Characterising the molecular basis of WDR82 binding in multiple contexts

Jessica Rose Kelley

Wolfson college, University of Oxford

A thesis submitted in partial fulfilment of the requirements for

the degree of Doctor of Philosophy

Trinity term 2022

Abstract

Precise regulation of gene expression is essential for the development of multicellular organisms. The progression of RNAPII transcription through initiation, elongation and termination is accompanied by mRNA processing events such as capping and splicing, and requires numerous regulatory factors. The extended C-terminal domain (CTD) of the largest RNAPII subunit, RBP1, consists of YSPTSPS heptapeptide repeats that can be post-translationally modified in a number of ways, most notably by phosphorylation. These different phosphorylation states occur in distinctive patterns across genes and have been implicated in regulating various aspects of transcription and mRNA processing. WDR82 is a small WD40-repeat protein which has been shown to bind the RNAPII CTD when it is phosphorylated on serine 5 (S5P), a modification associated with transcription initiation. Removal of WDR82 from cells has profound effects on transcription genome-wide. However, interpreting these results is complicated by the inclusion of WDR82 in three independent complexes that regulate transcription in different ways; the SET1 complexes act at promoters to support gene transcription, the PNUTS-PP1 complex promotes cleavage and polyadenylation-coupled transcription termination, and the ZC3H4 complex mediates premature transcription termination. Whilst the subunit compositions of these complexes have been characterised to varying extents, the molecular basis of WDR82 incorporation into each complex is unknown. Furthermore, the molecular function of WDR82 in these different contexts has not been well characterised. It has been proposed that WDR82 mediates SET1A and ZC3H4 binding to S5P CTD, whilst its role in the PNUTS complex is less clear.

To characterise WDR82 incorporation into these different complexes, I use endogenously tagged cell lines to purify SET1A, ZC3H4, PNUTS, and WDR82 from mouse embryonic stem cells and identify their interactors by mass spectrometry. I define the constitutive components of each complex and identify additional interactors which could provide functional specificity. I then use protein structure prediction combined with *in vivo* validation to characterise the molecular basis of SET1A, ZC3H4, and PNUTS binding to WDR82. I discover that SET1A and ZC3H4 bind WDR82 via remarkably similar interfaces, whereas PNUTS employs a distinct binding mode which nevertheless shares some features with SET1A and ZC3H4. Finally, I begin to characterise the molecular function of WDR82 in each complex. I propose that WDR82 provides an RNAPII CTD binding adapter function to the SET1A and ZC3H4 complexes and contributes to PP1 phosphatase regulation in the PNUTS complex. Together these results characterise the different binding activities of WDR82 and its function in multiple contexts, and provide key tools for further dissection of its role in regulating transcription.

Acknowledgements

First and foremost, I would like to thank Rob Klose for his encouragement, guidance, and support throughout my time in the lab. I am incredibly grateful for your mentorship and all the opportunities you have given me over the last six years.

Thank you to everyone who has helped me with experimental advice and guidance, in particular Emilia Dimitrova for teaching me all the basics, Ed Lowe for help with crystallography, and Kasia Kowalczyk for help with mass spec sample preparation. Thank you to Marjorie Fournier and Aygul Malone for mass spectrometry services. Thank you to Emilia for reading parts of this thesis

Thank you to Amy, Miles, Emma and the rest of the Klose and Brockdorff labs for creating a friendly and welcoming environment to work in. Thank you to Tatyana and Emma for your tireless efforts to keep the lab environment running so smoothly.

Special thanks go to Amy, my partner in crime in all things SET1, WDR82, and beyond. I am incredibly grateful to have shared this research adventure with you.

Thank you to all my friends for the love, laughter and adventures. A special thank you to my housemates Jake, Anna, Tom, Ben, and Emily for your companionship through the good times and the lockdowns alike. Thank you to Oli for proofreading parts of this thesis.

Most of all, thank you to my family for your love and support in everything I do.

Declaration of Authorship

I declare that all work presented in this thesis is my own, unless otherwise acknowledged.

This thesis has not been submitted, either partially or in full, for another degree, diploma, certificate or other qualification at this University or at any other institution.

Jessica Rose Kelley

October 2022

Table of Contents

List of figures

List of tables

List of common abbreviations

bp	base pair
cDNA	coding DNA
CGI	CpG Island
ChIP	Chromatin immunoprecipitation
ChIP seq	ChIP followed by massively parallel sequencing
CPA	Cleavage and polyadenylation
CpG	Cytosine-phosphate-guanine
CPSF	Cleavage and polyadenylation specificity factor
CTD	RPB1 C-terminal domain
CTR	Spt5 C-terminal region
DRE	Distal regulatory element
DSIF	DRB sensitivity inducing factor
dTAG	$FKBP12^{F36V}$
EC	Elongation complex
ESC	Embryonic stem cell
FL	Full-length
GST	Glutathione-S-Transferase
GTFs	General transcription factors
H3K4me3	Histone H3 Lysine 4 trimethylation
HMT	Histone methyltransferase
HRP	Horseradish peroxidase
IP	Immunoprecipitation
MSA	Multiple sequence alignment
NELF	Negative elongation factor
nt	Nucleotide
NTD	N-terminal domain
PAE	Predicted alignment error
PAS	Polyadenylation signal
PBS	Phosphate buffered saline
PBST	PBS + 0.01% Tween-20
PHD	Plant homeodomain
PIC	Pre-initiation complex
PIP	PP1 interacting protein
pLDDT	Predicted local distance difference test
PNUTS	PP1 Nuclear Targeting Subunit
PP1	Protein phosphatase 1
P-TEFb	Positive transcription elongation factor b
PTM	Post-translational modification
PTT	Premature transcription termination
RMSD	Root-mean square deviation
RNAPII	RNA Polymerase II
RRM	RNA recognition motif
RT	Room temperature

S2P	RNAPII CTD phosphorylated on serine 2
S5P	RNAPII CTD phosphorylated on serine 5
S7P	RNAPII CTD phosphorylated on serine 7
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
SET	Su(var)3-9, Enhancer-of-zeste and Trithorax
ST7	Triple-T7 and Twin-Strep tag
TAF	TBP-associated factor
TBP	TATA-binding protein
TES	Transcription end site
TSS	Transcription Start site
TTseq	Isolation of the transient transcriptome followed by massively parallel sequencing
UNT	Untreated
WRAD	WDR5, RBBP5, ASH2L, DPY30
WT	Wild type
ZF	Zinc finger
ZF-CxxC	Zinc finger - CxxC type

1 Introduction

1.1 Regulation of eukaryotic gene expression

Precise regulation of gene expression is a key process which enables cells to respond to external stimuli and to specialise their functions. In complex multicellular organisms, establishment of differential gene expression patterns enables cell type differentiation and hence organism development. As the first stage of gene expression, transcription is an important step at which regulation can occur.

1.1.1 Transcription

Transcription of eukaryotic protein-coding genes is carried out by RNA Polymerase II (RNAPII), a DNA-templated RNA polymerase (Roeder and Rutter, 1969; Sentenac, 1985). In the broadest model of the transcription cycle, RNAPII is loaded onto genes at a promoter, where it opens the DNA duplex and begins RNA synthesis (Cramer, 2019). Following transcription initiation and escape from the promoter, RNAPII elongates through the gene until it reaches a termination signal, which causes it to release the newly synthesised RNA and the DNA template (Proudfoot, 2016). The polymerase is then free to be re-loaded at the promoter to catalyse further rounds of transcription. The number of gene transcripts produced in a certain time period is therefore dependent on each of these stages: the number of initiation events, the speed of elongation, and the efficiency of termination. Furthermore, the process of transcription is closely coupled with mRNA processing steps such as capping, splicing, and 3' end processing (Bentley, 2014). Regulation of transcription and its coupling to these processes can thereby significantly influence downstream events such as mRNA export, stability, and translation, which rely on correct mRNA processing (Buccitelli and Selbach, 2020; Jurado et al., 2014).

This cycle of RNAPII initiation, elongation, termination and recycling is highly regulated at each stage by association of factors with RNAPII, and by features of the DNA template (Cramer, 2019). Specific DNA sequences are bound by transcription factors which can regulate the process of transcription positively or negatively (Lambert et al., 2018). Furthermore, the template DNA is not free in the nucleus, but is packaged into chromatin (Kornberg and Lorch, 1999). Nucleosomes act as a barrier to the passage of RNAPII and must be removed and re-established as transcription proceeds (Knezetic and Luse, 1986; Lorch et al., 1987). Alterations to nucleosome positioning or structure, or binding of additional factors to histone post-translational modifications (PTMs), can affect how easily RNAPII can pass through a gene, modulating its expression (Clapier and Cairns, 2009; Morgan and Shilatifard, 2020; Rothbart and Strahl, 2014; Workman and Kingston, 1998).

Whilst the majority of RNAPII transcriptional activity is confined to protein-coding genes, it has also been found to transcribe some non-coding RNAs and extragenic regions (Carninci et al., 2005; de Santa et al., 2010; Kapranov et al., 2007; Katayama et al., 2005). For example, low levels of transcription are found at enhancers (de Santa et al., 2010; Kim et al., 2010; Koch et al., 2011). Furthermore, transcription is in general bidirectional: RNAPII can be loaded at the TSS in either orientation, leading to the production of antisense transcripts upstream of promoters (Barman et al., 2019; Katayama et al., 2005; Preker et al., 2008). These transcriptional activities are also subject to close regulation (Barman et al., 2019; Henriques et al., 2018; Porrua and Libri, 2015).

1.1.2 Post-translational modification of RNAPII

RNAPII is a large multiprotein complex, the largest subunit of which, RPB1, contains a long unstructured C-Terminal domain (CTD) (Corden, 1990; Jasnovidova and Stefl, 2013; Meinhart et al., 2005). The RPB1 CTD (herein referred to as the RNAPII CTD, or simply the CTD) is not required for RNAPII catalysis, but is required for the regulation of transcription

and its coupling to cotranscriptional processes (Buratowski, 2009, 2003; Cossa et al., 2021; Cramer, 2019; Harlen and Churchman, 2017; Parua and Fisher, 2020). The CTD consists of a large number of heptapeptide repeats of the consensus sequence YSPTSPS, ranging from 26 repeats in yeast to 52 in humans (Corden, 1990; Corden et al., 1985; P. Liu et al., 2010). These repeats can be post-translationally modified in a variety of ways, but most notably by phosphorylation of the non-proline residues (Buratowski, 2009, 2003; Cossa et al., 2021; Harlen and Churchman, 2017). RNAPII is recruited to gene promoters in the unmodified state, but is highly phosphorylated during transcription (Buratowski, 2009; Lu et al., 1991). These different phosphorylations are found to occur in distinctive patterns across genes (Figure 1.1) (Cossa et al., 2021; Harlen and Churchman, 2017). The activities of kinases and phosphatases at different stages in the transcription cycle give rise to this so-called 'CTD code', which distinguishes, for example, initiating or elongating RNAPII (Buratowski, 2003).

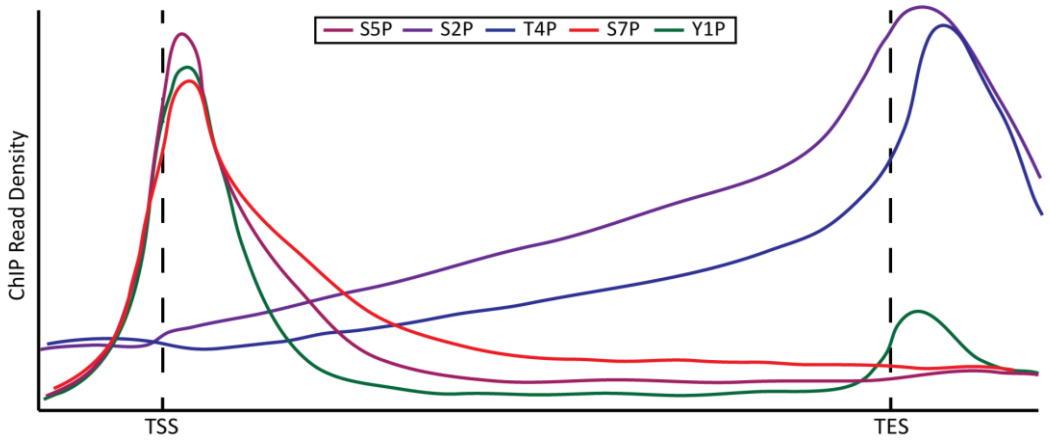


Figure 1.1 Schematic of average ChIP profiles of phosphorylated RNAPII**.** Schematic metaplots showing typical chromatin immunoprecipitation (ChIP) densities for the mammalian RNAPII CTD phosphorylated on tyrosine 1 (Y1P), serine 2 (S2P), threonine 4 (T4P), serine 5 (S5P) and serine 7 (S7P). Figure courtesy of Amy Hughes, adapted from Harlen and Churchman, 2017.

During initiation, the CTD is phosphorylated on serine 5 and serine 7 (S5P and S7P) by Cdk7 (Buratowski, 2009; Core and Adelman, 2019; Glover-Cutter et al., 2009; Kim et al., 2009). The prevalence of these modifications peaks shortly after the TSS then tails off following the transition to elongation due to dephosphorylation by phosphatases including PP1 and PP2A. During elongation, serine 2 and threonine 4 become more heavily

phosphorylated (S2P and T4P), peaking shortly after the transcription end site (TES) (Cossa et al., 2021; Harlen and Churchman, 2017). S2P is initially deposited by Cdk9 during the transition to productive elongation and is believed to be amplified by Cdk12 and Cdk13 as elongation progresses (Booth et al., 2018; Jasnovidova and Stefl, 2013; Parua et al., 2020; Parua and Fisher, 2020). Importantly, following transcription termination the CTD is dephosphorylated by a number of phosphatases to return it to the unmodified state required for new initiation events (Cossa et al., 2021).

Whilst phosphorylation is believed to alter the physical properties of the CTD, such as inducing its decompaction, the primary function of the CTD is as a protein-protein interaction interface, the specificity of which is altered by its post-translational modification (Harlen and Churchman, 2017; Jasnovidova and Stefl, 2013; Meinhart et al., 2005). The CTD provides a 'landing pad' for proteins involved in regulating transcription, mRNA processing, and the chromatin environment, and the repertoire of CTD binders is dictated at different stages of the transcription cycle by its phosphorylation (Harlen and Churchman, 2017; Srivastava and Ahn, 2015). The CTD is therefore a critical component in the coordination of transcription itself and its coupling with processes that have downstream effects on gene expression.

1.1.3 The process of transcription

1.1.3.1 Initiation

The start site from which transcription begins is embedded within a DNA sequence known as the core promoter (Haberle and Stark, 2018). RNAPII alone has low affinity for the core promoter and cannot efficiently initiate transcription. Instead, the core promoter is bound by the general transcription factors (GTFs), which enable RNAPII loading onto DNA, and the initiation of transcription (Figure 1.2) (reviewed in Schier and Taatjes, 2020).

Extensive biochemical and structural studies have revealed how transcription is initiated through a stepwise process involving the six main GTFs (Buratowski et al., 1989; Schier and Taatjes, 2020; Van Dyke et al., 1988). The first of these is TFIID, a large complex of TATA binding protein (TBP) and 13 TBP-associated factors (TAFs) which nucleates formation of the pre-initiation complex (PIC) (Figure 1.2A) (Patel et al., 2020, 2018). TBP and several TAFs recognise DNA sequences within the core promoter which direct the correct positioning of RNAPII with respect to the TSS (Louder et al., 2016; Patel et al., 2018). TBP is of particular importance during this initial phase of transcription initiation. TBP binds 25-30bp upstream of the TSS where it distorts the DNA, inducing a 90° kink which releases repressive interactions within TFIID to enable further PIC assembly (Patel et al., 2018; Schier and Taatjes, 2020). TFIIA and TFIIB bind to the TFIID-DNA complex, stabilising the distorted DNA conformation and allowing association of TFIIF and RNAPII (Hieb et al., 2007; Imbalzano et al., 1994). The RNAPII CTD is unmodified at this stage, and TFIIF has been proposed to promote CTD dephosphorylation to help maintain this state (Schier and Taatjes, 2020). The final GTFs, TFIIE and TFIIH, then associate with the PIC (Figure 1.2B).

A key function of the PIC is to enable the conformational changes in the DNA template and RNAPII which underpin promoter opening (Schier and Taatjes, 2020). The XPB subunit of TFIIH uses the energy from ATP hydrolysis to unwind the template DNA and propel it into the RNAPII active site, allowing RNA synthesis to begin (He et al., 2016; Kim et al., 2000; Tirode et al., 1999). Once more than 10 nucleotides have been synthesised, RNAPII undergoes ‘promoter escape’. This occurs due to reannealing of the DNA template, which provides sufficient energy to propel the polymerase forward, disrupting upstream contacts with the promoter and initiation factors (Kostrewa et al., 2009; X. Liu et al., 2010). These conformational changes in RNAPII result in formation of a stable DNA:RNA hybrid in the active site and a transition to elongation (Bernecky et al., 2016).

Transcription initiation is a key stage at which regulation can occur to modulate the timing and frequency of transcription events (Cramer, 2019). Whilst sufficient to stimulate PIC assembly, core promoters support only a low level of basal transcription and the activity of numerous additional factors is required to support more robust promoter activity (Haberle and Stark, 2018; Kadonaga, 2012; Lambert et al., 2018). Binding of DNA-sequence-specific transcription factors (TFs) to cis-regulatory elements in close proximity to gene promoters can stimulate the activity and/or stability of the PIC (Frietze and Farnham, 2011). For example, some TFs modulate the binding of the GTFs to the core promoter sequences to directly stimulate PIC assembly (Horikoshi et al., 1988). In addition to binding of factors directed by the DNA sequence, PIC formation can also be regulated through changes in the local chromatin environment. Histone modifications can directly alter the physical properties of the chromatin or enable binding of additional factors which promote or inhibit PIC formation (Bannister and Kouzarides, 2011; Hughes et al., 2020a; Hyun et al., 2017; Rothbart and Strahl, 2014; Tropberger and Schneider, 2010). Furthermore, chromatin remodellers can alter nucleosome positioning to modulate the accessibility of core promoter DNA (Knezetic and Luse, 1986; Lorch et al., 1987; Workman and Kingston, 1998). The presence of multiple regulatory mechanisms at or near promoters allows for combinatorial control and fine-tuning of promoter outputs (Cramer, 2019).

In addition to promoter-proximal regulatory elements, distal regulatory elements (DREs) such as enhancers, positioned up to kilobases away from the promoter, can also influence PIC assembly (Shlyueva et al., 2014; Spitz and Furlong, 2012). This effect is mediated by the Mediator complex, which bridges between DREs and core promoters (Allen and Taatjes, 2015; Fournier et al., 2016; Quevedo et al., 2019; Soutourina, 2018). Mediator makes numerous contacts with the PIC, in particular via TFIIF and by binding to the unmodified RNAPII CTD (Schier and Taatjes, 2020). Mediator stabilises the PIC and facilitates entry into productive elongation by stimulating the TFIIH kinase subunit Cdk7

(Figure 1.2C) (Robinson et al., 2016; Soutourina, 2018). Cdk7 phosphorylates the RNAPII CTD on serine 5 and serine 7, which disrupts mediator binding and facilitates promoter escape (Figure 1.2D) (Wong et al., 2014).

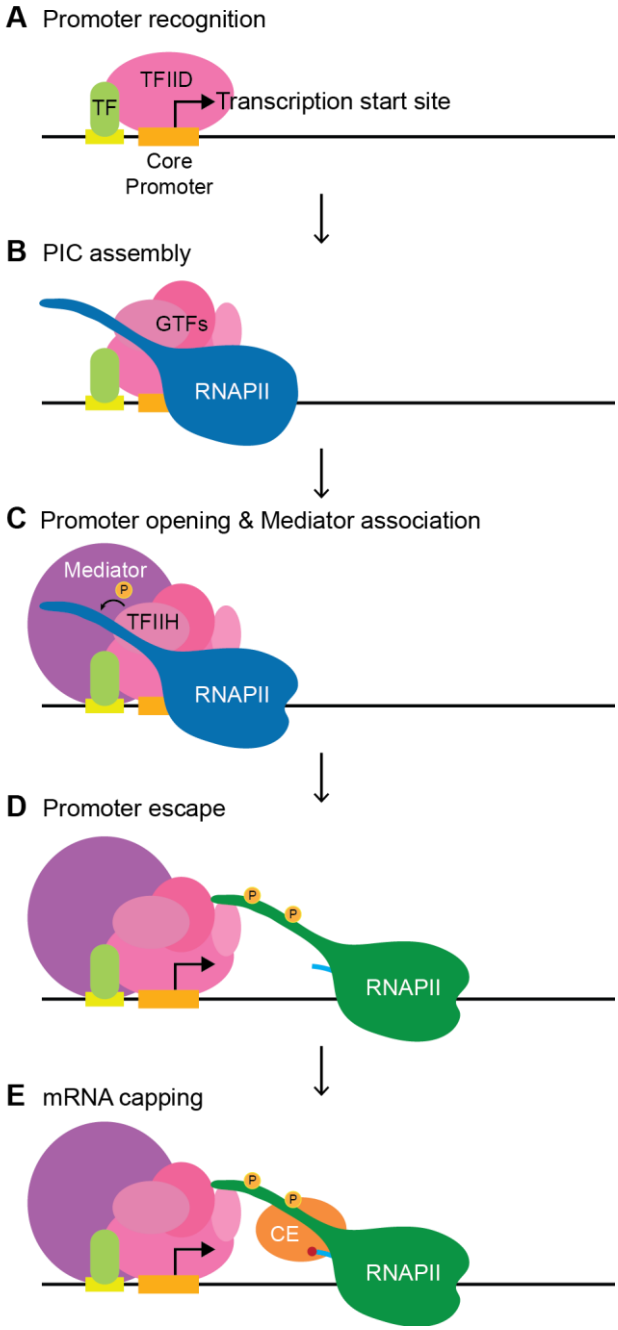


Figure 1.2 Model of transcription initiation. **A:** The core promoter (orange), which contains the transcription start site, is recognised by TFIID. Transcription factors (TFs) can bind regulatory elements (light green) and influence PIC assembly. **B:** Pre-initiation complex (PIC), consisting of the general transcription factors (pink) and RNAPII, assembles on the core promoter. **C:** Promoter opening allows RNAPII to engage with the template DNA. Mediator associates with the PIC and stimulates RNAPII CTD phosphorylation on serine 5 by TFIIH. **D:** Synthesis of the first 10-12nt of RNA (cyan) induces conformational changes in RNAPII which propel it forward and disrupt upstream contacts with the promoter and initiation factors. Serine 5 phosphorylation on the CTD disrupts mediator binding and facilitates promoter escape. **E:** The capping enzyme (CE) binds the nascent RNA and S5P CTD and catalyses addition of the mRNA cap (red). Figure adapted from Core & Adelman, 2019

The deposition of S5P and S7P by TFIIH is a key feature of transcription initiation and facilitates a number of associated processes by enabling specific binding of regulatory factors to the CTD (Buratowski, 2003; Harlen and Churchman, 2017). In particular, S5P CTD is recognised by the mRNA capping enzyme and allosterically stimulates its activity (Figure

1.2E) (Buratowski, 2009). The CTD extends from RNAPII close to the RNA exit channel and therefore positions the capping complex to act as soon as the nascent transcript emerges, when it is approximately 19-22nt in length (Core and Adelman, 2019). S5P CTD can also be bound by factors which modify the promoter chromatin environment in response to transcription (Srivastava and Ahn, 2015). For example, the SET1A/B histone methyltransferases associate with S5P CTD and deposit histone H3K4me3 at promoters, a histone modification associated with active transcription (Lee and Skalnik, 2008; Ng et al., 2003). This has been proposed to create a chromatin environment that is more permissive to future rounds of transcription (Hughes et al., 2020a). The role of S7P around transcription initiation is less certain, although it has been proposed to be important as a 'priming' event to enable subsequent phosphorylations of the CTD, ensuring a sequential progression in the pattern of CTD modifications through the transcription cycle (Buratowski, 2009; Core and Adelman, 2019; Parua and Fisher, 2020).

