among the four complexes are indicated with the red arrowheads. The α1 helix of IntS11 is missing in one of the subunits. **(b)**. The two Cys542 residues from neighboring IntS11 subunits in the crystal are involved in a disulfide bond, and the two complexes are related by a non-crystallographic two-fold axis. One complex is shown in color and the other in gray. From (Wu et al., 2017).

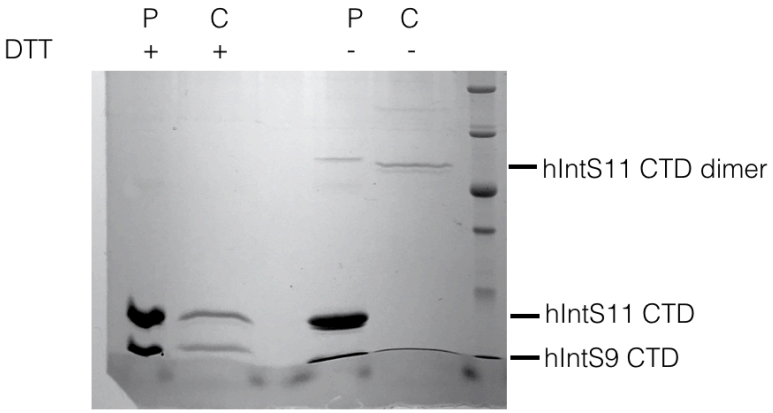


Figure 23 SDS-PAGE analysis of IntS9-IntS11 CTD complex in solution and in crystals. P, protein solution; C, crystal. +, 10 mM DTT; -, no DTT.

Section 2.3.6 Biochemical Studies of IntS9-IntS11 CTD complex

To assess the structural observations on the IntS9-IntS11 CTD complex, we collaborated with the Wagner Lab in UTMB, where biochemcial and functional studies were conducted.

Yeast two-hybrid assays

Truncation mutants of IntS9 and IntS11 CTDs were made to map the regions that are required and sufficient to mediate the interaction. Residues 500-600 of IntS11 interacted strongly with IntS9 CTD, whereas residues 510-600 showed no interactions (Figure 24A). Residue 510 is located in the middle of strand β1 (Figure 20), indicating that this strand is important for the interaction. The helix α1 is outside the residues 500-600 range and was demonstrated to be dispensable for the interaction, which was consistent with its being located at the periphery of the interface. Deleting only 10 residues from the C terminus of IntS11 (the mutant with residues 500-590) abolished the interaction (Figure 24A), indicating the importance of the last helix α3 for the IntS9-IntS11 interaction.

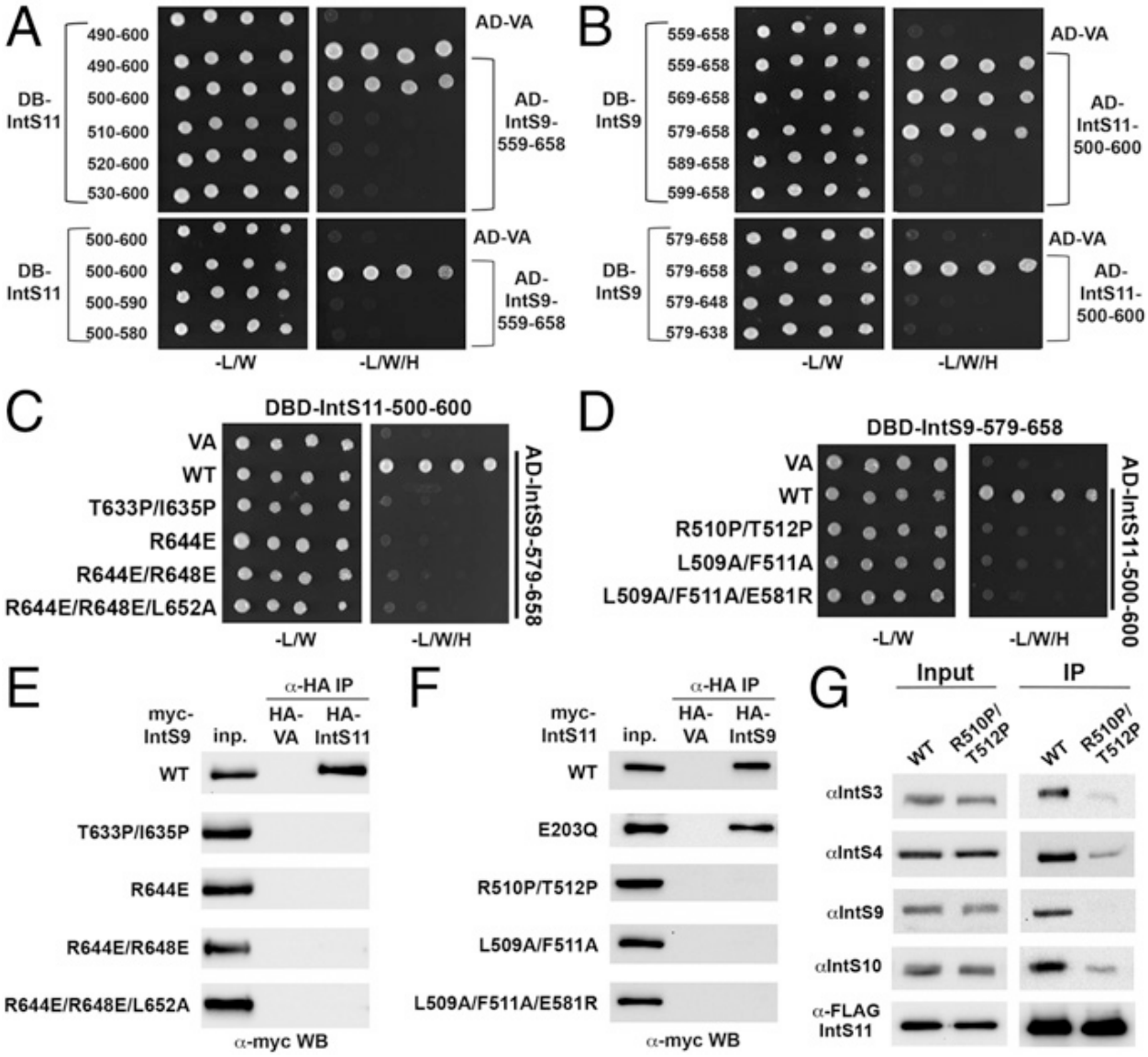


Figure 24 Biochemical studies of the IntS9-IntS11 CTD complex. From (Wu et al., 2017). (A) Yeast two-hybrid assay to define the minimal region of IntS11 sufficient to bind IntS9. AD, activation domain; BD, DNA-binding domain; VA, vector alone control. (B) Yeast two-hybrid assay to define the minimal region of IntS9 sufficient to bind IntS11. (C)Yeast two-hybrid assay using minimal regions of IntS9 and IntS11 with structure-based mutations in IntS11. (D) Yeast two-hybrid assay using minimal regions of IntS9 and IntS11 with structure-based mutations in IntS9. (E) Coimmunoprecipitation of full-length myc-tagged wild-type and mutant IntS9 with full-length HA-tagged wild-type IntS11. Proteins bound to HA affinity resin were probed with anti-myc antibody by Western blot (WB). (F) Coimmunoprecipitation of full-length myc-tagged wild-type and mutant IntS11 with full-length HA-tagged wild-type IntS9. (G)

Purification of endogenous INT from stable 293T cells expressing either wild-type FLAG-IntS11 or the interface mutant (R510P/T512P) using FLAG affinity resin.

For IntS9, residues 579-658 interacted strongly with the IntS11 CTD, whereas residues 589-658 showed no interactions with IntS11 CTD, indicating the requirement of strand β1 of IntS9 (Figure 24B). Deleting of 10 residues from the C terminus of IntS9 (residues 579-648) also abolished the interactions (Figure 24B). Residue 648 is located in the middle of helix α2, which was shown to interact with IntS11 in the structure. Overall, the minimal regions of IntS9 and IntS11 for sufficient interaction defined by yeast two-hybrid assays are fully consistent with the structural observations.

We next designed a series of point mutations that are expected to perturb the IntS9-IntS11 interaction based on the crystal structure. We mutated two residues in the middle of the each strand to prolines to disrupt the hydrogen bonds between the two β-strands at the center of the interface, i.e., the T633P/I635P double mutant for IntS9 (Figure 20b) and the R510P/T512P double mutant for IntS11 (Figure 20a). Mutations were also designed to disrupt interactions among side chains. We designed the R644E single mutant, the R644E/R648E double mutant, and the R644E/R648E/L652A triple mutant in helix α2 of IntS9, the L509A/F511A double mutant in strand β1 of IntS11, and the L509A/F511A/E581R triple mutant in strand β1 and helix α3 of IntS11 (Figure 20b). Most of these residues are strictly conserved among the homologs, but Leu652 of IntS9 and Leu509 of IntS11 show conservative variations to other hydrophobic residues (Figure 10). We introduced these mutations into the minimal CTDs of IntS9 or IntS11 for yeast two-hybrid assays. The results showed a complete loss of interaction (Figure 24 C, D).

Coimmunoprecipitation Assays

The point mutations mentioned above were also introduced into the full-length cDNAs encoding IntS9 and IntS11 to test their impact on the interaction using a coimmunoprecipitation assay. Previous studies showed that IntS9-IntS11 heterodimer could withstand rigorous washing with detergent and high salt (T. R. Albrecht & Wagner, 2012). The same conditions were tested for mutants. Various myc-tagged IntS9 cDNAs with HA-tagged wild-type IntS11 were transfected into 293T cells. The cell lysates were subjected to anti-HA immunoaffinity matrix and probed with anti-myc antibodies using Western blot analysis. While the wild-type IntS9 was able to coimmunoprecipitate with IntS11, none of the mutants was detected in the immunoprecipitate (Figure 24E). As shown in the figure, all IntS9 mutants were expressed to the similar level to wide-type IntS9 which suggests that these proteins are folded properly and the lack of coimmunoprecipitation is caused by the disruption of the interaction by the mutations.

The reciprocal coimmunoprecipitation were also performed where HA-tagged wild-type IntS9 with several IntS11 mutants were transfected into 293T cells. A catalytic mutant (E203Q, one of the conserved residue in the active site of the metallo-β-lactamase) was included as a control. Both the wide-type and the E203Q mutants were able to interact with IntS9, but the interface mutants did not show interaction (Figure 24F).

To test whether disruption of IntS9-IntS11 interface could inhibit their ability to incorporate into the endogenous Integrator complex, cell lines stably expressing FLAG-tagged wild-type IntS11 or the mutant R510P/T512P IntS11 were created and the INT from cell extracts were purified by pulling down FLAG-tagged IntS11. The levels of IntS3, IntS4, IntS9 and IntS10 purified by mutant IntS11 were significantly lower than the levels of proteins purified by wild-type IntS11 (Figure 24G). These results demonstrate that the CTD interaction between

IntS9 and IntS11 is important to incorporate the IntS11 endonuclease into the endogenous INT.

Section 2.3.7 Functional Importance of the IntS9-IntS11 Complex

Currently there is no *in vitro* assay to assess INT function. To address the functional relevance of IntS9-IntS11 interactions observed in the crystal structure, *in vivo* UsnRNA 3'-end formation was measured by using a cell-based fluorescence reporter, the U7 snRNA gene reporter. This reporter consists of the human U7 snRNA promoter, the U7 snRNA gene body, a 3'-box sequence for snRNA 3'-end processing, and the GFP coding gene followed by a strong polyadenylation signal (Figure 25A). When transfected into untreated cells, the reporter gives rise to no GFP signals, because the reporter is cleaved after the U7 snRNA gene due to normal 3'-end processing by INT. If the INT subunits were knocked down by siRNA, the reporter is unable to be properly processed, resulting in transcriptional read-through and the production of a GFP mRNA containing the U7 snRNA as its 5' UTR. The GFP mRNAs are then translated using the native start codon of GFP because U7 lacks an AUG sequence.

To do the RNAi rescue experiment, siRNAs targeting IntS9 or IntS11 were first transfected into HeLa cells to knock down endogenous IntS9 or IntS11. cDNA constructs of wild-type or interface mutant IntS9 or IntS11, containing silent point mutations that are RNAi-resistant, were then transfected into IntS9- or IntS11-depleted cells. The expression level of GFP were detected to determine the effectiveness of the constructs in restoring INT activity. In control siRNA-treated cells, nearly no GFP was detected after transfection of U7-GFP reporter (Figure 25 B, C). In cells treated with siRNA targeting either IntS9 or IntS11, robust GFP proteins were detected. Expression of RNAi-resistant wild-type IntS9 or IntS11 was able to reduce the expression of GFP to the levels observed in control siRNA transfected cells. As

expected, expression of IntS11 catalytic mutant E203Q was unable to rescue U7 snRNA processing. The interface mutants also failed to rescue INT 3'-end processing activity (Figure 25 B, C).

Our collaborator also conducted rescue experiment on endogenous U2 and U4 snRNAs. They created stable cell lines expressing either RNAi-resistant wild-type IntS11 or a subset of IntS11 CTD interface mutants and assessed the levels of misprocessed U2 or U4 snRNA. About 25-fold increase in the level of misprocessed snRNA was observed upon depletion of IntS11 (Figure 25D). The levels of misprocessed snRNA present in IntS11 knockdown cells could be rescued by stably expressing wild-type IntS11, but the IntS11 CTD interface mutants could not reduce the levels of misprocessed snRNA.

Collectively, these results demonstrate that interactions observed in the structure of IntS9-IntS11 CTD complex are critical to INT activity in snRNA processing.

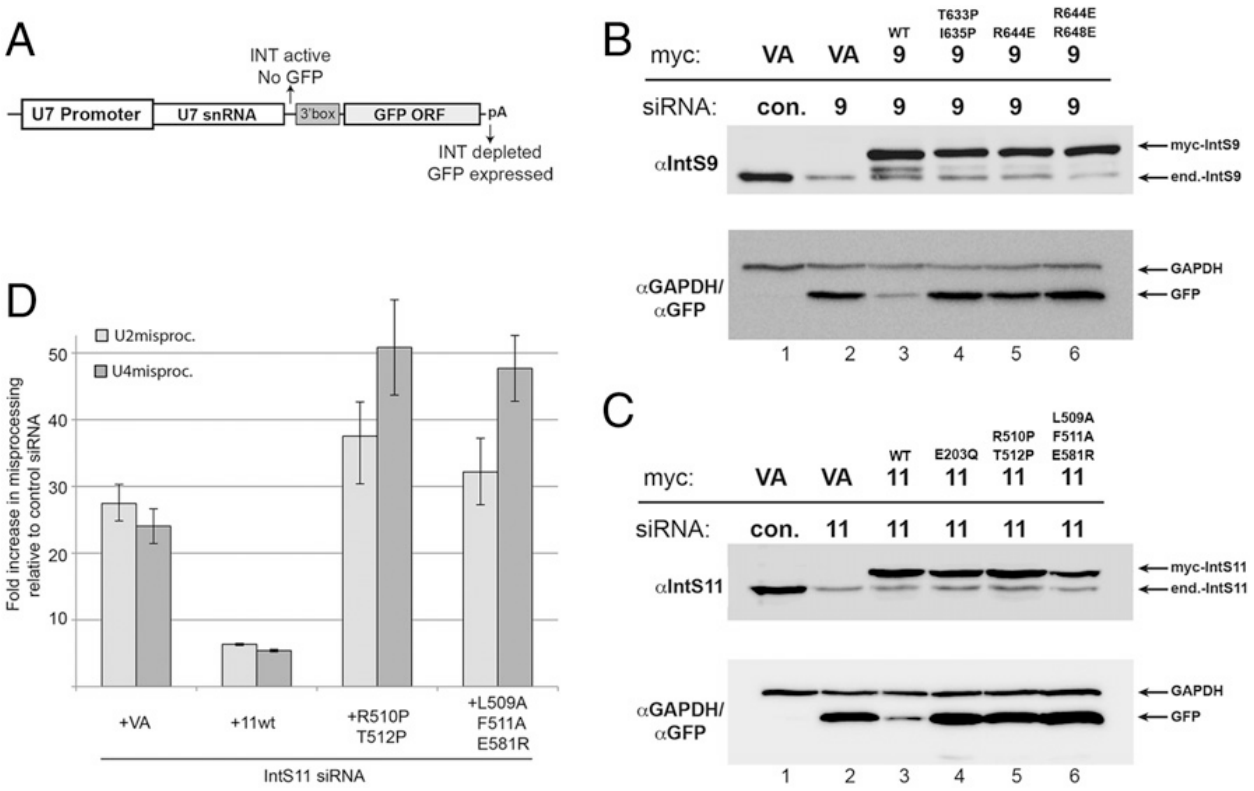


Figure 25 Functional importance of the IntS9-IntS11 interactions for snRNA 3'-end processing. (A) Schematic diagram of U7-GFP reporter. (B) Western blot analysis of lysates from IntS9 depletion Hela cells that were transfected with myc-tagged RNAi-resistant wild-type or mutant IntS9. The U7-GFP reporter was transfected into all the cells. (C) Western blot analysis of lysates from IntS11 depletion Hela cells that were transfected with myc-tagged RNAi-resistant wild-type or mutant IntS11. The U7-GFP reporter was transfected into all the cells. (D) Quantitative RT-PCR analysis of misprocessed endogenous U2 or U4 snRNA. The bar graph represents the fold increase in the levels of misprocessed snRNA; error bars represent the SD from the mean. From (Wu et al., 2017).

Section 2.4 Discussion

The crystal structure of IntS9-IntS11 CTD complex reveals an extensive molecular interface mediated by numerous interactions which explains the high binding affinity that has been reported for IntS9 and IntS11. Although IntS11 has been identified as the endonuclease in INT, how IntS11 is recruited to the RNA substrates is not known. Based on the structure of its

analog CPSF-73, the endonucleolytic activity of IntS11 seems to be regulated by other protein factors from INT. The structure indicates a role for the catalytically inactive IntS9, which provides a distinct structural surface established only through the heterodimerization with IntS11. This specific interface may facilitate the recognition and recruitment of the active IntS11 by other INT subunits. Such a mechanism might also be operative for CPSF-73 and CPSF-100 in the pre-mRNA 3'-end processing machinery.

The homologous enzyme RNase J also has a CTD, but it is substantially smaller with one three-stranded β sheet and two facing α helices (de la Sierra-Gallay, Zig, Jamalli, & Putzer, 2008). The CTD mediates homodimerization of RNase J in solution. Depletion of its CTD makes the enzyme monomeric and also abrogates all its activity *in vitro* even though its metallo-β-lactamase and β-CASP domains remain intact. Based on the studies using U7-GFP reporter, it is clear the mutations specifically disrupting IntS9-IntS11 CTD heterodimerizations have effects equivalent to those of the catalytic mutation (E203Q) of IntS11. This finding demonstrates that the binding of IntS9 is essential for IntS11 function in cells. It also suggested that homo- or hetero-dimerization of β-CASP RNA endonucleases either plays an important role in the recruitment to RNA substrates or impacts the activity of the catalytic domain. One possible explanation is that formation of IntS9-IntS11 CTD complex would induce conformational changes in IntS11 that would allow access to and cleavage of the RNA substrates. This structural requirement would ensure that any IntS11 not associated with IntS9 would be inactive. Our results could not exclude the possibility that IntS11 may require additional factors for endonucleolytic activity. It is known that for 3'-end processing of either pre-mRNA or histone pre-mRNA, the cleavage activity of CPSF73 requires interaction with CPSF-100 as well as a scaffold protein called Symplekin (Michalski & Steiniger, 2015). Recent studies has reported

the interaction between INT subunit 4 (IntS4) and IntS9-IntS11 heterodimer which is crucial for proper snRNA 3'-end processing (T. R. Albrecht et al., 2018). The efforts we made to understand the molecular basis for this interaction will be discussed in Chapter 4.

IntS9 and IntS11 interaction are likely to be significant for assembly of INT. The CTD dimerization may provide an essential surface that is recognized by other members in the INT and facilitate the recruitment of the cleavage factor into the complex. This mechanism could regulate the cleavage event by ensuring that only the authentic IntS9-IntS11 heterodimer is incorporated into INT. This model is supported by our experiments that a heterodimer-deficient IntS11 failed to pull down other members of INT in addition to IntS9 (Figure 24G).

We also noticed that between the metallo-β-lactamase domain and the CTD in IntS9 and IntS11 there are some unknown regions with secondary structure elements. These linkers are likely to be organized structurally. In the structure of RNase J, there is also a linker containing secondary structures between metallo-β-lactamase domain and the CTD. This linker interacts with both domains in RNase J (de la Sierra-Gallay et al., 2008). Similarly, the linkers in IntS9 and IntS11 may also interact with both domains and communicate heterodimerization to the active site to activate cleavage. The exact arrangement of the catalytic domains, the linkers and the CTD within IntS9 and IntS11 remains to be determined.

CHAPTER THREE：

Structural Studies of RNA deNADding enzyme Nudt12

Section 3.1 Introduction

In eukaryotes, the 5’-end of mRNA is modified by the addition of a 5',5'-triphosphate-linked 7-methylguanosine (m^7G) (cap 0 structure, Figure 26). The m^7G cap plays a major role in various functional processes, including mRNA processing, nuclear export, cytoplasmic translation initiation (Ramanathan, Robb, & Chan, 2016). The m^7G cap also functions as a protective group from 5' to 3' exonuclease cleavage. Recently, a novel cap structure with 2'O methylation of +1 nucleotide (cap 1 structure, Figure 26) has been reported as an identifier of self RNA in the innate immune system against foreign RNA (Ramanathan et al., 2016). Some other derivatives of the canonical m^7G cap have also been reported, including $m^{2,2,7}G$-capped small U-rich noncoding RNAs with two additional methyl moieties to the guanosine (Mattaj, 1986). Collectively , the different caps of RNA provide a layer of regulation of 5’-end decay of eukaryotic RNAs.

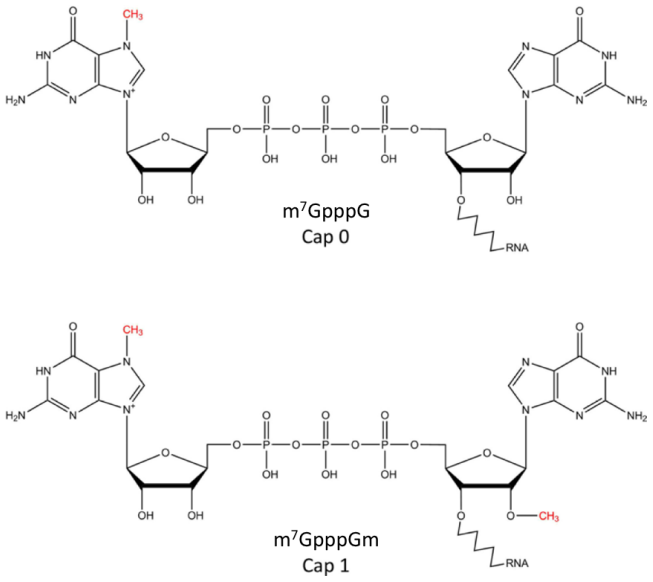


Figure 26 mRNA caps in eukaryotes.

RNA capping has been thought to occur only in eukaryotes for a long time. However, recent studies showed that the cofactor nicotinamide adenine dinucleotide (NAD) covalently links to the 5'-end of small regulatory RNA in bacteria in a cap-like manner and can protect the RNA from 5'-end decay by the bacterial RppH and RNaseE nuclease in vitro (Cahova, Winz, Hofer, Nubel, & Jaschke, 2015).

In eukaryotes, the m^7G cap is added to the 5'-end of mRNA cotranscriptionally when the RNA reaches a length of ~25 nucleotides(Moteki & Price, 2002). Addition of the NAD cap to bacterial RNA appears to occur during transcription initiation (Bird et al., 2016). NAD can be used by bacterial RNA polymerase in place of ATP as the first transcribed nucleotide (Bird et al., 2016). In vitro assay showed that eukaryotic RNA polymerase II (RNAPII) can also use NAD as an initiating nucleotide (Bird et al., 2016), which raised the possibility of existence of NAD-capped RNA in eukaryotes. Indeed, recent studies confirmed that NAD-capped RNAs are also present in eukaryotes, including *Saccharomyces cerevisiae* (Walters et al., 2017) and mammalian cells (Jiao et al., 2017). These findings demonstrate that NAD caps are broadly distributed in diverse organisms.

In mammals, the m^7G cap is removed by a subclass of Nudix (nucleoside diphosphate linked to another moiety X) hydrolase family of proteins, such as Dcp2, Nudt3 and Nudt16 (Grudzien-Nogalska & Kiledjian, 2017), to initiate RNA 5'-end decay. The presence of NAD-capped RNA in mammalian cells suggested that eukaryotic cells likely harbor enzymes that can remove NAD cap from an RNA. More recently, the mammalian non-canonical decapping enzyme, DXO, has been demonstrated to possess deNADding activity on NAD capped RNA (Jiao et al., 2017). It can decap the RNA by removing the entire NAD moiety to produce a 5'-end monophosphate RNA. The NAD-capped RNAs were more stable in cells lacking DXO,

indicating that this enzyme can modulate the levels of NAD-capped RNAs in cells (Jiao et al., 2017). And different from prokaryotes, the NAD cap in mammalian cells seems to promote rather than prevent RNA decay since a synthetic NAD-capped RNA transfected into cells was less stable than m^7G-capped RNA (Jiao et al., 2017).

In prokaryotes, the NAD cap is removed by a nudix hydrolase protein named NudC (Cahova et al., 2015). NudC was initially described as a NAD(H) pyrophosphohydrolase. It can remove the NAD cap by hydrolyzing the diphosphate linkage to produce nicotinamide mononucleotide (NMN) and 5' monophosphate RNA (Cahova et al., 2015). *In vitro* competition experiments showed that NudC prefers NAD-capped RNAs over NAD(H) by several orders of magnitude, suggesting that NAD-capped RNAs are its primary biological targets (Hofer et al., 2016). The crystal structure of *E. coli* NudC showed that it contains three domains, an N-terminal domain (NTD), a zinc-binding motif and a C-terminal domain (CTD) (Figure 27). The conserved Nudix sequence elements are located in CTD.

Nudt12 is a mammalian Nudix hydrolase and a close homolog of bacterial NudC. The catalytic domain of Nudt12 shares 29% amino acid sequence identity with *E. coli* NudC. A unique feature of Nudt12 is the possess of an Ankyrin repeat domain in the N-terminal region (Figure 27). Nudt12 was identified as an NADH diphosphatase with a substrate preference for the reduced nicotinamide adenine dinucleotide (Abdelraheim, Spiller, & McLennan, 2003). At the beginning, Nudt12 was shown to localize to peroxisomes when fused to a C-terminal GFP (Abdelraheim et al., 2003), but recent immunocytochemistry reveals its presence in the cytoplasm in kidney cells (Carreras-Puigvert et al., 2017), which indicates its cytoplasmic function. Together, Nudt12 is a presumable deNADding enzyme in mammals. In order to demonstrate the deNADding activity of Nudt12 and understand its catalytic mechanism, we

determined the crystal structure of the catalytic domain of mouse Nudt12 in complex with the hydrolyzed product AMP. Together with biochemical and functional studies, we prove that Nudt12 is a mammalian deNADding enzyme distinct from DXO. Our manuscript of this study is under review.

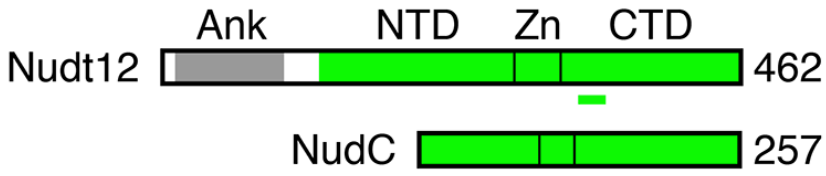


Figure 27 Domains organization of mouse Nudt12 and E. coli NudC. The Ankyrin repeat domain of Nudt12 is indicated in gray, and the catalytic domain in green. The conserved Nudix motif in Nudt12 is indicated with a bar (green).

Section 3.2 Experimental Procedures

Protein Expression and Purification

Residues 126-462 of wild-type mouse Nudt12 were sub-cloned into the pET28a vector (Novagen), and the recombinant protein carried an N-terminal His-tag. The purified protein did not produce any crystallization hits. The E219A/E220A/E221A triple mutation was introduced using Transfer-PCR method (Erijman, Dantes, Bernheim, Shifman, & Peleg, 2011). The mutant protein was expressed in *Escherichia coli* BL21(DE3) Star cells at 20℃ for 18 h. The cells were lysed by sonication in a buffer containing 20 mM Tris (pH 8.0), 250 mM NaCl, and 5% (v/v) glycerol. The lysate was loaded onto an Ni-NTA (Qiagen) column. The eluted protein was treated overnight with thrombin at 4℃ to remove the His-tag and was further purified by gel

filtration chromatography (Sephacryl S-300; GE healthcare). The purified protein was concentrated to 23 mg/ml in a solution containing 20 mM Tris (pH 8.0), 200 mM NaCl, and 5 mM DTT before being flash-frozen in liquid nitrogen and stored at –80℃.

Protein Crystallization

The protein at 8 mg/ml concentration was incubated with 1.5 mM NAD (Sigma) in a buffer containing 20 mM Tris (pH 8.0), 200 mM NaCl, 5 mM DTT, and 2 mM $MgCl_2$ at 4℃ for 30 min. Crystals were obtained at 20℃ using the sitting-drop vapor-diffusion method. The reservoir solution contained 0.1 M HEPES (pH 7.5) and 24% (w/v) PEG2000MME. The crystals were cryo-protected by the reservoir solution supplemented with 15% (v/v) ethylene glycol and were flash-frozen in liquid nitrogen for data collection at 100K.

Data Collection and Structure Determination

X-ray diffraction data were collected using the Pilatus-6M detector at the Advanced Photon Source (APS) beamline 24-ID-C. The X-ray wavelength was 0.9791 Å. The diffraction images were processed and scaled using the XDS program (Kabsch, 2010). The crystal belonged to space group *P*1 with unit cell dimensions of a = 56.2 Å, b = 58.6 Å, c = 61.7 Å, α = 102.4°, β = 115.2°, γ = 104.7°. There is a Nudt12 homodimer in the crystallographic asymmetric unit.

The catalytic domain of Nudt12 shares 29% sequence identity with NudC, and a molecular replacement solution could readily be found using the structure of NudC as the search model (Hofer et al., 2016) with the program Phaser (McCoy et al., 2007). However, the resulting electron density map was not of sufficient quality to rebuild the model. To obtain separate phase information, the CTD of Nudt12 was located using the molecular replacement method. The

phase information from this model was used to calculate an anomalous difference electron density map, which clearly revealed the positions of the two zinc atoms (with 11 **?** peak heights), even though the diffraction data were collected far above the zinc absorption edge. Phase information from anomalous scattering was combined with that from the model, and the CTD could be rebuilt automatically using the resulting map with PHENIX (Adams et al., 2002a). Several ☆-strands in the NTD and segments near the zinc could be built manually with the program Coot (Emsley & Cowtan, 2004a) and were included in the subsequent structure refinement with PHENIX, which led to an improved $2F_o–F_c$ map. The entire structure was obtained after several rounds of manual model building followed by refinement. The crystallographic information is summarized in Table 4.

Table 4 Summary of crystallographic information

Data collection	
Space group	*P*1
Cell dimensions	
a, *b*, *c* (Å)	56.2, 58.6, 61.7
α, β, γ(°)	102.4, 115.2, 104.7
Resolution (Å)[1]	53-1.6 (1.69-1.6)
R_{merge} (%)	4.3 (65.3)
$CC_{1/2}$	0.999 (0.716)
I/σI	11.5 (1.2)
Completeness	89.6 (76.3)
No. of observations	214,665 (27,577)
Redundancy	2.8 (2.6)
Refinement	
Resolution (Å)	53-1.6
No. of reflections	76,425
R_{work}	18.0
R_{free}	21.9
No. of atoms	
Protein	4,969
Ligand/Ion	54
Water	290
B-factors	
Protein	39.3
Ligand/Ion	28.4
Water	38.8
R.m.s. deviations	
Bond lengths (Å)	0.015
Bond angles (°)	1.4
The numbers in parentheses are for the highest resolution shell.	

Section 3.3 Results and Discussion

Section 3.3.1 Initial Constructs Design and Soluble Protein Expression

Initially, three constructs were made using mouse Nudt12 gene (Table 5). The two truncated constructs were designed based on the sequence alignment with *E. coli* NudC, the structure of which has been solved. The truncated constructs include only the catalytic domain which is aligned with *E. coli* NudC. The Ankryin repeat domain was absent in the truncated constructs (Figure 28).

Table 5 Initial mouse Nudt12 constructs

Name	Tag	Vector	Start	End	Total Residues	M.W. (kDa)
Nudt12-1	N-His	pET28a	1	462	462	52
Nudt12-2	N-His	pET28a	126	462	337	39
Nudt12-3	N-His	pET28a	147	462	316	37

Two of the constructs, Nudt12-1 and Nudt12-2, could produce soluble proteins. Nudt12-3 did not express maybe because it lacks an α helix that is important for protein folding. Both Nudt12-1 and Nudt12-2 proteins showed good behaviors and purity on gel filtration and SDS-PAGE analysis and were used for crystallization screening (Figure 29).

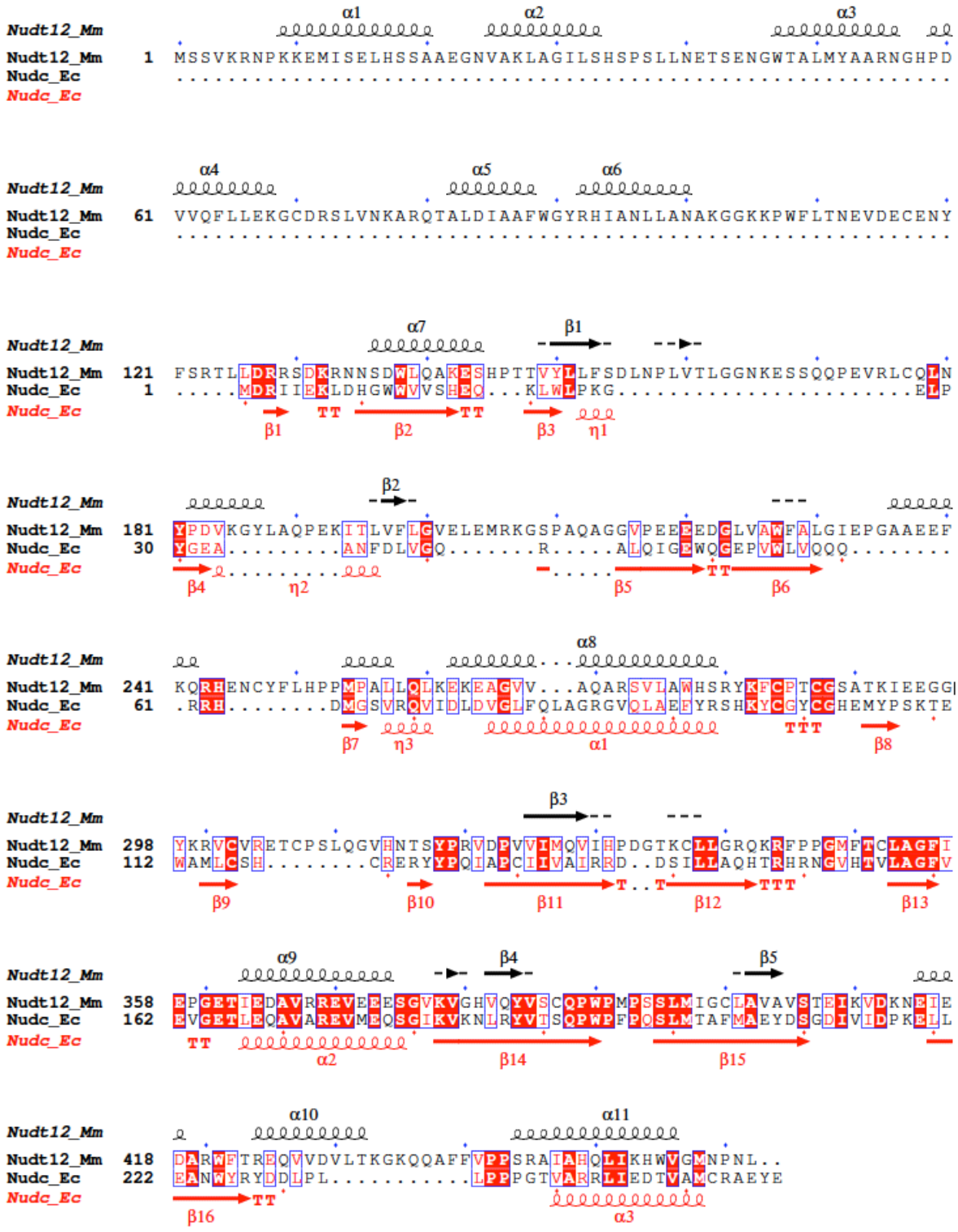


Figure 28 Sequence alignment of mouse Nudt12 and E. coli Nudc. Mm, *Mus musculus*; Ec, *Escherichia coli.* Figure made by ESPript (Gouet et al., 1999).

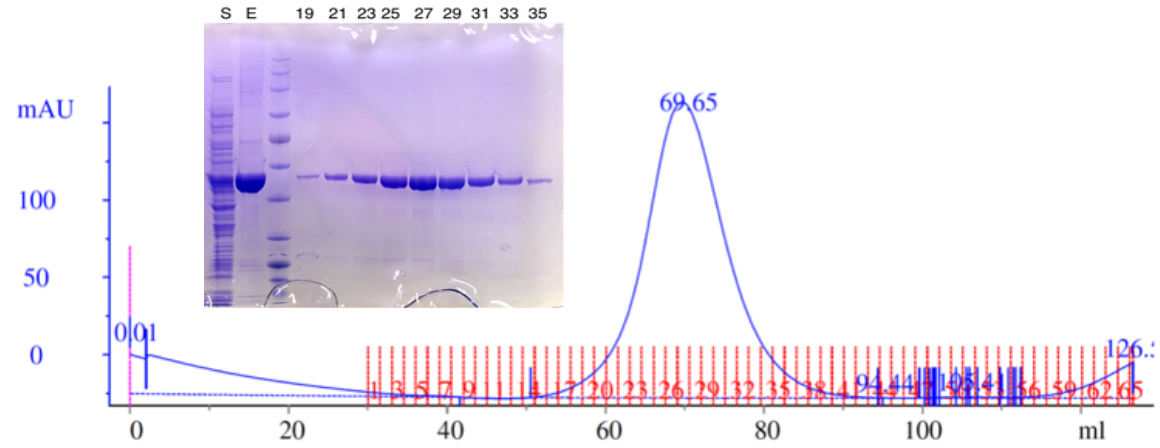


Figure 29 Gel filtration profile of full-length mouse Nudt12.

Section 3.3.2 Crystal Screening and Optimization

The purified Nudt12 full-length protein and catalytic domain (residue 126-462) were screened for crystallization conditions with several commercial crystallization kits (Cudney et al., 1994; Jancarik et al., 1991) by sitting drop vapor diffusion at 20 ℃. The protein samples were incubated with 3' NADP with a molar ratio 1:5 (enzyme: substrate) before being used for crystallization. The full-length protein was crystallized from condition containing 4% v/v Tacsimate pH 7.0 and 12% w/v PEG 3350 (Figure 30). However, these tiny crystals did not diffract and were unable to be optimized. The truncated protein (residue 126-462) was crystallized from a different condition containing 1.4 M sodium malonate pH 6.0 (Figure 31). After optimization, we got bigger crystals from a condition with slightly different pH. But diffraction of these crystals could only go to 9-Å.

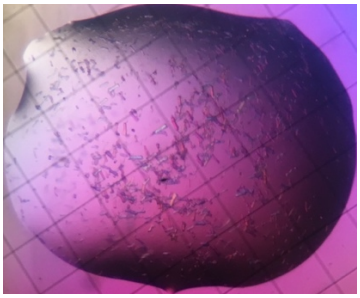
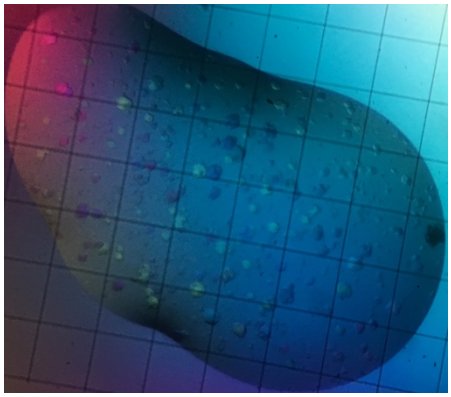


Figure 30 Crystals of full-length Nudt12.