1.1.3.2 Promoter-proximal pausing

Following initiation, RNAPII transcribes for a short distance before pausing approximately 50nt downstream of the TSS prior to entering fully productive elongation (Core and Adelman, 2019; Fraser et al., 1978). This promoter-proximal pausing was first identified at heatshock genes in *Drosophila* and was believed to poise genes for rapid reactivation, however it has since been identified as a pervasive feature of protein-coding genes (Giardina et al., 1992; Gilmour and Lis, 1986; Jonkers et al., 2014; Muse et al., 2007; Rahl et al., 2010; Rasmussen and Lis, 1993; Rougvie and Lis, 1988; Zeitlinger et al., 2007). Pausing is a highly regulated process, with levels of paused polymerase dependent on relative levels of initiation, entry into the paused state, and release from pausing either by entry into productive elongation, or by termination and RNAPII eviction from the template DNA (Core and Adelman, 2019). An accumulation of paused RNAPII detected at most genes

suggest its recruitment and initiation is typically faster than its release from pausing (Muse et al., 2007; Zeitlinger et al., 2007). Pausing therefore represents a significant bottleneck in transcription and its regulation can have profound consequences for the ultimate transcriptional output of genes (Cramer, 2019). Pausing has been proposed to primarily regulate transcription by sterically blocking the promoter to prevent re-initiation events and hence limit transcriptional output (Shao and Zeitlinger, 2017). Alternatively, it has been suggested that pausing may contribute to the maintenance of a nucleosome depleted region at the promoter to enable future rounds of transcription, or it may serve to slow the early phase of elongation to allow for mRNA capping to occur prior to extensive transcript elongation (Adelman and Lis, 2012). Furthermore, pausing of RNAPII may allow sufficient time for integration of additional signals that regulate transcription, either positively or negatively (Core and Adelman, 2019).

Paused RNAPII is stabilised by two key factors, DRB-sensitivity inducing factor (DSIF) and negative elongation factor (NELF) (Figure 1.3A) (Core and Adelman, 2019; Yamaguchi et al., 1999). DSIF is a dimeric complex of Spt4 and Spt5, which binds stably to RNAPII (Ivanov et al., 2000; Wada et al., 1998). Its binding requires the release of initiation factors including TFIIE, which would otherwise occlude its binding surface (Bernecky et al., 2017; Core and Adelman, 2019; Larochelle et al., 2012; Vos et al., 2018b, 2018a). Spt5 also binds the nascent RNA, which emerges from the exit channel after transcription of 18-20nt, ensuring Spt5 binding shortly following promoter escape (Core and Adelman, 2019; Missra and Gilmour, 2010; Qiu and Gilmour, 2017). These direct contacts with the nascent RNA may also dictate where pausing occurs (Qiu and Gilmour, 2017). Spt5 also associates with the capping complex to stimulate rapid and efficient capping of the nascent transcript (Pei and Shuman, 2002; Wen and Shatkin, 1999). NELF is a four subunit complex consisting of NELFA, NELFB, NELFC or NELFD, and NELFE (Yamaguchi et al., 1999). It binds at the RNAPII 'funnel', where NTPs enter the active site, and also contacts the RNAPII-Spt5 interface

(Cheng and Price, 2007; Vos et al., 2018b). Whilst NELF is not strictly required to initiate pausing, it stabilises and extends the lifetime of paused complexes (Adelman and Henriques, 2018; Cheng and Price, 2007; Core and Adelman, 2019; Vos et al., 2018b). Mechanistically, NELF stabilises paused RNAPII and prevents reactivation of its catalytic site by blocking entry of NTPs to the funnel and stabilising distortions of the active site which restrict its mobility and prevent translocation (Adelman and Henriques, 2018; Vos et al., 2018a, 2018b). NELF also prevents paused RNAPII from being erroneously restarted by blocking the binding site of TFIIS, which rescues stalled and backtracked polymerases by stimulating RNA cleavage (Cheung and Cramer, 2011; Core and Adelman, 2019; Kettenberger et al., 2003; Palangat et al., 2005; Vos et al., 2018b). Despite this stabilisation of pausing, it has recently been shown that a large proportion of paused RNAPII is turned over or terminated rather than entering into productive elongation (Erickson et al., 2018; Kamieniarz-Gdula and Proudfoot, 2019; Krebs et al., 2017; Steurer et al., 2018).

The small fraction of paused RNAPII that does continue to the elongation phase undergoes a highly regulated release from pausing (Core and Adelman, 2019). Pause release is triggered by Positive Transcription Elongation Factor b (P-TEFb), in particular the kinase activity of its Cdk9 subunit (Chao and Price, 2001; Marshall and Price, 1995). P-TEFb phosphorylates Spt5, NELF and the RNAPII CTD on serine 2 (Figure 1.3B) (Fujinaga et al., 2004; Marshall et al., 1996; Yamada et al., 2006; Zhou et al., 2012). Phosphorylation of Spt5 on its C-terminal repeat region (CTR) triggers NELF dissociation and allows RNAPII to release into productive elongation (Figure 1.3C) (Cheng and Price, 2007; Vos et al., 2018a; Wu et al., 2003; Yamada et al., 2006). Whilst P-TEFb deposits serine 2 phosphorylation on the RNAPII CTD, this modification is not required for pause release (Cheng and Price, 2007; Guo et al., 2000; Lu et al., 2016; Yamada et al., 2006). Instead, S2P CTD is believed to function as a marker of RNAPII that has entered into productive elongation, allowing for coordination of related activities by specific recognition of this modification (Harlen and Churchman,

2017; Srivastava and Ahn, 2015). Interestingly, serine 2 phosphorylation by P-TEFb has been proposed to require prior phosphorylation of serine 7 by Cdk7 to 'prime' the substrate, ensuring S2P is only deposited after initiation has occurred (Buratowski, 2009; Core and Adelman, 2019; Czudnochowski et al., 2012).

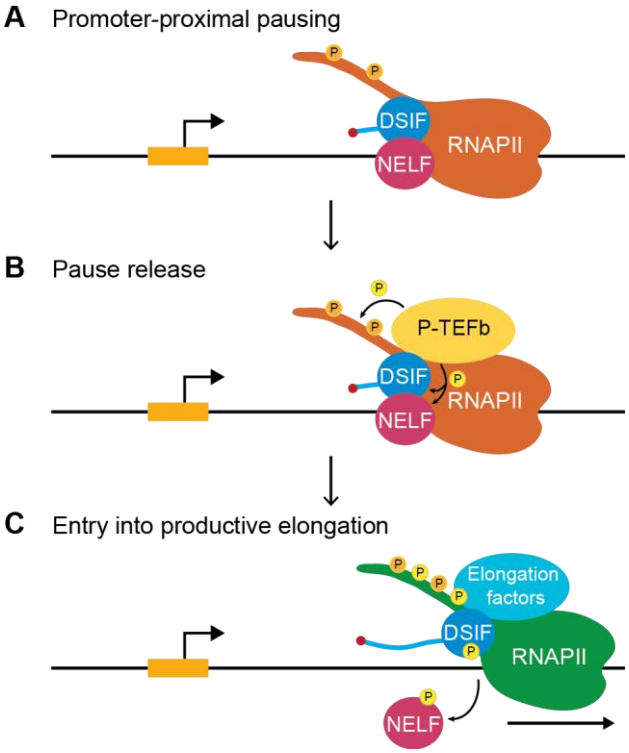


Figure 1.3 Model of promoter-proximal pausing and release into productive elongation. **A:** RNAPII transcription pauses approximately 50nt downstream of the transcription start site. The paused polymerase is stabilised by DSIF and NELF. **B:** P-TEFb phosphorylates DSIF, NELF, and the RNAPII CTD on serine 2, stimulating pause release. **C:** Phosphorylation of DSIF causes NELF dissociation, allowing RNAPII to enter productive elongation. Elongation factors associate with phosphorylated DSIF and S2P CTD. Figure adapted from Core & Adelman, 2019.

1.1.3.3 Elongation

Following pause release RNAPII enters into productive elongation, extending the nascent RNA in a processive manner (Cramer, 2019; Zhou et al., 2012). Spt5, which is phosphorylated by P-TEFb to trigger pause release, remains associated with the elongating polymerase, promoting RNAPII processivity and stabilising the elongating complex (Bernecky et al., 2017; Cheng and Price, 2007; Shetty et al., 2017; Vos et al., 2018a, 2018b; Wada et al., 1998). Furthermore, the phosphorylated CTR of Spt5 enables association of a number of factors which stimulate RNAPII activity and RNA processing, in a similar manner to the RNAPII CTD (Harlen and Churchman, 2017; Hartzog and Fu, 2013; Yamada et al., 2006). In particular, phosphorylation of Spt5 and the RNAPII CTD stimulates association of elongation factors such as the PAF1 complex and Spt6, which help elongating RNAPII pass

through inherent barriers, such as tightly wrapped nucleosomes, that could cause it to stall (Core and Adelman, 2019; Harlen and Churchman, 2017; Kwak et al., 2013; Weber et al., 2014). The PAF1 complex binds the core RNAPII enzyme at a site previously occupied by NELF and, along with Spt6, may aid transcription through the first few nucleosomes, which slow RNAPII and render it susceptible to termination (Adelman et al., 2006; Core and Adelman, 2019; Shi et al., 1996; Van Oss et al., 2016; Vos et al., 2018a). Spt6 and the FACT complex, which also associates with elongating RNAPII, are histone chaperones which facilitate disassembly and reassembly of nucleosomes as the elongation complex (EC) passes through, promoting processive elongation and maintaining chromatin organisation (Ardehali et al., 2009; Belotserkovskaya et al., 2004; Core and Adelman, 2019; Guo et al., 2000; Kaplan et al., 2003).

During pause release, P-TEFb also deposits serine 2 phosphorylation on the RNAPII CTD, a mark of productive elongation which increases gradually towards the 3' end of genes (Harlen and Churchman, 2017). The contribution of P-TEFb towards total S2P is, however, believed to be low and the bulk of this modification is instead deposited by Cdk12 and Cdk13, which associate with RNAPII as it elongates downstream (Bartkowiak et al., 2010; Blazek et al., 2011). It has been reported that, similarly to the requirement for prior S7P for efficient serine 2 phosphorylation by P-TEFb, Cdk12 also requires prior phosphorylation to 'prime' the substrate (Bösken et al., 2014; Czudnochowski et al., 2012; Greifenberg et al., 2016). This mechanism may explain the gradual increase in S2P across gene bodies, as phosphorylation-dependent phosphorylation by Cdk12/13 could amplify this modification as elongation progresses (Buratowski, 2009). These conditional phosphorylation events ensure patterns of CTD modification occur only in a strict sequence that is coupled to transcription. Interestingly, S2P does not seem to be required for elongation *per se*, but instead coordinates events that occur alongside elongation, such as splicing (Buratowski, 2009; Parua and Fisher, 2020).

The rate of transcription elongation is highly variable both within and between genes, ranging from 0.5kb/min to more than 5kb/min (Danko et al., 2013; Fuchs et al., 2014; Jonkers et al., 2014). The speed of transcription can have profound effects on cotranscriptional processes, in particular splicing (Bentley, 2014). Assembly of the splicing machinery on a splice site takes time, and a slower-moving polymerase provides a larger 'window of opportunity' for the splicing machinery to assemble on a weaker splice site before the transcription of downstream RNA elements which could outcompete the weaker upstream splice site (Bentley, 2014; Dujardin et al., 2013). Various factors have been shown to affect elongation rate including chromatin structure, properties of the elongating RNAPII itself, and factors which associate with the elongation complex (Jonkers et al., 2014; Zhou et al., 2012). Interestingly, Spt5 has been shown to affect both the processivity of transcription – i.e. the fraction of elongation complexes that complete a full-length transcript – and the speed of elongation (Hartzog et al., 1998; Shetty et al., 2017; Wada et al., 1998). Phosphorylation of the Spt5 CTR is associated with the promotion of elongation and increased RNAPII speed, whilst its dephosphorylation by the PNUTS-PP1 phosphatase complex is proposed to restrict RNAPII speed throughout gene bodies (Cortazar et al., 2019).

1.1.3.4 Termination

The final stage of transcription is termination, which releases RNAPII and the newly synthesised RNA transcript after transcription of anywhere from tens to over one million nucleotides (Eaton and West, 2020). Canonically this occurs at a polyadenylation signal (PAS) at the 3' end of protein-coding genes, however RNAPII is vulnerable to termination as soon as transcription begins, and premature transcription termination can occur almost anywhere on genes (Porrua and Libri, 2015). Termination is also a key mechanism by which

extragenic transcription, such as promoter upstream antisense and enhancer transcription, is restricted.

1.1.3.4.1 Termination at 3' ends

Most protein-coding genes have a PAS which defines the 3' end of the mRNA and directs termination of RNAPII transcription (Connelly and Manley, 1988; Proudfoot, 2011). Termination at the ends of genes delineates each transcriptional unit to prevent interference between neighbouring genes and enables appropriate mRNA processing to facilitate the stability and cellular localisation of transcripts (Porrua and Libri, 2015). The cleavage and polyadenylation (CPA) machinery associates with the elongation complex via contacts with S2P CTD and the nascent RNA (Proudfoot, 2011; Sun et al., 2018). Following its transcription, the CPA machinery recognises the PAS, cleaves the pre-mRNA and catalyses addition of a long poly(A) tail (Casañal et al., 2017; Shi et al., 2009; Sun et al., 2018). RNAPII remains associated with, and actively transcribing, the template DNA throughout this process and must then be terminated (Eaton et al., 2020; Eaton and West, 2020).

Various models have been proposed for how transcription is terminated after the PAS, however the current prevailing view is a unified model which encompasses both the previously proposed 'allosteric' and 'torpedo' models (Eaton et al., 2020; Proudfoot, 2016). In this 'sitting duck torpedo' model of transcription termination (Figure 1.4), an allosteric switch decelerates RNAPII elongation towards the 3' end of genes, slowing it from approximately 2kb/min to as little as 0.1kb/min. This deceleration is coupled to PAS recognition and relies on dephosphorylation of the Spt5 CTR by the PNUTS-PP1 phosphatase complex (Cortazar et al., 2019; Parua et al., 2018). Following cleavage and polyadenylation of the pre-mRNA, the Xrn2 5'→ 3' exonuclease is loaded onto the newly exposed 5' end of the RNA exiting the still-transcribing RNAPII (Kim et al., 2004; West et al.,

2004). Xrn2 progresses along this RNA until it 'catches up' with RNAPII, inducing termination of transcription (Cortazar et al., 2019). In order for Xrn2 to effectively terminate transcription, it must outpace RNAPII. Rapid transcription will prolong this 'race', whilst slow transcription will lead to more proximal termination (Eaton and West, 2020). The CPA-associated changes in RNAPII which slow its elongation facilitate more proximal termination by leaving it a 'sitting duck' for the Xrn2 'torpedo' to catch (Cortazar et al., 2019).

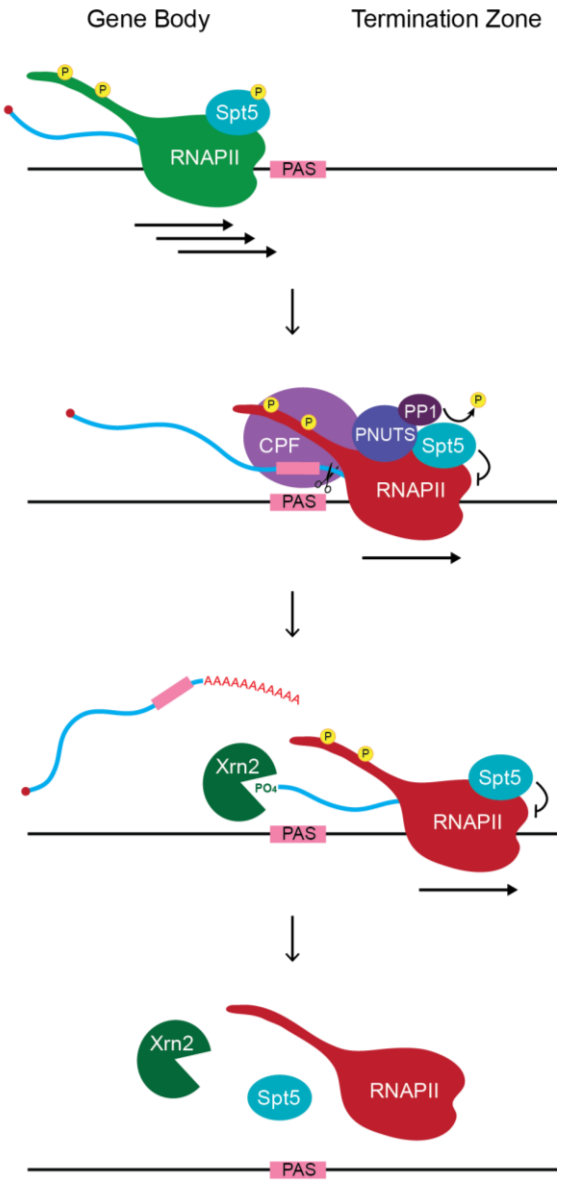


Figure 1.4 Model of the sitting duck torpedo mechanism of transcription termination. Elongating polymerase carries S2P CTD, and Spt5 is phosphorylated to enable a higher rate of transcription. When the PAS has been transcribed, the cleavage and polyadenylation factor (CPF) associates with the nascent RNA and S2P CTD, stimulating RNA cleavage. The PNUTS-PP1 complex concurrently associates with RNAPII and dephosphorylates Spt5, inducing a slowing of transcription. Following mRNA cleavage and polyadenylation, RNAPII continues to transcribe slowly and Xrn2 is loaded onto the free 5' end of the RNA transcript. The slow elongation speed of RNAPII allows Xrn2 to rapidly 'catch up' with RNAPII and induce termination. Figure adapted from Cortazar et al., 2019.

1.1.3.4.2 Premature transcription termination

Whilst the 'canonical' transcription termination mechanisms at the 3' end of genes are reasonably well-characterised, premature transcription termination (PTT) is emerging as a potentially important mechanism for transcriptional regulation (Kamieniarz-Gdula and Proudfoot, 2019). Indeed, it has been found that a large fraction of promoter-proximally paused RNAPII undergoes PTT rather than entering productive elongation (Erickson et al., 2018; Krebs et al., 2017; Steurer et al., 2018). PTT can also occur further into genes when RNAPII encounters the stable nucleosomes at the boundary of promoter-associated CGIs, and the presence of cryptic poly(A) sites throughout genes can induce PTT and production of alternative transcripts (Almada et al., 2013; Chiu et al., 2018; Vlaming et al., 2022). Early termination can thereby have profound consequences for final gene expression outputs by dictating how many polymerases reach the end of genes and produce complete mature transcripts (Eaton and West, 2020). In addition, PTT is also an important process in limiting non-genic transcription. Suppression of antisense, enhancer, and intergenic transcription by premature termination is important to prevent a build-up of non-functional RNAs (Kamieniarz-Gdula and Proudfoot, 2019). This is facilitated by the nuclear exosome, which rapidly degrades the short transcripts produced following PTT (Schmid and Jensen, 2018).

The mechanisms by which RNAPII can undergo PTT are not well understood and may be context-specific. The Integrator complex binds to RNAPII and has been implicated in promoter-proximal PTT through cleavage of the nascent RNA by the INTS11 endonuclease, whilst the CPA machinery has been found to aid PTT further into genes at cryptic poly(A) sites (Andersen et al., 2012; Beckedorff et al., 2020; Elrod et al., 2019; Kamieniarz-Gdula et al., 2019; Li et al., 2015; Lykke-Andersen et al., 2021; Ntini et al., 2013; Richard and Manley, 2009; Tatomer et al., 2019). A more recently identified factor, ZC3H4, has been shown to attenuate transcription of extragenic and lncRNA transcripts, however the precise mechanism of this effect is unclear (Austenaa et al., 2021; Estell et al., 2021). There also

exists a number of processes which can prevent PTT, particularly within genes, to promote continued elongation. For example, the U1 snRNP binds to 5' splice sites in nascent RNA and inhibits CPA activity at downstream cryptic poly(A) sites (Berg et al., 2012; Kaida et al., 2010; Oh et al., 2017). SCAF4 and SCAF8, which interact with early elongating RNAPII by binding to CTD which carries both S2P and S5P, have been found to supress use of poly(A) sites within genes (Gregersen et al., 2019). TFIIS is also a key factor in preventing premature termination by restarting stalled, backtracked RNAPII (Zatreanu et al., 2019). Whilst the mechanisms of PTT are poorly understood at present, these examples demonstrate how control of termination away from the 3' end of genes can provide an additional layer of transcriptional regulation.

1.2 WDR82 and its role in transcriptional regulation

The majority of proteins which act to regulate eukaryotic transcription do not do so alone, but rather as part of large multiprotein complexes. These complexes bring together different activities, such as enzymes, protein or nucleic acid binding modules, and regulatory subunits to allow for highly spatiotemporally specific activity. Importantly, some individual protein subunits are found in multiple such complexes, where they can perform similar molecular functions, for example binding to a specific substrate, in different contexts. One protein that is a core component of multiple complexes is WDR82, a small (35.1kDa) WD40-repeat protein that is highly conserved across all eukaryotes from yeast to humans (Cheng et al., 2004; Lee and Skalnik, 2005). WD40-repeat proteins are typically protein-protein interaction modules that are involved in a wide range of cellular functions, including signalling, cell cycle control, and transcriptional regulation (Ma et al., 2019; Stirnimann et al., 2010).The yeast WDR82 homolog SWD2 is essential for cell viability, and WDR82 is essential for mouse embryonic development and cell proliferation (Bi et al., 2011; Cheng et al., 2004; Miller et al., 2001; Soares and Buratowski, 2012) (Sylvia Mahara,

unpublished data). Immunoprecipitation and mass spectrometry experiments have revealed that WDR82 is a component of three key complexes; the SET1A/B histone methyltransferase complexes, the PNUTS-PP1 phosphatase complex, and a more recently characterised complex with the zinc finger protein ZC3H4 (Figure 1.5A) (Austenaa et al., 2021; Brewer-Jensen et al., 2016; Lee et al., 2010; Lee and Skalnik, 2005; Miller et al., 2001; van Nuland et al., 2013). Importantly, all three complexes have been shown to regulate aspects of transcription and WDR82 has been proposed to mediate interaction of these three complexes with the transcription machinery by binding to S5P RNAPII CTD (Lee and Skalnik, 2008; Park et al., 2022).

As would be expected given its presence in multiple complexes involved in transcriptional regulation, loss of WDR82 is associated with significant changes in gene expression. Depletion of WDR82 by RNAi in macrophages resulted in several hundred genes increasing in expression, and a more pervasive increase in expression in extragenic regions that was consistent with a defect in transcription termination (Austenaa et al., 2015). Recent work in the Klose lab used calibrated transient transcriptome sequencing (cTTseq) to profile changes in transcription following acute removal of WDR82 with the dTAG inducible degron system, and found that a remarkable 40% of genes had significantly changed transcription following loss of WDR82 (Figure 1.5B). Around twice as many genes were significantly decreased in expression (5469) as were increased (2764). Heatmaps of transcription across genes (Figure 1.5C) reveal that for genes with overall decreased transcription, transcription within the gene body was reduced, however transcription downstream of the TES and in the upstream antisense direction was increased. A similar pattern is also observed for many genes whose transcription was not statistically significantly affected. Genes with overall increased transcription tended to display increased transcription both within the gene body and downstream of the TES. These results are consistent with a role for WDR82 in regulating transcription termination, but

indicate that it also contributes to regulation of transcription at promoters and throughout gene bodies. The overall effect on transcription following removal of WDR82 represents the integration of its activities in multiple complexes, which may not be equally abundant and could have opposing activities, making interpretation challenging. In order to understand how WDR82 contributes to transcriptional regulation, we must understand its integration into different complexes and how it contributes to their activities.