A 

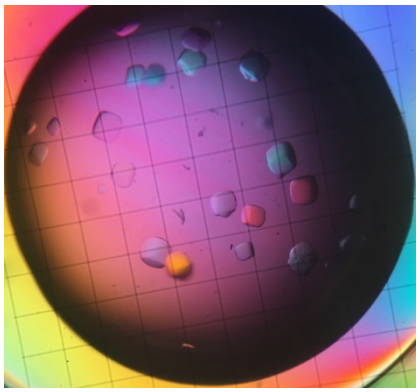
B 

Figure 31 Crystals of Nudt12 catalytic domain (residue 126-462). A. A sitting drop from original screening. B. A sitting drop from optimization screening.

To further optimize the crystals, we tried several methods including removing His-tag during purification, screening additives for crystallization, incubating the sitting drops under different temperature, and controlled dehydration. Unfortunately, the diffraction of the crystals could not go beyond 8-Å. We also made additional constructs around the initial constructs (Table 6). Although most of these constructs produced soluble proteins, none of them could give crystals with good diffraction.

We also tried to express and purify frog Nudt12 which shares 67% sequence identity with the mouse Nudt12, but the frog protein did not produce good quality crystals, either.

Table 6 Additional Nudt12 constructs

Name	Tag	Vector	Start	End	Total Residues	M.W. (kDa)
Nudt12-4(E374Q)	N-His	pET28a	1	462	462	52
Nudt12-5	N-His	pET28a	9	462	454	51
Nudt12-6	N-His	pET28a	102	462	361	41
Nudt12-7	N-His	pET28a	44	462	419	47
Nudt12-8	N-His	pET28a	76	462	387	43
Nudt12-9	C-His	pET26b	126	462	337	39
Nudt12-10(E374Q)	N-His	pET28a	126	462	337	39
Nudt12-11	N-His	pET28a	115	462	348	40
Nudt12-12	N-His	pET28a	121	462	342	40
Nudt12-13	N-His	pET28a	130	462	333	37
Nudt12-14	N-His	pET28a	135	462	328	37
Nudt12-15	N-His	pET28a	126	459	334	39
Nudt12-16	N-His	pET28a	115	459	345	40
Nudt12-17	N-His	pET28a	121	459	339	40
Nudt12-18	N-His	pET28a	130	459	330	37
Nudt12-19	N-His	pET28a	135	459	325	37
Nudt12-20	N-His	pET28a	279	462	184	21
XlNudt12-1	N-His	pET28a	1	458	458	53
XlNudt12-2	N-His	pET28a	122	458	377	38
XlNudt12-3	C-His	pET26b	1	458	458	53
XlNudt12-4	C-His	pET26b	122	458	337	38

Section 3.3.3 Surface engineering of protein for crystallization

Since the methods we mentioned above all failed to improve the quality of the crystals, we looked for other ways to modify the crystallization behavior of the protein.

Chemical Modification -- Reductive methylation of lysine residues

Lysine is a basic residue with a long side chain. This kind of flexible, solvent-exposed amino acid side chain can be disruptive to a well-ordered crystal lattice. Reductive methylation of lysines that are on the exposed surface of the protein can reduce surface entropy and improve crystal packing (Walter et al., 2006). However, in our case, this method did not help improve diffraction of the crystals.

Surface Entropy-reduction Genetic/Mutagenesis Approach

This approach is to mutate large, charged residues on the surface of proteins in order to reduce the entropic cost of forming ordered intermolecular crystal contacts and thus enhancing the crystallizability of proteins. Typically, large and charged side chains, such as lysine and glutamate are mutated to small and uncharged side chains, like alanine (Derewenda & Vekilov, 2006). The sites designed for mutagenesis were predicted by sequence analysis and the webserver, called SERp (Surface Entropy Reduction Prediction Server) (Goldschmidt, Cooper, Derewenda, & Eisenberg, 2007). The mutants made for this purpose are listed in Table 7. Most of the mutants did not effect the protein folding and produced soluble proteins except for Nudt12-24 (E294A/E295A, residue 126-462), which was completely insoluble. All of the soluble constructs were purified and used for crystallization screening. Although most of the constructs made no difference in crystallization, Nudt12-25 (E219A/E220A/E221A, residue 126-462) crystallized in a distinct condition from the previous ones that gave crystals. The condition contained 0.1 M HEPES pH 7.5 and 25% w/v PEG 2000MME. The crystal had a new shape which indicated a different packing. This crystal could diffract to 3-Å at home X-ray source. The reproduced crystals were able to diffract to 1. 6-Å at the Advanced Photon Source. The crystal

belonged to space group P1 with unit cell dimensions of a = 56.2 Å, b = 58.6 Å, c = 61.7 Å, α=102.4°, β = 115.2°, γ = 104.7°. We were able to collect a 1.6-Å dataset for structure determination.

Table 7 Surface entropy-reduction mutants

Name	Tag	Vector	Mutation Site	Start	End	Total Residues
Nudt12-21	N-His	pET28a	K166A/E167A	126	462	337
Nudt12-22	N-His	pET28a	E192A/K193A	126	462	337
Nudt12-23	N-His	pET28a	K261A/E262A	126	462	337
Nudt12-24	N-His	pET28a	E294A/E295A	126	462	337
Nudt12-25	N-His	pET28a	E219A/E220A/E221A	126	462	337
Nudt12-27	N-His	pET28a	K261A/E262A	1	462	462
Nudt12-28	N-His	pET28a	E294A/E295A	1	462	462
Nudt12-29	N-His	pET28a	E219A/E220A/E221A	1	462	462
Nudt12-30	N-His	pET28a	K105A/K106A	1	462	462

Section 3.3.4 Structure Determination

Although we collected a data set at 1.6 Å, the structure was not readily solvable. At the beginning, we tried molecular replacement using NudC as the search model. Using the program Phaser, we were able to find a solution. However, when we tried to build the model into the resulting electron density, we were only able to build in the CTD of Nudt12. The rest of the electron density is of poor quality maybe because those regions of Nudt12 share lower similarity with NudC.

In the structure of NudC, there is a zinc-binding motif where a zinc ion coordinates with four conserved cysteine residues (Cys98, Cys101, Cys116, and Cys119). Based on sequence alignment, we found that there are also four conserved cysteine residues (Cys284, Cys287, Cys302, and Cys307) in the corresponding region of Nudt12, indicating a zinc motif. To

determine whether Nudt12 binds a zinc ion, the zinc content of purified Nudt12 was measured using PAR (4-(2-Pyridylazo) resorcinol disodium salt). The absorbance at 500 nm increased after adding Nudt12 to a solution containing PAR indicates the formation of a PAR_2Zn^{2+} complex and the presence of zinc in Nudt12. Based on this information, we tried MR-SAD method to obtain more phase information from the original dataset.

As mentioned above, the CTD of Nudt12 was able to be located using the molecular replacement method. Using the phase information of this model, we calculated an anomalous difference electron density map. Even though the diffraction data was collected far above the zinc absorption edge, the map still clearly showed the positions of two zinc atoms (with 11σ peak heights). Combining the phase information from anomalous scattering with that from the molecular replacement, we finally got the map good enough to rebuild the whole model.

Section 3.3.5 Structural Insights into Nudt12

The structure we determined includes the catalytic domain (residues 126-462) of mouse Nudt12 in complex with AMP and 3 Mg^{2+} ions (Figure 32). The refined atomic model has good agreement with the X-ray diffraction data and the expected geometric parameters. 98.7% of the residues are in favored region of the Ramachandran plot. Although the surface mutations (E219A/E220A/E221A) seem to be necessary for crystals to form, there is not much electron density around this region and we can not explain how these mutations affect the crystal packing.

The Nudt12 catalytic domain consists of an N-terminal sub-domain (NTD, residue 126-282), a zinc-binding motif (residues 283-318, where Zn is bound by four Cys side chains), and a C-terminal sub-domain (CTD, residues 319-462) (Figure 27). The conserved Nudix sequence elements are located in the CTD, whereas both NTD and CTD adopt a Nudix-fold. Like bacterial

NudC, the catalytic domain of Nudt12 forms a homodimer. All three sub-domains are involved in the dimerization and forms an extensive interface.

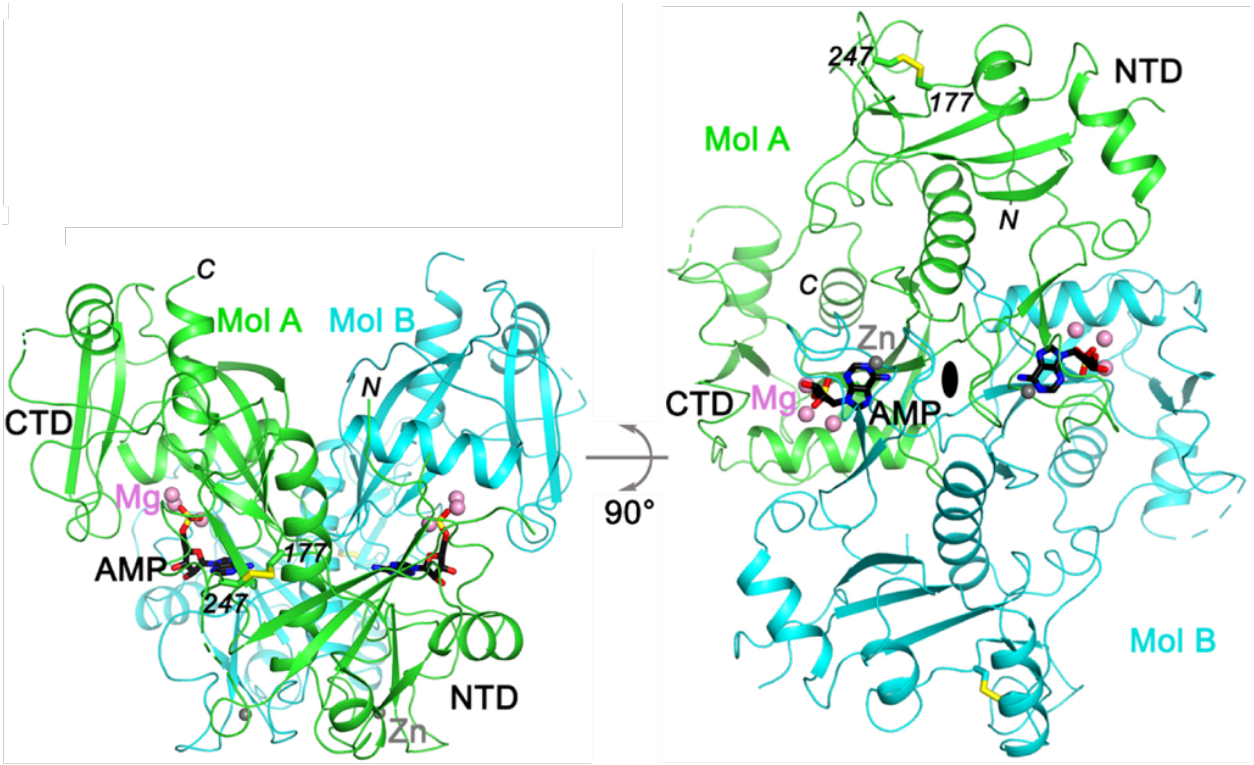


Figure 32 Structure of Nudt12 catalytic domain. Two views of the crystal structure at 1.6 Å resolution of mouse Nudt12 catalytic domain in complex with AMP and 3 Mg^{2+} ions. The two monomers are colored green and cyan, respectively. AMP is shown as stick models in black, Mg^{2+} ions as pink spheres, and Zn as gray spheres. A disulfide bond between residues 177 and 247, formed during crystallization, is indicated as stick models.

The conformations of the two monomers are essentially the same, with rms distance of 0.46-Å for 303 aligned Cα atoms (Figure 33) . The overall structure of Nudt12 is similar to that of NudC, but there are also substantial differences between them, especially for the NTD (Figure 33). The rms distance in Cα atoms for 226 aligned residues between Nudt12 and NudC is 2.0-Å.

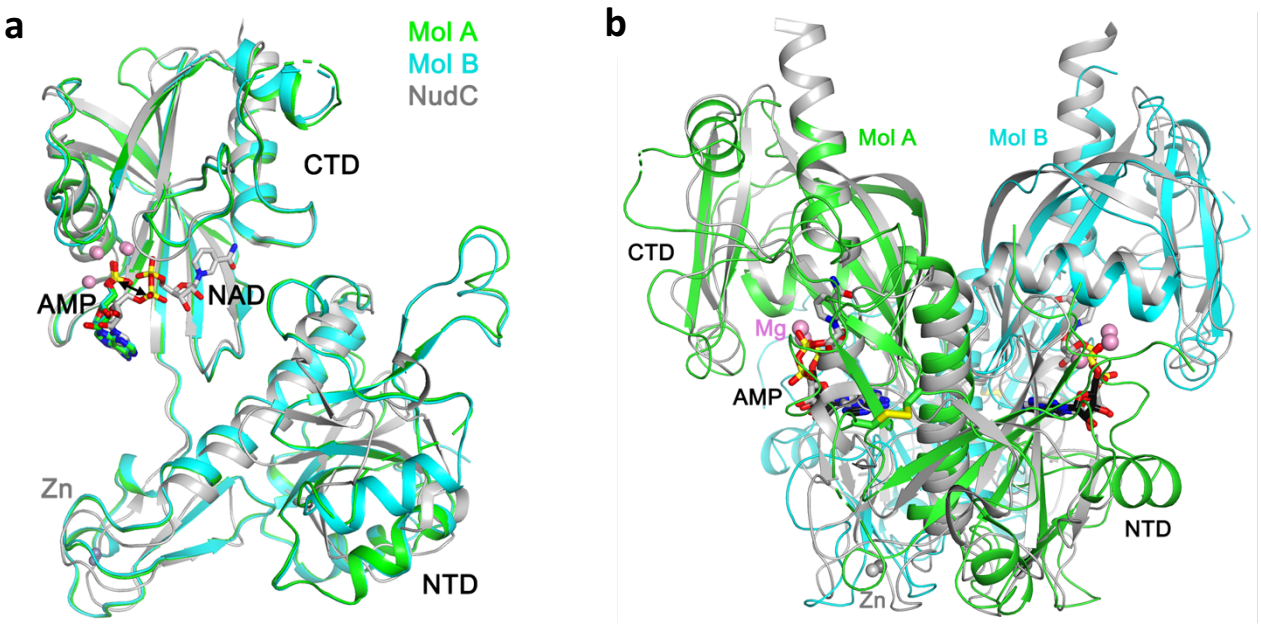


Figure 33 (a) Overlay of the structures of the Nudt12 monomers in complex with AMP (green and cyan) and Mg^{2+} (pink spheres) as well as NudC in complex with NAD (gray). (b) Overlay of the structures of the Nudt12 dimer (green and cyan for the two monomers) and the NudC dimer (gray).

Although Nudt12 was co-crystallized with NAD, the solved structure showed clear density for AMP instead of NAD at the active site (Figure 34). AMP is the hydrolysis product of NAD by Nudt12. Our structure demonstrates that the construct (E219A/E220A/E221A, residues 126-462) of Nudt12 we used for crystallization possesses catalytic activity. The adenosine is in the syn configuration and the adenine base is π-stacked with the side chain of Phe356 on one face and on the other face with the side chain of Tyr318 in the loop connecting the zinc-binding motif and the CTD of the other protomer (Figure 34). The hydroxyls of AMP are exposed to the surface, which leaves substantial space to accommodate the RNA body in the NAD-capped RNA.

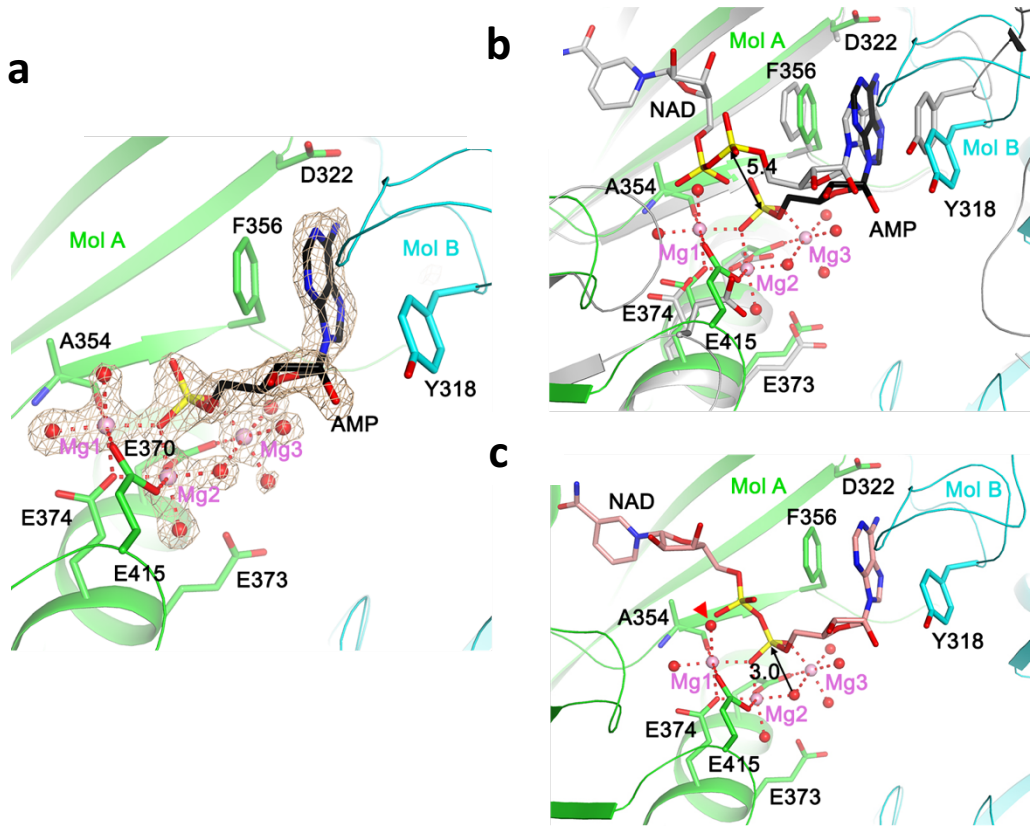


Figure 34 (a) Detailed binding mode of AMP and Mg^{2+} ions in Nudt12. The coordination sphere of each metal ion is indicated with the red dashed lines. Water molecules are shown as red spheres. The omit F_o-F_c electron density at 1.6 Å resolution for AMP, Mg^{2+} ions and their water ligands is shown in wheat color, contoured at 5 σ. (b) Comparison to the binding mode of NAD in NudC. Overlay of the structure of Nudt12 (in color) in complex with AMP (black) and the three Mg^{2+} ions (pink) with that of NudC in complex with NAD (gray). The position of the phosphate of AMP is separated by 5.4 Å from the equivalent phosphate of NAD, indicated with the black arrow. (c) Molecular mechanism of the deNADding reaction. A model of the binding mode of NAD is shown (salmon color). The AMP portion of the model is identical to the crystal structure, and the NMN portion is based on that in the NudC complex. The nucleophilic attack by the bridging ligand of Mg2 and Mg3 on the phosphate is indicated by the black arrow, which initiates the deNADding reaction. The water molecule that is displaced in the NAD complex is indicated with arrowhead (red).

There is a large network of interactions between the phosphate group of AMP and Nudt12, which is mediated through three Mg^{2+} ions (named Mg1, Mg2 and Mg3) (Figure 34a).

One of the terminal oxygen atoms of the phosphate is a bridging ligand to Mg1 and Mg2, while a second terminal oxygen atom is coordinated to Mg3. Glu370, Glu374 and Glu415 in the Nudix motif are ligands of the Mg^{2+} ions. The main-chain carbonyl of Ala354 is coordinated to Mg1. Each Mg^{2+} is coordinated octahedrally, and the coordination spheres are completed by seven water molecules or hydroxide ions. One water molecule is a bridging ligand between Mg2 and Mg3 (Figure 34a).

The binding mode of AMP in our structure has significant differences compared to that of the AMP portion of NAD in the complex with NudC (Figure 34b) (Hofer et al., 2016). The distance between the phosphorus atoms in the two structure is 5.4 Å (Figure 34b). This difference is probably due to the absence of metal ions in the NudC complex, so that the phosphate group is pushed away from the Glu side chains in the active site. In addition, the adenine is in the *anti* configuration in the NudC complex, and many residues in the active site region have different conformations as well, especially the residue equivalent to the Glu415 ligand in Nudt12 (Figure 34b). A model for the binding mode of NAD to Nudt12 could be built based on the NMN portion of NudC-NAD structure (Figure 34c). In this model, one of the water molecules in the current structure can be displaced by the terminal oxygen atom on the phosphate in NMN, and the oxygen atom can be directly coordinated to Mg1.

The model of Nudt12-NAD complex provides clear molecular insights into the deNADding mechanism. The bridging hydroxide ion between Mg2 and Mg3 is located directly beneath the phosphate group of AMP, with a distance of 3.0 Å to the phosphorus atom (Figure 34c). This hydroxide ion attacks the phosphorus atom and causes breakage of the pyrophosphate bond. The oxygen atom of AMP that is connected to the NMN is the leaving group. The oxyanion of the leaving group does not seem to be stabilized by a general acid in either Nudt12

or NudC. A solvent molecule might become bound in the presence of NAD and stablize the oxyanion of the leaving group.

Section 3.3.6 *In vitro* Assay for Nudt12 deNADding Activity

To test whether Nudt12 possesses deNADding activity, we collaborated with the Kiledjian Lab at Rutgers University. In humans, both Nudt12 and Nudt13 are reported with hydrolysis activity on free NAD. They both contain the SQPWPFPxS sequence motif found in the conserved Nudix domain common in NADH diphosphatases (Abdelraheim et al., 2003; Frick & Bessman, 1995). To test their deNADding activity in vitro, mouse Nudt12 or Nudt13 was incubated with either ^{32}P labeled NAD (N_{ic}pp*A; asterisk denotes ^{32}P label) or in vitro transcribed ^{32}P labeled RNA capped with NAD (N_{ic}pp*A-RNA). Both Nudt12 and Nudt13 can hydrolyze free NAD into NMN (N_{ic}p) and AMP (p*A) which can be detected by thin-layer chromatography (TLC) (Figure 35a). To determine whether these proteins can hydrolyze NAD-RNA, the reaction products were further treated with nuclease P1, which cleaves all phosphodiester bonds within an RNA and release the detectable labeled p*A. Based on the results, Nudt12 possessed NAD cap deNADding activity in vitro, whereas Nudt13 could not hydrolyze NAD-capped RNA (Figure 35b). The results showed that Nudt12 cleaved the NAD-capped RNA between the diphosphate linkage, similar to the activity observed with NudC (Cahova et al., 2015) and distinct from that of DXO (Jiao et al., 2017), which is consistent with the molecular insights observed from the structure. The Nudt12 decapping activity on m^{7}G-capped RNA was also compared to the deNADding activity on NAD-capped RNA. The result showed that Nudt12 has greater deNADding activity than decapping activity (Figure 35c). As the

structure shows, the glutamate residues in the active site of Nudt12 are crucial for the hydrolysis. Mutating two glutamate residues in the active site into glutamine abolished the deNADding activity (Figure 35c).

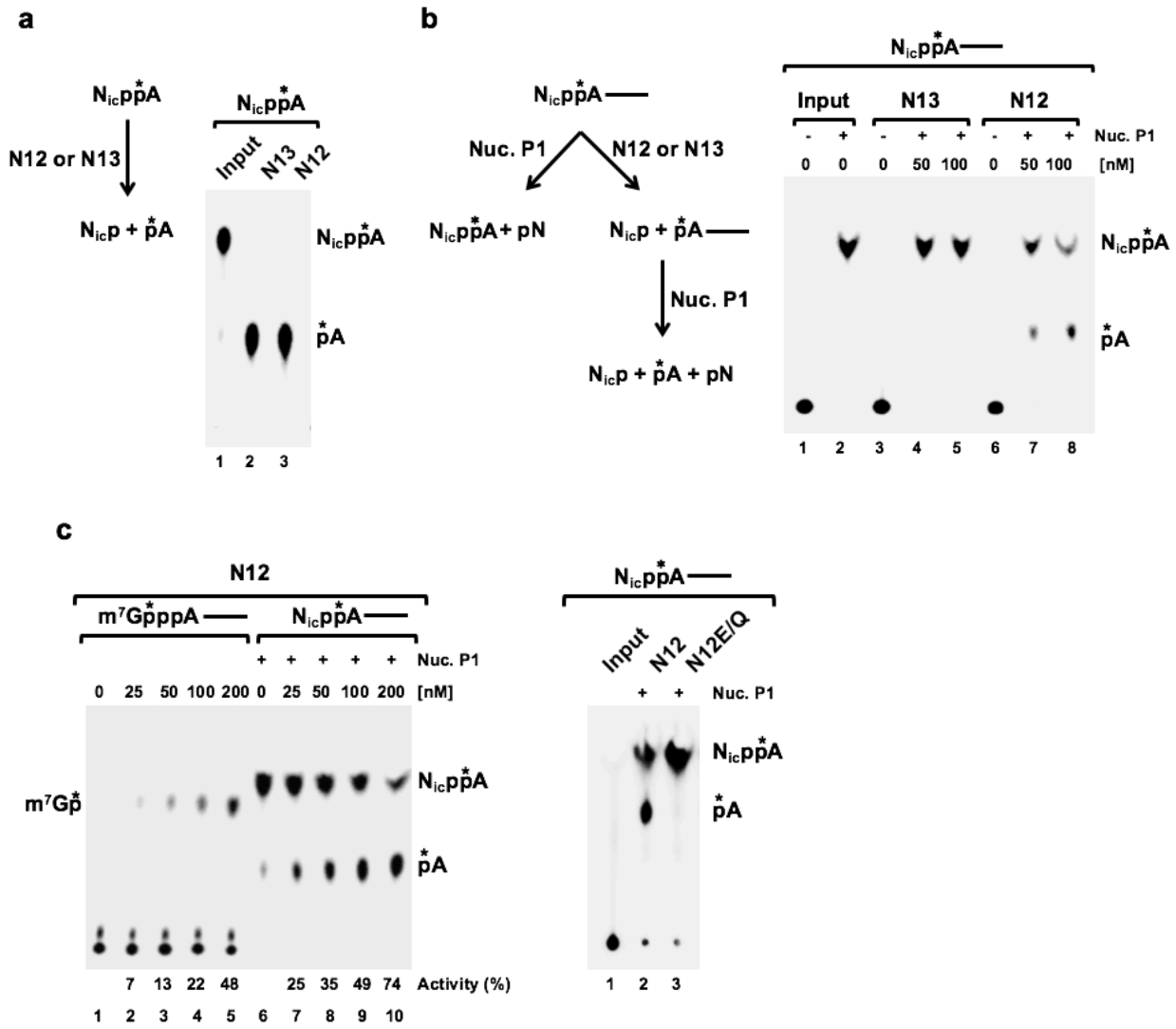


Figure 35 Mouse Nudt12 deNADding activity in vitro. N_{ic} denotes nicotinamide. The line represents the RNA. (a) Mouse Nudt12 and Nudt13 enzymatic activity on ^{32}P-labeled free NAD^+ (N_{ic}pp*A). (b) Mouse Nudt12 and Nudt13 enzymatic activity on ^{32}P-labeled NAD^+ -capped RNA (N_{ic}pp*A------). (c) In vitro decapping/deNADding assays with Mouse Nudt12 and indicated ^{32}P-cap-labeled substrates. N12E/Q represents the catalytically inactive double-mutant Nudt12 E373Q/E374Q.

Section 3.3.7 Comparison of the Active Site of *E. coli* RppH with that of Nudt12

We compared the active site of Nudt12 with that of E. coli RppH (RNA pyrophosphohydrolase, (Deana, Celesnik, & Belasco, 2008; Vasilyev & Serganov, 2015)) and found that the binding mode of the Mg^{2+} ions and the reaction mechanism of the two proteins are generally similar. For RppH, an Arginine residue (Arg8) stabilizes the oxyanion of the leaving group (Figure 36). The RppH Arg8 is equivalent to Asp322 in Nudt12 (Figure 34a), and its guanidinium group would clash with the side chain of Phe356 in Nudt12 (Figure 36). Unlike Nudt12, RppH is monomeric. It covers only the CTD of Nudt12 and lacks the NTD and the zinc-binding motif.

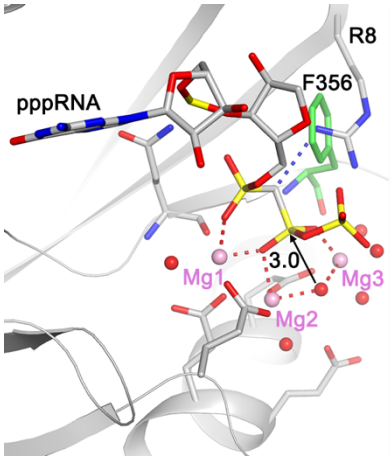


Figure 36 Comparison to the substrate binding mode and reaction mechanism of RppH. The structure of RppH in complex with pppRNA is shown (Serganov, 2015), in the same orientation as that for Nudt12 (Figure 34c). Residue Phe356 in Nudt12 is shown (green), clashing with the side chain of Arg8 in RppH.

Section 3.3.8 *E. coli* RppH Possesses RNA deNADding Activity *in vitro*

E. coli RppH was reported to have no hydrolytic activity on NAD-capped RNA (Cahova et al., 2015). So RppH was initially used as a negative control in the assays to determine Nudt12

and NudC *in vitro* hydrolysis activity on NAD, NAD-capped RNA and m^7G-capped RNA. Surprisingly, although the bacterial RppH could not hydrolyze free NAD, it showed robust deNADding activity on NAD-capped RNA (Figure 37a). The disagreement with previous reports may be caused by different *in vitro* parameters and/or different substrate RNA sequences employed in the two studies. RppH removes the NAD cap by hydrolyzing within the diphosphate bond, which is similar to the activity of Nudt12 and NudC. The *in vitro* assay indicates that RppH is a potential deNADding enzyme in bacteria.

The deNADding activity of RppH might be interpreted by modeling NAD into RppH active site. The binding mode for NAD in RppH would not be the same as the one in Nudt12 because the nicotinamide portion would clash with the protein. The binding mode of pppRNA to RppH suggested that a folded conformation of NAD (Figure 37b) could fit in the RppH active site. The adenine nucleotide of NAD would assume a conformation similar to that of the first nucleotide of pppRNA. The nicotinamide of NAD could have cation-π interactions with Arg8 of RppH. The RNA body could bind RppH in the same way for deNADding as for PPH activity. It is reported that the pyrophosphohydrolase acitivity of *E. coli* RppH has a preference for a guanine at the second position of the substrate (Vasilyev & Serganov, 2015), suggesting that the deNADding activity would prefer a guanine nucleotide at the first position of the substrate RNA. This model also suggests that RppH is only active on NAD-RNA rather than NAD alone because the RNA body is important for substrate binding to RppH.

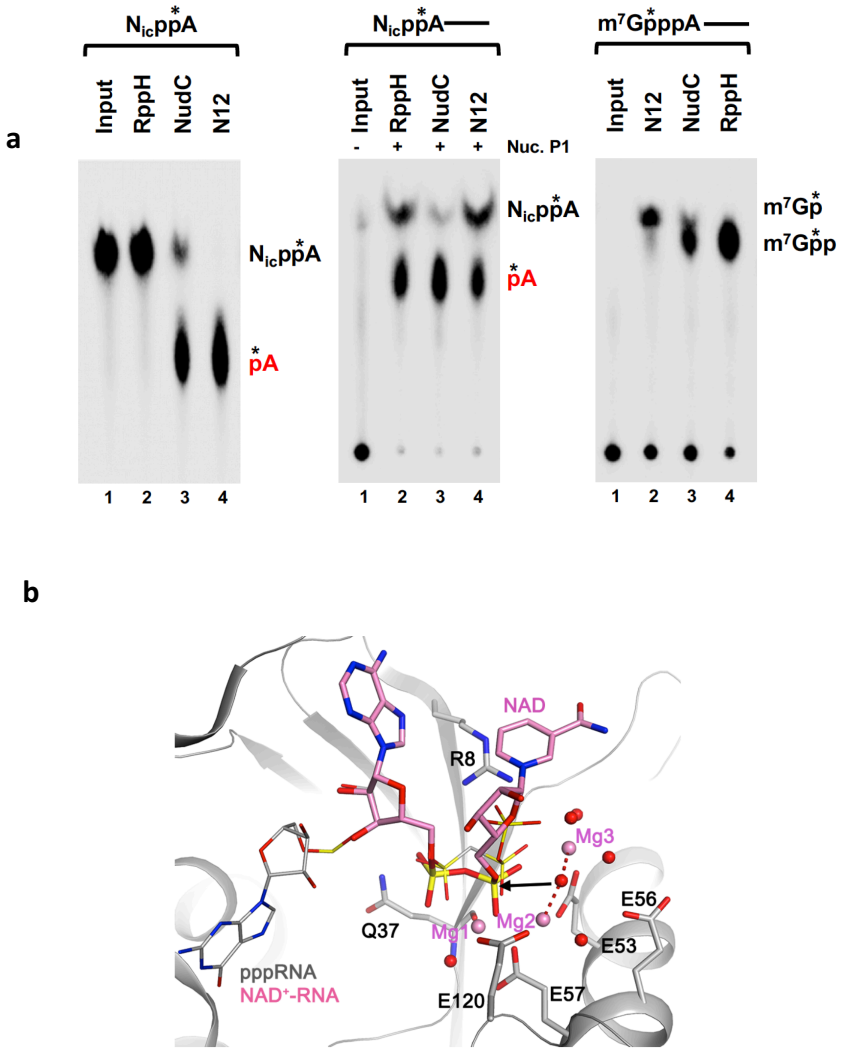


Figure 37 RppH has RNA deNADding activity in vitro. (a) In vitro decapping assays of RppH, NudC and Nudt12 with ^{32}P-labeled substrates: free NAD (left panel), NAD-capped RNA (middle panel) and m^{7}G-capped RNA (right panel). (b) A model for the binding mode of NAD to RppH and the molecular mechanism for its deNADding activity. The AMP portion is based on the first nucleotide of pppRNA (gray) in RppH. The amide group of nicotinamide could be recognized by hydrogen-bonding interactions (dashed lines in red). Arg8 could have cation-π interactions with the adenine base as well as stabilize the leaving group.

Section 3.3.9 *In vivo* Assays for Nudt12 deNADding Activity

The first assay to determine Nudt12 function in mammalian cells is to test whether the stability of NAD-capped RNAs is altered in the absence of Nudt12. HEK293T cell lines with individual or double knockout of *Nudt12* or *DXO gene* were generated by CRISPR/Cas9n (Figure 38a). ^{32}P labeled NAD-capped transcripts or m^{7}G-capped transcripts were transfected into the knockout HEK293T cell lines. After transfection, RNAs were harvested and detected over the time course. As shown in Figure 38b, NAD$^+$-capped RNAs were more stable in N12-KO (Nudt12 knockout) than in control cells. In contrast, m^{7}G-capped RNA showed similar half life in both cell line backgrounds (Figure 38c), which indicates that Nudt12 modulates the stability of NAD-capped RNAs rather than that of m^{7}G-capped RNAs transfected in HEK293T cells.

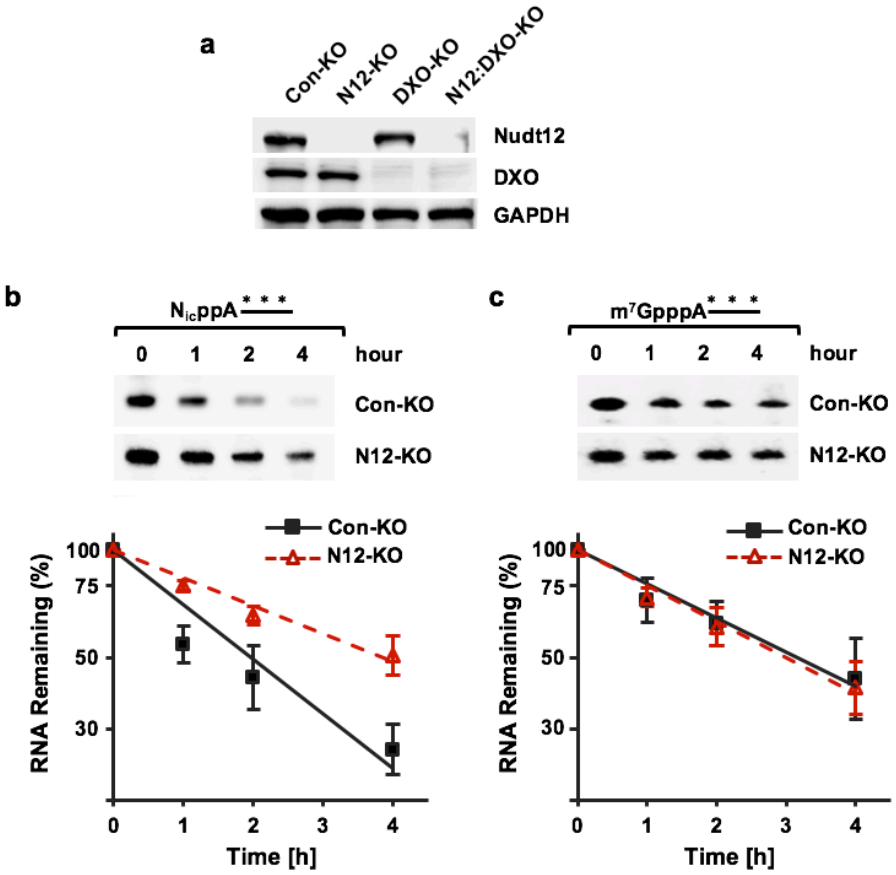


Figure 38 The stability of NAD^+-capped RNA in Nudt12 knockout cells. (a) Western blot of Nudt12 and DXO protein levels in HEK293T control knockout (Con-KO), Nudt12 knockout (N12-KO), DXO knockout (DXO-KO) or double knockout (N12:DXO-KO) cell lines. GAPDH was used as an internal control. (b) Remaining ^{32}P labeled NAD+-capped RNAs, or (c) m^7G-capped RNAs after transfection into knockout HEK293T cells.

The second assay is to determine whether Nudt12 regulates expression of endogenous NAD-capped RNA in cells. The approach used here to detect en masse NAD caps is called NAD-CapQ (Figure 39a). This approach combines enzymatic properties of Nuclease P1 with a colorimetric NAD/NADH Quantitation to detect total NAD and NADH levels. RNA from Con-KO or N12-KO cells were measured by NAD-CapQ. In N12-KO cells, a 1.5-fold increase in total NAD-capped RNA was detected compared to the control (Figure 39b). A similar increase

was observed in DXO-KO cells as well (Figure 39b). The cell line harboring a double knockout of both Nudt12 and DXO (N12: DXO-KO; Figure 38a) resulted a 2.7-fold higher level of NAD caps compared to Con-KO cells (Figure 39b). These data demonstrated that in addition to DXO, Nudt12 is also a deNADding enzyme in cells. Since the double knockout of both Nudt12 and DXO resulted in the highest level of NAD-capped RNA, the two enzymes appear to function on distinct NAD-capped RNA substrates.

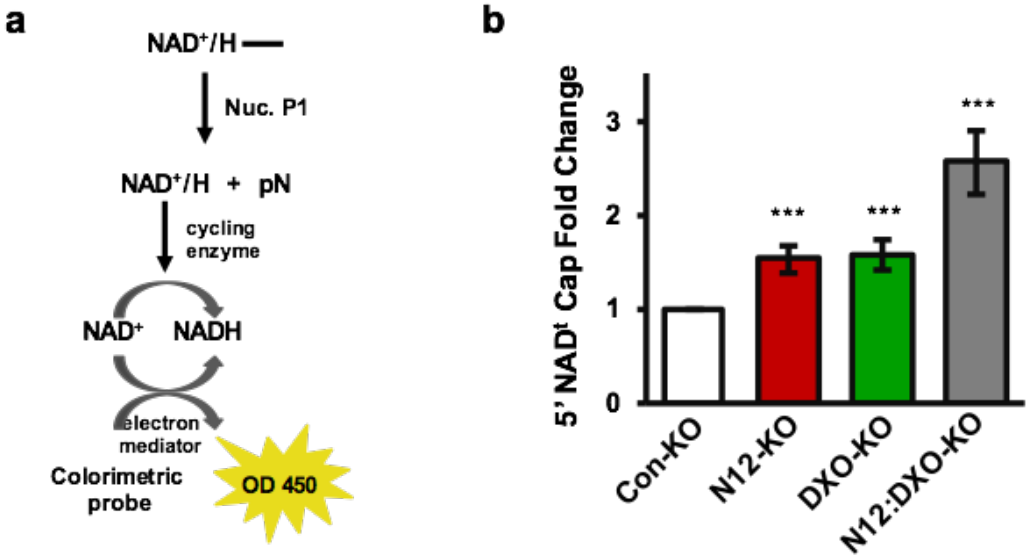


Figure 39 Nudt12 regulates the level of endogenous NAD-capped RNA in cells. (a) Schematic of NAD-CapQ assay. (b) NAD-capped RNA levels in different knockout cells.