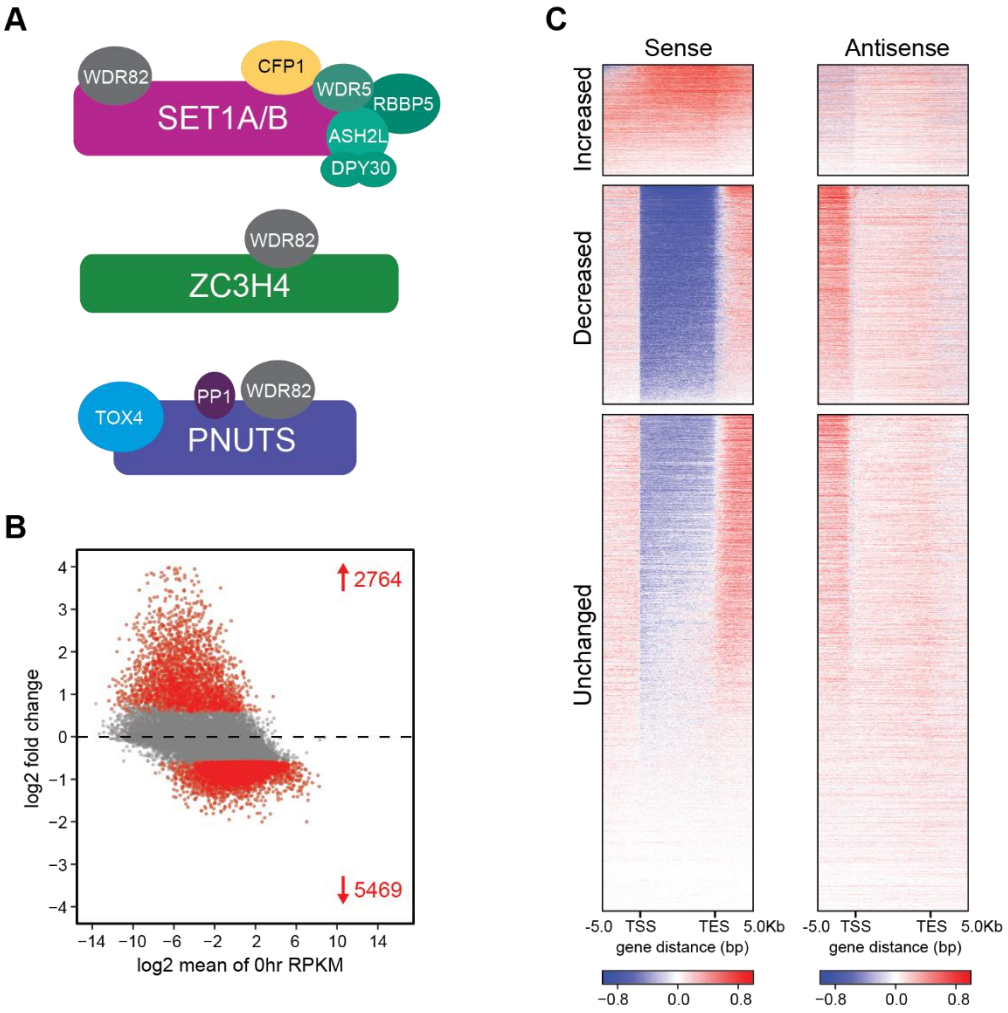


Figure 1.5 WDR82 regulates transcription. **A:** Schematic representation of the subunit composition of WDR82-containing complexes. **B:** An MA-Plot showing log_2 fold change in transcription (cTT-seq) in WDR82-dTAG cells following 2h of dTAG13 treatment to remove WDR82 (20633 genes). Significant changes in transcription (p-adj<0.05 and >1.5-fold) are coloured red and the number of significantly changed genes is indicated. **C:** Heatmaps of log_2 fold change in cTT-seq signal in antisense and sense direction across all gene bodies following WDR82 removal. Genes are categorised based on their expression changes in B. Intervals are sorted by total cTT-seq signal in untreated cells. Parts B and C courtesy of Amy Hughes.

1.2.1 The SET1 complexes

SET1A and SET1B, collectively referred to as the SET1 proteins, are essential mammalian H3K4 histone methyltransferases (HMTs) that are closely related to the yeast H3K4 HMT ySet1, and Drosophila dSet1 (Lee et al., 2007; Lee and Skalnik, 2005). The SET1 complexes, and in particular SET1A, have been proposed to deposit the majority of H3K4me3 in mammalian cells (Bledau et al., 2014; Thomson et al., 2010). The SET1 proteins exist in complex with a number of regulatory factors which are required for their catalytic activity and genomic localisation. The core SET1 complex has been generally defined as SET1A or SET1B with WDR5, RBBP5, ASH2L, DPY30, CFP1 and WDR82 (Figure 1.5A) (Lee et al., 2007; Lee and Skalnik, 2005; van Nuland et al., 2013; Wu et al., 2008). HCF1 has also been found to interact substoichiometrically with both SET1A and SET1B, whilst BOD1L1 is SET1A-specific and BOD1 is SET1B-specific (van Nuland et al., 2013; Wang et al., 2017; Wysocka et al., 2003). The WRAD complex, consisting of WDR5, RBBP5, ASH2L and DPY30, associates with the catalytic SET domain of SET1A/B to form the core catalytic module of the SET1 complexes, and is required for proper catalytic activity (Dehé et al., 2006; Dharmarajan et al., 2012; Haddad et al., 2018; Odho et al., 2010; Shinsky et al., 2015; Worden et al., 2020; Zhang et al., 2014, 2015, 2012). Structural studies of SET1 complexes on nucleosome substrates have revealed how the WRAD complex generically recognises substrate nucleosomes and positions the SET domain in a productive orientation for H3K4 methylation (Hsu et al., 2019; Worden et al., 2020). Whilst substrate nucleosome interactions by the WRAD complex contribute to chromatin binding by the SET1 complexes, the genomic localisation of the SET1 proteins is predominantly determined by CFP1, which binds to the core catalytic module (Brown et al., 2017; Dehé et al., 2006; Kim et al., 2013; Lee and Skalnik, 2005; Yang et al., 2020). In addition to binding SET1A/B, CFP1 binds non-methylated CpG islands via a ZF-CxxC domain, and H3K4me3, the modification deposited by the SET1 complexes, via a PHD domain (Clouaire et al., 2012; He et al., 2019; Lee et al.,

2001; Murton et al., 2010; Voo et al., 2000; Xu et al., 2011). H3K4me3 levels at promoters are highly correlated with transcriptional activity, so these multivalent interactions with chromatin serve to preferentially direct SET1A/B occupancy to actively transcribed CpG island promoters, which are enriched for H3K4me3 (Barski et al., 2007; Brown et al., 2017; Guenther et al., 2007; Hughes et al., 2020b; Mikkelsen et al., 2007; Sze et al., 2020, 2017). This has been proposed to create a simple feedback mechanism by which H3K4me3 is amplified at the promoters of actively transcribed genes through the binding of SET1A/B to pre-existing H3K4me3-containing chromatin (Brown et al., 2017).

If SET1A/B enrichment at actively transcribed genes relies on pre-existing H3K4me3, how is H3K4me3 initially enriched at actively transcribed CGI-associated gene promoters? Interestingly, in the absence of CFP1, SET1A is partially retained at the promoters of the most highly expressed genes, suggesting an additional transcription-related mechanism for its residence on chromatin (Brown et al., 2017). It has been proposed that WDR82, which binds SET1A/B away from the catalytic SET domain, could mediate this process (Lee and Skalnik, 2008). WDR82 has been shown to bind S5P RNAPII CTD when in complex with SET1A, and has been suggested to mediate SET1A recruitment to promoters in certain contexts (Ebmeier et al., 2017; Franks et al., 2017; Lee and Skalnik, 2008). The relevance of this potential SET1A/B recruitment mechanism for H3K4me3 genome-wide is however uncertain. There is a widespread requirement for WDR82 in the acquisition of H3K4me3 during lipopolysaccharide-induced gene induction in macrophages, however the removal of WDR82 does not result in a global loss of H3K4me3 from CGI-associated gene promoters in the steady state, suggesting that WDR82 is not required for maintenance of H3K4me3 at CGI promoters (Austenaa et al., 2015).

Whilst H3K4me3 at promoters is highly correlated with active transcription, it is unclear to what extent the modification and its deposition by SET1A/B directly regulates gene

expression (Howe et al., 2017; Hughes et al., 2020b; Morgan and Shilatifard, 2020). Loss of SET1A/B complex components is associated with widespread changes in gene expression, but these are poorly correlated with changes in H3K4me3 (Austenaa et al., 2015; Brown et al., 2017; Clouaire et al., 2014, 2012; Franks et al., 2017). Furthermore, cells lacking SET1A methyltransferase activity are viable, whilst complete removal of SET1A causes cell and embryonic lethality, suggesting its role in regulating transcription may be at least partially independent of its catalytic activity (Bledau et al., 2014; Sze et al., 2017). Recent work from the Klose lab has shown that tethering of SET1A to the promoter is sufficient to enable reporter gene transcription and this activity is independent of its catalytic activity. Instead, gene activation was reliant on SET1A binding to WDR82. Given the previously identified interaction between WDR82-SET1A complexes and S5P RNAPII CTD, we hypothesised that SET1A could regulate transcription through direct effects on RNAPII, however the mechanism for this effect remains uncertain (Hughes et al., 2022).

Hence, WDR82 has been proposed to serve two subtly different functions within the SET1A/B complexes based on its ability to interact with the RNAPII CTD. The first is as a binding module which contributes to SET1A/B occupancy at actively transcribed gene promoters to direct H3K4me3 deposition. Second, it could link SET1A/B to RNAPII to allow for direct regulation of transcription through a mechanism independent of H3K4me3. Further examination of how WDR82 integrates into SET1 complexes and its additional binding behaviours in this context will be important to enable dissection of these potential functions.

1.2.2 ZC3H4

ZC3H4 (formerly referred to as C19orf7) was identified in early mass spectrometry experiments as an interactor of WDR82 (Lee et al., 2010; van Nuland et al., 2013), however its function has only recently been investigated. The only folded domains identifiable from

the ZC3H4 sequence are a group of three C3H1-type zinc fingers, whilst the remainder of the protein is expected to be unstructured. ChIP-seq analysis shows ZC3H4 occupies genes with a binding profile that is broadly similar to RNAPII, but is enriched at the shoulders of RNAPII peaks where early transcription elongation complexes predominate (Austenaa et al., 2021; Estell et al., 2021; Hughes et al., 2022; Park et al., 2022). The close association of ZC3H4 with RNAPII could be explained by its binding to WDR82, which has been shown to bind specifically to S5P RNAPII CTD *in vitro* when in complex with ZC3H4 (Park et al., 2022). ZC3H4 has also been suggested to have RNA-binding activity via its zinc finger domains (Austenaa et al., 2021), and has been identified as an interactor of the cap-binding complex component ARS2 (Schulze et al., 2018). These observations together suggest a possible multivalent mechanism by which ZC3H4 specifically recognises and binds to early elongating RNAPII through simultaneous interactions with S5P CTD, the cap binding complex, and the nascent RNA.

Multiple recent reports have implicated ZC3H4 in regulating transcription termination away from the 3' end of genes, in particular enhancer transcription and upstream antisense transcription at promoters (Austenaa et al., 2021; Brewer-Jensen et al., 2016; Estell et al., 2021; Hughes et al., 2022; Park et al., 2022). Depletion of ZC3H4 by RNAi in HCT116 cells was associated with increased and extended enhancer and antisense promoter transcripts that was consistent with a failure to properly terminate their transcription. In support of this, tethering of ZC3H4 was sufficient to restrict transcription and cause RNA degradation by the nuclear exosome (Estell et al., 2021). Furthermore, depletion of ZC3H4 in HeLa cells resulted in increased extragenic transcription that was similar to effects seen following WDR82 removal, suggesting the activity of ZC3H4 may be related to its interaction with WDR82 (Austenaa et al., 2021). Whilst it has been predominantly implicated in the regulation of extragenic and non-coding transcription, recent work from the Klose lab has shown that ZC3H4 also affects genic transcription, particularly at low-to-moderately

transcribed genes (Hughes et al., 2022). Importantly, we discovered that SET1A/B enable gene transcription through counteracting premature transcription termination by ZC3H4, an activity we propose to be mediated by the shared subunit WDR82.

The mechanism by which the WDR82-ZC3H4 complex could mediate transcription termination is unclear, however the essentiality of ZC3H4 for embryonic development and cell proliferation suggest its activity is crucial for normal gene expression (Su et al., 2021). Despite its large size (140kDa), few proteins other than WDR82 have been found to interact with ZC3H4. No direct interactions have been identified with termination factors or exosome components, however a recent study has proposed that a complex of ZC3H4, WDR82 and casein kinase 2 supresses antisense transcription through phosphorylation of an N-terminal region of Spt5 (Park et al., 2022). The precise molecular identity of the ZC3H4 complex, and how this translates to its transcriptional regulatory activity, requires further characterisation.

1.2.3 PNUTS-PP1

PNUTS (PP1 NUclear Targeting Subunit, also referred to as PPP1R10) is a large (94kDa) scaffold protein that forms a conserved core complex with PP1, WDR82, and TOX4 (Figure 1.5A) (Allen et al., 1998; Jagiello et al., 1995; Kieft et al., 2020; Kreivi et al., 1997; Lee et al., 2010, 2009; van Nuland et al., 2013). ChIP-seq reveals PNUTS occupancy across gene bodies that largely reflects the distribution of RNAPII, with peaks around both the promoter and TES (Cortazar et al., 2019; Verheyen et al., 2015)(Amy Hughes, unpublished data). Mutation of PNUTS to abolish its interaction with PP1 results in a shift in PNUTS occupancy that reflects changes in RNAPII distribution, suggesting the primary determinant of PNUTS occupancy may be PP1-independent interaction with RNAPII (Cortazar et al., 2019; Verheyen et al., 2015). Multiple studies have reported interactions between PNUTS and RNAPII, and proteomics studies of the RNAPII CTD interactome identified the PNUTS

complex binding to both unphosphorylated and S5P CTD, suggesting PNUTS binding to RNAPII may be independent of its phosphorylation state (Carminati et al., 2022; Ciurciu et al., 2013; Ebmeier et al., 2017; Jerebtsova et al., 2011; Lee et al., 2010). The presence of WDR82 in the PNUTS-PP1 complex could provide a link between PNUTS and RNAPII through specific binding to the CTD, however the CTD-binding behaviour of WDR82 in the context of PNUTS complexes has not been examined. Furthermore, WDR82 has been shown to bind specifically to S5P CTD, whereas the PNUTS complex binds both phosphorylated and unphosphorylated CTD. This suggests either a difference in WDR82 specificity when in complex with PNUTS or a WDR82-independent mechanism of interaction with RNAPII. It has also been shown that PNUTS can bind RNA and ssDNA, and TOX4 contains an HMG box DNA-binding domain, however the relevance of these interactions for PNUTS binding or activity is unknown (Kim et al., 2003; Lee et al., 2009; Xing et al., 2014).

PNUTS is essential for development and normal gene expression (Ciurciu et al., 2013). Its best-characterised function is in promoting transcription termination. The PNUTS-PP1 complex has been shown to interact with the 3' pre-mRNA processing machinery and depletion of either WDR82 or PNUTS caused increased read-through transcription at the 3' end of genes, consistent with a role in promoting transcription termination (Austenaa et al., 2015; Benjamin et al., 2021; Cortazar et al., 2019; Shi et al., 2009; Vanoosthuyse et al., 2014). Expression of PNUTS carrying a mutation in the PP1 binding site was associated with global suppression of transcription termination, suggesting its role in regulating termination is PP1-dependent. Poly(A) site-dependent dephosphorylation of the Spt5 CTR by PNUTS-PP1 was shown to slow RNAPII elongation, and this was required for efficient termination of transcription via the 'sitting duck torpedo' mechanism (Figure 1.4) (Cortazar et al., 2019). The role of WDR82 in this mechanism is unclear, however the fact that WDR82 depletion results in similar transcription termination defects suggests that this activity may require WDR82 (Austenaa et al., 2015).

In addition to its role in regulating transcription termination at the 3' end of genes, PNUTS is also present at gene promoters and throughout gene bodies. However, its function at these sites is less well characterised. The PNUTS-PP1 complex component TOX4 was found to enhance expression of a luciferase reporter and loss of PNUTS, WDR82 or TOX4 has been associated with decreases in gene expression, suggesting the PNUTS complex possesses pro-transcriptional activity (Austenaa et al., 2015; Ciurciu et al., 2013; Cortazar et al., 2019; Lee et al., 2009; Liu et al., 2022)(Amy Hughes, unpublished data). Indeed, there is evidence that the PNUTS complex may regulate RNAPII at several stages of transcription. Similar to its role in slowing RNAPII elongation during termination, PNUTS-PP1 mediated dephosphorylation of the Spt5 CTR restricts RNAPII elongation speed throughout the genome (Cortazar et al., 2019). Knockout of TOX4 was also associated with increased RNAPII elongation rate, particularly during early elongation, supporting a role for the PNUTS complex in regulating the speed of transcription throughout genes (Liu et al., 2022). How this activity could influence gene expression is unclear, however elongation rate has been implicated in the regulation of cotranscriptional processes such as splicing. PNUTS has also been proposed to promote spliceosome activity downstream of TSSs through regulation of spliceosome phosphorylation (Cossa et al., 2020). Interestingly, interactions have been identified between components of the PNUTS complex and the PAF1 complex (PAF1C), a key factor involved in transcription elongation (Cermakova et al., 2021; Ding et al., 2015; Landsverk et al., 2020, 2019). In addition to physical interactions between the two complexes, systems analyses have revealed closely related roles in maintaining ESC identity for TOX4 and WDR82 and the PAF1C components CTR9, RTF1, and WDR61, suggesting these two complexes also have a functional relationship (Ding et al., 2015). The precise nature of the interactions between the PNUTS and PAF1 complexes, and how they could regulate transcription, is unclear.

A key functional subunit of the PNUTS complex is the Ser/Thr protein phosphatase PP1. PP1 has very broad substrate specificity and is estimated to dephosphorylate around one third of eukaryotic proteins (Bollen et al., 2010; Rebelo et al., 2015). Its activity is however tightly controlled *in vivo* through binding to a diverse family of regulatory proteins known as PIPs (PP1 Interacting Proteins), which restrict PP1 activity, localise it within the cell, and specify its substrate binding capabilities (Bollen et al., 2010; Rebelo et al., 2015; Verbinnen et al., 2017). PP1 affinity for its substrates is generally low and it often relies on additional binding by PIPs to increase overall affinity for substrates and enhance phosphatase activity. PNUTS is one of the most abundant nuclear PIPs and accounts for 2.5-5.5% of total cellular PP1 complexes (Jagiello et al., 1995; Mehta et al., 2022). *In vitro* phosphatase activity assays with recombinant PNUTS fragments have shown that it is capable of strongly inhibiting PP1 activity, however PNUTS complexes purified from cells dephosphorylate RNAPII CTD *in vitro*, suggesting this inhibitory activity is not absolute (Kim et al., 2003; Kreivi et al., 1997; Landsverk et al., 2020; Lee et al., 2010; Wu et al., 2018). Substrates identified for the PNUTS-PP1 complex include Spt5, the transcription factor Myc, and S5P RNAPII CTD (Ciurciu et al., 2013; Cortazar et al., 2019; Dingar et al., 2018; Landsverk et al., 2020; Lee et al., 2010; Washington et al., 2002; Wei et al., 2022; Wu et al., 2018). Whilst WDR82 does not seem to be required for PNUTS-PP1 to dephosphorylate RNAPII CTD *in vitro*, WDR82 depletion in HeLa cells was associated with increased RNAPII phosphorylation, suggesting it may be required for proper activity *in vivo* (Landsverk et al., 2020). Dephosphorylation of RNAPII CTD by the PNUTS-PP1 complex has been shown on multiple occasions, however the contribution of this activity to the regulation of gene expression is unclear. S5P CTD is most closely associated with promoter-proximal regions of genes (Figure 1.1) and its dephosphorylation by PNUTS-PP1 has been suggested to promote RNAPII degradation on chromatin during transcription-replication conflicts (Landsverk et al., 2020).

There are therefore two potential functions for WDR82 in the PNUTS-PP1 complex. The first, similar to the SET1 and ZC3H4 complexes, is as an adapter module to mediate PNUTS recruitment to RNAPII through binding the RNAPII CTD. This could then allow PNUTS to regulate transcription through PP1, or TOX4. The second potential function also relates to the ability of WDR82 to bind RNAPII CTD, but rather than this enabling stable binding of PNUTS to RNAPII, WDR82 may instead act as a substrate binding platform to enhance PP1 activity. Whilst there is a significant correlation between the effects of WDR82 and PNUTS depletion *in vivo*, a detailed dissection of the molecular function of WDR82 in the PNUTS complex is lacking.

1.3 Aims of thesis

WDR82 is essential for cell viability and organism development, and its removal results in widespread changes to transcription. However, understanding the role of WDR82 in transcriptional regulation is complicated by its incorporation into three distinct complexes which affect transcription in different ways. The SET1 complexes act at promoters to support transcriptional activity, ZC3H4 promotes premature transcription termination at promoters and extragenic regions, and the PNUTS-PP1 complex is required for termination at the 3' end of genes. The effects observed following WDR82 removal will therefore represent the combined consequences of its loss from all three complexes. The central aims of this thesis are to therefore understand how WDR82 integrates into each complex and its molecular function in each context. This will then enable dissection of its roles in regulating transcription.

Whilst the core SET1 and PNUTS complexes are well characterised, there exists little data regarding the composition of the ZC3H4 complex. Furthermore, previous proteomics studies to define the molecular identity of the SET1 and PNUTS complexes have relied on exogenous expression of tagged proteins in various different cell types. I therefore first

seek to biochemically define the composition of the SET1A, ZC3H4 and PNUTS complexes in ESCs using endogenously tagged cell lines (Chapter 3). I then go on to use biochemical and structural methods to characterise the molecular basis of WDR82 interaction with SET1A, ZC3H4 and PNUTS, and characterise mutations which disrupt these interactions (Chapter 4). Finally, I use *in vitro* and *in vivo* assays to address the molecular function of WDR82, in particular its ability to mediate RNAPII CTD binding in different contexts, and its contribution to overall RNAPII phosphorylation (Chapter 5). Together, this work characterises the molecular basis of WDR82 binding in multiple contexts, and begins to address how it could contribute to the regulation of transcription.

2 Materials & methods

2.1 DNA and cloning

2.1.1 cDNAs

Mouse SET1A cDNA was codon optimised for expression in *E. coli* with internal restriction sites removed and synthesised by Invitrogen. cDNA coding for mouse SET1B$^{1\text{-}209}$ was synthesised by Invitrogen. Synthesised sequences are given in Table A.1. Mouse cDNA clones for WDR82 (RIKEN clone ID 9430088P09) and PNUTS (IMAGE clone ID 5702171) were purchased from Source Bioscience. Partial mouse ZC3H4 cDNA (MGC clone ID 6822470) was purchased from Horizon Discovery.

2.1.2 Mammalian expression constructs

To express FLAG- or FLAG-GST tagged proteins in mammalian cells, coding sequences were assembled by either Gibson assembly, restriction digestion and ligation, or ligation independent cloning (LIC) (Aslanidis and de Jong, 1990) into a modified pcDNA3 plasmid which contained an N-terminal FLAG or FLAG-GST tag and LIC sites. Mutations were generated using Quikchange II XL kit (Agilent), according to manufacturer's protocol. A list of mutagenesis primers is given in Table 2.1.

Table 2.1 List of mutagenesis primers

Mutation	Forward Primer	Reverse Primer
SET1A D62A P63A R64A	tttgctacgaacatgacaagccgcagcctgcagatcttcaaccgg	ccggttgaagatctgcaggctgcggcttgtcatgttcgtagcaaa
SET1A L76D	gtttgaatttcggaaccggatcgctaaaatcacgtgctttgctacg	cgtagcaaagcacgtgattttagcgatccggttccgaaattcaaac
SET1A P79A	cagtttgaatttcgcaaccggcaggctaaaatcac	gtgattttagcctgccggttgcgaaattcaaactg
SET1A F81A	caatgtaaaattcatccagtttggctttcggaaccggcaggctaaaa	ttttagcctgccggttccgaaagccaaactggatgaattttacattg

SET1A L76D P79A F81A	gaccaatgtaaaattcatccagtttggctttcgcaaccggatcgctaaaatcacgtgctttgctacgaa	ttcgtagcaaagcacgtgattttagcgatccggttgcgaaagccaaactggatgaattttacattggtc
ZC3H4 L975D	gagcaaagctcggcttgctatcggtcacatccttcttgatg	catcaagaaggatgtgaccgatagcaagccgagctttgctc
ZC3H4 P978A	cgagcaaagctcgccttgctcagggtc	gaccctgagcaaggcgagctttgctcg
ZC3H4 F980A	agcactgtgcgagcagcgctcggcttgctcag	ctgagcaagccgagcgctgctcgcacagtgct
ZC3H4 L975D P978A F980A	cagagcactgtgcgagcagcgctcgccttgctatcggtcacatccttcttgat	atcaagaaggatgtgaccgatagcaaggcgagcgctgctcgcacagtgctctg
ZC3H4 D958A P959A R960A	ctgttgcagctgagaggcagcggcccgtagtgggtgtcc	ggacacccactacgggccgctgcctctcagctgcaacag
ZC3H4 I946D	cagagggtccaggggatcgttcaccgccttttca	tgaaaaggcggtgaacgatcccctggaccctctg
ZC3H4 I946D P947A L948A D949A	gggtgtccaggcagaggggccgcggcatcgttcaccgccttttcacg	cgtgaaaaggcggtgaacgatgccgcggccctctgcctggacaccc
PNUTS W461A	cctgggacacaccgcaggcactttctcctcc	ggaggagaaagtgcctgcggtgtgtcccagg
PNUTS P464A	gaggcctggcacacacccaaggcactt	aagtgccttgggtgtgtgccaggcctc
PNUTS L467D	gaggtgagggcagaacatcaggcctgggacacacc	ggtgtgtcccaggcctgatgttctgccctcacctc
PNUTS W461A P464A L467D	ggtgagggcagaacatcaggcctggcacacaccgcaggcactttctcct	aggagaaagtgcctgcggtgtgtgccaggcctgatgttctgccctcacc
PNUTS I520D P521A L522A D523A	ctcatccatggcacattcctcagccgcggcgtcgagttttgggggaatgggttca	tgaacccattcccccaaaactcgacgccgcggctgaggaatgtgccatggatgag
WDR82 K63A Y64A	tgtatctgatgaggtccacaccggccgccttactgtacagggttctctttg	caaagagaaccctgtacagtaaggcggccggtgtggacctcatcagataca
WDR82 R137A	tggcagttaggagacgcgagatcccagagtcg	cgactctgggatctcgcgtctcctaactgcca
WDR82 D205A K207A	ttggtggagatgagtatcaatgcgccagcattgctgaacttaagtcctg	caggacttaagttcagcaatgctggcgcattgatactcatctccaccaa
WDR82 F184A	tactgcatcttaaatgttgcagctggtcccttatcaaaagaacgaagg	ccttcgttcttttgataagggaccagctgcaacatttaagatgcagta
WDR82 F224A	gcatcaccacacccttggctgcgtcaatcagtcgga	tccgactgattgacgcagccaagggtgtggtgatgc
WDR82 K181A	tgttgcaaatggtcccgcatcaaaagaacgaaggtcataaagtttgacc	ggtcaaactttatgaccttcgttcttttgatgcgggaccatttgcaaca
WDR82 D41A	gtcatctctagcagcgctgatgactccatcgtg	cacgatggagtcatcagcgctgctagagatgac

2.1.3 Recombinant expression constructs

pNIC plasmids for expression of recombinant proteins in *E. coli* were generated by LIC. MultiBac donor and acceptor plasmids were cloned by either restriction digestion and ligation cloning or by LIC into pAceBac1 vectors which had been modified to carry N-

terminal FLAG or Twin-Strep tags and LIC sites. *In vitro* Cre-LoxP recombination of MultiBac vectors was performed as described (Berger and Craig, 2010).