Section 3.3.10 Nudt12 Preferentially Targets a Subset of mRNAs for deNADding

The knockout results shown in Section 3.3.9 indicates that Nudt12 may preferentially target a subset of NAD-capped mRNA for deNADding. To further test this hypothesis, our collaborator utilized the NAD captureSeq approach to identify NAD-capped RNAs enriched in HEK293T N12-KO cells. The result showed that 188 NAD-capped RNAs were specifically enriched in the N12-KO cells. These enriched RNAs belong to five major pathways, three

biological process (BP) terms and two cellular compartment (CC) terms (Figure 40). One of these is associated with rRNA processing while the remaining are linked with mitochondrial metabolism or translation. The mRNA enriched in N12-KO cells are primarily involved in respiratory functions, which suggests an interaction between cell metabolism and Nudt12 decapping of NAD-capped transcripts.

To further validate Nudt12 targeted RNAs, 12 RNAs (5 RNAs in oxidative phosphorylation, 7 random RNAs) were chosen for direct quantitative reverse-transcription (qRT)-PCR (Figure 41). Among the 12 RNAs, 11 were validated by direct qRT-PCR. CKS2 mRNA was not a target of Nudt12 based on NAD captureSeq, and qRT-PCR showed no change of CKS2 RNA level in N12-KO cells compared to WT cells. Seven of the qRT-PCR validated genes (COX17, MRPL15, MRPS23, NDUFAF4, NDUFB2, NDUFB9, and NDUFS3) were found in the enriched GO terms in Figure 40. All together, we verified Nudt12 as a deNADding enzyme in cells and the RNAs enriched in N12-KO cells suggest a role for Nudt12 in NAD-capped transcripts encoding proteins with mitochondrial functions.

mRNAs enriched in N12-KO cells were relatively distinct from those enriched in DXO-KO cells (Jiao et al., 2017). Only 13 out of the 188 mRNAs found enriched in N12-KO cells overlap the 67 mRNAs enriched in DXO-KO cells. The hierarchical clustering of all 242 transcripts by replicate samples indicates different patterns of transcripts enriched in Nudt12 or DXO (Figure 41). The Nudt12-enriched GO terms are similarly distinct from DXO-enriched GO terms, highlighted in the left colorbar (Figure 41). A major class of NAD-capped RNAs enriched exclusively in N12-KO cells is involved in mitochondrial metabolism. In contrast, the most prominent NAD-capped RNAs elevated in DXO-KO cells are the small nucleolar RNAs (snoRNAs) and the related small Cajal body RNAs (scaRNAs) (Jiao et al., 2017) (Figure 41),

which did not increase in cells lacking Nudt12. NAD-capped mRNA involved in oxidative phosphorylation were also specifically enriched in N12-KO cells (Figure 41). These analyses revealed that there are NAD-capped RNAs regulated by Nudt12 or DXO in mammalian cells. Each deNADding enzyme targets a distinct subset of mRNAs.

GO	Term	Abbr	Count	%	p-value	Fold Enrichment	FDR
BP	GO:0070125~mitochondrial translational elongation	MTE	11	5.85	1.76E-08	12.42	2.72E-05
BP	GO:0070126~mitochondrial translational termination	MTT	11	5.85	1.98E-08	12.27	3.05E-05
BP	GO:0006364~rRNA processing	RIB	14	7.45	3.64E-07	6.28	5.62E-04
CC	GO:0005739~mitochondrion	Mito	53	28.19	1.00E-18	4.05	1.30E-15
CC	GO:0005743~mitochondrial inner membrane	Mito	32	17.02	4.18E-18	7.39	5.41E-15

Figure 40 Top gene ontology (GO)-biological process (BP) and cellular component (CC) terms enriched with 188 genes increased in Nudt12-KO over control (FDR<5%; > 2-fold increased, and > 1 FPKM in Nudt12-KO). GO terms were filtered for those with <5% FDR and at least 10 genes per term.

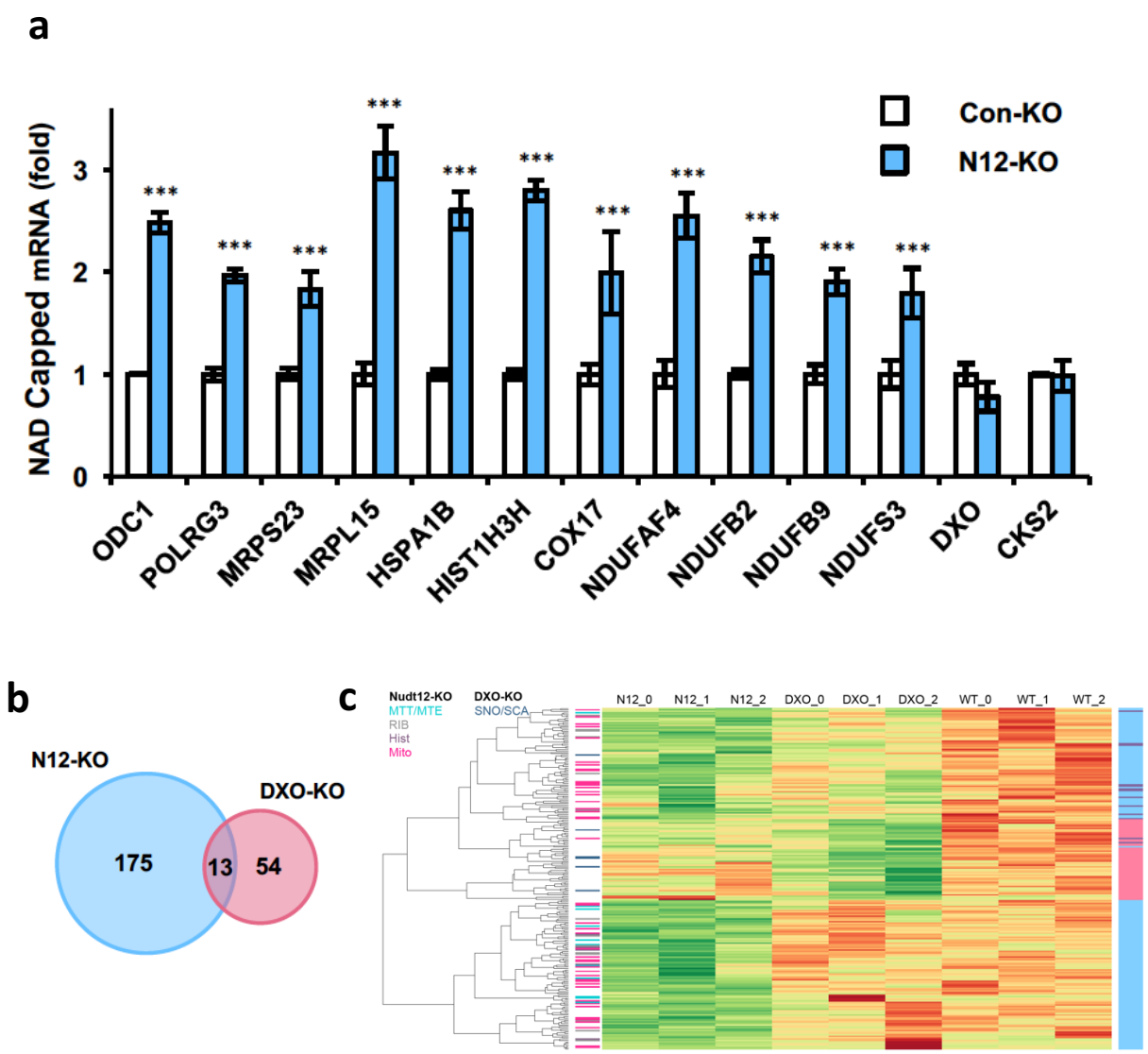


Figure 41 Nudt12 preferentially targets a subset of mRNAs for deNADding. (a) qRT-PCR validation of NAD-capped mRNAs in N12-KO cells. Data are presented relative to the HEK293T Con-KO cells and set to 1. Error bars represent ± SD. p values are denoted by asterisks; (***) p< 0.001 (Student's t test). (b) A Venn Diagram of NAD-capped RNAs enriched in N12-KO and DXO-KO cells (≥2-fold, ≤5% FDR, >1 FPKM in HEK293T WT cells). (c) Heatmap of mRNAs enriched in either N12-KO or DXO-KO. The color bar at left indicates enriched gene groups, either from top GO biological processes (Figure 40) or presence of major gene families as indicated (Hist: histones, SNO/SCA: snoRNAs or scaRNAs). The color bar at right indicates membership in components of Venn diagram (Figure 41b). For each mRNA in the heatmap, green indicates relative enrichment, red indicates relative depletion, with expression normalized for each mRNA across all samples to indicate relative differences. Individual replicates samples from each group are indicated by a trailing number (" _0", "_1", etc.).

Section 3.4 Discussion

The NAD-capped RNA has been demonstrated to exist not only in prokaryotes but also eukaryotes. The evolutionary conservation of the noncanonical NAD cap implies a function in RNA metabolism. In mammalian cells, there are multiple m^7G cap decapping enzymes (Grudzien-Nogalska & Kiledjian, 2017). Our studies demonstrated there are also multiple deNADding enzymes in cells. In mammalian cells, in addition to DXO, Nudt12 Nudix hydrolase is also a deNADding enzyme and targets a distinct subset of mRNA for deNADding. The crystal structure of mouse Nudt12 catalytic domain (residue 126-462) defines the binding mode of the AMP product and the three coordinated magnesium ions. The structure provides detailed insights into the molecular mechanism of deNADding reaction. Different from DXO that releases an intact NAD, Nudt12 hydrolyzes the diphosphate bond within the NAD to release NMN from NAD-capped RNA. A major class of NAD-capped RNA that were regulated by Nudt12 are involved in mitochondrial metabolism which supports a correlation between Nudt12 deNADding activity and cellular energetics.

The identification of Nudt12 and *E. coli* RppH as deNADding enzymes indicates that there are at least three classes of deNADding enzymes. The first is the DXO family of proteins that remove the intact NAD from the 5'-end of the capped RNA (Jiao et al., 2017). The second is Nudt12/NudC-like proteins that cleave within the pyrophosphate of both NAD and NAD-capped RNA. The third one contains RppH, which does not hydrolyze NAD, but can cleave NAD-capped RNA. In mammalian cells, DXO and Nudt12 target different subsets of RNA substrates for deNADding. The question how these enzymes differentiate between different RNAs for

specific deNADding remains unknown. One possibility is that deNADding enzymes can selectively bind to the 5'-end of specific RNAs. The other possibility is that NAD-RNA is specifically recognized by different RNA-binding proteins. The deNADding enzyme might later be recruited to particular substrates by different RNA-binding proteins.

REFERENCES

Abdelraheim, S. R., Spiller, D. G., & McLennan, A. G. (2003). Mammalian NADH diphosphatases of the Nudix family: cloning and characterization of the human peroxisomal NUDT12 protein. *Biochemical Journal, 374*, 329-335. doi:10.1042/Bj20030441

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., . . . Terwilliger, T. C. (2002a). PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr, 58*(Pt 11), 1948-1954.

Adams, P. D., Grosse-Kunstleve, R. W., Hung, L. W., Ioerger, T. R., McCoy, A. J., Moriarty, N. W., . . . Terwilliger, T. C. (2002b). PHENIX: building new software for automated crystallographic structure determination. *Acta Crystallographica Section D-Biological Crystallography, 58*, 1948-1954. doi:10.1107/S0907444902016657

Akhtar, M. S., Heidemann, M., Tietjen, J. R., Zhang, D. W., Chapman, R. D., Eick, D., & Ansari, A. Z. (2009). TFIIH Kinase Places Bivalent Marks on the Carboxy-Terminal Domain of RNA Polymerase II. *Molecular Cell, 34*(3), 387-393. doi:10.1016/j.molcel.2009.04.016

Albrecht, R., & Zeth, K. (2011). Structural Basis of Outer Membrane Protein Biogenesis in Bacteria. *Journal of Biological Chemistry, 286*(31), 27792-27803. doi:10.1074/jbc.M111.238931

Albrecht, T. R., Shevtsov, S. P., Wu, Y. X., Mascibroda, L. G., Peart, N. J., Huang, K. L., . . . 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, 46*(8), 4241-4255. doi: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. doi:10.1128/Mcb.06511-11

Amodeo, G. A., Rudolph, M. J., & Tong, L. (2007). Crystal structure of the heterotrimer core of Saccharomyces cerevisiae AMPK homologue SNF1. *Nature, 449*(7161), 492-U413. doi:10.1038/nature06127

Aravind, L. (1999). An evolutionary classification of the metallo-beta-lactamase fold proteins. *In Silico Biol, 1*(2), 69-91.

Armon, A., Graur, D., & Ben-Tal, N. (2001). ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of Molecular Biology, 307*(1), 447-463. doi:DOI 10.1006/jmbi.2000.4474

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

Baillat, D., & Wagner, E. J. (2015). Integrator: surprisingly diverse functions in gene expression. *Trends in Biochemical Sciences, 40*(5), 257-264. doi:10.1016/j.tibs.2015.03.005

Bird, J. G., Zhang, Y., Tian, Y., Panova, N., Barvik, I., Greene, L., . . . Nickels, B. E. (2016). The mechanism of RNA 5 ' capping with NAD(+), NADH and desphospho-CoA. *Nature, 535*(7612), 444-+. doi:10.1038/nature18622

Cahova, H., Winz, M. L., Hofer, K., Nubel, G., & Jaschke, A. (2015). NAD captureSeq indicates NAD as a bacterial cap for a subset of regulatory RNAs. *Nature, 519*(7543), 374-+. doi:10.1038/nature14020

Callebaut, I., Moshous, D., Mornon, J. P., & de Villartay, J. P. (2002a). Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Res, 30*(16), 3592-3601.

Callebaut, I., Moshous, D., Mornon, J. P., & de Villartay, J. P. (2002b). Metallo-beta-lactamase fold within nucleic acids processing enzymes: the beta-CASP family. *Nucleic Acids Research, 30*(16), 3592-3601. doi:DOI 10.1093/nar/gkf470

Carreras-Puigvert, J., Zitnik, M., Jemth, A. S., Carter, M., Unterlass, J. E., Hallstrom, B., . . . Helleday, T. (2017). A comprehensive structural, biochemical and biological profiling of the human NUDIX hydrolase family. *Nature Communications, 8*. doi:ARTN 1541
10.1038/s41467-017-01642-w

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

Chen, J. D., & Wagner, E. J. (2010). snRNA 3 ' end formation: the dawn of the Integrator complex. *Biochemical Society Transactions, 38*, 1082-1087. doi:10.1042/Bst0381082

Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., . . . Gingeras, T. R. (2005). Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science, 308*(5725), 1149-1154. doi:10.1126/science.1108625

Cudney, B., Patel, S., Weisgraber, K., & Newhouse, Y. (1994). Screening and Optimization Strategies for Macromolecular Crystal-Growth. *Acta Crystallographica Section D-Biological Crystallography, 50*, 414-4123.

de la Sierra-Gallay, I. L., Zig, L., Jamalli, A., & Putzer, H. (2008). Structural insights into the dual activity of RNase J. *Nature Structural & Molecular Biology, 15*(2), 206-212. doi:10.1038/nsmb.1376

Deana, A., Celesnik, H., & Belasco, J. G. (2008). The bacterial enzyme RppH triggers messenger RNA degradation by 5' pyrophosphate removal. *Nature, 451*(7176), 355-U314. doi:10.1038/nature06475

Derewenda, Z. S., & Vekilov, P. G. (2006). Entropy and surface engineering in protein crystallization. *Acta Crystallographica Section D-Biological Crystallography, 62*, 116-124. doi:10.1107/S0907444905035237

Dominski, Z., Yang, X. C., & Marzluff, W. F. (2005). The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing. *Cell, 123*(1), 37-48. doi:DOI 10.1016/j.cell.2005.08.002

Dominski, Z., Yang, X. C., Purdy, M., Wagner, E. J., & Marzluff, W. F. (2005). A CPSF-73 homologue is required for cell cycle progression but not cell growth and interacts with a protein having features of CPSF-100. *Molecular and Cellular Biology, 25*(4), 1489-1500. doi:10.1128/Mcb.25.4.1489-1500.2004

Egloff, S., O'Reilly, D., & Murphy, S. (2008). Expression of human snRNA genes from beginning to end. *Biochemical Society Transactions, 36*, 590-594. doi:10.1042/Bst0360590

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. doi: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. doi:10.1016/j.molcel.2011.11.006

Emsley, P., & Cowtan, K. (2004a). Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr, 60*(Pt 12 Pt 1), 2126-2132. doi:10.1107/S0907444904019158

Emsley, P., & Cowtan, K. (2004b). Coot: model-building tools for molecular graphics. *Acta Crystallographica Section D-Biological Crystallography, 60*, 2126-2132. doi:10.1107/S0907444904019158

Erijman, A., Dantes, A., Bernheim, R., Shifman, J. M., & Peleg, Y. (2011). Transfer-PCR (TPCR): a highway for DNA cloning and protein engineering. *J Struct Biol, 175*(2), 171-177. doi:10.1016/j.jsb.2011.04.005

Frick, D. N., & Bessman, M. J. (1995). Cloning, purification, and properties of a novel NADH pyrophosphatase. Evidence for a nucleotide pyrophosphatase catalytic domain in MutT-like enzymes. *Journal of Biological Chemistry, 270*(4), 1529-1534.

Gaertner, B., & Zeitlinger, J. (2014). RNA polymerase II pausing during development. *Development, 141*(6), 1179-1183. doi:10.1242/dev.088492

Gardini, A., Baillat, D., Cesaroni, M., Hu, D. Q., Marinis, J. M., Wagner, E. J., . . . Shiekhattar, R. (2014). Integrator Regulates Transcriptional Initiation and Pause Release following Activation. *Molecular Cell, 56*(1), 128-139. doi:10.1016/j.molcel.2014.08.004

Goldschmidt, L., Cooper, D. R., Derewenda, Z. S., & Eisenberg, D. (2007). Toward rational protein crystallization: A Web server for the design of crystallizable protein variants. *Protein Science, 16*(8), 1569-1576. doi:10.1110/ps.072914007

Gouet, P., Courcelle, E., Stuart, D. I., & Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics, 15*(4), 305-308. doi:DOI 10.1093/bioinformatics/15.4.305

Grudzien-Nogalska, E., & Kiledjian, M. (2017). New insights into decapping enzymes and selective mRNA decay. *Wiley Interdisciplinary Reviews-Rna, 8*(1). doi:ARTN e1379 10.1002/wrna.1379

Guiro, J., & Murphy, S. (2017). Regulation of expression of human RNA polymerase II-transcribed snRNA genes. *Open Biology, 7*(6). doi:ARTN 170073 10.1098/rsob.170073

Hata, T., & Nakayama, M. (2007). Targeted disruption of the murine large nuclear KIAA1440/Ints1 protein causes growth arrest in early blastocyst stage embryos and eventual apoptotic cell death. *Biochimica Et Biophysica Acta-Molecular Cell Research, 1773*(7), 1039-1051. doi:10.1016/j.bbamcr.2007.04.010

Hendrickson, W. A., Horton, J. R., & Lemaster, D. M. (1990). Selenomethionyl Proteins Produced for Analysis by Multiwavelength Anomalous Diffraction (Mad) - a Vehicle for Direct Determination of 3-Dimensional Structure. *Embo Journal, 9*(5), 1665-1672.