2.1.4 Bacmid DNA purification

Assembled shuttle vectors were transformed into DH10Bac *E. coli* and recombination was screened for by blue-white selection on LB agar plates containing 50µg/ml kanamycin, 7µg/ml gentamycin, 10µg/ml tetracycline, 80µg/ml X-Gal and 40µg/ml IPTG. Single white colonies were picked into 2ml overnight culture containing 50µg/ml kanamycin, 7µg/ml gentamycin and 10µg/ml tetracycline. 1.5ml of overnight culture was pelleted at 12,000x*g* for 3 minutes to pellet cells. Cell pellet was resuspended in 300µl Solution A (15mM Tris pH 8.0, 10mM EDTA, 100µg/ml RNase A), then 300µl solution B (0.2M NaOH, 1% SDS) was added and lysis was allowed to proceed for 5 minutes at room temperature. 300µl 3M sodium acetate pH 5.5 was added to neutralise the reaction and samples were incubated on ice for 10 minutes then centrifuged at 14,000x*g* for 10 minutes at 4°C. Supernatant was mixed with 800µl isopropanol and incubated on ice for 10 minutes. Precipitated DNA was pelleted by centrifugation at 14,000x*g* for 20 minutes, washed with 500ul 70% ethanol, then air-dried. The DNA pellet was resuspended in 40µl buffer EB (Qiagen) and incubated briefly at 37°C to dissolve. Presence of desired coding sequences was verified by PCR.

2.2 Cell culture methods

2.2.1 Cell culture conditions

Mouse embryonic stem cells were grown in Dulbecco's Modified Eagle Medium (Gibco) supplemented with 10% foetal bovine serum (FBS, Sigma), 1x non-essential amino acids (Gibco), 2mM L-glutamine (Gibco), 1x penicillin/streptomycin (Gibco), 0.5mM β-mercaptoethanol (Gibco), and 10ng/ml leukaemia inhibitory factor. ESCs were grown on

gelatinised plates at 37°C and 5% CO_2. Cells were passaged using TrypLE Express Enzyme (1X, Gibco). Human HEK293T cells were grown at 37°C and 5% CO_2 in Dulbecco's Modified Eagle Medium, supplemented with 10% FBS, 1x penicillin/streptomycin, 2mM L-glutamine, and 0.5mM β-mercaptoethanol. Cells were harvested either by trypsinisation or, for large-scale experiments, by scraping into ice-cold PBS containing 1x cOmplete protease inhibitor cocktail (Roche).

Sf9 cells were grown at 27°C in suspension in Sf900 II medium (Gibco) supplemented with 0.5x penicillin/streptomycin with shaking at 100rpm.

2.2.2 dTAG treatments

Cell lines expressing dTAG fusion proteins were treated with 100nM dTAG-13 (Tocris) to induce protein depletion. Media was pre-warmed for at least 1 hr before treatment.

2.2.3 Embryonic stem cell lines

All ESC lines used in this study are listed in Table 2.2

Table 2.2 List of embryonic stem cell lines.

Name	Genotype	Origin
dTAG SET1AB	*FKBP12^{F36V}-Setd1A; FKBP12^{F36V}-Setd1B*	Hughes *et al.* 2022
ST7-dTAG-SET1A	*Triple T7-TwinStrep-FKBP12^{F36V}-Setd1A*	Hughes *et al.* 2022
ZC3H4-ST7	*Zc3h4-TwinStrep-Triple T7*	Hughes *et al.* 2022
PNUTS-ST7	*Pnuts-TwinStrep-Triple T7*	Amy Hughes
WDR82-ST7	*Wdr82-TwinStrep-Triple T7*	Amy Hughes
ZC3H4-dTAG	*Zc3h4-FKBP12^{F36V}*	Hughes *et al.* 2022
PNUTS-dTAG	*Pnuts-FKBP12^{F36V}*	Amy Hughes
WDR82-dTAG	*Wdr82-FKBP12^{F36V}*	Amy Hughes

2.2.4 Transient transfections

293T cells were transfected in 10cm dishes at approximately 70% confluency using Lipofectamine 2000 (Invitrogen) and 5µg plasmid, according to manufacturer's protocol. The following day, cells were passaged onto 15cm dishes and allowed to grow for a further 24hrs before harvesting by trypsinisation. Cell pellets were snap frozen and stored at -80°C until required.

2.2.5 Transfection of Sf9 cells and baculovirus amplification

Sf9 cells were transfected with bacmid DNA using Cellfectin II according to manufacturer's protocol. P1 viruses were harvested 96 hours post-transfection. To amplify viruses, 20ml suspension cultures of Sf9 cells were seeded at $1.5x10^6$ cells/ml, inoculated with 750µl virus stock, and allowed to grow for 96hrs. Cultures were centrifuged at 2000rpm for 5 minutes and the supernatant taken as the amplified virus stock. This amplification was repeated to yield final P3 stocks used for protein expression.

2.2.6 Recombinant protein expression

To express recombinant proteins for purification, 500ml cultures of Sf9 cells were seeded at $1x10^6$ cells/ml and allowed to grow at 27°C with shaking at 120rpm until density reached $1.5\text{-}2x10^6$ cells/ml. Cultures were inoculated with P3 virus stocks and allowed to grow for 72hrs before harvesting by centrifugation at 2000rpm for 10 minutes. Cell pellets were snap frozen and stored at -80°C until required.

2.3 Protein Methods

2.3.1 Preparation of nuclear extract

Unless otherwise stated, all centrifugation steps involving protein samples were performed at 4°C. Cell pellets were resuspended in 10 cell volumes of buffer A (10mM

HEPES pH 7.9, 1.5mM $MgCl_2$, 10mM KCl, 0.5mM DTT, 0.5mM PMSF, 1x cOmplete protease inhibitor cocktail (Roche)) and incubated for 10 min on ice. After centrifugation at 1500x*g* for 5 min, the cell pellet was resuspended in 3 cell volumes of buffer A supplemented with 0.1% NP-40 and incubated on ice for 10 min. Nuclei were pelleted at 1500x*g* for 5 min then resuspended in 1 cell volume of buffer C (250mM NaCl, 5mM HEPES pH 7.9, 26% glycerol, 1.5mM MgCl2, 0.2mM EDTA, 0.5mM DTT, 1x protease inhibitor cocktail). The volume of the nuclear suspension was measured and the NaCl concentration increased to 400mM by dropwise addition of 5M NaCl, assuming a 250mM starting concentration. Nuclei were incubated at 4°C for 1hr with gentle inversion to extract nuclear proteins. After centrifugation at 18,000x*g* for 20 min, nuclear proteins were recovered in the supernatant. Protein concentration was determined by Bradford assay (BioRad).

Where blotting for RNAPII phospho-modifications, 10mM NaF was added to all buffers.

2.3.2 Immunoprecipitation

Immunoprecipitation using anti-ZC3H4 antibody was performed using 1mg nuclear extract. Extracts were diluted with 950µl BC150 (150mM KCl, 10% glycerol, 50mM HEPES pH 7.9, 0.5mM EDTA) supplemented with 1x protease inhibitors. 250U Benzonase nuclease (Sigma) was added and extract were incubated for 30 minutes at 4°C with gentle agitation then centrifuged for 5 minutes at 21,000x*g* and the supernatant taken as input. Antibodies were added to cleared extracts and incubated overnight at 4°C with gentle agitation. 20µl ZC3H4 antibody was used for specific pulldown, and 6µl anti-FLAG antibody was used for the negative control samples. Protein A beads (Repligen) were blocked in BC150 supplemented with 1% Fish gelatin and 0.2mg/ml BSA overnight at 4°C. Blocked beads were rinsed 3 times in BC150 and 80µl slurry was used to precipitate antibody-bound protein at 4°C for 2hrs with gentle agitation. Beads were pelleted for 3 minutes at 1000x*g*, and washed 5 times for 10 minutes with BC150 + 0.02% NP-40. For subsequent western

blotting, beads were resuspended in 2x SDS loading buffer, boiled at 95°C for 5 min, and the supernatant taken as the immunoprecipitate.

2.3.3 FLAG Immunoprecipitation

Immunoprecipitations were performed using 600µg nuclear extract. Extracts were diluted in nuclear extraction buffer C without NaCl to give a final NaCl concentration of 150mM, 125U Benzonase nuclease was added, and extract were incubated for 30 minutes at 4°C with gentle agitation. Cleared extracts were centrifuged for 5 minutes at 21,000x*g* and the supernatant taken as input. 25ul anti-FLAG M2 resin (Sigma) per IP was washed 3 times in BC150 and added to precleared extract. Extracts were incubated with beads for 4hrs at 4°C with gentle agitation. Beads were pelleted at 1000x*g* and washed 3 times in 1ml BC150 + 0.02% NP-40 with 10 minute incubation. For subsequent western blotting, beads were resuspended in 2x SDS loading buffer, boiled at 95°C for 5 min, and the supernatant taken as the immunoprecipitate.

2.3.4 SDS-PAGE

Where necessary, protein extracts were mixed with SDS loading buffer to 1x and boiled at 95°C for 5 min. Gels (0.1% SDS, 0.1% ammonium persulphate (Sigma), 0.1% TEMED (Sigma), 400mM Tris-HCl pH 8.8) were cast using the Mini-Protean Tetra Cell system (BioRad). Depending on the size of the protein of interest, resolving gels containing 8-12% acrylamide/bis-acrylamide (BioRad) were used. Stacking gel contained 5% acrylamide/bis-acrylamide and 125 mM Tris-HCl pH 6.8. For silver staining, 4-20% Mini-Protean TGX precast gels (BioRad) were used. Gels were run at 200V in 1x SDS-PAGE running buffer (25mM Tris, 192mM glycine, 0.1% SDS). For small proteins (<25kDa), gels were run at 125V in 1x Tris-Tricine running buffer (25mM Tris, 25mM Tricine, 0.05% SDS). Alternatively, when analysing proteins running >180kDa, 3-8% pre-cast NuPAGE Tris-Acetate gels (Invitrogen)

were used. Proteins were visualised in-gel with coomassie blue or silver staining, or transferred to nitrocellulose membranes for western blotting.

2.3.5 Coomassie blue staining

SDS-PAGE gels were briefly washed in water then incubated for at least 20 minutes in coomassie blue stain (50% Methanol, 10% Acetic Acid, 0.1% Brilliant Blue R250 Dye). Excess stain was removed with water and the gel was immersed in de-stain solution consisting of 30% ethanol and 10% acetic acid. The de-stain was allowed to proceed for several hours until the background staining was no longer visible. Gels were then soaked overnight in water and imaged the following morning.

2.3.6 Silver staining

Silver staining was performed using SilverQuest Silver Staining Kit (Invitrogen), according to manufacturer's protocol.

2.3.7 Western blotting

Protein extracts were resolved using SDS-PAGE and transferred to nitrocellulose membrane by semi-dry transfer using the Trans-Blot Turbo Transfer System (BioRad). Transfer was performed as per the manufacturer's guidelines, depending on the size of the proteins being transferred. All blocking and antibody incubation steps were performed in 1x PBS with 0.1% Tween 20 (Fisher) and 5% milk. Membranes were blocked for 1hr at RT and primary antibody incubations were carried out overnight at 4°C (See Table 2.3 for antibody dilutions). Membranes were washed for 10 min at RT, and then incubated with infrared-dye-conjugated secondary antibody (1: 15,000) for at least 1 hr. Membranes were washed twice for 10 minutes in PBST and once for 10 minutes in PBS prior to imaging using an Odyssey Fc system (LI-COR). Quantification was performed using ImageStudio Lite.

For FLAG IP samples, HRP (horseradish peroxidase)-conjugated Anti-FLAG primary antibody (Sigma) was used to avoid cross-reactivity with denatured IgG. Following primary antibody incubation, membranes were washed twice with PBST and once with PBS for 10 minutes each and developed by chemiluminescence (Solution 1: 100mM Tris HCl, pH 8.5, 2.5mM Luminol in DMSO, 396µM p-Coumaric acid in DMSO; Solution 2: 100mM Tris HCl, pH 8.5, 5.6mM H_20_2). Solutions were mixed in a 1:1 ratio, applied to membranes for 1 min and membranes were then exposed to X-ray film (GE Healthcare).

2.3.8 Antibodies

2.3.8.1 Production of anti-SET1A antibody

Antigenic protein expression and purification

pNIC expression plasmid encoding His_6-tagged antigenic SET1A fragment (residues 1101-1378, designed by Amy Hughes) was transformed into BL21 DE3 (pLysS) *E. coli* cells and plated with kanamycin and chloramphenicol selection. 650ml expression cultures of 2xTY media supplemented with 50µg/ml kanamycin and 34µg/ml chloramphenicol were set up and allowed to grow at 37°C until OD600 was ~0.7. Expression was induced with 1mM IPTG for ~16hrs at 18°C. Cells were harvested by centrifugation and stored at -80°C.

Cells were thawed and resuspended in lysis buffer (20mM Tris-HCl pH 8.0, 500mM NaCl, 0.1% NP-40 and cOmplete EDTA-free protease inhibitor cocktail (Roche)) and sonicated on ice at 60% amplitude for 6 cycles of 30 seconds on, 30 seconds off. Lysates were centrifuged at 12,500rpm for 20 minutes to pellet insoluble material. Lysate was supplemented with imidazole to a final concentration of 10mM and incubated with 500µl Ni^{2+}-charged IMAC Sepharose resin (GE healthcare) for 2 hrs at 4°C with gentle agitation. Lysate and beads were then poured into an empty gravity flow column (BioRad) and

allowed to flow through. Column was washed with 20-40 volumes of wash buffer 1 (50mM NaH_2PO_4 pH 8.0, 300mM NaCl, 20mM Imidazole) then 10 column volumes wash buffer 2 (50mM NaH_2PO_4 pH 8.0, 300mM NaCl, 30mM Imidazole). Purified proteins were eluted with elution buffer (50mM NaH_2PO_4 pH 8.0, 300mM NaCl, 250mM Imidazole) in 500µl fractions until no protein could be observed in eluate with Bradford reagent. Samples from elution fractions were subject to SDS-PAGE and fractions containing the desired protein were pooled. TEV protease was added at 1:100 w/w and protein was dialysed overnight into TEV cleavage buffer (50mM Tris-HCl pH 8.0, 150mM NaCl, 5mM β-mercaptoethanol). Dialysed protein was incubated with 250µl Ni^{2+}-Sepharose resin for 2hrs at 4°C with gentle agitation to capture contaminants and TEV protease. Purified protein was concentrated in 10kDa MWCO spin concentrator (Sartorius) and glycerol was added to 10% v/v. Protein concentration was quantified by Bradford assay.

Immunisation

Antibody production was performed by Eurogentec. Rabbits were immunised on days 0, 7, 10, and 18 with 200µg antigen, and final bleed was taken on day 28.

Antibody purification

SET1A antigen (6mg) was conjugated to 500µl Affi-Gel 15 (Bio-Rad) overnight at 4°C with gentle agitation. Resin was washed with 1ml 0.2M ethanolamine then blocked in a further 1ml of 0.2M ethanolamine for 2hrs at 4°C. Resin was washed with 25ml 1M NaCl then 25ml PBS. Washed resin was incubated with 10ml final bleed rabbit serum for 3hrs at RT with gentle agitation. Resin and serum were run through an empty gravity flow column, and resin was washed with 25ml 0.5M NaCl followed by 25ml PBS. Antibody was eluted in 500µl fractions of 0.1M Glycine pH 3.0 directly into tubes containing 50µl 1M Tris pH 8.0 and mixed immediately to neutralise. Eluted antibody fractions were analysed by SDS-PAGE

and desired fractions were pooled. NaCl was added to a final concentration of 50mM and glycerol was added to 10% v/v. Antibody specificity was verified by western blot of nuclear extract from untreated and dTAG treated dTAG-SET1A cells.

2.3.8.2 List of antibodies

Antibodies used are listed in Table 2.3

Table 2.3 List of antibodies

Antigen	Source	Dilution for western blotting
WDR82	Cell Signalling (D2I3B)	1:1000
SET1A	This thesis (see section 2.3.8.1)	1:500
PNUTS	Cell Signalling (14171)	1:1000
ZC3H4	Atlas Antibodies (HPA040934)	1:500
HDAC1	Abcam (ab109411)	1:1000
T7	Cell Signalling (D9E1X)	1:1000
CFP1	Klose lab	1:2000
ASH2L	Cell Signalling (D93F6)	1:1000
PAF1	Bethyl (A300-172A)	1:1000
LEO1	Bethyl (A300-175A)	1:1000
PP1α	Invitrogen (43-8100)	1:1000
RBBP5	Cell Signalling (D3I6P)	1:1000
FLAG	Sigma (F1804)	1:1000
FLAG (HRP)	Sigma (A8592)	1:500
RNAPII NTD	Cell Signalling (D8L4Y)	1:1000
RNAPII S2P	Cell Signalling (E1Z3G)	1:1000
RNAPII S5P	Cell Signalling (D9N5I)	1:1000
Anti-mouse secondary (IRDye800)	LiCOR 32210	1:15000
Anti-mouse secondary (IRDye680)	LiCOR 68070	1:15000
Anti-rabbit secondary (IRDye800)	LiCOR 32211	1:15000

Anti-rabbit secondary (IRDye680)	LiCOR 68071	1:15000

2.3.9 Purification of Twin-Strep tagged proteins from nuclear extracts

For small-scale purifications followed by western blot, 1mg nuclear extract was used. For large-scale purifications for mass spectrometry, 6mg nuclear extract was used. Extracts were diluted in nuclear extraction Buffer C without NaCl to give a final NaCl concentration of 150mM, 250U Benzonase nuclease and 20µg avidin per mg nuclear extract was added, and extracts were incubated for 30 minutes at 4°C with gentle agitation. Cleared extracts were centrifuged for 5 minutes at 21,000x*g* and the supernatant taken as input. 10µl Strep-Tactin XT 4Flow high capacity (iba) was used per mg extract. Beads were washed 3 times in nuclear extract buffer C containing 150mM NaCl and added to precleared extracts. Extracts were incubated with beads for 2.5hrs at 4°C with gentle agitation. Beads were pelleted at 1000x*g* and washed 5 times in 1ml wash buffer (100mM Tris pH 8.0, 150mM NaCl, 1mM EDTA, 1mM DTT) with 5 minute incubation. For subsequent western blotting, beads were resuspended in 2x SDS-PAGE loading buffer, boiled at 95°C for 5 min, and the supernatant taken as the immunoprecipitate. For mass spectrometry, proteins were eluted by incubation for 15 minutes with 85µl 1x buffer BXT (iba) with occasional agitation, beads were then pelleted by centrifugation for 3 minutes at 1000x*g* and the supernatant collected as the elution fraction. This process was repeated four times in total.

2.3.10 Preparation of samples for mass spectrometry

Elution fractions 1 and 2 from Strep purifications were pooled and 100µl of the pool was used as input for digestion following a modified SP3 protocol (Hughes et al., 2019). Speedbeads (Cytiva) were prepared by washing twice in H_2O and resuspended in 5x original volume H_2O. Input protein was denatured by addition of 8M urea in 50mM ammonium bicarbonate to final concentration of 6M urea. TCEP (Thermo Scientific BondBreaker) was

added to a final concentration of 10mM and samples were incubated at room temperature for 10 minutes with shaking at 1000rpm. Freshly prepared 0.5M 2-chloro-acetamide was added to a final concentration of 50mM and samples were incubated in the dark at RT for 30 minutes with shaking at 1000rpm. 10µl prepared bead stock was added and samples were vortexed briefly to mix. 100% acetonitrile was added to a final concentration of 77% and samples were incubated for 30 minutes on a flywheel at room temperature. Beads were concentrated using a magnetic rack and supernatant removed. Beads were washed twice with 900µl 70% ethanol and once with 900µl 100% acetonitrile. This three-step washing cycle was repeated a further four times. Washed beads were resuspended in 34µl 50mM ammonium bicarbonate and 1µl Trypsin Gold (Promega, resuspended at 100ng/µl in 50mM acetic acid) and 5µl Lys-C (Fujifilm Wako, resuspended at 20ng/µl in 2mM Tris pH 8.0) was added. Samples were allowed to digest overnight at 37°C with shaking at 1000rpm. Beads were concentrated using a magnetic rack and supernatant collected. Beads were resuspended in 10µl 10% formic acid and incubated at 37°C for 15 minutes with shaking at 1000rpm. Beads were concentrated using a magnetic rack and supernatant was pooled with overnight digest fraction. This formic acid wash step was repeated a further 3 times. Digested samples were then desalted using C18 tips prior to mass spectrometry analysis.

2.3.11 Mass Spectrometry

Dried peptides were resuspended into 5% acetonitrile / 5% formic acid for LC-MS/MS analysis. Peptides were separated by nano liquid chromatography (Thermo Scientific Easy-nLC 1000 or Ultimate RSLC 3000) coupled in line with a QExactive mass spectrometer equipped with an Easy-Spray source (ThermoFisher Scientific). Peptides were trapped onto a C18 PepMac100 precolumn (300µm i.d.x5mm, 100Å, ThermoFisher Scientific) using Solvent A (0.1% Formic acid, HPLC grade water). The peptides were further separated onto an Easy-Spray RSLC C18 column (75um i.d., 50cm length, ThermoFisher Scientific) using a

30-minute linear gradient (15% to 35% solvent B (0.1% formic acid in acetonitrile)) at a flow rate of 200nl/min. The raw data were acquired on the mass spectrometer in a data-dependent acquisition mode (DDA). Full-scan MS spectra were acquired in the Orbitrap (Scan range 350-1500m/z, resolution 70,000; AGC target, 3e6, maximum injection time, 50ms). The 10 most intense peaks were selected for higher-energy collision dissociation (HCD) fragmentation at 30% of normalized collision energy. HCD spectra were acquired in the Orbitrap at resolution 17,500, AGC target 5e4, maximum injection time of 120ms with fixed mass at 180m/z. Charge exclusion was selected for unassigned and 1+ ions. The dynamic exclusion was set to 5s.

Data Processing

Tandem mass (MS/MS) spectra were searched using Sequest HT in Proteome discoverer software version 1.4 as follows: MS/MS data from whole cell lysate samples were searched against a protein sequence database containing 17,745 protein entries, including 17,462 *Mus musculus* proteins (Uniprot release from 2021-12-16) in which protein sequences for PNUTS, SET1A, WDR82 and ZC3H4 were replaced by their respective tagged protein sequences, and 283 common contaminants. During database searching cysteines (C) were considered to be fully carbamidomethylated (+57,0215, statically added), methionine (M) to be fully oxidised (+15,9949, dynamically added), all N-terminal residues to be acetylated (+42,0106, dynamically added). Two missed cleavages were permitted. Peptide mass tolerance was set at 50ppm on the precursor and 0.6 Da on the fragment ions. Data was filtered at FDR below 1% at PSM level. Label free quantification was performed by Normalised Spectral Abundance (NSAF) as previously described (Florens et al., 2006), as the number of spectral counts (PSM) that identify a protein, divided by the protein length (L), the PSM/L value represents the SAF, which is then divided by the sum of PSM/L for all proteins in the experiment. NSAF values were calculated after common

contaminants being removed. For better visualization of the data, NSAF values were multiplied by 100 (NSAF·100). The statistical analysis was performed on NSAF values from four replicates experiments using t-test in Perseus (Tyanova et al., 2016) and scatter plots were generated in GraphPad Prism 9.2.0.

2.3.12 Size exclusion chromatography of nuclear extracts

Nuclear extract was Benzonase treated (250U Benzonase per mg nuclear extract) and dialysed overnight into BC200 buffer (50mM HEPES pH 7.9, 200mM KCl, 10% Glycerol, 1mM DTT). 2mg dialysed nuclear extract was loaded on a Superose 6 Increase 10/300 GL column (Cytiva, precalibrated with dextran blue, Mix 1 (Ferritin, 440kDa; Conalbumin, 75kDa) and Mix 2 (Thyroglobulin, 669kDa; Aldolase, 158kDa; Ovalbumin, 43 kDa)) and run in BC200 buffer at 0.2ml/min. Eluate was collected in 250µl fractions. Protein fractions were trichloroacetic acid precipitated and 15% of the fraction was loaded onto SDS-PAGE gel for analysis by western blot.

2.3.13 Purification of recombinant Twin-Strep tagged proteins for crystallisation

Sf9 cells were resuspended in approximately 10 volumes lysis buffer S (50mM HEPES pH 8.0, 200mM NaCl, 2mM $MgCl_2$, 15% Glycerol) supplemented with 1x cOmplete protease inhibitors, 0.5mM PMSF, 1mM DTT, 1250U Supernuclease (Stratech) and 100ug avidin (iba), and sonicated at 80% amplitude for 1 minute in 5 second on/off pulses. Lysates were centrifuged at 20,000rpm for 30 minutes at 4°C, then passed through a column containing a suitable volume of Strep-Tactin XT 4Flow high capacity resin. Flowthrough was passed back over the column a further two times. Column was washed three times with at least 10 volumes wash buffer S (100mM Tris pH 8.0, 150mM NaCl) and eluted with 1x buffer BXT (iba) until no further protein was detectable by Bradford reagent. Eluted fractions were

analysed by SDS-PAGE and desired fractions were pooled. If necessary, pooled protein was concentrated to suitable volume in a 10kDa MWCO spin concentrator (Sartorius) and protein concentration was determined by Bradford assay. TEV protease was added 1:50 w/w and protein was dialysed overnight into dialysis buffer (50mM HEPES pH 7.0, 150mM NaCl, 10% Glycerol, 0.5mM DTT).

Dialysed protein was centrifuged for 10 minutes at 21,000x*g* to pellet precipitates, then concentrated to <500µl in a 10kDa MWCO spin concentrator (Millipore). Concentrated protein was centrifuged for 5 minutes then loaded onto a Superdex 75 10/300 column (Cytiva) at 0.3ml/min in SEC Buffer (50mM HEPES pH 7.0, 150mM NaCl, 0.5mM DTT) and collected in 150µl fractions. Fractions were analysed by SDS-PAGE and desired fractions were pooled. Glycerol was added to 5% and protein was concentrated in a 10kDa MWCO spin concentrator (Millipore) immediately prior to use for crystallisation.

2.3.14 Crystallisation

Crystallisation screens were set up as sitting drops using a Mosquito robotic system and commercial screens as outlined in Table A.2 and Figure A.1. All protein-only drops were 200nl volume with protein: precipitant ratio as given. Additive screen drops were 250nl with 50nl seed stock and protein: precipitant ratio as given. Microseeded drops were 250nl, with 50nl seed stock and protein: precipitant ratio as given. Seed stocks were prepared using PFTE Seed Beads (Hampton Research) according to manufacturer protocol and diluted with 50-100µl water prior to use.

2.3.15 Purification of recombinant FLAG-tagged proteins

Sf9 cells were resuspended in 4 volumes Lysis Buffer F (10mM Tris pH 8.0, 500mM NaCl, 4mM $MgCl_2$, 20% Glycerol, 0.4mM PMSF, 1x cOmplete protease inhibitors) and lysed by sonication at 80% amplitude for 1 minute in 5 second on/off pulses. Lysates were diluted

with 2 volumes dilution buffer (10mM Tris pH 8.0, 10% glycerol, 0.02% NP-40) and incubated in ice for 30 minutes before centrifugation at 20,000 rpm for 30minutes at 4°C. Anti-FLAG M2 resin (Sigma) was added to cleared lysate and incubated for 4hrs at 4°C with gentle agitation. Beads were pelleted by centrifugation at 1000x*g* for 3minutes, and washed three times for 10 minutes with 10ml Wash buffer FS (50mM HEPES pH 7.0, 150mM NaCl, 2mM $MgCl_2$, 15% glycerol). Protein was eluted in 1-2CV fractions using 0.2mg/ml FLAG peptide in wash buffer FS with 20 minute incubation until no protein was detectable with Bradford reagent. Elution fractions were analysed by SDS-PAGE and desired fractions were pooled. Protein was concentrated as required in a 10kDa MWCO spin concentrator (Millipore).

2.3.16 CTD binding assays

Biotinylated 4-repeat CTD peptides (GL Biochem) were conjugated to Dynabeads MyOne Streptavidin T1 in conjugation buffer (25mM Tris pH 8.0, 50mM NaCl, 5% Glycerol, 0.03% Triton X-100, 1mM DTT) for 30 minutes at 4°C with gentle agitation. Beads were captured on a magnetic rack and washed 3 times for 10 minutes in conjugation buffer and once with binding buffer (25mM Tris pH 8.0, 150mM NaCl, 5% Glycerol, 0.03% Triton X-100, 1mM DTT) then resuspended in 2 volumes binding buffer. 1µg protein samples were diluted to 45µl in wash buffer FS (as above) and 60µl bead suspension was added. Samples were incubated for 2hrs at 4°C with gentle agitation then washed 3 times in 1ml wash buffer (25mM Tris pH 8.0, 150mM NaCl, 5% Glycerol, 0.1% Triton X-100, 1mM DTT). Beads were resuspended in 20µl wash buffer and protein was eluted by addition of 4x SDS-PAGE loading dye to 1x and incubation at 95°C for 5 minutes. Beads were captured using a magnetic rack and supernatant taken as pulldown sample. 5µl of input protein sample and 50% of the pulldown sample was analysed by western blot.

2.4 Computational Methods

2.4.1 Sequence alignments

Protein sequence alignments were generated using T-Coffee (Notredame et al., 2000) and visualised with Jalview.

2.4.2 Protein structure prediction using ColabFold

Structure prediction using Colabfold ("ColabFold,"; Mirdita et al., 2022) was performed using default settings with AMBER relaxation and without use of templates, except for WDR82-PNUTS-PP1 prediction, which used templates detected from pdb70.

2.4.3 Protein Structure Prediction using AlphaFold

Structure prediction using the AlphaFold Colab ("AlphaFold,"; Jumper et al., 2021) was performed using default settings and without relaxation.

2.4.4 Structure visualisation & analysis

Structures were visualised using Pymol. Structure alignment was performed in Pymol using the cealign function. Analysis of interfaces was performed using the PISA server (Krissinel and Henrick, 2007).

3 Characterising the composition of WDR82-containing complexes in ESCs

3.1 Introduction

WDR82 is a highly conserved eukaryotic protein that is essential in all organisms from yeast to humans and its loss is associated with significant changes to transcription (Austenaa et al., 2015; Bi et al., 2011; Cheng et al., 2004; Lee and Skalnik, 2005; Miller et al., 2001; Soares and Buratowski, 2012). Previous work has identified WDR82 as a core component of a number of protein complexes involved in transcriptional regulation: the SET1 histone H3K4 methyltransferase complexes, the PNUTS-PP1 phosphatase complex, and more recently a complex with the zinc finger protein ZC3H4 (Austenaa et al., 2021; Brewer-Jensen et al., 2016; Lee et al., 2010; Lee and Skalnik, 2005; Miller et al., 2001; van Nuland et al., 2013). Interestingly, these different complexes seem to have opposing roles in regulating transcription. The SET1 complexes act at gene promoters to support transcriptional activity, whereas the PNUTS-PP1 phosphatase complex is best characterised as promoting transcription termination at the 3' end of genes (Cortazar et al., 2019; Hughes et al., 2022). ZC3H4 also supports transcription termination, but is believed to primarily promote termination away from the 3' end of genes, for example enhancer transcripts or upstream antisense transcripts (Austenaa et al., 2021; Brewer-Jensen et al., 2016; Estell et al., 2021). Recent work from the Klose lab has discovered that the SET1 complexes can promote transcription independently of their methyltransferase activity, and instead regulate gene expression through interaction with WDR82 (Hughes et al., 2022). Interestingly, it was found that the SET1 complexes enable gene expression by specifically

counteracting premature transcription termination by ZC3H4. Given that WDR82 is a common subunit amongst the SET1, PNUTS, and ZC3H4 complexes, one could envisage a role for WDR82 in enabling coordination of these opposing activities. However, the molecular function of WDR82 within each complex is poorly understood.

An important aspect of understanding the function of WDR82 in these different contexts is a detailed picture of the molecular identity of each complex. Previous work to biochemically define the composition of WDR82-containing complexes has employed a number of methods to exogenously express tagged proteins in a variety of cell types for purification and identification of interactors by mass spectrometry (Lee et al., 2010; Park et al., 2022; van Nuland et al., 2013). These studies defined the core components of the SET1 and PNUTS complexes, however there does not exist directly comparable datasets identifying interactors of WDR82, SET1A, PNUTS, and ZC3H4 in the same cell type. I therefore set out to define the repertoire of WDR82-containing complexes in ESCs, in particular their constitutive components, relative abundances, and any additional interactors which could confer functional specificity.

3.2 Results

3.2.1 Defining the molecular identity of WDR82-containing complexes

In order to examine the composition of WDR82-containing complexes in mESCs, I set out to purify WDR82, SET1A, ZC3H4, and PNUTS by Strep affinity under near-native conditions, preserving protein-protein interactions for identification by mass spectrometry. To do this, I made use of cell lines previously generated in the Klose lab ((Hughes et al., 2022) and Amy Hughes) in which WDR82, SET1A, PNUTS, and ZC3H4 had been endogenously tagged with triple-T7 and Twin-Strep tags (herein referred to as ST7). The tagged WDR82, PNUTS, and ZC3H4 cell lines were generated in a dTAG-SET1AB background,

so the ST7-tagged SET1A also carried an N-terminal dTAG and the parental dTAG-SET1AB cell line was used as the negative control. Western blot analysis shows that levels of the tagged proteins are comparable to the parental cell line (Figure 3.1A). SET1A levels are slightly reduced in the tagged line but are similar to the ZC3H4-ST7 line, suggesting this may be due to general variability in SET1A levels rather than a consequence of tagging SET1A. Direct comparison of SET1A, PNUTS, and ZC3H4 levels using an anti-T7 antibody revealed that PNUTS is approximately four-fold more highly expressed that SET1A and ZC3H4, which are expressed at approximately equal levels (Figure 3.1B, C). The relative level of WDR82 could not be examined in parallel as it is significantly smaller (45.2kDa including tags) than SET1A, PNUTS and ZC3H4 so could not be visualised on the same gel, however its expression is believed to be in excess of its binding partners (Cho et al., 2022).

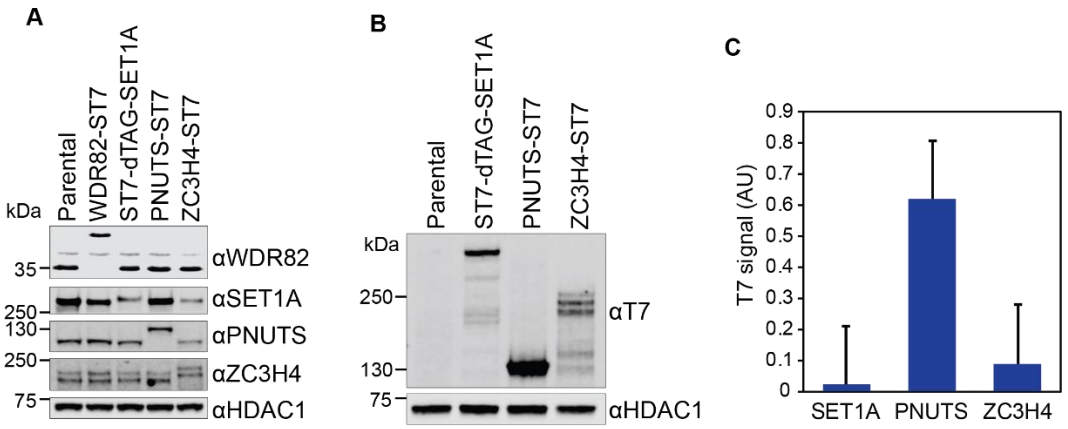


Figure 3.1 Endogenously tagged cell lines. **A:** Western blot of nuclear extract from parental and ST7-tagged cell lines showing comparable levels of tagged and WT proteins. HDAC1 is included as a loading control. **B:** Western blot of nuclear extracts from ST7-tagged SET1A, PNUTS and ZC3H4 cell lines probed with antibody specific to T7 tag to compare relative expression levels. HDAC1 is included as a loading control. **C:** Quantification of T7 signal from western blots of nuclear extract from ST7-tagged SET1A, PNUTS, and ZC3H4 cell lines. T7 signal was normalised to HDAC1. Error bars show SEM from three biological replicates.

Given that WDR82, SET1A, PNUTS, and ZC3H4 are localised to the nucleus and I am interested primarily in their roles as transcriptional regulators, I chose to purify each protein from nuclear extracts. A workflow for the pulldown and mass spec experiments is presented in Figure 3.2A. Extracts were precleared with Benzonase to digest any nucleic

acids and thus ensure I am only capturing direct protein-protein interactions and not interactions mediated by RNA or chromatin. Pulldown wash conditions were mild (150mM NaCl) to minimise disruption of less stable interactions. Silver staining of eluted proteins (Figure 3.2B) shows a distinct profile of bands for each bait protein, indicating a different set of copurified factors. The strongest bands seen in the SET1A and ZC3H4 pulldowns are also present in the control, suggesting relatively low enrichment of specific factors over nonspecific background binding, however some unique bands are distinguishable. Both the PNUTS and WDR82 pulldowns show strong enrichment of different bands, indicating a good signal over background that is reflective of their higher expression levels.

In order to identify the proteins present in each pulldown, samples were subjected to proteolytic digestion and liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS). Four biological replicates were performed to allow for statistical analysis. SET1A exhibited relatively poor enrichment in the pulldown samples, with a fold change of 16.42 over the control (Figure 3.2C, Table 3.1). This may contribute to the limited range of interactors identified as statistically significant. Of the core complex components, only CFP1 (Cxxc1) and RBBP5 were found to be above the significance threshold of fold change (FC) >2 relative to the control and *p* value <0.05. WDR82 was highly enriched (FC 8.03), however its *p* value of 0.019 was below the significance threshold. Similarly, ASH2L (FC 3.95, *p* 0.15) and WDR5 (FC1.82, *p* 0.15) were also present but not considered statistically significant hits. These proteins are highly abundant in cells and were also present in negative control samples, meaning the modest enrichment of SET1A and its interactors was not sufficient to consistently identify these proteins with high confidence. Both HCF1 and BOD1L, which have been previously identified as components of SET1A complexes, were also identified in my SET1A pulldown samples (van Nuland et al., 2013; Wang et al., 2017). In summary, the limited enrichment of SET1A resulted in identification of a relatively limited set of interactors that are nevertheless consistent with previously published experiments. SET1A

was the least abundant of the bait proteins I examined (Figure 3.1), and this likely contributed to its poor enrichment in pulldown samples. If this experiment were to be repeated, scaling up the amount of input material to account for lower endogenous protein levels may allow for more detailed interrogation of the associated proteins.

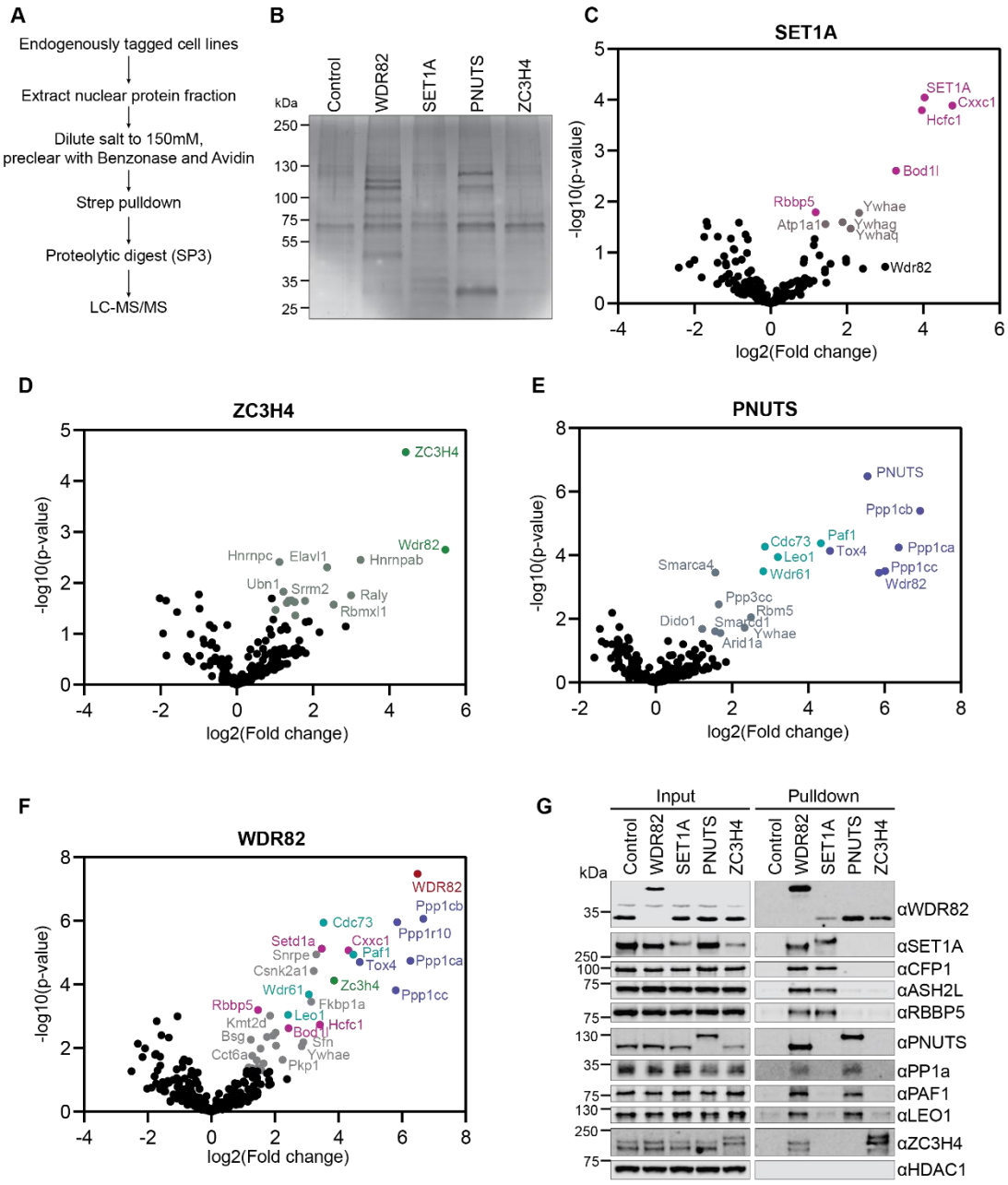


Figure 3.2 Characterising the composition of WDR82-containing complexes. **A:** Flowchart of experimental procedure for identification of protein interactors by Strep pulldown and mass spectrometry. **B:** Silver-stained SDS-PAGE gel of Strep pulldown samples as indicated. Control cell line was dTAG-SET1AB. **C:** Volcano plot of SET1A- interacting proteins. SET1A complex components are highlighted in pink and other statistically significant proteins in grey. WDR82 was highly enriched

with *p* below significance threshold and is labelled in black. Data from four biological replicates. Significance threshold FC>2, *p*<0.05. See also Table 3.1. **D:** Volcano plot of ZC3H4- interacting proteins. Known ZC3H4 complex components are highlighted in green and other statistically significant proteins in grey. Significance threshold FC>2, *p*<0.05. See also Table 3.2. **E:** Volcano plot of PNUTS- interacting proteins. Known PNUTS complex components are highlighted in blue, PAF1 complex components are highlighted in teal, and other statistically significant proteins in grey. Significance threshold FC>2, *p*<0.05. See also Table 3.3 and 3.4. **F:** Volcano plot of WDR82-interacting proteins. SET1 complex components are highlighted in pink, ZC3H4 complex components in green, PNUTS complex components in blue, PAF1 complex components in teal, and other statistically significant proteins in grey. Significance threshold FC>2, *p*<0.05. See also Table 3.5. **G:** Western blot of input nuclear extracts and Strep pulldown samples from tagged cell lines as indicated. HDAC1 is a loading control for inputs and negative control for pulldown experiments.

Gene name	Protein	Fold change	*p* value	Molecular function
SET1A (bait)	Histone-lysine N-methyltransferase SETD1A (SET1A)	16.42	9.00E-05	SET1 histone methyltransferase complex
Cxxc1	CXXC-type zinc finger protein 1 (CFP1)	27.15	1.30E-04	SET1 histone methyltransferase complex
Hcfc1	Host cell factor 1 (HCF1)	15.62	1.61E-04	SET1 histone methyltransferase complex
Bod1l	Biorientation of chromosomes in cell division protein 1-like 1 (BOD1L)	9.77	0.0025	SET1 histone methyltransferase complex
Rbbp5	Retinoblastoma-binding protein 5 (RBBP5)	2.27	0.0162	SET1 histone methyltransferase complex
Ywhae	14-3-3 protein ε	4.99	0.0167	Signalling
Ywhag	14-3-3 protein γ	3.70	0.0254	Signalling
Atp1a1	Sodium/potassium-transporting ATPase subunit alpha-1 (ATP1A1)	2.71	0.0278	Ion transport
Ywhaq	14-3-3 protein θ	4.28	0.0340	Signalling

Table 3.1 List of statistically significant (FC>2, *p*<0.05) SET1A interactors.

ZC3H4 showed slightly better enrichment than SET1A in pulldown samples, with 21.6-fold enrichment over the control (Figure 3.2D, Table 3.2). As expected, WDR82 was also highly enriched. Interestingly, many of the other statistically significant hits were RNA-binding proteins, including a number of proteins associated with splicing such as the

spliceosome component SRRM2. Whilst ZC3H4 does also bind RNA, nuclear extracts were treated with Benzonase prior to performing the pulldowns, so these interactions are unlikely to be mediated by RNA. These results suggest ZC3H4 may act as an interaction 'hub' for RNA-binding proteins and this could underpin its role in regulating termination by specifying RNAs to be terminated. Interestingly, XRN2, the 5' to 3' exonuclease involved in canonical transcription termination, was identified just below the significance threshold (FC 1.95, *p* 0.052), suggesting a possible interaction with ZC3H4. This interaction could also contribute to transcription termination by ZC3H4. In summary, ZC3H4 interacts with WDR82 as previously reported, as well as a number of RNA-binding proteins. Given nuclear extracts were treated with Benzonase prior to ZC3H4 purification, it is unlikely that these interactions are mediated by RNA.

Table 3.2 List of statistically significant (FC>2, *p*<0.05) ZC3H4 interactors.

Gene name	Protein	Fold change	*p* value	Molecular Function
ZC3H4 (bait)	Zinc finger CCCH domain-containing protein 4 (ZC3H4)	21.60	2.71E-05	ZC3H4-WDR82 complex
Wdr82	WD-repeat containing protein 82 (WDR82)	44.27	0.0022	ZC3H4-WDR82 complex
Hnrnpab	Heterogeneous nuclear ribonucleoprotein A/B	9.50	0.0035	RNA binding
Hnrnpc	Heterogeneous nuclear ribonucleoproteins C1/C2	2.17	0.0039	Binds pre-mRNA, possible roles in splicing
Elavl1	ELAV-like protein 1 (ELAVL1)	5.17	0.0049	RNA binding
Ubn1	Ubinuclein 1	2.34	0.0149	Chromatin regulation
Raly	RNA-binding protein RALY	7.98	0.0176	RNA binding
Ddx3y	ATP-dependent RNA helicase DDX3Y	2.71	0.0221	Probable ATP-dependent RNA helicase
Srrm2	Serine/arginine repetitive matrix protein 2 (SRRM2)	2.57	0.0223	Spliceosome component, RNA binding
Hadha	Trifunctional enzyme subunit alpha, mitochondrial	3.47	0.0224	Mitochondrial β-oxidation pathway enzyme

Arid1a	AT-rich interactive domain-containing protein 1A (ARID1A)	2.87	0.0237	SWI/SNF chromatin remodelling complex
Smarcd1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily D member 1 (SMARCD1)	2.48	0.0246	SWI/SNF chromatin remodelling complex
Rbmxl1	RNA binding motif protein, X-linked-like-1 (RBMXL1)	5.82	0.0267	RNA binding, possible roles in splicing
L1td1	LINE-1 type transposase domain-containing protein 1 (L1TD1)	2.03	0.0338	Possible RNA binding
Utp14a	U3 small nucleolar RNA-associated protein 14 homolog A (UTP14A)	2.90	0.0436	rRNA processing

The PNUTS pulldown showed good enrichment of PNUTS (46.9 fold over control) as well as the previously identified core complex components WDR82, TOX4 and PP1 (Figure 3.2E, Table 3.3). The three PP1 enzymes – α, β, and γ were all represented, consistent with previous reports that PNUTS does not preferentially bind a specific type (Choy et al., 2014). Interestingly, a number of PAF1 complex (PAF1C) components were also identified. PAF1, CDC73, LEO1 and WDR61 were all significantly enriched, whilst CTR9 was also present but below the statistical significance threshold (FC 1.13, *p* 0.098). This interaction between PNUTS and the elongation factor complex PAF1C suggests a possible role for PNUTS in regulating transcription away from its best-characterised role promoting termination at the 3' end of genes. Both the PNUTS and PAF1 complexes can interact with RNAPII, however no RNAPII components were significantly enriched in the PNUTS pulldown, suggesting PNUTS and PAF1C interact directly and independently of RNAPII (Carminati et al., 2022; Ciurciu et al., 2013; Ebmeier et al., 2017; Jerebtsova et al., 2011; Lee et al., 2010; Vos et al., 2018a). Whilst not statistically significant, a number of 3' end processing factors were also identified in the PNUTS pulldown, including Symplekin, several CPSF proteins, and XRN2

(Table 3.4), suggesting potential transient interactions between these complexes, as has been previously described (Benjamin et al., 2021; Shi et al., 2009; Vanoosthuyse et al., 2014). In summary, these results suggest PNUTS may regulate transcription away from the 3' end of genes via the PAF1 complex, however the precise details of these effects are unknown.