Henras, A. K., Plisson-Chastang, C., O'Donohue, M. F., Chakraborty, A., & Gleizes, P. E. (2015). An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdisciplinary Reviews-Rna, 6*(2), 225-242. doi:10.1002/wrna.1269

Hofer, K., Li, S., Abele, F., Frindert, J., Schlotthauer, J., Grawenhoff, J., . . . Jaschke, A. (2016). Structure and function of the bacterial decapping enzyme NudC. *Nat Chem Biol, 12*(9), 730-734. doi:10.1038/nchembio.2132

Holm, L., Kaariainen, S., Rosenstrom, P., & Schenkel, A. (2008). Searching protein structure databases with DaliLite v.3. *Bioinformatics, 24*(23), 2780-2781. doi:10.1093/bioinformatics/btn507

Jancarik, J., Scott, W. G., Milligan, D. L., Koshland, D. E., & Kim, S. H. (1991). Crystallization and Preliminary-X-Ray Diffraction Study of the Ligand-Binding Domain of the Bacterial Chemotaxis-Mediating Aspartate Receptor of Salmonella-Typhimurium. *Journal of Molecular Biology, 221*(1), 31-34. doi:Doi 10.1016/0022-2836(91)90798-B

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. doi:10.1016/j.bbagrm.2008.04.001

Jiao, X., Doamekpor, S. K., Bird, J. G., Nickels, B. E., Tong, L., Hart, R. P., & Kiledjian, M. (2017). 5 ' End Nicotinamide Adenine Dinucleotide Cap in Human Cells Promotes RNA Decay through DXO-Mediated deNADding. *Cell, 168*(6), 1015-+. doi:10.1016/j.cell.2017.02.019

Jodoin, J. N., Sitaram, P., Albrecht, T. R., May, S. B., Shboul, M., Lee, E., . . . Lee, L. A. (2013). Nuclear-localized Asunder regulates cytoplasmic dynein localization via its role in the Integrator complex. *Molecular Biology of the Cell, 24*(18), 2954-2965. doi:10.1091/mbc.E13-05-0254

Jurado, A. R., Tan, D. Z., Jiao, X. F., Kiledjian, M., & Tong, L. (2014). Structure and Function of Pre-mRNA 5 '-End Capping Quality Control and 3 '-End Processing. *Biochemistry, 53*(12), 1882-1898. doi:10.1021/bi401715v

Kabsch, W. (2010). Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr D Biol Crystallogr, 66*(Pt 2), 133-144. doi:10.1107/S0907444909047374

Kapp, L. D., Abrams, E. W., Marlow, F. L., & Mullins, M. C. (2013). The Integrator Complex Subunit 6 (Ints6) Confines the Dorsal Organizer in Vertebrate Embryogenesis. *Plos Genetics, 9*(10). doi:ARTN e1003822

10.1371/journal.pgen.1003822

Kheirallah, A. K., de Moor, C. H., Faiz, A., Sayers, I., & Hall, I. P. (2017). Lung function associated gene Integrator Complex subunit 12 regulates protein synthesis pathways. *Bmc Genomics, 18*. doi:ARTN 248

10.1186/s12864-017-3628-3

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

Lee, Y., & Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry, Vol 84, 84*, 291-323. doi:10.1146/annurev-biochem-060614-034316

Lin, C. Q., Smith, E. R., Takahashi, H., Lai, K. C., Martin-Brown, S., Florens, L., . . . Shilatifard, A. (2010). AFF4, a Component of the ELL/P-TEFb Elongation Complex and a Shared Subunit of MLL Chimeras, Can Link Transcription Elongation to Leukemia. *Molecular Cell, 37*(3), 429-437. doi:10.1016/j.molcel.2010.01.026

Luo, Z. J., Lin, C. Q., & Shilatifard, A. (2012). The super elongation complex (SEC) family in transcriptional control. *Nature Reviews Molecular Cell Biology, 13*(9), 543-547. doi:10.1038/nrm3417

Luse, D. S. (2013). Promoter clearance by RNA polymerase II. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms, 1829*(1), 63-68. doi:10.1016/j.bbagrm.2012.08.010

Mandel, C. R., Kaneko, S., Zhang, H. L., 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. doi:10.1038/nature05363

Matera, A. G., Terns, R. M., & Terns, M. P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature Reviews Molecular Cell Biology, 8*(3), 209-220. doi:10.1038/nrm2124

Matera, A. G., & Wang, Z. F. (2014). A day in the life of the spliceosome (vol 15, pg 108, 2014). *Nature Reviews Molecular Cell Biology, 15*(4). doi:10.1038/nrm3778

Mattaj, I. W. (1986). Cap Trimethylation of U-Snrna Is Cytoplasmic and Dependent on U-Snrnp Protein-Binding. *Cell, 46*(6), 905-911. doi:Doi 10.1016/0092-8674(86)90072-3

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). Phaser crystallographic software. *J Appl Crystallogr, 40*(Pt 4), 658-674. doi:10.1107/S0021889807021206

Michalski, D., & Steiniger, M. (2015). In vivo characterization of the Drosophila mRNA 3 ' end processing core cleavage complex. *Rna-a Publication of the Rna Society, 21*(8), 1404-1418. doi:10.1261/rna.049551.115

Millevoi, S., & Vagner, S. (2010). Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Research, 38*(9), 2757-2774. doi:10.1093/nar/gkp1176

Moravcevic, K., Mendrola, J. M., Schmitz, K. R., Wang, Y. H., Slochower, D., Janmey, P. A., & Lemmon, M. A. (2010). Kinase Associated-1 Domains Drive MARK/PAR1 Kinases to Membrane Targets by Binding Acidic Phospholipids. *Cell, 143*(6), 966-977. doi:10.1016/j.cell.2010.11.028

Moteki, S., & Price, D. (2002). Functional coupling of capping and transcription of mRNA. *Molecular Cell, 10*(3), 599-609. doi:Doi 10.1016/S1097-2765(02)00660-3

Mura, C., Phillips, M., Kozhukhovsky, A., & Eisenberg, D. (2003). Structure and assembly of an augmented Sm-like archaeal protein 14-mer. *Proceedings of the National Academy of Sciences of the United States of America, 100*(8), 4539-4544. doi:10.1073/pnas.0538042100

Murphy, S. (1997). Differential in vivo activation of the class II and class III snRNA genes by the POU-specific domain of Oct-1. *Nucleic Acids Res, 25*(11), 2068-2076.

Murphy, S., Yoon, J. B., Gerster, T., & Roeder, R. G. (1992). Oct-1 and Oct-2 Potentiate Functional Interactions of a Transcription Factor with the Proximal Sequence Element of Small Nuclear-Rna Genes. *Molecular and Cellular Biology, 12*(7), 3247-3261. doi:Doi 10.1128/Mcb.12.7.3247

O'Reilly, D., Kuznetsova, O. V., Laitem, C., Zaborowska, J., Dienstbier, M., & Murphy, S. (2014). Human snRNA genes use polyadenylation factors to promote efficient transcription termination. *Nucleic Acids Research, 42*(1), 264-275. doi:10.1093/nar/gkt892

Ohno, M., Segref, A., Bachi, A., Wilm, M., & Mattaj, I. W. (2000). PHAX, a mediator of U snRNA nuclear export whose activity is regulated by phosphorylation. *Cell, 101*(2), 187-198. doi:Doi 10.1016/S0092-8674(00)80829-6

Otani, Y., Nakatsu, Y., Sakoda, H., Fukushima, T., Fujishiro, M., Kushiyama, A., . . . Asano, T. (2013). Integrator complex plays an essential role in adipose differentiation. *Biochemical and Biophysical Research Communications, 434*(2), 197-202. doi:10.1016/j.bbrc.2013.03.029

Otwinowski, Z., & Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. *Macromolecular Crystallography, Pt A, 276*, 307-326. doi:Doi 10.1016/S0076-6879(97)76066-X

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. *Nature Reviews Genetics, 14*(4), 288-295. doi:DOI 10.1038/nrg3458

Peterlin, B. M., & Price, D. H. (2006). Controlling the elongation phase of transcription with P-TEFb. *Molecular Cell, 23*(3), 297-305. doi:10.1016/j.molcel.2006.06.014

Ramanathan, A., Robb, G. B., & Chan, S. H. (2016). mRNA capping: biological functions and applications. *Nucleic Acids Research, 44*(16), 7511-7526. doi:10.1093/nar/gkw551

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

Romeo, V., & Schumperli, D. (2016). Cycling in the nucleus: regulation of RNA 3' processing and nuclear organization of replication-dependent histone genes. *Curr Opin Cell Biol, 40*, 23-31. doi:10.1016/j.ceb.2016.01.015

Schaukowitch, K., Joo, J. Y., Liu, X. H., Watts, J. K., Martinez, C., & Kim, T. K. (2014). Enhancer RNA Facilitates NELF Release from Immediate Early Genes. *Molecular Cell, 56*(1), 29-42. doi:10.1016/j.molcel.2014.08.023

Schmid, E. M., Ford, M. G. J., Burtey, A., Praefcke, G. J. K., Peak-Chew, S. Y., Mills, I. G., . . . McMahon, H. T. (2006). Role of the AP2 beta-appendage hub in recruiting partners for clathrin-coated vesicle assembly. *Plos Biology, 4*(9), 1532-1548. doi:ARTN e262
10.1371/journal.pbio.0040262

Shi, Y. S., & Manley, J. L. (2015). The end of the message: multiple protein-RNA interactions define the mRNA polyadenylation site. *Genes & Development, 29*(9), 889-897. doi:10.1101/gad.261974.115

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

Stadelmayer, B., Micas, G., Gamot, A., Martin, P., Malirat, N., Koval, S., . . . Benkirane, M. (2014). Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nature Communications, 5*. doi:ARTN 5531
10.1038/ncomms6531

Stivala, A., Wybrow, M., Wirth, A., Whisstock, J. C., & Stuckey, P. J. (2011). Automatic generation of protein structure cartoons with Pro-origami. *Bioinformatics, 27*(23), 3315-3316. doi:10.1093/bioinformatics/btr575

Sullivan, K. D., Steiniger, M., & Marzluff, W. F. (2009). A Core Complex of CPSF73, CPSF100, and Symplekin May Form Two Different Cleavage Factors for Processing of Poly(A) and Histone mRNAs. *Molecular Cell, 34*(3), 322-332. doi:10.1016/j.molcel.2009.04.024

Takata, H., Nishijima, H., Maeshima, K., & Shibahara, K. (2012). The integrator complex is required for integrity of Cajal bodies. *Journal of Cell Science, 125*(1), 166-175. doi:10.1242/jcs.090837

Tao, S. J., Cai, Y., & Sampath, K. (2009). The Integrator subunits function in hematopoiesis by modulating Smad/BMP signaling. *Development, 136*(16), 2757-2765. doi:10.1242/dev.034959

Townley, R., & Shapiro, L. (2007). Crystal structures of the adenylate sensor from fission yeast AMP-activated protein kinase. *Science, 315*(5819), 1726-1729. doi:10.1126/science.1137503

Vasilyev, N., & Serganov, A. (2015). Structures of RNA Complexes with the Escherichia coli RNA Pyrophosphohydrolase RppH Unveil the Basis for Specific 5 '-End-dependent mRNA Decay. *Journal of Biological Chemistry, 290*(15), 9487-9499. doi:10.1074/jbc.M114.634824

Wagner, E. J., Burch, B. D., Godfrey, A. C., Salzler, H. R., Duronio, R. J., & Marzluff, W. F. (2007). A genome-wide RNA interference screen reveals that variant histones, are necessary for replication-dependent histone pre-mRNA processing. *Molecular Cell, 28*(4), 692-699. doi:10.1016/j.molcel.2007.10.009

Walter, T. S., Meier, C., Assenberg, R., Au, K. F., Ren, J. S., Verma, A., . . . Grimes, J. M. (2006). Lysine methylation as a routine rescue strategy for protein crystallization. *Structure, 14*(11), 1617-1622. doi:10.1016/j.str.2006.09.005

Walters, R. W., Matheny, T., Mizoue, L. S., Rao, B. S., Muhlrad, D., & Parker, R. (2017). Identification of NAD(+) capped mRNAs in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America, 114*(3), 480-485. doi:10.1073/pnas.1619369114

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 of the United States of America, 114*(17), 4394-4399. doi:10.1073/pnas.1616605114

Xiao, B., Heath, R., Saiu, P., Leiper, F. C., Leone, P., Jing, C., . . . Gamblin, S. J. (2007). Structural basis for AMP binding to mammalian AMP-activated protein kinase. *Nature, 449*(7161), 496-U414. doi:10.1038/nature06161

Yamaguchi, Y., Shibata, H., & Handa, H. (2013). Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond. *Biochim Biophys Acta, 1829*(1), 98-104. doi:10.1016/j.bbagrm.2012.11.007

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. doi:10.4161/trns.19783

Zhang, F., Ma, T., & Yu, X. C. (2013). A core hSSB1-INTS complex participates in the DNA damage response. *Journal of Cell Science, 126*(21), 4850-4855. doi:10.1242/jcs.132514

Zhou, Q., Li, T. D., & Price, D. H. (2012). RNA Polymerase II Elongation Control. *Annual Review of Biochemistry, Vol 81, 81*, 119-143. doi:10.1146/annurev-biochem-052610-095910

APPENDIX

Wu, Y., Albrecht, T.R., Baillat, D., Wagner, E.J. and Tong, L. (2017) Molecular basis for the interaction between Integrator subunits IntS9 and IntS11 and its functional importance. *P Natl Acad Sci USA*, 114, 4394-4399.

Molecular basis for the interaction between Integrator subunits IntS9 and IntS11 and its functional importance

Yixuan Wu[a,1], Todd R. Albrecht[b,1], David Baillat[b], Eric J. Wagner[b,2], and Liang Tong[a,2]

[a]Department of Biological Sciences, Columbia University, New York, NY 10027; and [b]Department of Biochemistry and Molecular Biology, The University of Texas Medical Branch, Galveston, TX 77550

Edited by Michael Sattler, Helmholtz Zentrum München, Neuherberg, Germany, and accepted by Editorial Board Member Dinshaw J. Patel, March 16, 2017 (received for review October 5, 2016)

The metazoan Integrator complex (INT) has important functions in the 3′-end processing of noncoding RNAs, including the uridine-rich small nuclear RNA (UsnRNA) and enhancer RNA (eRNA), and in the transcription of coding genes by RNA polymerase II. The INT contains at least 14 subunits, but its molecular mechanism of action is poorly understood, because currently there is little structural information about its subunits. The endonuclease activity of INT is mediated by its subunit 11 (IntS11), which belongs to the metallo-β-lactamase superfamily and is a paralog of CPSF-73, the endonuclease for pre-mRNA 3′-end processing. IntS11 forms a stable complex with Integrator complex subunit 9 (IntS9) through their C-terminal domains (CTDs). Here, we report the crystal structure of the IntS9–IntS11 CTD complex at 2.1-Å resolution and detailed, structure-based biochemical and functional studies. The complex is composed of a continuous nine-stranded β-sheet with four strands from IntS9 and five from IntS11. Highly conserved residues are located in the extensive interface between the two CTDs. Yeast two-hybrid assays and coimmunoprecipitation experiments confirm the structural observations on the complex. Functional studies demonstrate that the IntS9–IntS11 interaction is crucial for the role of INT in snRNA 3′-end processing.

snRNA processing | protein complex | integrator complex

The Integrator complex (INT) was originally characterized as an important factor for U-rich small nuclear RNA (UsnRNA) 3′-end processing and as a binding partner for phosphorylated RNA polymerase II (Pol II) (1–3). Since then it also has been found to participate in Pol II transcription initiation, pause release, elongation, and termination at protein-coding genes (4–6), thereby broadening the scope of INT function. Most recently INT was shown to be important for enhancer RNA (eRNA) 3′-end processing as well (7). The physiological importance of INT is reflected by its involvement in other cellular processes, such as embryogenesis (8), ciliogenesis (9), adipose differentiation (10), human lung function (11), and maturation of herpesvirus microRNA 3′ ends (12).

The INT contains at least 14 subunits (IntS1 through IntS14), ranging from 49 to 244 kDa, and the mass of the full complex is more than 1 million daltons. Despite its functional importance, INT is poorly understood at the molecular level, and currently there is very little structural information for any of its subunits. Moreover, the majority of the INT subunits share little sequence homology with other proteins and have few recognizable domains, making it difficult to predict the function and the mechanism of the subunits (2).

Two INT subunits that do share significant sequence conservation with other proteins are IntS9 and IntS11 (1, 2). Most importantly, IntS11 is a close homolog of CPSF-73 (cleavage and polyadenylation specificity factor, 73 kDa), the endonuclease for the pre-mRNA 3′-end processing machinery (13–16). Similarly, IntS11 is the endonuclease subunit for INT, responsible for the cleavage reaction at the 3′ end of target RNAs (1, 2). CPSF-73 and IntS11 belong to the metallo-β-lactamase superfamily of proteins (17). The metallo-β-lactamase domain of each protein contains conserved amino acids that coordinate metal ions for catalysis (Fig. 1*A* and Fig. S1) (18). CPSF-73, IntS11, and their close homologs also have a β-CASP domain (named after its founding members: CPSF73, Artemis, SNM, and PSO), which is an insert in the metallo-β-lactamase domain (Fig. 1*A*). The active site is located at the interface between the two domains, and therefore the β-CASP domain likely regulates substrate access to the active site. The metallo-β-lactamase and the β-CASP domains of human IntS11 (residues 1–450) (Fig. S1) and CPSF-73 share 40% amino acid sequence identity, indicating their high degree of sequence conservation and underscoring the importance of these two domains.

The CPSF complex also contains an inactive homolog of CPSF-73, CPSF-100, in which several of the active-site residues have been changed during evolution. Likewise, IntS9 is the inactive homolog of IntS11 in INT (Fig. 1*B* and Fig. S2). The sequence conservation between IntS9 and CPSF-100 is weaker, and IntS9 has two inserted segments in the metallo-β-lactamase domain (Fig. 1*B*). The functional roles of IntS9 and CPSF-100 in their respective complexes remain to be clarified.

IntS9, IntS11, CPSF-73, and CPSF-100 all contain a C-terminal domain (CTD) beyond their metallo-β-lactamase and β-CASP domains (Fig. 1). However, the sequence conservation for these domains is much poorer. We showed previously that the CTDs of IntS9 and IntS11 are required and sufficient to mediate their association, suggesting that these domains ensure specific

Significance

The Integrator complex (INT) has important functions in the 3′-end processing of noncoding RNAs and RNA polymerase II transcription. The INT contains at least 14 subunits, but its molecular mechanism of action is still poorly understood. The endonuclease activity of INT is mediated by its subunit 11 (IntS11), which forms a stable complex with Integrator complex subunit 9 (IntS9) through their C-terminal domains (CTDs). Here, we report the crystal structure of the IntS9–IntS11 CTD complex at 2.1-Å resolution and detailed, structure-based biochemical and functional studies. Highly conserved residues are located in the extensive interface between the two CTDs. Yeast two-hybrid assays and coimmunoprecipitation experiments confirm the structural observations. Functional studies demonstrate that the IntS9–IntS11 interaction is crucial for INT in snRNA 3′-end processing.

Author contributions: Y.W., T.R.A., D.B., E.J.W., and L.T. designed research; Y.W., T.R.A., D.B., E.J.W., and L.T. performed research; Y.W., T.R.A., D.B., E.J.W., and L.T. analyzed data; and Y.W., T.R.A., E.J.W., and L.T. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. M.S. is a Guest Editor invited by the Editorial Board.

Data deposition: Crystallography, atomic coordinates, and structure factors reported in this paper have been deposited in the Protein Data Bank (accession code 5V8W).

[1]Y.W. and T.R.A. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: ltong@columbia.edu or ejwagner@utmb.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1616605114/-/DCSupplemental.