Table 3.3 List of statistically significant (FC>2, p<0.05) PNUTS interactors.

Gene name	Protein	Fold change	*p* value	Molecular function
PNUTS (bait)	PP1 nuclear targeting subunit (PNUTS)	46.88	3.27E-07	PNUTS-PP1 phosphatase complex
Ppp1cb	Protein phosphatase 1 β (PP1β)	121.95	4.03E-06	PNUTS-PP1 phosphatase complex
Paf1	RNAPII -associated factor 1 (PAF1)	20.16	4.21E-05	PAF1 complex
Cdc73	Cell division control protein 73 (CDC73)	7.29	5.33E-05	PAF1 complex
Ppp1ca	Protein phosphatase 1 α (PP1α)	82.82	5.73E-05	PNUTS-PP1 phosphatase complex
Tox4	TOX HMG box family member 4 (TOX4)	23.73	7.34E-05	PNUTS-PP1 phosphatase complex
Leo1	RNA Polymerase-associated protein LEO1 (LEO1)	9.22	1.13E-04	PAF1 complex
Ppp1cc	Protein phosphatase 1 γ (PP1γ)	64.68	3.12E-04	PNUTS-PP1 phosphatase complex
Wdr61	WD-repeat containing protein 61 (WDR61)	7.06	3.21E-04	PAF1 complex
Smarca4	Transcription activator BRG1 (BRG1)	2.95	3.50E-04	SWI/SNF chromatin remodelling complex
Wdr82	Wd-repeat containing protein 82 (WDR82)	57.69	3.53E-04	PNUTS-PP1 phosphatase complex
Ppp3cc	Serine/threonine-protein phosphatase 2B catalytic subunit gamma isoform (PP2Bγ)	3.14	0.0035	Phosphatase

Rbm5	RNA-binding protein 5 (RBM5)	5.65	0.0090	Spliceosome
Ywhae	14-3-3 protein ε	5.03	0.0190	Signalling
Dido1	Death-inducer obliterator 1 (DIDO1)	2.33	0.0206	Putative transcription factor
Smarcd1	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily D member 1 (SMARCD1)	2.95	0.0245	SWI/SNF chromatin remodelling complex
Arid1a	AT-rich interactive domain-containing protein 1A (ARID1A)	3.25	0.0277	SWI/SNF chromatin remodelling complex

Table 3.4 List of 3' end processing factors identified in PNUTS pulldown.

Gene name	Protein	Fold change	*p* value
Sympk	Symplekin	2.04	0.0939
Cpsf2	Cleavage and polyadenylation specificity factor subunit 2 (CPSF2)	1.91	0.3119
Cpsf1	Cleavage and polyadenylation specificity factor subunit 1 (CPSF1)	1.76	0.3145
Cpsf3	Cleavage and polyadenylation specificity factor subunit 3 (CPSF3)	2.51	0.3189
Wdr33	WD-repeat containing protein 33 (WDR33)	1.34	0.3627
Xrn2	5'-3' exoribonuclease 2 (XRN2)	1.40	0.3966

In order to further validate and characterise the precise constituents of the SET1A, ZC3H4, and PNUTS complexes, I also identified interactors of WDR82 (Figure 3.2F, Table 3.5). Of the SET1 complex components, SET1A, CFP1, RBBP5, HCF1 and BOD1L were all significantly enriched in the WDR82 pulldown, whilst ASH2L was also present but slightly below the significance threshold (FC 3.65, *p* 0.13). SET1B was not detected, most likely due to its very low expression in ESCs. ZC3H4 was significantly enriched, however none of the RNA-binding proteins identified from the ZC3H4 pulldown were found to be statistically

significant interactors of WDR82, suggesting the 'core' complex consists of just ZC3H4 and WDR82. Rbmxl1, which was identified as a ZC3H4 interactor, was enriched in the WDR82 pulldown (FC 3.52), but was not statistically significant (*p* 0.099), suggesting a possible transient or substoichiometric interaction. The core PNUTS complex components were well represented in the WDR82 pulldown, as were the PAF1C components identified in the PNUTS pulldown. This suggests a stable association between the PNUTS and PAF1 complexes and confirms that this complex does include WDR82. In addition to the SET1A, ZC3H4 and PNUTS/PAF1 complexes, the WDR82 purification also included a number of T-complex proteins. These are components of the TRiC type II chaperonin complex, which is known to assist folding of WD40-domain proteins such as WDR82 (Lopez et al., 2015).

Table 3.5 List of statistically significant (FC>2, *p*<0.05) WDR82 interactors

Gene name	Protein	Fold change	*p* value	Molecular function
WDR82 (bait)	WD-repeat containing protein 82 (WDR82)	89.93	3.35E-08	WDR82
Ppp1cb	Protein phosphatase 1 β (PP1β)	102.01	8.69E-07	PNUTS-PP1 phosphatase complex
Ppp1r10	PP1 nuclear targeting subunit (PNUTS)	57.64	1.10E-06	PNUTS-PP1 phosphatase complex
Cdc73	Cell division control protein 73 (CDC73)	11.55	1.14E-06	PAF1 complex
Setd1a	Histone-lysine N-methyltransferase SETD1A (SET1A)	11.11	7.49E-06	SET1 histone methyltransferase complex
Cxxc1	CXXC-type zinc finger protein 1 (CFP1)	19.91	8.52E-06	SET1 histone methyltransferase complex
Snrpe	Small nuclear ribonucleoprotein E (SNRPE)	9.86	1.13E-05	Spliceosome
Paf1	RNAPII -associated factor 1 (PAF1)	22.20	1.16E-05	PAF1 complex
Ppp1ca	Protein phosphatase 1 α (PP1α)	76.87	1.81E-05	PNUTS-PP1 phosphatase complex
Tox4	TOX HMG box family member 4 (TOX4)	25.51	1.99E-05	PNUTS-PP1 phosphatase complex

Csnk2a1	Casein kinase II subunit alpha (CK2α)	9.34	3.78E-05	Kinase
Zc3h4	Zinc finger CCCH domain-containing protein 4 (ZC3H4)	14.58	7.50E-05	ZC3H4-WDR82 complex
Ppp1cc	Protein phosphatase 1 γ (PP1γ)	55.89	1.52E-04	PNUTS-PP1 phosphatase complex
Wdr61	WD-repeat containing protein 61 (WDR61)	8.41	2.07E-04	PAF1 complex
Fkbp1a	Peptidyl-prolyl cis-trans isomerase FKBP1A	8.83	3.50E-04	Proline isomerase
Rbbp5	Retinoblastoma-binding protein 5 (RBBP5)	2.77	6.37E-04	SET1 histone methyltransferase complex
Leo1	RNA Polymerase-associated protein LEO1 (LEO1)	5.35	9.00E-04	PAF1 complex
Kmt2d	Histone-lysine N-methyltransferase 2D (KMT2D, MLL4)	3.61	9.56E-04	H3K4 histone methyltransferase
Hcfc1	Host cell factor 1 (HCF1)	10.70	0.0018	SET1 histone methyltransferase complex
Bod1l	Biorientation of chromosomes in cell division protein 1-like 1 (BOD1L)	5.40	0.0024	SET1 histone methyltransferase complex
Med4	Mediator of RNA polymerase II transcription subunit 4 (MED4)	4.06	0.0033	Mediator complex
Rbm5	RNA-binding protein 5 (RBM5)	3.84	0.0041	Spliceosome
Dido1	Death-inducer obliterator 1 (DIDO1)	3.37	0.0045	Putative TF
Bsg	Basigin	2.37	0.0054	Receptor
Sfn	14-3-3 protein σ	7.45	0.0066	Signalling
Tubb4b	Tubulin beta-4b chain	4.10	0.0084	Microtubule component
Ywhae	14-3-3 protein ε	7.18	0.0088	Signalling
Tcp1	T-complex protein 1 subunit alpha (TCPα)	2.93	0.0104	Chaperonin-containing T-complex (TRiC)
Cct6a	T-complex protein 1 subunit zeta (TCPζ)	2.44	0.0176	Chaperonin-containing T-complex (TRiC)
Pkp1	Plakophilin 1	4.73	0.0234	Cell adhesion

Cct5	T-complex protein 1 subunit epsilon (TCPε)	2.70	0.0244	Chaperonin-containing T-complex (TRiC)
Cct3	T-complex protein 1 subunit gamma (TCPγ)	3.12	0.0305	Chaperonin-containing T-complex (TRiC)
Cct2	T-complex protein 1 subunit beta (TCPβ)	2.27	0.0410	Chaperonin-containing T-complex (TRiC)
Tecr	Very-long-chain enoyl-CoA reductase	2.74	0.0422	Fatty acid synthesis
Atp2a2	Sarcoplasmic/endoplasmic reticulum calcium ATPase 2	2.41	0.0428	Ion transport
Macf1	Microtubule-actin cross-linking factor 1	2.99	0.0501	Microtubule component

To confirm the interactions identified by mass spectrometry I performed Strep pulldowns followed by western blot for various components of each complex (Figure 3.2G). This confirms that the SET1A, PNUTS and ZC3H4 complexes each interact with WDR82 in a mutually exclusive manner. I also confirmed the presence of several complex-specific components: RBBP5, CFP1 and ASH2L are specific to SET1A complexes, whilst PP1, PAF1, and LEO1 are specific to PNUTS complexes. Importantly, the presence of these proteins in the WDR82 pulldown confirms that they are not mutually exclusive with WDR82.

3.2.2 Further characterising WDR82-containing complexes

To further characterise WDR82-containing complexes in ESCs I wanted to understand the distribution of WDR82 and its interactors in different complexes. I therefore performed size-exclusion chromatography of Benzonase-treated nuclear extract to fractionate the complexes by apparent molecular weight (MW) (Figure 3.3). WDR82 is present in a wide range of high MW fractions, however a significant pool is present in monomeric form. The distribution of WDR82 in high MW fractions approximately reflects the combined distributions and relative abundances of SET1A, PNUTS and ZC3H4, consistent with it being a core component of these three complexes. SET1A exists only in very high (>700kDa)

apparent MW complexes that are larger than would be expected from the previously characterised core complex components (SET1A, WDR82, CFP1, and the WRAD complex), which total 484kDa. This shift in apparent MW could be accounted for by stable association with HCF1 (209kDa) and BOD1L1 (327kDa) or by dimerisation of the core SET1 complex via Dpy30, as has been described for the yeast Set1 complex (Choudhury et al., 2019). The distribution of CFP1 closely reflects that of SET1A, with very little protein detectable in lower MW fractions, consistent with previous observations that CFP1 is unstable when not in complex with SET1A/B. ASH2L, which is a component of six different H3K4 HMT complexes, shows a bimodal distribution; a large proportion is present in high MW fractions corresponding to the H3K4 HMT complexes, whilst another pool exists in lower MW fractions representing either monomeric ASH2L or smaller subcomplexes such as WRAD or ASH2L-DPY30 that can exist independently of the larger HMT complex assemblies (van Nuland et al., 2013).

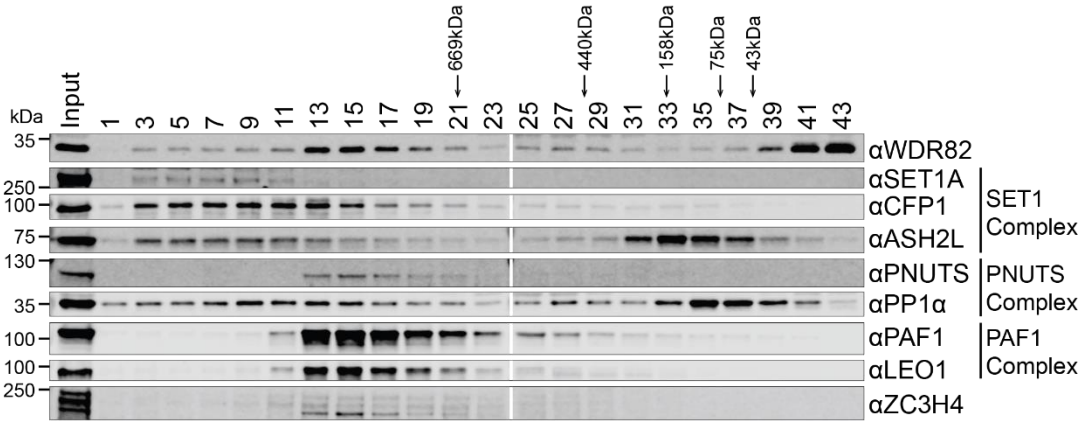


Figure 3.3 Fractionation of nuclear extract. Size exclusion chromatography analysis of wild-type ESC nuclear extract probed by western blot with antibodies indicated. Elution positions of known MW standards are marked.

PNUTS eluted exclusively in high (>669kDa) MW fractions, suggesting it exists constitutively in complex with other factors. The expected MW of the previously defined PNUTS-PP1-WDR82-TOX4 complex is only 233kDa, strongly suggesting that PNUTS stably

interacts with additional factors which increase the apparent MW of the complex. My previous experiments identified components of the PAF1 complex as specific interactors of the PNUTS complex and both PAF1 and LEO1 elute in a very similar profile to PNUTS, suggesting that the interaction between the two complexes may be relatively stable. The combined MW of the PNUTS-PP1 and PAF1 complexes is 678kDa, a figure more consistent with their elution profiles, further supporting the conclusion that the PNUTS and PAF1 complexes exist in a previously uncharacterised stable complex. Comparative elution profiles of recombinant PNUTS and PAF1 complexes or of nuclear extract following PNUTS depletion would be required to shed further light on this unexpected observation.

The elution profile of PP1α shows almost no protein in the monomeric form (38kDa). Instead, it exists in a variety of higher MW complexes that are consistent with previous observations that all PP1 in cells exists in complex with a regulatory subunit to prevent spurious unregulated activity by the free enzyme (Jagiello et al., 1995; Verbinnen et al., 2017). A large pool of PP1 eluted in fractions around 75kDa, which is likely to represent PP1 in complex with NIPP1 (Nuclear Inhibitor of PP1, 38kDa), a complex which has been shown to sequester more than a third of the nuclear pool of PP1 (Mehta et al., 2022; O'Connell et al., 2012). The remainder of the nuclear PP1 is distributed relatively broadly across high MW fractions, including overlapping with PNUTS, reflecting the variety of PP1 regulatory complexes that exist in the nucleus.

Surprisingly, ZC3H4 shows a very similar elution profile as PNUTS and PAF1, however my pulldown experiments did not detect any interaction between these proteins, suggesting this is a coincidental phenomenon. Interestingly, ZC3H4 also migrates at a significantly higher apparent MW than expected. The combined MW of a 1:1 WDR82-ZC3H4 complex is only 175kDa, suggesting potential stable interactions with additional proteins, however I did not identify any candidate proteins from my pulldown experiments. This

suggests that ZC3H4 may instead form a homo-oligomeric complex, the stoichiometry of which is not clear from this experiment.

Overall, these results show that the population of WDR82 is present both in a large monomeric pool and distributed across multiple complexes proportionally to their abundance, suggesting it is a constitutive component of these complexes. I have also shown that PNUTS and PAF1 elute at an apparent MW that is consistent with a previously uncharacterised stable association between these two complexes. Further experiments will be required to confirm and characterise this larger complex.

3.3 Summary and Discussion

WDR82 is an essential, highly conserved eukaryotic protein that is part of multiple complexes known to regulate transcription. The SET1, PNUTS, and ZC3H4 complexes have been identified as interactors of WDR82, however comprehensive characterisation of all three complexes in the same cell type has been lacking. I therefore sought to define the repertoire of WDR82-containing complexes in ESCs by purification of endogenously tagged proteins followed by LC-MS/MS. The SET1A H3K4 HMT complex is well characterised in ESCs and my data was consistent with previous observations. I found that the constitutive ZC3H4 core complex is likely to consist of only ZC3H4 and WDR82, potentially in a homo-oligomeric assembly. Nevertheless, a number of RNA-binding proteins were found to bind ZC3H4 substoichiometrically. Interestingly, I found that the PNUTS-PP1 complex interacts with the PAF1 complex, an elongation factor which promotes PolII processivity and elongation speed (Hou et al., 2019).

My SET1A interactome data was limited due to poor enrichment of the SET1A bait protein during pulldowns. SET1A was the lowest expressed of the bait proteins (Figure 3.1B, C), and therefore yielded lower recovery from the equal amounts of input nuclear extract

used for all pulldowns. Nevertheless, I was able to identify the majority of previously defined SET1A complex components, namely CFP1, WDR82, RPPB5, ASH2L, WDR5, HCF1, and BOD1L (Figure 3.2C). I validated the presence of ASH2L and WDR82, which were below the statistical significance threshold for MS, by western blot (Figure 3.2G). Size exclusion chromatography of nuclear extract found that almost all SET1A is present in high apparent MW complexes which were consistent with association of HCF1 and/or BOD1L in addition to the 'canonical' core complex components. The presence of HCF1 and BOD1L in complex with SET1A has been noted previously, however their roles in this context are poorly understood. HCF1 has been proposed to physically link the SET1A complex with the Sin3 histone deacetylase complex, however the relevance of this interaction for chromatin regulation or transcription is unclear (Wysocka et al., 2003). BOD1L has been shown to protect stalled DNA replication forks in a manner dependent on the catalytic activity of SET1A, suggesting a context-specific function for the SET1A-BOD1L interaction (Higgs et al., 2018, 2015).

Consistent with previous reports, WDR82 was highly enriched in the ZC3H4 pulldown and ZC3H4 was enriched in the WDR82 pulldown, indicating a stable interaction. No other proteins were found in both the ZC3H4 and WDR82 pulldowns, suggesting the 'core' ZC3H4 complex consists of only ZC3H4 and WDR82. Interestingly, the apparent molecular weight of the ZC3H4 complex, as determined by size exclusion chromatography, was significantly higher than would be expected from a 1:1 complex of ZC3H4 and WDR82. This suggests that ZC3H4 may form a homo-oligomeric complex, which could form via several predicted coiled-coil regions in ZC3H4. Others have identified a complex of ZC3H4 with WDR82 and casein kinase 2 (CK2), however I did not identify any CK2 subunits in my ZC3H4 pulldown (Park et al., 2022). CK2α was a significant hit in the WDR82 pulldown, so I cannot rule out the possibility of a substoichiometric ZC3H4-WDR82-CK2 complex in ESCs. ZC3H4 has also been previously identified as an interactor of the mRNA cap-binding proteins ARS2 (Schulze

et al., 2018). I did not identify any cap-binding components in my pulldowns. This may be because these interactions are wholly or partially mediated by RNA, and I treated the nuclear extracts with Benzonase prior to purification, disrupting any interactions which require RNA. The majority of proteins identified in the ZC3H4 pulldown were RNA-binding proteins, including a number of proteins linked with splicing. Furthermore, XRN2 was identified as a potential interactor of ZC3H4. The mechanism by which ZC3H4 enables premature transcription termination is unknown, however these interactions suggest it could involve regulation of the splicing machinery or recruitment of the XRN2 exonuclease. The presence of many RNA-binding proteins in the ZC3H4 interactome suggest a function that is closely related to RNA binding and ZC3H4 itself has also been shown to bind RNA (Austenaa et al., 2021). A repeat experiment without Benzonase treatment, to enable identification of RNA-mediated interactions, may be informative to understand further functional associations of ZC3H4.

PNUTS was the most abundant of the three WDR82-interacting proteins I examined (Figure 3.1B, C) and showed accordingly good yield in pulldown experiments. This higher abundance of PNUTS complexes is consistent with previous reports (van Nuland et al., 2013) and is important to note when considering the effects of WDR82 loss, as each WDR82-containing complex may not contribute equally to the observed transcriptional changes. Indeed, the most widely reported transcriptional change following WDR82 loss is a termination defect consistent with the disruption of PNUTS complexes (Austenaa et al., 2015). Interestingly, I identified a stable interaction between the PNUTS-PP1 complex and the PAF1 complex. PNUTS has been previously identified in PAF1C interactome experiments, but PAF1C components were absent from many PNUTS and WDR82 pulldowns (Cermakova et al., 2021; Ding et al., 2015; Landsverk et al., 2020, 2019). The reason for this discrepancy is unclear, however it could represent cell-type specific differences in interactions. In addition to their identification in PNUTS pulldowns, I also

identified PAF1C components in WDR82 pulldowns by both MS and western blot, indicating the interaction is not independent of WDR82. Furthermore, size exclusion chromatography of nuclear extract showed very similar elution profiles for PNUTS and PAF1 complex components at an apparent MW that is consistent with a combined complex, suggesting stable interaction. Additional experiments comparing these elution profiles with recombinant complexes or nuclear extracts from PNUTS-depleted cells will be required to validate this. Importantly, a motif has been identified in PAF1 which interacts with TFIIS N-terminal-like domains (TNDs) in a number of proteins, including PNUTS (Cermakova et al., 2021). *In vitro*, the PNUTS TND was found to interact in a mutually exclusive manner with motifs of IWS1, Spt6 and PAF1. Of these, only PAF1 was identified in my experiments, suggesting some mechanism for specification of the PNUTS TND binding partner *in vivo*. Interestingly, this region of PNUTS has also been shown to bind TOX4 (Lee et al., 2010). PAF1C components have been identified in TOX4 immunoprecipitation experiments, suggesting PNUTS binding to PAF1 and TOX4 is not mutually exclusive (Ding et al., 2015). Further investigation of the molecular basis of these interactions may shed light on the relationship between the PNUTS and PAF1 complexes.

The PAF1 complex is an elongation factor which has been shown to promote RNAPII processivity and efficient elongation (Francette et al., 2021; Hou et al., 2019; Vos et al., 2018a; Wang et al., 2022). Whilst PNUTS is best characterised in its role promoting termination at the 3' end of genes, its genomic distribution closely reflects that of RNAPII, suggesting it also acts to regulate transcription throughout gene bodies (Cortazar et al., 2019; Verheyen et al., 2015)(Amy Hughes, unpublished data). The interaction I have identified between PNUTS and PAF1 therefore provides a possible mechanism for transcriptional regulation separate to termination. Interestingly, PAF1 has been shown to be important for RNAPII to attain optimal speed and processivity, especially through the first 20-30kb of genes (Hou et al., 2019). In contrast, PNUTS was found to restrict

elongation speed throughout genes (Cortazar et al., 2019) and TOX4 has specifically been implicated in the restriction of early elongation (Liu et al., 2022). These observations suggest an intriguing mechanism in which the interaction between PNUTS and PAF1 brings together two opposing activities, perhaps to allow fine-tuning of elongation speed. The connection between PNUTS and PAF1 identified here represents an exciting avenue for further research into the role of PNUTS in transcription regulation.

Size exclusion chromatography of nuclear extracts to examine the distribution of WDR82 across these different complexes revealed a significant pool of monomeric protein, suggesting it is sufficiently abundant to constitutively incorporate into all SET1, ZC3H4, and PNUTS complexes. This is consistent with measurements of absolute protein levels in 293T cells, which show that WDR82 is present in excess of the combined amounts of SET1A/B, PNUTS, and ZC3H4 (Cho et al., 2022). It is therefore unlikely that competition for binding to WDR82 itself mediates the functional competition between, for example, SET1A/B and ZC3H4. If competition for binding to WDR82 itself does not underpin the functional relationship between these complexes, the function imparted by WDR82 itself may instead be the defining factor. In order to investigate this functional competition, we need to first understand how WDR82 integrates into each complex and its molecular function in those contexts. In the next chapter, I will examine the molecular basis of WDR82 binding to SET1A, ZC3H4, and PNUTS with a view to designing specific mutations which can be used to dissect these different complexes.

4 Understanding the molecular basis of WDR82 protein-protein interactions

4.1 Introduction

WDR82 is a small (35.1kDa) protein that is highly conserved across all eukaryotes. Sequence analysis suggests it consists of seven WD40 repeats with no other folded domains and little additional sequence at either terminus. The predicted structure of WDR82 from the AlphaFold database (Jumper et al., 2021; Varadi et al., 2022) reveals a classical 7 bladed β propeller structure with a short additional helix at the N-terminus (Figure 4.1A). Each blade of the propeller consists of a single WD40 repeat forming a four stranded antiparallel β sheet. As is common in WD40 repeat proteins, the outer strand of the final blade (7) is formed by the most N-terminal β strand of the protein. This 'velcro' closure is thought to enhance the stability of the fold (Xu and Min, 2011). WD40 domains are amongst the most common eukaryotic protein domains and are involved in a wide range of cellular functions, including signalling, cell cycle control, vesicular trafficking, and transcriptional regulation (Ma et al., 2019; Stirnimann et al., 2010). No WD40 domain has yet been identified with intrinsic enzymatic activity. Instead, these domains act as scaffolds for protein-protein interactions, often binding multiple proteins simultaneously to enable the coordination of downstream processes (Jain and Pandey, 2018; Stirnimann et al., 2010; Xu and Min, 2011). For example, FBW (F-box-WD40) proteins are substrate adapters for SCF (Skp1-Cul1-Fbox) ubiquitin ligases (Skaar et al., 2013, 2009), whilst the ability of the G_{β} subunit of the heterotrimeric G proteins to bind multiple partners is a key feature of signal transduction from G-protein coupled receptors (GPCRs) (Syrovatkina et al., 2016; Xu and Min, 2011).

WD40 proteins are also found in many transcription regulatory complexes, including Mediator, Polycomb Repressive Complex 2 (PRC2), and the Cleavage and Polyadenylation Specificity Factor (CPSF) complex (Chammas et al., 2020; Chen et al., 2021; Zhang et al., 2020). The majority of structurally characterised interactions with WD40 proteins are mediated via small peptide motifs which dock into the central pocket on the top of the β propeller, although binding has been observed on all faces of the domain (Stirnimann et al., 2010; Xu and Min, 2011). For example, FBW7 binds Cyclin E via its top face, whereas the clathrin-box motif of β-arrestin 2 binds the circumference of the clathrin WD40 domain, and G_{γ} binds the bottom face of G_{β} (Figure 4.1B)(Hao et al., 2007; Ter Haar et al., 2000; Wall et al., 1995).

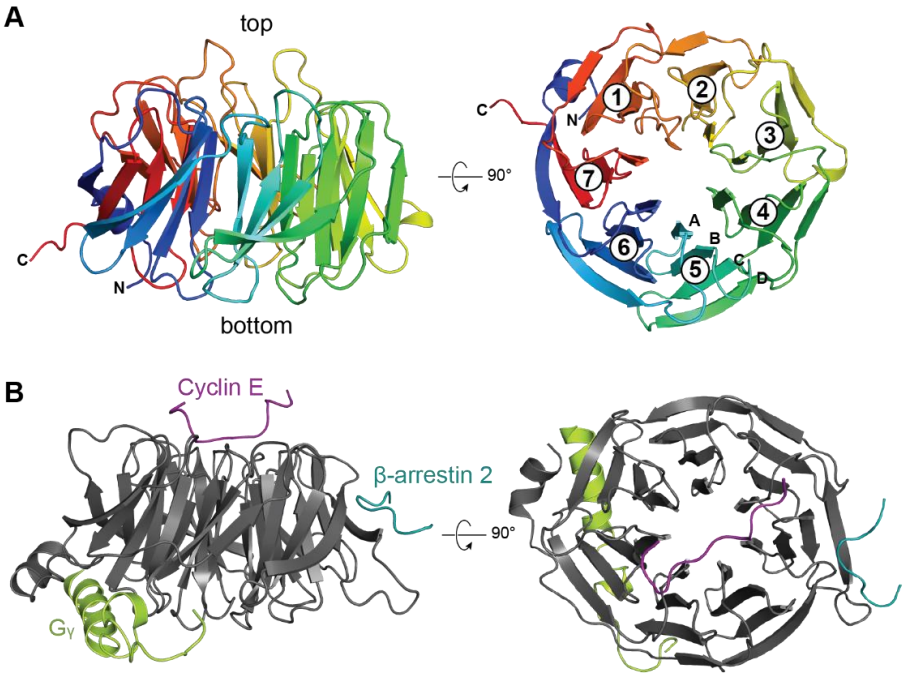


Figure 4.1 WD40 domains are protein-protein interaction scaffolds. **A:** Structure of WDR82. Side and top views of the predicted structure of mouse WDR82 (Uniprot ID Q8BFQ4) from the AlphaFold database. The chain is coloured rainbow from N-terminus (blue) to C-terminus (red), each blade is numbered 1-7, and strands within each blade are designated A-D. **B:** Side and top views of overlaid WD40 domain-peptide complexes showing interactions on different surfaces. Cyclin E (purple, 2OVQ (Hao et al., 2007)) binds the top face of the FBW7 β propeller, the clathrin-box motif of β-arrestin 2 (teal, 1C9I (Ter Haar et al., 2000)) binds the circumference of clathrin, and G_{γ} (lime, 1GP2 (Wall et al., 1995)) binds the bottom face of G_{β}. For clarity, only the WD40 domain of G_{β} is shown.

In the previous chapter I defined the interactome of WDR82 in mESCs and confirmed that its binding to SET1A, ZC3H4, and PNUTS is mutually exclusive. Given their mutual exclusivity, I hypothesised that SET1A, ZC3H4, and PNUTS would have overlapping binding interfaces on WDR82. Previous work has defined the regions of SET1A, ZC3H4, and PNUTS which bind WDR82 with varying degrees of refinement. SET1A has been shown to bind WDR82 within its first 247 amino acids, a region which encompasses an RRM domain and an N-terminal domain that is predicted to be largely unstructured (Lee and Skalnik, 2008). In yeast, the SET1A homolog ySET1 has been shown to bind the WDR82 homolog SWD2 specifically via the region N-terminal to the RRM (Bae et al., 2020; Kim et al., 2013). The WDR82 binding sites in ZC3H4 and PNUTS have been mapped to regions of approximately 200 amino acids, neither of which is predicted to harbour any significant structure (Austenaa et al., 2021; Lee et al., 2010; Park et al., 2022). Whilst these general WDR82-binding domains have been identified in SET1A, ZC3H4, and PNUTS, the precise molecular features that enable these interactions have not been characterised. I therefore sought to investigate the molecular basis of SET1A, ZC3H4, and PNUTS binding to WDR82. In particular, I aimed to understand the basis for their mutual exclusivity, and to identify key hotspot residues I could mutate to specifically disrupt each protein binding to WDR82.

4.2 Results

4.2.1 SET1A and SET1B bind WDR82 via their N-terminal domains

The WDR82-binding region of SET1A has been mapped to a 247 amino acid N-terminal region which includes both an RRM and unstructured NTD (Lee and Skalnik, 2008). However, the yeast SET1A homolog, ySET1, has been shown to bind the WDR82 homolog SWD2 specifically via its NTD (Bae et al., 2020; Kim et al., 2013). I therefore wanted to determine whether the NTD of SET1A is also responsible for binding to WDR82. To do this I

transiently transfected 293T cells with plasmids expressing FLAG-tagged full length SET1A, or SET1A in which specific domains had been deleted (Figure 4.2A), and assessed the capacity of these proteins to bind WDR82 by FLAG IP and western blot. SET1A lacking the RRM, central unstructured region ('linker'), or catalytic SET domain co-immunoprecipitated WDR82 to a similar extent as the full-length protein, indicating that these domains are dispensable for the interaction (Figure 4.2B). Deletion of the NTD, however, was associated with complete loss of WDR82 interaction, confirming that the SET1A NTD is necessary for binding to WDR82. The SET1A ΔNTD construct expressed poorly, which is consistent with the previously observed destabilisation of SET1A upon removal of WDR82 (Wu et al., 2008). However, the low expression of this construct may mean the WDR82 pulldown was simply below the detection limit of the western blot. I therefore decided to confirm that the NTD of SET1A is sufficient to bind WDR82 by testing whether the NTD-RRM, NTD only, and RRM only fragments of SET1A were able to interact with WDR82 (Figure 4.2C). The NTD and NTD-RRM fragments of SET1A co-purified WDR82, demonstrating that the NTD is sufficient for binding WDR82. The RRM-only fragment of SET1A did not express to detectable levels so I could not directly assess its ability to bind WDR82. However, IPs of the equivalent fragments of SET1B, which is highly similar to SET1A, demonstrate that the NTD alone, but not the RRM alone, is sufficient to bind WDR82 (Figure 4.2C, 2D). In conclusion, the NTD of SET1A is both necessary and sufficient to bind WDR82.

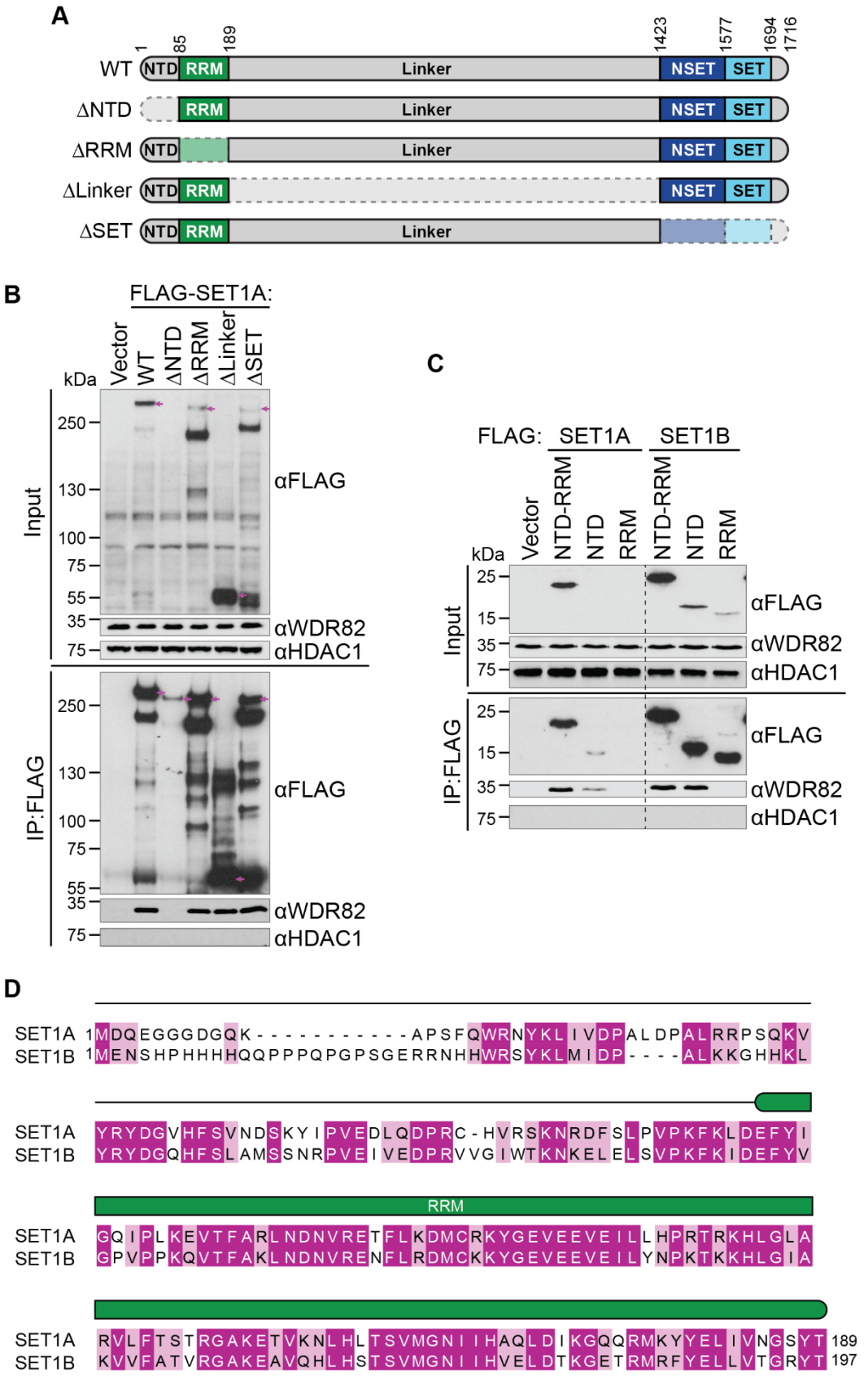


Figure 4.2 The SET1A NTD binds WDR82 **A**: Schematic representation of the domain organisation of SET1A and domain deletion constructs used in B. Numbers indicate amino acid positions. NTD: N-terminal Domain, RRM: RNA Recognition Motif, NSET: N-terminal to SET, SET: Su(var)3-9, Enhancer-

of-zeste and Trithorax. **B**: Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing N-terminally FLAG-tagged full-length (WT) SET1A or SET1A with specific domains deleted. Bands representing each fragment are marked with magenta arrows. HDAC1 is a loading control for inputs and negative control for IP experiments. **C**: Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing FLAG-tagged fragments of SET1A or SET1B. HDAC1 is a loading control for inputs and negative control for IP experiments. A dashed line indicates where the blot has been cropped for clarity. **D:** Sequence alignment of the NTD and RRM regions of SET1A and SET1B, shaded by similarity. A schematic above indicates the position of the RRM.

4.2.2 Predicting the structure of WDR82 complexes

Whilst the regions within SET1A, ZC3H4, and PNUTS that bind WDR82 have been crudely mapped, there is little information regarding the molecular basis of these interactions. Experimental interrogation of these interfaces would require either structure determination or significant biochemical interrogation to map smaller regions or motifs which mediate binding. However, recent developments in protein structure prediction, in particular the DeepMind AlphaFold2 pipeline, have proved to be extremely powerful, enabling highly accurate prediction of protein complex structures *in silico* within a few hours (Jumper et al., 2021). The AlphaFold2 neural network exploits the information in multiple sequence alignments (MSAs) of related proteins to directly predict the 3D structure of the query protein. Remarkably, evaluation of structure prediction by AlphaFold2 at CASP14 (Critical Assessment of protein Structure Prediction, round 14) found many structures were accurate to within the error margin of experimental structure determination methods. The original AlphaFold2 has been further optimised for prediction of protein-protein complex structures as AlphaFold2-multimer (Evans et al., 2022). ColabFold is an implementation of AlphaFold2 and AlphaFold2-multimer available through Google Colab (Mirdita et al., 2022). The most computationally intensive stage of structure prediction using the 'canonical' Alphafold2 pipeline is the generation of an input MSA using the highly sensitive HMMer and HHblits to search extensive environmental databases. The size and speed of these searches and the handling of the large MSAs they produce make

them computationally impractical for the average user. ColabFold modifies the AlphaFold2 pipeline by replacing this highly sensitive but ultimately very slow homology search with the 40-60-fold faster MMseqs2. Direct comparison has found that ColabFold matches AlphaFold2 and AlphaFold2-multimer in prediction quality (Mirdita et al., 2022). I therefore used ColabFold to interrogate the molecular basis of WDR82 protein-protein interactions by predicting the structure of its complexes with SET1A, ZC3H4, and PNUTS.

Computational capacity provided by Google Colab puts an upper limit of the length of sequence that can be predicted using ColabFold at around 1200 amino acids. SET1A (1716 amino acids), ZC3H4 (1304 amino acids), and PNUTS (888 amino acids) are all too large to input as full-length proteins alongside full length WDR82 (313 amino acids). I therefore used the shorter regions that had been previously identified to bind WDR82 as my input sequences for structure prediction.

4.2.2.1 Prediction of the WDR82-SET1A structure

To predict the interface between SET1A and WDR82, I input the sequences of full-length WDR82 and the region of SET1A comprising the NTD and RRM (residues 1-189) (Figure 4.3A) into ColabFold. Whilst I have shown that the SET1A NTD alone is sufficient for binding to WDR82, I also included the RRM as it is in close proximity to the WDR82-binding region and may form some interactions with the NTD or with WDR82 which are not strictly necessary for the complex to form but may be functionally relevant. Figure 4.3B shows the sequence coverage of the MSA generated during the prediction. There is a notable difference in the extent of MSA coverage between the folded domains in the query sequence (WDR82 and the SET1A RRM) and the SET1A NTD, which is largely unstructured. The reduced coverage and presence of fewer sequences with low sequence identity to the SET1A NTD is consistent with the features of intrinsically unstructured sequences, which lack selection pressure to maintain a specific fold so are more likely to vary between

species, and are often repetitive or of low complexity, which makes homology searches and accurate MSA generation more difficult. Coverage of at least 100 sequences is considered the minimum for a good prediction, and almost all regions of the input sequence were covered by an order of magnitude more than this. ColabFold was able to predict the structure of the WDR82-SET1A$^{1-189}$ complex with high confidence across almost the entire input sequence (Figure 4.3C). pLDDT (predicted Local Distance Difference Test) is a measure of confidence in the local protein structure, with a value above 70 considered to be a generally good backbone prediction and above 90 to be high accuracy even for side chains. Any value below 50 is a reasonably strong predictor of disorder (Tunyasuvunakool et al., 2021). A few regions of the SET1A$^{1-189}$ fragment, such as the termini, are predicted with low pLDDT, suggesting a more unstructured conformation.

ColabFold also provides an assessment of the overall structure as a plot of the Predicted Alignment Error (PAE) (Figure 4.3D). Unlike pLDDT, which measures local confidence for each residue, the PAE is a measure of confidence in the position of every residue relative to every other residue. A low PAE (blue) therefore represents a high confidence in the position of a residue within the overall structure. Low intra-chain error, as seen for the entirety of WDR82, indicates high confidence in the overall fold of a protein or domain, whilst low inter-chain error indicates high confidence in the relative positions of each protein and hence the structure of a complex. The PAE plot for the WDR82-SET1A$^{1-189}$ complex shows that the SET1A$^{1-189}$ fragment contains two discrete domains predicted with high confidence (low error), corresponding to the NTD and RRM. The interaction between SET1A and WDR82 is predicted to be mediated by the NTD, with very low PAE scores between the NTD and WDR82. The position of the SET1A RRM relative to the WDR82-SET1A NTD complex is less well defined with much more modest PAEs, perhaps indicating some conformational flexibility between the two SET1A domains.

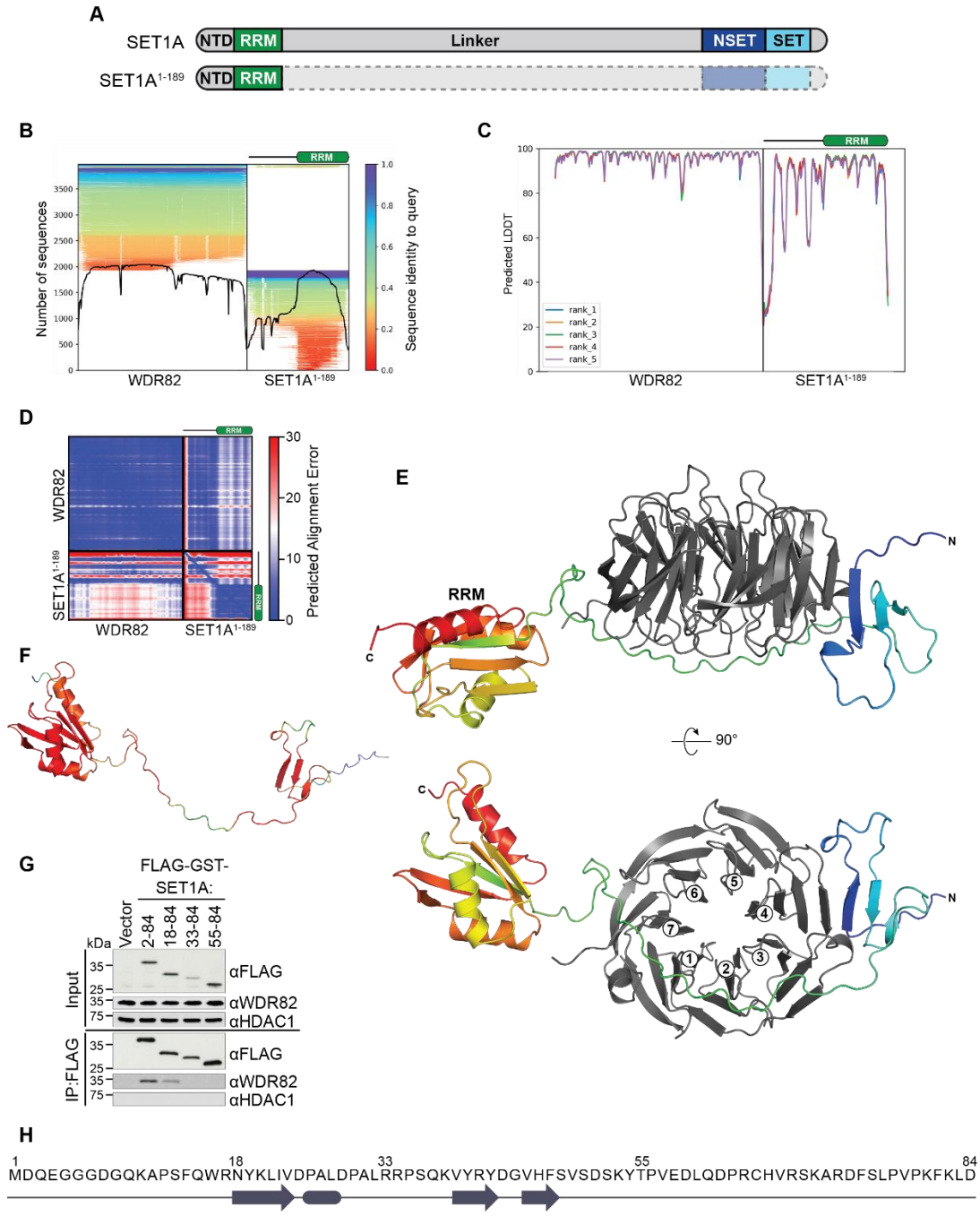


Figure 4.3 Prediction of the WDR82-SET1A complex structure. **A**: Schematic representation of the domain organisation of SET1A indicating the region used as input for structure prediction. **B**: Graphical representation of the depth and diversity of the MSA generated by ColabFold structure prediction for WDR82 and SET1A$^{1-189}$. A schematic above indicates the position of the SET1A RRM. **C**: Graph showing the per-position pLDDT for the 5 predicted structures of the WDR82-SET1A$^{1-189}$ complex generated by ColabFold. A schematic above indicates the position of the SET1A RRM. The 'rank 1' structure was used for all further analysis. **D**: Plot of Predicted Alignment Error (PAE) for the rank 1 WDR82-SET1A$^{1-189}$ structure. A schematic above indicates the position of the SET1A RRM. **E**: Side and bottom views of a cartoon representation of the rank 1 predicted WDR82-SET1A$^{1-189}$ structure generated by ColabFold. WDR82 is in grey and SET1A$^{1-189}$ is coloured rainbow from N-terminus (blue) to C-terminus (red). Numbers indicate the blades of WDR82. **F**: Cartoon representation of SET1A$^{1-189}$ from the rank 1 predicted WDR82-SET1A$^{1-189}$ complex structure

coloured by pLDDT on a spectrum from red (high) to blue (low). **G:** Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing FLAG-GST-tagged fragments of SET1A. HDAC1 is a loading control for inputs and negative control for IP experiments. **H**: Sequence of the SET1A NTD. Secondary structure elements from the predicted structure are indicated below, with helices as rounded rectangles and β strands as arrows. Numbers above are residue positions.

The SET1A NTD is predicted to bind WDR82 via an extensive interface across the bottom face of the β propeller (Figure 4.3E). Analysis of this predicted interface with PISA (Proteins, Interfaces, Structures and Assemblies) (Krissinel and Henrick, 2007; "PDBe < PISA < EMBL-EBI,") suggests it buries 1782.8Å^2 of the WDR82 surface, corresponding to 12.9% of the total surface area. The N-terminal region of the NTD forms a three-stranded β sheet which extends blade 4 of the WDR82 β propeller. The polypeptide chain then continues in an extended conformation across the bottom face of WDR82, contacting the loops of blades 3, 2, 1 and 7, and exiting in a groove between blades 7 and 6. The C-terminal RRM is a small folded domain which is not predicted to make direct contact with the WDR82-NTD complex. Visualisation of the SET1A$^{1\text{-}189}$ predicted structure isolated from the complex coloured by pLDDT shows that the residues which contact WDR82 are predicted with high confidence (red), whilst the regions predicted with lower confidence (blue, green) are limited to the termini and loop regions (Figure 4.3F). Interestingly, it is the most N-terminal strand of the SET1A NTD β sheet which is predicted to directly contact the outermost strand of WDR82 blade 4. To validate this predicted interaction experimentally, I tested the WDR82-binding ability of SET1A fragments which had been N-terminally truncated at various positions (Figure 4.3G, H). SET1A truncated just prior to the first β strand (fragment 18-84) retained interaction with WDR82, however deletion of the first β strand (fragment 33-84) or the entire β sheet structure (fragment 55-84) abolished WDR82 binding, demonstrating the importance of this small folded domain for SET1A binding to WDR82 and supporting the accuracy of the predicted structure.