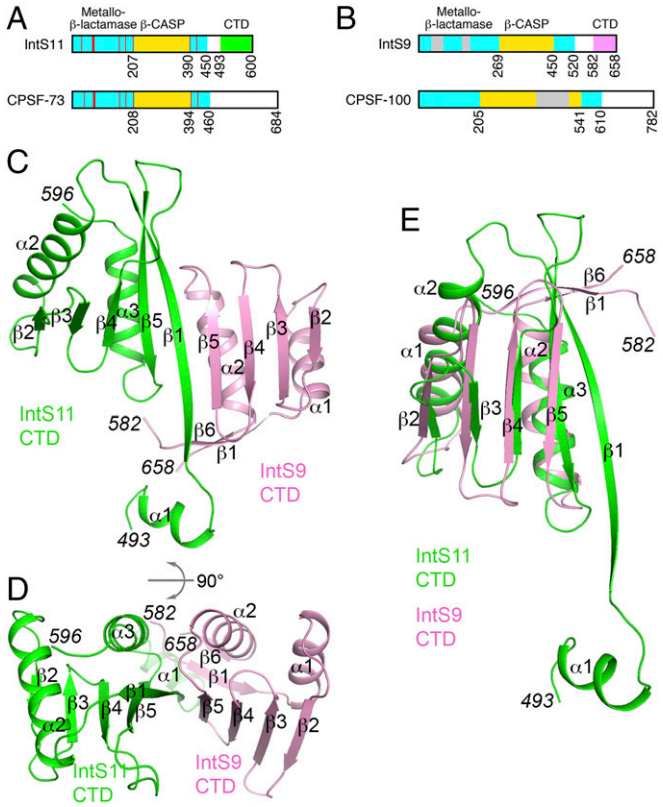


Fig. 1. Crystal structure of the human IntS9–IntS11 CTD complex. (*A*) Domain organizations of human IntS11 and CPSF-73. The metallo-β-lactamase and β-CASP domains are shown in cyan and yellow, respectively. The conserved residues in the active site are indicated by red lines. The CTD of IntS11 is shown in green. CPSF-73 also has a CTD, but its sequence is highly divergent from that of IntS11, and its exact boundary is not known. (*B*) Domain organizations of human IntS9 and CPSF-100. The CTD of IntS9 is shown in pink. An insert in the β-CASP domain of CPSF-100 and two inserts in the metallo-β-lactamase domain of IntS9 are shown in gray. (*C*) Structure of the human IntS9–IntS11 CTD complex. The IntS9 CTD is in pink, and the IntS11 CTD is in green. (*D*) Structure of the human IntS9–IntS11 CTD complex, viewed after 90° rotation around the horizontal axis. (*E*) Overlay of the structure of the IntS9 CTD (pink) with the structure of the IntS11 CTD (green). The structure figures were produced with PyMOL (www.pymol.org).

heterodimerization of the two subunits, thereby compartmentalizing INT from CPSF (19). However, the molecular basis for this association is not known. Currently available dimer structures of other β-CASP homologs (20–23) do not provide any insights into the IntS9–IntS11 heterodimer, nor are structures available for the CPSF-73/CPSF-100 CTDs.

We have determined the crystal structure of the IntS9–IntS11 CTD complex at 2.1-Å resolution. The structure is composed of a continuous nine-stranded β-sheet, with four strands from IntS9 and five from IntS11. Four helices cover one face of this β-sheet, and the other face is exposed to solvent. Highly conserved residues are located in the extensive interface between the two CTDs formed by the two neighboring strands and two helices. We designed truncation and site-specific mutants based on the structure, and both our yeast two-hybrid assays with the CTDs and coimmunoprecipitation experiments with the full-length proteins confirm the structural observations on the complex. Finally, we demonstrated that mutations that disrupt the IntS9–IntS11 interaction also abolish U7 snRNA 3′-end processing, indicating that this interaction is crucial for the function of the Integrator complex.

Results

Structures of the IntS9 and IntS11 CTDs. The crystal structure of the complex of IntS9 and IntS11 CTDs has been determined at 2.1-Å resolution. The atomic model has good agreement with the X-ray diffraction data and the expected geometric parameters (Table S1). Of the residues, 98.1% are in the favored region of the Ramachandran plot, and no residues are in the disallowed region.

The structure of the IntS9 CTD (covering residues 582–658) (Fig. 1*B*) contains a four-stranded, antiparallel β-sheet (β2–β5) (Fig. 1*C*). Two helices (α1–α2), formed by residues just before and after the β-sheet, cover one of its faces (Fig. 1*D*). A two-stranded antiparallel β-sheet (β1, β6) formed by residues near the beginning and end of the domain likely provides further stability to this domain.

The structure of the IntS11 CTD (covering residues 493–596) (Fig. 1*A*) contains a five-stranded antiparallel β-sheet (β1–β5) with two helices on one of its faces (α2-α3) (Fig. 1*C*). A short helix (α1) precedes the first β-strand and is partly stabilized by interactions with IntS9 (see next section).

The up–down organization of the last four strands (β2–β5) of the IntS11 CTD is similar to that for the β-sheet in IntS9 CTD. In fact, the two structures can be superposed with an rmsd of 2.4 Å for 64 equivalent Cα atoms (Fig. 1*E*), although the sequence identity between the two proteins in this region is only 16%. The two helices covering the β-sheet are located at similar positions in the two structures as well. A unique feature of IntS11 is strand β1, which is the longest strand in the structure and is located in the center of the interface with IntS9.

Close structural homologs for IntS9 CTD include the CTD of an atypical Sm-like archaeal protein (24) and the platform subdomain of the AP-2 complex β subunit (Fig. S3) (25), based on a DaliLite search (26). Close structural homologs for IntS11 CTD include the kinase associated-1 domain (KA1 domain) at the C terminus of yeast septin-associated kinases and human MARK/PAR1 kinases (27), the C-terminal domain of the catalytic subunit of AMP-activated protein kinase (AMPK, SNF1 in yeast) that mediates heterotrimer formation (28–30), and the N-terminal domain of BamC, part of the β-barrel assembly machinery (Fig. S3) (31). These structural homologs do not offer much insight into the functions of the two CTDs.

Crystal Structure of the IntS9–IntS11 CTD Complex. The complex of IntS9–IntS11 CTDs is formed by juxtaposing the β-sheets of the two domains, such that strand β5 of IntS9 forms a parallel β-sheet with strand β1 of IntS11 (Fig. 1*C*). This juxtaposition creates a nine-stranded, mostly antiparallel β-sheet in the IntS9–IntS11 CTD heterodimer, with only the two strands at the subunit interface being in parallel. The four flanking helices cover the same face of this β-sheet, and the other face of the β-sheet is open to the solvent (Fig. 1*D*). Seven hydrogen bonds are formed between the two β-strands at the center of the interface (Fig. 2*A*). In addition to these interactions, many side chains mediate the formation of this heterodimer as well, and ∼1,200 Å^2 of the surface area of each subunit is buried in this interface (Fig. 3 *A*–*C*). The neighboring side chains of the two strands on the exposed face of the β-sheet are in contact with each other. In addition, the N-terminal helix (α1) of IntS11 contacts the N-terminal segment of IntS9, likely stabilizing both proteins in this region of the interface (Fig. 2*A*).

On the other face of the β-sheet, helix α2 of IntS9 and helix α3 of IntS11 are positioned next to each other, allowing favorable interactions among some of their side chains as well as the side chains of the two β-strands in the center of the interface (Fig. 2*B*). Residues from the two β-strands are mostly hydrophobic in this part of the interface, whereas those from the two helices are mostly hydrophilic or charged. Most of the residues at this interface are highly conserved among IntS11 (Fig. 3*D* and Fig. S1) and IntS9 (Fig. 3*E* and Fig. S2) homologs, especially near the center of the interface.

BIOCHEMISTRY

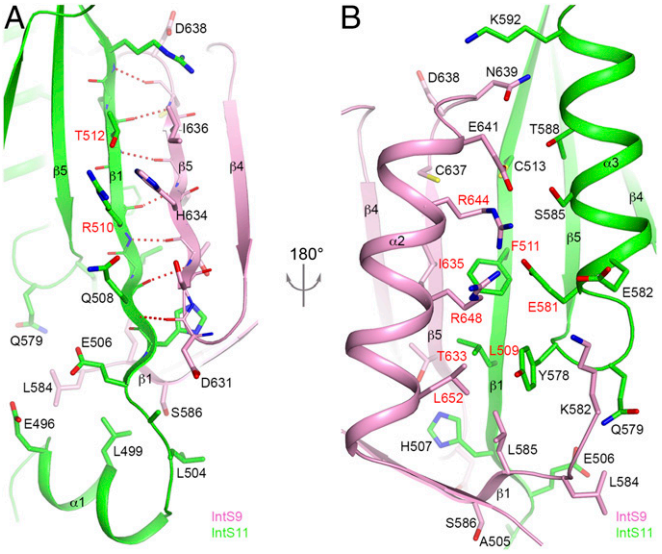


Fig. 2. Detailed interactions at the interface of the IntS9–IntS11 CTD complex. (*A*) Hydrogen-bonding interactions between strand β1 of IntS11 (green) and strand β5 of IntS9 (pink) are indicated by the dashed lines in red. The side chains of the two β-strands are placed next to each other. Interactions between helix α1 of IntS11 and residues in IntS9 are also shown. Residues selected for mutagenesis studies are labeled in red. (*B*) Interactions between residues in helix α3 of IntS11 (green) and residues in helix α2 of IntS9 (pink). Residues in strand β1 of IntS11 and strand β5 of IntS9 also contribute to this part of the interface.

There are four copies of the IntS9–IntS11 CTD complex in the crystallographic asymmetric unit. The overall structures of the two subunits in the four complexes are similar, with rmsds of ∼0.5 Å for equivalent Cα atoms between any pair of them. The overall structures of the four complexes are similar as well, especially for the β-sheet and the four flanking helices (Fig. S4). However, there are large differences in the conformations of several of the loops, suggesting that these regions are somewhat flexible. In addition, the Cys542 residues from two IntS11 subunits in neighboring complexes form a disulfide bond covalently linking two complexes (Fig. S4), and this disulfide linkage is likely a crystallization artifact. The other cysteine residues in the structure are in the fully reduced state. Cys542 is located just before strand β2 in IntS11, and some conformational differences in this strand are observed among the four complexes (Fig. S4). It is unlikely that this disulfide bond affects the overall structure of the complex, because it creates only a relatively small region of contact between two complexes (Fig. S4).

Biochemical Studies Confirm the Structural Observations. To assess the structural observations on the IntS9–IntS11 CTD complex, we carried out yeast two-hybrid assays to evaluate interactions between different variants of the two CTDs as well as coimmunoprecipitation experiments with the full-length proteins. We previously demonstrated that using yeast two-hybrid assay to characterize the interaction between IntS9 and IntS11 is remarkably robust, because binding could be detected even when 3-amino-1,2,4-triazole is present at a 100-mM concentration (32). We carefully mapped the regions within the CTDs of IntS9 and IntS11 that are both required and sufficient to mediate their interaction. We created a series of truncation mutants, removing 10 amino acid residues at a time, and found that residues 500–600 of IntS11 interacted strongly with IntS9 CTD, whereas residues 510–600 showed no interaction (Fig. 4*A*). Residue 500 is located near the end of helix α1, and residue 510 is in the middle of strand β1 (Fig. 1*C*), indicating the importance of β1. This assay also determined that helix α1 of IntS11 is not required for the interaction, as is consistent with its being located at the periphery of the interface. On the other hand, deleting only 10 residues from the C terminus of IntS11 (resulting in a variant with residues 500–590) abolished the interaction (Fig. 4*A*). Residue 590 is located in the last turn of helix α3, confirming its importance for the IntS9–IntS11 interaction.

Similarly, we found that residues 579–658 of IntS9 interacted strongly with the IntS11 CTD, whereas residues 589–658 showed no interactions (Fig. 4*B*). Residue 579 is before the first residue observed in the current structure, and residue 589 is just after strand β1 (Fig. 1*C*). Deleting 10 residues from the C terminus of IntS9 (579–648) also abolished the interaction (Fig. 4*B*). Residue 648 is located in the middle of helix α2. Overall, the results from the truncation mutants define the minimal regions of IntS9 and IntS11 that are important for their interactions; these results are fully consistent with the structural observations.

We next designed a series of point mutations that are expected to perturb the IntS9–IntS11 interaction based on the structural observations. To perturb the hydrogen-bonding interactions between the two β-strands in the dimer interface, we mutated two residues in the middle of each strand to proline, i.e., the T633P/I635P double mutant for IntS9 (Fig. 2*B*) and the R510P/T512P double mutant for IntS11 (Fig. 2*A*). We also designed mutations to disrupt interactions among the side chains, including the R644E single mutant, the R644E/R648E double mutant, and the R644E/R648E/L652A triple mutant in helix α2 of IntS9, the L509A/F511A double mutant in strand β1 of IntS11, and the L509A/F511A/E581R triple mutant in strand β1 and helix α3 of IntS11 (Fig. 2*B*). Most of these residues are strictly conserved among the homologs, whereas Leu652 of IntS9 and Leu509 of IntS11 show conservative variations to other hydrophobic residues (Figs. S1 and S2). We introduced these mutations into the

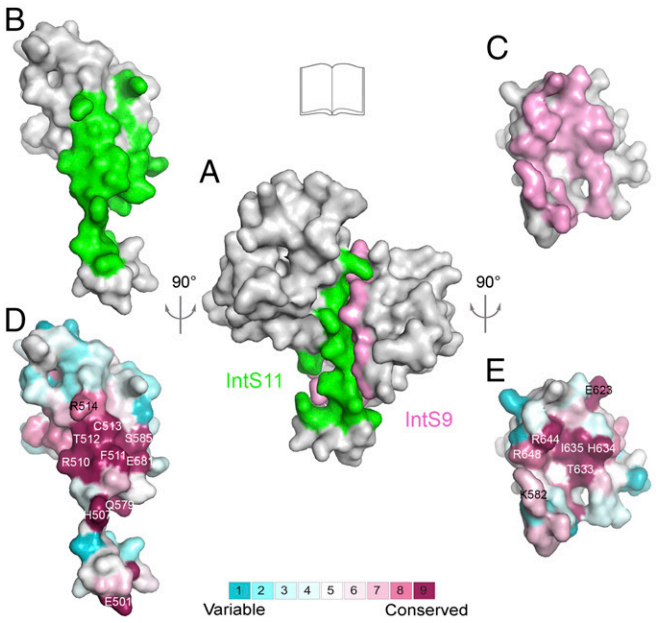


Fig. 3. (*A*) Molecular surface of the IntS9–IntS11 CTD complex. Residues in IntS9 that contribute to the interface with IntS11 are colored in pink, and those in IntS11 that contact IntS9 are in green. The other residues are in gray. (*B*) An "open-book" view of the IntS9–IntS11 interface showing the surface area of IntS11 in contact with IntS9 after 90° rotation around the vertical axis. (*C*) An open-book view of the IntS9–IntS11 interface showing the surface area of IntS9 in contact with IntS11 after 90° rotation around the vertical axis. (*D*) Molecular surface of IntS11 colored by sequence conservation, produced by ConSurf (40). Highly conserved residues are labeled. The color scheme runs from dark red (highly conserved) to cyan (poorly conserved) (color bar at bottom). The view is the same as in *B*. (*E*) Molecular surface of IntS9 colored by sequence conservation.

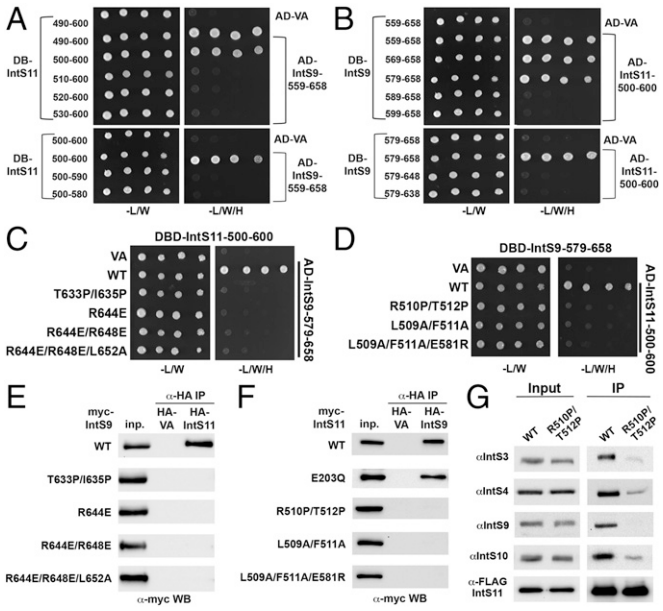


Fig. 4. Biochemical studies confirm the structural observations on the IntS9–IntS11 CTD complex. (*A*) Yeast two-hybrid assay to define the minimal region of IntS11 sufficient to bind IntS9. (*Upper*) Ten amino acid deletions starting from residue 490. (*Lower*) Ten amino acid deletions starting from the C terminus of IntS11. AD, activation domain; BD, DNA-binding domain; VA, vector alone control. (*B*) Yeast two-hybrid assay to define the minimal region of IntS9 sufficient to bind IntS11. (*C*) Yeast two-hybrid assay using the minimal regions of IntS9 and IntS11 sufficient for their interaction, with structure-based mutations in IntS11 CTD. (*D*) Yeast two-hybrid assay using the minimal regions of IntS9 and IntS11 sufficient for their interaction, with structure-based mutations in IntS9 CTD. (*E*) Coimmunoprecipitation of full-length myc-tagged wild-type and mutant IntS9 with full-length wild-type HA-tagged IntS11. Proteins bound to HA affinity resin were probed with anti-myc antibody by Western blot (WB). (*F*) Coimmunoprecipitation of full-length myc-tagged wild-type and mutant IntS11 with full-length wild-type HA-tagged IntS9. (*G*) Purification of endogenous INT from stable 293T cells expressing either wild-type FLAG-IntS11 or the heterodimeric mutant (R510P, T512P) using FLAG affinity resin.

minimal CTDs of IntS9 or IntS11 and observed a complete loss of interaction based on the yeast two-hybrid assays (Fig. 4 *C* and *D*).

To extend upon these data, we also introduced the point mutations into the full-length cDNAs encoding IntS9 and IntS11 and tested their impact on the interaction using a coimmunoprecipitation assay. We had established previously that the IntS9–IntS11 heterodimer could withstand rigorous washing with detergent and high salt (19) and therefore tested these same conditions here. We transfected various myc-tagged IntS9 cDNAs with HA-tagged wild-type IntS11 into 293T cells and subjected the lysates to anti-HA immunoaffinity matrix followed by probing with anti-myc antibodies using Western blot analysis. Only the wild-type IntS9 was able to coimmunoprecipitate with IntS11; none of the mutants was detected in the immunoprecipitate (Fig. 4*E*). Importantly, all four mutants tested were expressed at levels similar to the wild type, suggesting that these proteins are folded properly and that the lack of coimmunoprecipitation is caused by the disruption of the interaction by the mutations.

We then performed the reciprocal coimmunoprecipitation in which we transfected HA-tagged wild-type IntS9 with several mutants of IntS11. We also included a catalytic mutant (E203Q, mutating one of the conserved residues in the active site of the metallo-β-lactamase domain) as a control; this mutation is not expected to disrupt interaction with IntS9. Both the wild type and the E203Q mutant of IntS11 were able to interact with IntS9, but the interface mutants did not show interaction (Fig. 4*F*). Finally, we extended these analyses to ask whether disruption of the IntS9–IntS11 interface inhibits the ability of the heterodimer to be incorporated into the endogenous Integrator complex. Previously, we established that purification of the intact INT could be achieved by pulling down a FLAG-tagged IntS11 from nuclear extracts derived from stable cells lines expressing FLAG-IntS11 (33). Therefore, we created cells stably expressing either wild-type IntS11 or a mutant IntS11 and purified the INT from nuclear extracts made from these cells lines. Almost no IntS9 was associated with the mutant IntS11, as expected, but we also observed a significant reduction in the levels of copurifying IntS3, IntS4, and IntS10 relative to wild-type IntS11 (Fig. 4*G*). Taken together, these results strongly validate the structure of the IntS9–IntS11 CTD complex and demonstrate that this region is essential for the two Integrator subunits to interact and may play a significant role in the incorporation of the IntS11 endonuclease into the endogenous INT.

Functional Importance of the IntS9–IntS11 Complex. Currently an in vitro system is not available to assess INT function. Therefore, to address the functional relevance of the IntS9–IntS11 interactions observed in the crystal structure, we used a cell-based fluorescence reporter that assays for UsnRNA 3′-end formation (34). The U7-GFP reporter consists of the human U7 snRNA promoter, the U7 snRNA gene body, a 3′ box sequence for snRNA 3′-end processing, and then the coding sequence for GFP followed by a strong polyadenylation signal (Fig. 5*A*). When transfected into untreated cells, the reporter gives rise to no GFP expression because of the 3′-end processing activity of endogenous INT. If cells are transfected with siRNA targeting INT subunits, the efficiency of snRNA 3′-end formation is reduced, resulting in transcriptional read-through and the production of a GFP mRNA containing the U7 snRNA as its 5′ UTR using the native start codon of GFP because U7 lacks an AUG sequence.

We transfected cDNA constructs containing silent point mutations rendering them RNAi-resistant into IntS9- or IntS11-depleted cells to determine how effective these constructs are at restoring INT activity through the reduction of GFP expression. In control siRNA-treated cells, we detected little to no expression of GFP after transfection with the U7-GFP reporter (Fig. 5 *B* and *C*). In contrast, we could detect robust GFP expression in cells treated with siRNA targeting either IntS11 or IntS9. Expression of RNAi-resistant myc-tagged wild-type IntS11 or IntS9 could reduce GFP expression nearly back to the levels observed in control siRNA-transfected cells. We observed that, as expected, the E203Q catalytic mutant of IntS11 was unable to rescue U7 snRNA processing (Fig. 5*B*). Most importantly, mutations in IntS11 or IntS9 that disrupt the interactions between their CTDs also failed to rescue U7 snRNA processing, despite being expressed at levels similar to the wild-type protein (Fig. 5 *B* and *C*). Finally, we created stable cell lines expressing either RNAi-resistant wild-type IntS11 or a subset of the heterodimerization IntS11 mutants and assessed the levels of misprocessed, endogenously expressed U2 or U4 snRNA after knockdown. Upon depletion of IntS11 we observed an ∼25-fold increase in the levels of misprocessed snRNA relative to the levels observed in control siRNA-treated cells (Fig. 5*D*). This level of misprocessing was similar to what we observed in *Drosophila* S2 cells upon depletion of INT subunits (35). The levels of misprocessed snRNA present in IntS11 knockdown cells could be rescued upon stable expression of the wild-type IntS11, but, as in the U7-GFP reporter experiments, we did not observe any rescue of snRNA misprocessing in cells expressing RNAi-resistant IntS11 heterodimerization mutants (Fig. 5*D*). Collectively, these results demonstrate that interactions observed in the structure of the IntS9–IntS11 CTD complex are as critical to INT activity in snRNA processing as the residues within the active site of IntS11.

Discussion

Although a role for IntS11 in the cleavage of UsnRNA and eRNA has been established using mutations of the metal-coordinating residues within the active site, the molecular mechanism by which IntS11 is recruited to these substrates and

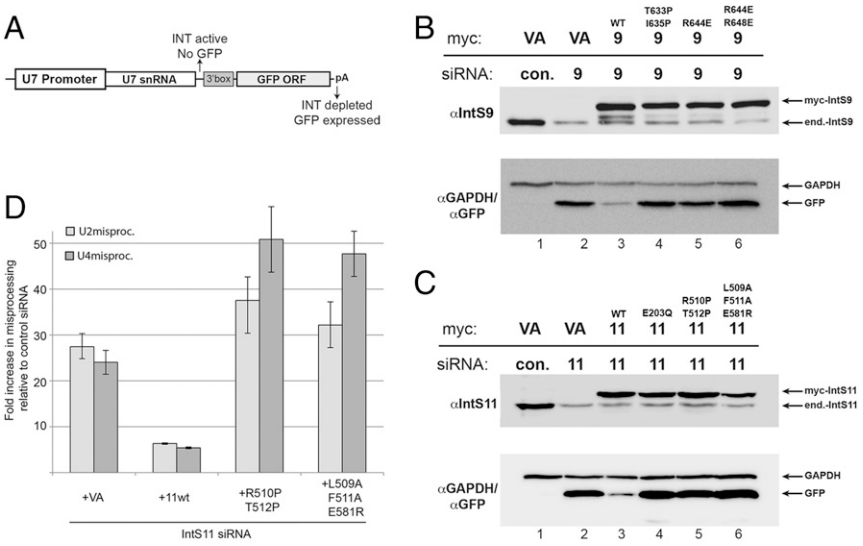


Fig. 5. Functional importance of the IntS9–IntS11 interactions for snRNA 3′-end processing. (*A*) Schematic of the U7-GFP reporter that is transfected into human cells. (*B*) Western blot analysis of lysates from HeLa cells transfected with either control siRNA or IntS9 siRNA that then were transfected with either empty vector or myc-tagged RNAi-resistant IntS9. All cells were also transfected with the U7-GFP reporter. (*C*) The same analysis as in *B*, except that cells were treated with siRNA targeting IntS11 rather than IntS9. (*D*) Quantitative RT-PCR analysis of misprocessed U2 or U4 snRNA that are endogenously expressed. The bar graph represents the fold increase in the levels of misprocessed snRNA; data show the results of biological triplicates; error bars represent the SD from the mean.

carries out specific endonucleolytic processing is not known. The structure of the IntS9–IntS11 CTD heterodimer reveals an extensive molecular interface mediated by numerous interactions and explains the high binding affinity that has been reported for the two proteins. The structure also indicates a role for the catalytically inactive IntS9, which provides a distinct structural surface established only through heterodimerization with IntS11. This specific interface could allow recognition of only the active cleavage factor by the other members of the INT complex. Such a mechanism also might be operative for CPSF-73 and CPSF-100 in the pre-mRNA 3′-end processing machinery.

The CTDs of IntS9 and IntS11 are substantially larger than that of the homologous enzyme RNase J, which is comprised of a three-stranded β-sheet and two facing α-helices (23). Deletion of the RNase J CTD makes the enzyme become monomeric in solution and also abrogates all catalytic activity in vitro, even though the ΔCTD RNase J retains structurally intact metallo-β-lactamase and β-CASP domains. Based on results from our studies using the U7-GFP reporter, it is clear the mutations that specifically disrupt the formation of the IntS9–IntS11 CTD heterodimer have effects equivalent to those of the mutation (E203Q) that disrupts the active site of IntS11. This finding demonstrates that the binding to IntS9 is essential for IntS11 function in cells and suggests that homo- or heterodimerization of β-CASP RNA endonucleases either plays an important role in the recruitment to RNA substrates or somehow impacts the activity of the catalytic domain. One potential explanation is that formation of the IntS9–IntS11 CTD complex induces obligatory conformational changes in IntS11, for example in the interface between IntS11 metallo-β-lactamase and β-CASP domains, to allow access to and cleavage of the RNA substrates. This structural requirement would ensure that any IntS11 not associated with IntS9 would be inactive, and, by analogy, the same would hold true for CPSF-73 and CPSF-100.

CTD heterodimerization may provide another important function in addition to modulating the catalytic activity of IntS11. In this model, the IntS9–IntS11 CTD complex produces an essential surface that is recognized by a different member of the Integrator complex to recruit the dimerized cleavage factor into the complex. This mechanism would provide an elegant way of ensuring that only the authentic IntS11–IntS9 heterodimer is incorporated into INT and could represent an additional layer of regulation to prevent spurious cleavage events. This model is supported by our experiments demonstrating that a heterodimer-deficient IntS11 failed to associate with other members of INT in addition to IntS9 (Fig. 4*G*). Indeed, a large scaffolding protein, symplekin, interacts with CPSF-73 and CPSF-100 and likely plays a critical role in mediating cleavage of pre-mRNA substrates (13–16). Such a protein is likely to exist within the Integrator complex, but currently there is no candidate based upon sequence comparison with symplekin.

The linkers between the metallo-β-lactamase domain and the CTD in IntS9 and IntS11 are expected to contain secondary structure elements (Figs. S1 and S2) and hence are likely to be organized structurally. In the structure of RNase J, the linker also contains secondary structure elements and has interactions with both domains (23). Exactly how the linkers in IntS9 and IntS11 connect the two parts of these proteins remains to be determined. It is possible that this region of the proteins functions to communicate heterodimerization to the active site to allow cleavage to take place.

Methods

Protein Expression, Purification, and Crystallization. The C-terminal domain of human IntS11 (residues 491–600) was subcloned into the pET28a vector (Novagen), which introduced an N-terminal His-tag. The C-terminal domain of human IntS9 (residues 582–658) was subcloned into pCDFDuet vector (Novagen) without any affinity tag. The two proteins were coexpressed in *Escherichia coli* BL21Star (DE3) cells at 23 °C for 16–20 h. The cells were lysed by sonication in a buffer containing 20 mM Tris (pH 8.5), 200 mM NaCl, and 5% (vol/vol) glycerol. The IntS9–IntS11 heterodimer was purified by Ni-NTA (Qiagen) chromatography. The eluted protein was treated overnight with thrombin at 4 °C to remove the His-tag and was further purified by gel filtration chromatography (Sephacryl S-300; GE Healthcare). The purified protein was concentrated to 30 mg/mL in a solution containing 20 mM Tris (pH 8.5), 200 mM NaCl, and 10 mM DTT before being flash-frozen in liquid nitrogen and stored at −80 °C.

Crystals of the IntS9–IntS11 complex were obtained at 20 °C using the sitting-drop vapor-diffusion method. The reservoir solution contained 0.1 M Bis-Tris (pH 6.5) and 21–24% (wt/vol) PEG 3350. The protein concentration was 10 mg/mL. Crystals took 2 wk to grow to full size. A heavy-atom derivative was prepared by soaking native crystals in the mother liquor with 1 mM HgCl for 3 h. All crystals were cryo-protected by the reservoir solution supplemented with 5% (vol/vol) ethylene glycol and were flash-frozen in liquid nitrogen for data collection at 100 K.

Data Collection and Structure Determination. X-ray diffraction data of native and heavy-atom–derivative crystals were collected at a wavelength of 0.979 Å on an ADSC Q315R CCD at the 5.0.1 beamline of Advanced Light Source (ALS). The diffraction images were processed with the HKL program (36). The crystals belonged to space group $P2_1$ with cell dimensions of a = 63.0 Å, b = 67.8 Å, c = 98.6 Å, and β = 100.6°. There are four copies of the IntS9–IntS11 complex in the crystallographic asymmetric unit.

A native dataset was collected to 2.1-Å resolution, and the derivative dataset was collected to 2.3-Å resolution. Four Hg atoms were located and used for phasing by the AutoSol routine in PHENIX (37), using the single isomorphous replacement (SIR) method. Most of the protein residues were

automatically built by the AutoBuild routine in PHENIX, and further manual building was carried out with the program Coot (38). The structure was refined using PHENIX. The crystallographic information is summarized in Table S1.

Yeast Two-Hybrid Assays. Yeast two-hybrid assays were carried out in PJ69-4a and PJ49-4alpha. Human IntS11 or IntS9 CTD fragments were cloned into either pOBD or pOAD vectors using conventional cloning. Clones were sequenced to verify identity; PCR primers are available upon request. pOBD plasmids were transformed into PJ69-4a yeast and were selected on tryptophan-dropout medium; pOAD plasmids were transformed into PJ49-4alpha yeast and were selected on leucine-dropout medium. Double transformants were created by mating the yeast strains followed by selection on medium lacking both tryptophan and leucine. Interactions were tested through serial dilution of diploid yeast followed by plating on medium lacking tryptophan and leucine or on medium lacking tryptophan, leucine, and histidine that also was supplemented with 1 mM 3-amino-1,2,4-triazole.

Coimmunoprecipitation. IntS11 and IntS9 cDNAs were cloned into pcDNA3 expression plasmids and were subjected to site-directed mutagenesis as described previously (10). All clones were sequenced to confirm identity. Approximately 5×10^5 293T cells (in one well of a six-well plate) were transfected with 1 μg of each plasmid encoding either HA-tagged or myc-tagged IntS9 or IntS11 using Lipofectamine 2000 according to the manufacturer's instructions (Thermo Fisher). Forty-eight hours after transfection, cells were lysed in 500 μL of denaturing lysis buffer (19), and 50 μL was removed for input lanes. To the remaining lysate, 20 μL of anti-HA affinity resin (Sigma) was added and incubated at 4 °C for 1 h with rotation. Following immunoprecipitation, beads were washed three times in lysis buffer and eluted with SDS loading buffer. Western blots were performed using SDS/PAGE as described previously (19). Affinity purification of FLAG-IntS11 was conducted essentially as described previously (33). Western blots were conducted using antibodies raised to IntS3 (PTGlab), IntS4 (Bethyl), IntS9 (Bethyl), IntS10 (PTGlab), and FLAG epitope (Sigma).

Cell Culture and RNAi Assays. RNAi-rescue experiments were performed using HeLa cells, which were grown under standard conditions using DMEM and 10% FBS. Cells were plated initially at 8.5×10^4 cells per well in a 24-well plate. Cells were transfected with control siRNA (GGUCCGGCUCCCCCA-AAUGdTdT), IntS9 siRNA (GAAAUGCUUUCUUGGACAAdTdT), or IntS11 siRNA (CAGACUUCCUGGACUGUGUdTdT) using a two-hit protocol (39). Twenty-four hours after the second siRNA transfection, cells were transfected a third time with 500 ng of the U7-GFP reporter (19) and with 200 ng of empty pcDNA-myc vector or with pcDNA-myc where either RNAi-resistant wild-type IntS9/IntS11 or mutant versions were cloned. Two days after the transfection, cells were lysed in denaturing buffer and probed using Western blot analysis with antibodies raised against GFP (Clontech), IntS11 (Bethyl), or GAPDH (Thermo). To monitor endogenous snRNA misprocessing, RNA was isolated from cells using TRIzol (Thermo Scientific) and was subjected to MMLV reverse transcription according to the manufacturer's instructions (Life Sciences). Real-time PCR was conducted using SYBR Green PCR mix on a CFX quantitative PCR machine (Bio-Rad), and fold calculation was done as described previously (35). Primers to measure U2snRNA misprocessing are 5′-CTTCGGGGAGAGAACAACC-3′ and 5′-GACACTCAAACACGCGTCA-3′. Primers to measure U4snRNA misprocessing are 5′-GCATTGGCAATTTTTGACAG-3′ and 5′-GAACCCCGGACATTCAATC-3′.

ACKNOWLEDGMENTS. We thank Marc Allaire and Nathan Smith for access to beamline 5.0.1 at the Advanced Light Source. This research was supported by NIH Grants R35GM118093 and S10OD012018 (to L.T.) and by Grants H1880 from the Welch Foundation and Cancer Prevention and Research Institute of Texas Grant RP140800 (to E.J.W.). The Berkeley Center for Structural Biology is supported in part by the NIH, the National Institute of General Medical Sciences, and the Howard Hughes Medical Institute. The Advanced Light Source is supported by the US Department of Energy under Contract DE-AC02-05CH11231.

1. Baillat D, et al. (2005) Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell* 123:265–276.
2. Baillat D, Wagner EJ (2015) Integrator: Surprisingly diverse functions in gene expression. *Trends Biochem Sci* 40:257–264.
3. Yamamoto J, et al. (2014) DSIF and NELF interact with Integrator to specify the correct post-transcriptional fate of snRNA genes. *Nat Commun* 5:4263.
4. Gardini A, et al. (2014) Integrator regulates transcriptional initiation and pause release following activation. *Mol Cell* 56:128–139.
5. Skaar JR, et al. (2015) The Integrator complex controls the termination of transcription at diverse classes of gene targets. *Cell Res* 25:288–305.
6. Stadelmayer B, et al. (2014) Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes. *Nat Commun* 5:5531.
7. Lai F, Gardini A, Zhang A, Shiekhattar R (2015) Integrator mediates the biogenesis of enhancer RNAs. *Nature* 525:399–403.
8. Kapp LD, Abrams EW, Marlow FL, Mullins MC (2013) The integrator complex subunit 6 (Ints6) confines the dorsal organizer in vertebrate embryogenesis. *PLoS Genet* 9:e1003822.
9. Jodoin JN, et al. (2013) The snRNA-processing complex, Integrator, is required for ciliogenesis and dynein recruitment to the nuclear envelope via distinct mechanisms. *Biol Open* 2:1390–1396.
10. Otani Y, et al. (2013) Integrator complex plays an essential role in adipose differentiation. *Biochem Biophys Res Commun* 434:197–202.
11. Obeidat M, et al. (2013) GSTCD and INTS12 regulation and expression in the human lung. *PLoS One* 8:e74630.
12. Xie M, et al. (2015) The host Integrator complex acts in transcription-independent maturation of herpesvirus microRNA 3′ ends. *Genes Dev* 29:1552–1564.
13. Millevoi S, Vagner S (2010) Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation. *Nucleic Acids Res* 38:2757–2774.
14. Yang Q, Doublié S (2011) Structural biology of poly(A) site definition. *Wiley Interdiscip Rev RNA* 2:732–747.
15. Jurado AR, Tan D, Jiao X, Kiledjian M, Tong L (2014) Structure and function of pre-mRNA 5′-end capping quality control and 3′-end processing. *Biochemistry* 53:1882–1898.
16. Romeo V, Schümperli D (2016) Cycling in the nucleus: Regulation of RNA 3′ processing and nuclear organization of replication-dependent histone genes. *Curr Opin Cell Biol* 40:23–31.
17. Pettinati I, Brem J, Lee SY, McHugh PJ, Schofield CJ (2016) The chemical biology of human metallo-b-lactamase fold proteins. *Trends Biochem Sci* 41:338–355.
18. Callebaut I, Moshous D, Mornon J-P, de Villartay J-P (2002) Metallo-beta-lactamase fold within nucleic acids processing enzymes: The beta-CASP family. *Nucleic Acids Res* 30:3592–3601.
19. Albrecht TR, Wagner EJ (2012) snRNA 3′ end formation requires heterodimeric association of integrator subunits. *Mol Cell Biol* 32:1112–1123.
20. Li de la Sierra-Gallay I, Pellegrini O, Condon C (2005) Structural basis for substrate binding, cleavage and allostery in the tRNA maturase RNase Z. *Nature* 433:657–661.
21. Mir-Montazeri B, Ammelburg M, Forouzan D, Lupas AN, Hartmann MD (2011) Crystal structure of a dimeric archaeal cleavage and polyadenylation specificity factor. *J Struct Biol* 173:191–195.
22. Silva AP, et al. (2011) Structure and activity of a novel archaeal β-CASP protein with N-terminal KH domains. *Structure* 19:622–632.
23. Li de la Sierra-Gallay I, Zig L, Jamalli A, Putzer H (2008) Structural insights into the dual activity of RNase J. *Nat Struct Mol Biol* 15:206–212.
24. Mura C, Phillips M, Kozhukhovsky A, Eisenberg D (2003) Structure and assembly of an augmented Sm-like archaeal protein 14-mer. *Proc Natl Acad Sci USA* 100:4539–4544.
25. Schmid EM, et al. (2006) Role of the AP2 beta-appendage hub in recruiting partners for clathrin-coated vesicle assembly. *PLoS Biol* 4:e262.
26. Holm L, Kääriäinen S, Rosenström P, Schenkel A (2008) Searching protein structure databases with DaliLite v.3. *Bioinformatics* 24:2780–2781.
27. Moravcevic K, et al. (2010) Kinase associated-1 domains drive MARK/PAR1 kinases to membrane targets by binding acidic phospholipids. *Cell* 143:966–977.
28. Townley R, Shapiro L (2007) Crystal structures of the adenylate sensor from fission yeast AMP-activated protein kinase. *Science* 315:1726–1729.
29. Amodeo GA, Rudolph MJ, Tong L (2007) Crystal structure of the heterotrimer core of Saccharomyces cerevisiae AMPK homologue SNF1. *Nature* 449:492–495.
30. Xiao B, et al. (2007) Structural basis for AMP binding to mammalian AMP-activated protein kinase. *Nature* 449:496–500.
31. Albrecht R, Zeth K (2011) Structural basis of outer membrane protein biogenesis in bacteria. *J Biol Chem* 286:27792–27803.
32. Dominski Z, Yang X-C, Purdy M, Wagner EJ, Marzluff WF (2005) A CPSF-73 homologue is required for cell cycle progression but not cell growth and interacts with a protein having features of CPSF-100. *Mol Cell Biol* 25:1489–1500.
33. Baillat D, Russell WK, Wagner EJ (2016) CRISPR-Cas9 mediated genetic engineering for the purification of the endogenous integrator complex from mammalian cells. *Protein Expr Purif* 128:101–108.
34. Peart N, Wagner EJ (2016) Gain-of-function reporters for analysis of mRNA 3′-end formation: Design and optimization. *Biotechniques* 60:137–140.
35. Ezzeddine N, et al. (2011) A subset of Drosophila integrator proteins is essential for efficient U7 snRNA and spliceosomal snRNA 3′-end formation. *Mol Cell Biol* 31:328–341.
36. Otwinowski Z, Minor W (1997) Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 276:307–326.
37. Adams PD, et al. (2002) PHENIX: Building new software for automated crystallographic structure determination. *Acta Crystallogr D Biol Crystallogr* 58:1948–1954.
38. Emsley P, Cowtan K (2004) Coot: Model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60:2126–2132.
39. Wagner EJ, Garcia-Blanco MA (2002) RNAi-mediated PTB depletion leads to enhanced exon definition. *Mol Cell* 10:943–949.
40. Armon A, Graur D, Ben-Tal N (2001) ConSurf: An algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* 307:447–463.
41. Gouet P, Courcelle E, Stuart DI, Métoz F (1999) ESPript: Analysis of multiple sequence alignments in PostScript. *Bioinformatics* 15:305–308.