Whilst ColabFold has been shown to match the original AlphaFold2 pipeline in prediction accuracy, I wanted to verify this was the case for the types of structures I am interested in. I therefore ran the same WDR82-SET1A complex through DeepMind's AlphaFold2 Colab, which is a slightly simplified version of the full pipeline that uses a smaller sequence database and no templates. The sequence coverage for WDR82 was much deeper using the Alphafold2 MSA protocol, with a maximum of 13,000 sequences for some regions compared to the 2,000 maximum from ColabFold (Figure 4.4A). The SET1A alignment, however, was of very similar depth to that generated by MMseqs2 for ColabFold, despite an approximately 10-fold longer search time (Figure 4.4B). Handling of the large WDR82 alignment during the prediction process was highly memory intensive, and side chain relaxation failed due to a lack of available memory. Comparison of the output prediction from AlphaFold2 run without amber relaxation to the equivalent unrelaxed top ranked structure from ColabFold reveals excellent agreement, with an RMSD for the aligned structures of 1.3Å across the full structure (excluding the unstructured N-terminal tail of SET1A), and 0.39Å for the WDR82-NTD only complex (Figure 4.4C). I am therefore confident that the results from ColabFold represent the best predictions available outside perhaps a full locally installed AlphaFold2-multimer pipeline.

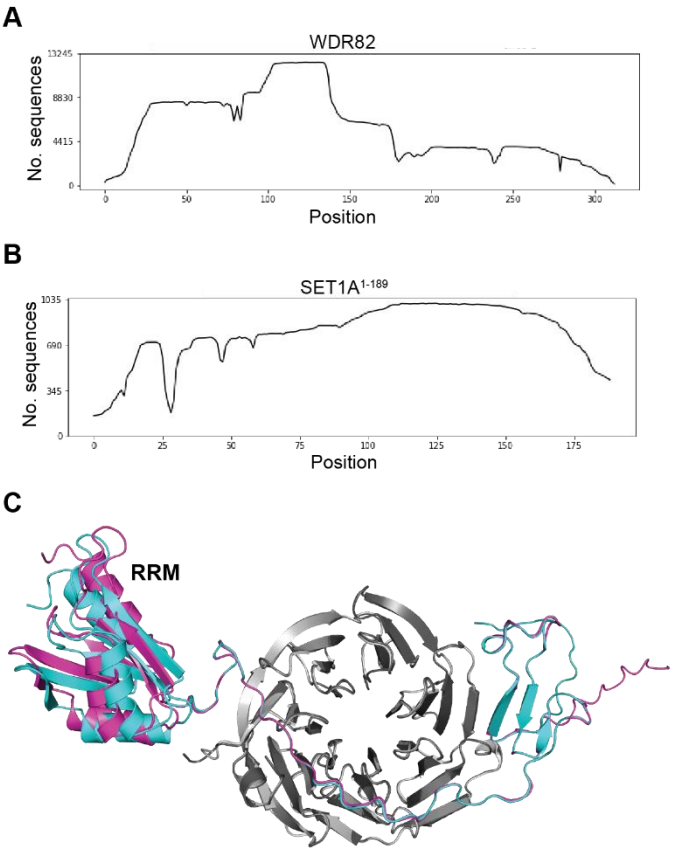


Figure 4.4 AlphaFold2 prediction of the WDR82-SET1A complex structure. **A**: Graph showing per-residue coverage in the MSA for WDR82 generated during prediction by AlphaFold2. **B**: Graph showing per-residue coverage in the MSA for SET1A$^{1-189}$ generated during prediction by AlphaFold2. **C**: Alignment of the WDR82-SET1A$^{1-189}$ structure predicted by AlphaFold2 (light grey/cyan) and the structure predicted by ColabFold (grey/magenta). Structural alignment was centred on WDR82.

4.2.2.2 Prediction of the WDR82-ZC3H4 structure

ZC3H4 is a large protein (140kDa, 1304 amino acids) that is predicted from sequence to be almost entirely unstructured other than an array of three C3H1-type zinc finger domains. The previously mapped WDR82-binding region is an approximately 200-amino acid stretch that lies C-terminal to the zinc fingers and contains no annotated structural features or sequence biases (Figure 4.5A) (Austenaa et al., 2021). To predict the structure of the ZC3H4-WDR82 complex, I used this region of ZC3H4 as the query sequence for ColabFold alongside full-length WDR82. The depth of coverage in the ZC3H4$^{831-1062}$ sequence alignment is well above the 100 sequence minimum required for a good prediction (Figure 4.5B). Interestingly, the ZC3H4 alignment contains fewer sequences with low identity to the query than the WDR82 or SET1A alignments. This may reflect a more evolutionarily recent protein, which has fewer divergent homologs, or may be due to the challenges of aligning sequences with little structure and low complexity. Remarkably, ColabFold was able to

confidently predict the structure of a small WDR82-binding region in ZC3H4. A region of approximately 70 amino acids was predicted with high confidence, whilst the surrounding regions were predicted with pLDDT of around 20, strongly suggesting an intrinsically disordered conformation (Figure 4.5C). The PAE plot (Figure 4.5D) shows low error in the predicted alignment of this region relative to WDR82, and hence high confidence in the predicted interaction.

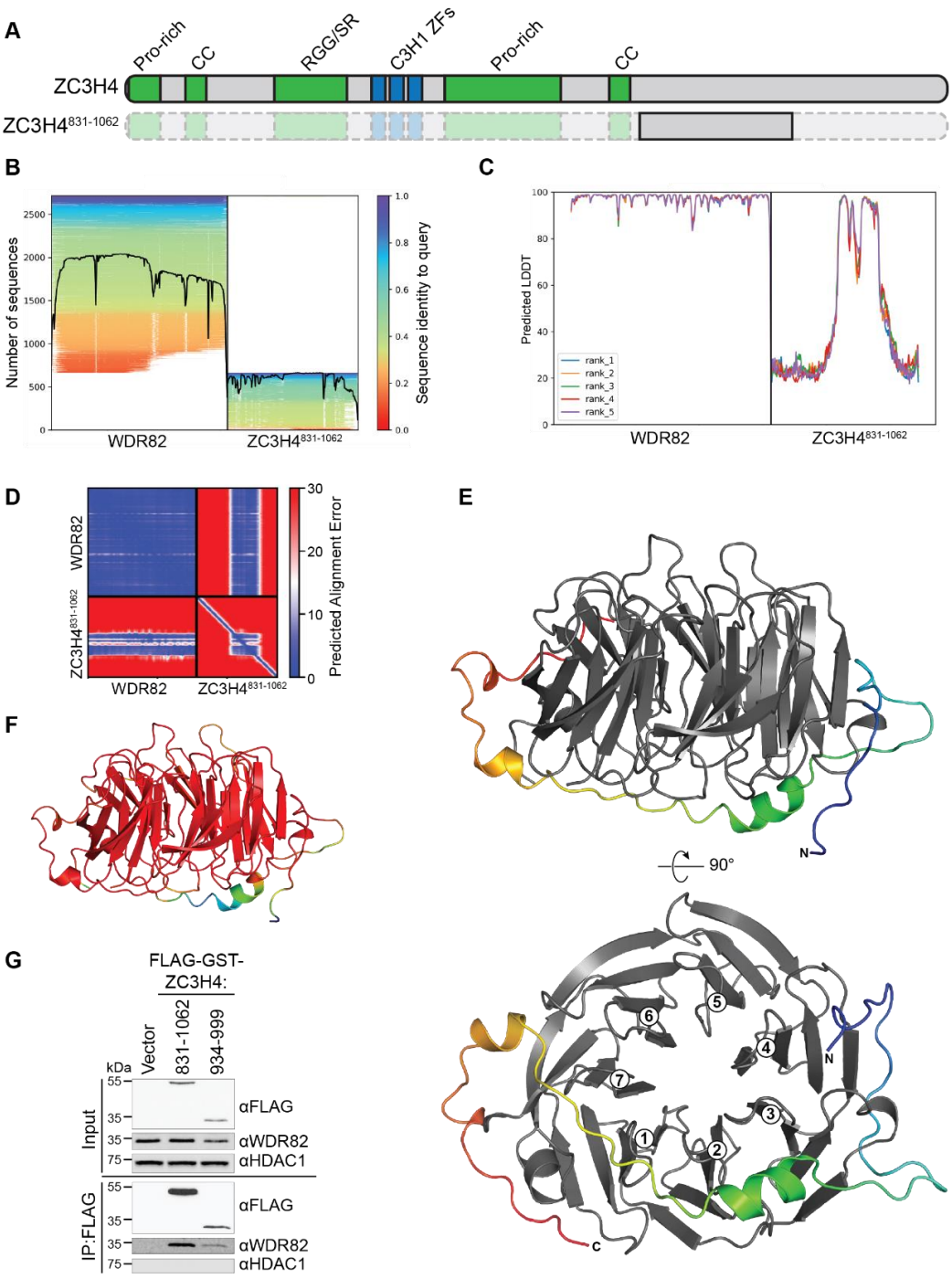


Figure 4.5 Prediction of the WDR82-ZC3H4 complex structure. **A**: Schematic of ZC3H4 indicating the region used as input for sequence prediction. Zinc finger domains are in blue and other sequence features in green. CC: coiled coil, RGG-SR: RGG repeat/serine-rich. **B**: Graphical representation of the depth and diversity of the MSA generated by ColabFold structure prediction for WDR82 and ZC3H4$^{831-1062}$. **C**: Graph showing the per-position pLDDT for the 5 predicted structures of the WDR82-ZC3H4$^{831-1062}$ complex generated by ColabFold. The 'rank 1' structure was used for all further analysis. **D**: Plot of Predicted Alignment Error for the rank 1 WDR82-ZC3H4$^{831-1062}$ structure. **E**: Side and bottom views of a cartoon representation of the rank 1 predicted WDR82-ZC3H4 structure generated by ColabFold, showing only the region of ZC3H4 predicted with high confidence (residues 934-999, pLDDT>80 except for internal loop). WDR82 is in grey, ZC3H4$^{934-999}$ is coloured rainbow from N-terminus (blue) to C-terminus (red). Numbers indicate the blades of WDR82. **F**: Side view of a

cartoon representation of the rank 1 predicted WDR82-ZC3H4 structure coloured by pLDDT on a spectrum from red (high) to blue (low), showing the same region as in E. **G**: Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing FLAG-GST-tagged fragments of ZC3H4. HDAC1 is a loading control for inputs and negative control for IP experiments.

The predicted interface (Figure 4.5E) follows a similar path to SET1A on the surface of WDR82, contacting the side of blade 4 and the loops of blades 3, 2, 1, 7 and 6 across the bottom face of the β propeller, burying a surface area of 2454Å^2 (18.6% of total WDR82 surface area). Visualisation of the predicted structure coloured by pLDDT demonstrates that the majority of the interacting region is predicted with high confidence (red), whilst the sections of lower confidence (blue, green) are limited to loop regions which do not contact WDR82 (Figure 4.5F). I was able to validate this interaction with IP experiments, where the ZC3H4$^{934-999}$ fragment was sufficient to bind WDR82, consistent with the predicted structure (Figure 4.5G). This domain of ZC3H4 is of very similar overall sequence conservation and complexity to the surrounding regions and hence would be difficult to predict from manual inspection of sequence alignments. This result therefore reveals a remarkably powerful ability of structure prediction to precisely define protein-protein interaction interfaces that would be otherwise difficult to identify without extensive mapping or mutagenesis experiments.

4.2.2.3 Prediction of the WDR82-PNUTS structure

The region of PNUTS previously found to bind WDR82 is a 200 amino acid section C-terminal to the PP1 binding motif which, similarly to the ZC3H4 WDR82-interacting region, lacks any annotated structural or sequence features (Figure 4.6A) (Lee et al., 2010). In order to predict the structure of PNUTS in complex with WDR82, I input this sequence alongside full length WDR82 into ColabFold. The depth of coverage for the PNUTS$^{418-620}$ MSA was again well above the 100 sequence minimum for a good prediction, with a wide range of similarities to the query sequence. A region of approximately 80 amino acids towards the

N-terminus of the PNUTS query sequence was predicted to bind WDR82 with high local confidence and low PAE (Figure 4.6C, D). The more C-terminal region of the PNUTS fragment was predicted with very low pLDDT (around 20), suggesting an intrinsically disordered conformation. Interestingly, unlike the SET1A and ZC3H4 structures, which were generally predicted with either very high or very low local confidence, the predicted structure of WDR82-PNUTS complex included a region at the N-terminus that was predicted with only medium confidence (pLDDT between 50 and 80 depending on model ranking) (Figure 4.6C). This region corresponds to a long α helix that runs across the top face of the WDR82 propeller; however, the PAE of this region is relatively high, indicating some uncertainty in its position (Figure 4.6D, E). Indeed, the position of this helix relative to WDR82 varies across the five ranked models provided by ColabFold (Figure 4.6F). The alpha-helical backbone conformation of this segment is likely to be correct, however due to the uncertainty in its position relative to the rest of the WDR82-PNUTS complex, I excluded it from my analysis of the complex.

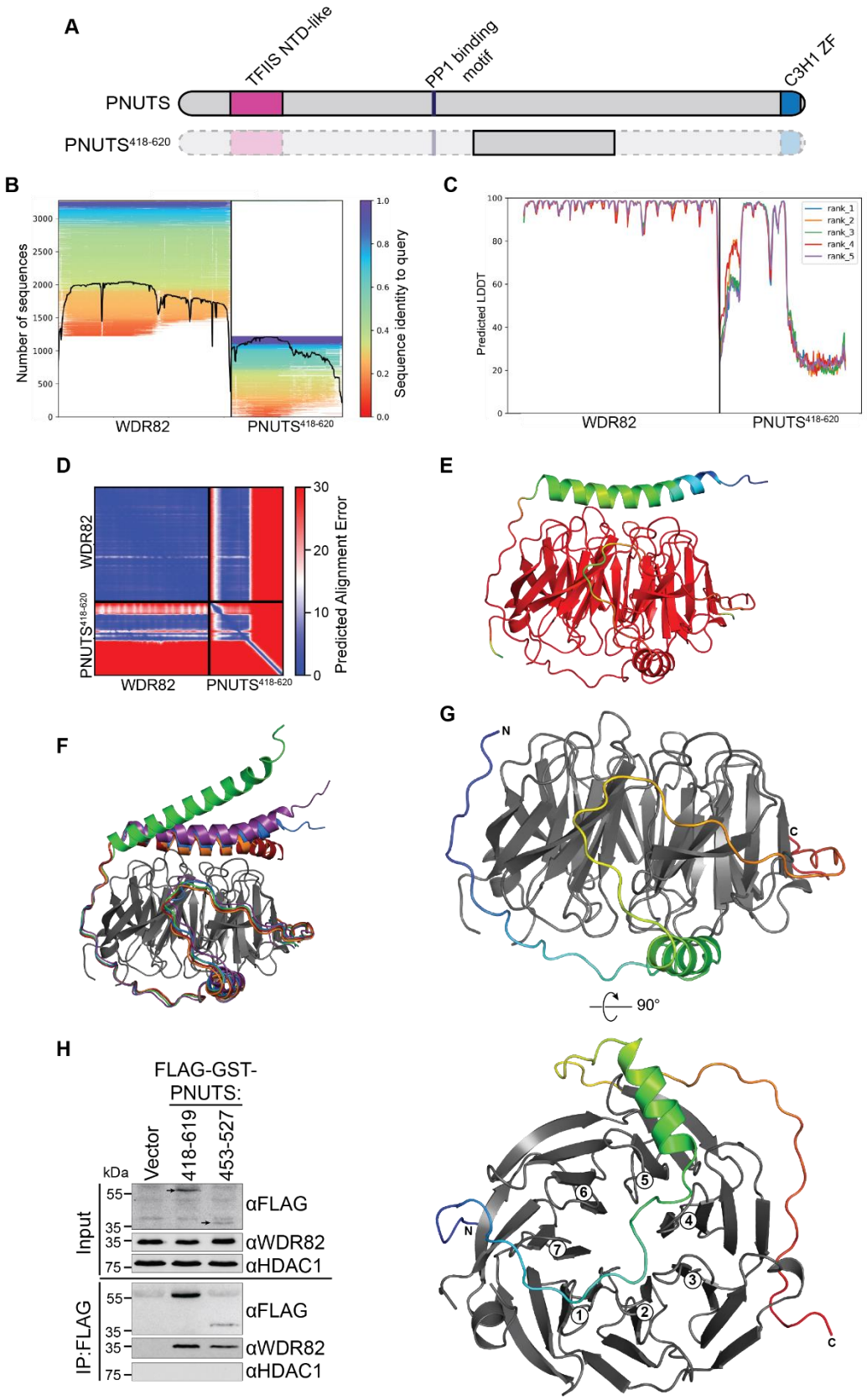


Figure 4.6 Prediction of the WDR82-PNUTS complex structure. **A**: Schematic representation of the domain organisation of PNUTS indicating the region used as input for structure prediction. **B**: Graphical representation of the depth and diversity of the MSA generated by ColabFold structure prediction for WDR82 and PNUTS$^{418-620}$. **C**: Graph showing the per-position pLDDT for the 5 predicted structures of the WDR82-PNUTS$^{418-620}$ complex generated by ColabFold. The 'rank 1' structure was

used for all further analysis. **D**: Plot of Predicted Alignment Error for the rank 1 WDR82-PNUTS$^{418-620}$ structure. **E**: Side view of a cartoon representation of the rank 1 predicted WDR82-PNUTS structure coloured by pLDDT on a spectrum from red (high) to blue (low). For clarity, the C-terminal region predicted with very low confidence (pLDDT<50, after residue 527) is not shown. **F**: Overlay of the 5 structures of WDR82-PNUTS complex produced by ColabFold, coloured by rank as in C. For clarity, only one structure of WDR82 is shown, and C-terminal regions of PNUTS (after residue 527) have been removed. **G**: Side and bottom views of a cartoon representation of the rank 1 predicted WDR82-PNUTS structure generated by ColabFold, showing only the region of PNUTS predicted with high confidence (residues 453-527, pLDDT>80 except for internal loop). WDR82 is in grey and PNUTS$^{453-527}$ is coloured rainbow from N-terminus (blue) to C-terminus (red). Numbers indicate the blades of WDR82. **H**: Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing FLAG-GST-tagged fragments of PNUTS. HDAC1 is a loading control for inputs and negative control for IP experiments. Arrows indicate PNUTS fragment bands in the input samples.

PNUTS is predicted to bind WDR82 via an interface that differs somewhat from that of SET1A and ZC3H4. The interacting region adopts a largely extended conformation which makes extensive contacts with WDR82, burying 3164Å^2 (23.4%) of the WDR82 surface area (Figure 4.6G). The polypeptide chain runs across the bottom face of the β propeller in the opposite direction to that predicted for SET1A and ZC3H4, starting in the groove between blades 6 and 7, then contacting the loops of blades 7, 1 and 2 before diverting across the central axis and over blades 4 and 5 in a short helix. The more C-terminal region of the interacting sequence is then predicted to wrap around the circumference of WDR82, contacting the sides of blades 5, 4 and 3. IP experiments show that this minimal region of PNUTS is indeed sufficient to bind WDR82 (Figure 4.6H), once again demonstrating the remarkable ability of ColabFold to predict small interaction domains within larger input sequences.

4.2.3 Comparison and validation of WDR82 binding motifs

Comparison of the predicted structures of SET1A, ZC3H4, and PNUTS in complex with WDR82 reveals both similarities and difference between their modes of binding. Whilst ZC3H4 and PNUTS are almost exclusively in an extended conformation with some regions of alpha helical secondary structure, the SET1A NTD binds WDR82 via an extended region

combined with a small three-stranded antiparallel β sheet (Figure 4.7A). All three proteins have interfaces across the bottom face of the WDR82 β propeller, as well as on the circumference around blade 4. The paths of SET1A and ZC3H4 across the face of WDR82 are very similar, whereas PNUTS loops around to cover more of the WDR82 surface in a distinct manner. Interestingly, when viewed in the same orientation, the polypeptide chains of SET1A and ZC3H4 run in the opposite direction across the face of WDR82 compared to that of PNUTS (Figure 4.7B, right to left versus left to right, respectively).

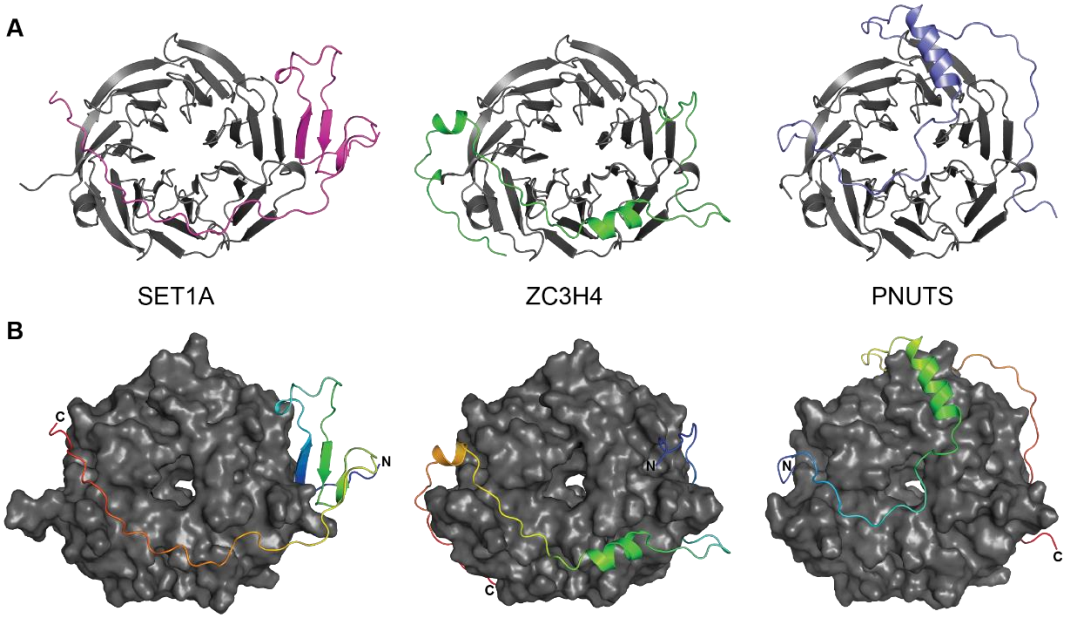


Figure 4.7 Comparison of SET1A, ZC3H4, and PNUTS in complex with WDR82. **A**: Cartoon representations of the predicted structures of WDR82 (grey) in complex with SET1A NTD (residues 1-84, magenta), ZC3H4$^{934-999}$ (green), and PNUTS$^{453-527}$ (slate). **B**: Complexes as for A with WDR82 shown in surface representation (grey) and SET1A, ZC3H4, and PNUTS shown as cartoons coloured from N terminus (blue) to C-terminus (red)

The regions of all three proteins that directly contact WDR82 were predicted with very high local confidence (pLDDT >90), and very low PAE, indicating an overall excellent level of confidence in the predicted interfaces. In addition, all predictions were subject to AMBER molecular dynamics to relax side chain conformations. I am confident, therefore, that these predictions are of sufficient quality to characterise the specific residue-level contacts which

mediate complex formation. Remarkably, despite no known evolutionary relationships, SET1A, ZC3H4, and PNUTS display a number of shared modes of interaction with WDR82.

4.2.3.1 SET1A, ZC3H4 and PNUTS share a hydrophobic anchor motif

Alignment of all three complex structures reveals a shared interface in a groove on the bottom face of WDR82 created by blades 6, 7 and 1 of the β propeller (Figure 4.8A). Closer inspection of the SET1A, ZC3H4 and PNUTS sequences in this region reveals a remarkable set of shared interactions in which conserved anchor leucine, proline and large hydrophobic (phenylalanine/tryptophan) residues dock into hydrophobic pockets on the surface of WDR82 (Figure 4.8B, C). Overlay of all three structures shows how the side chains of these 'hydrophobic motifs' occupy highly similar positions, despite the fact that the polypeptide chain of PNUTS runs in the opposite direction to that of SET1A and ZC3H4 (Figure 4.8D). Closer examination of each complex structure reveals both shared and subtly different sets of contacts which accommodate binding of these motifs into the same surface crevice on WDR82.

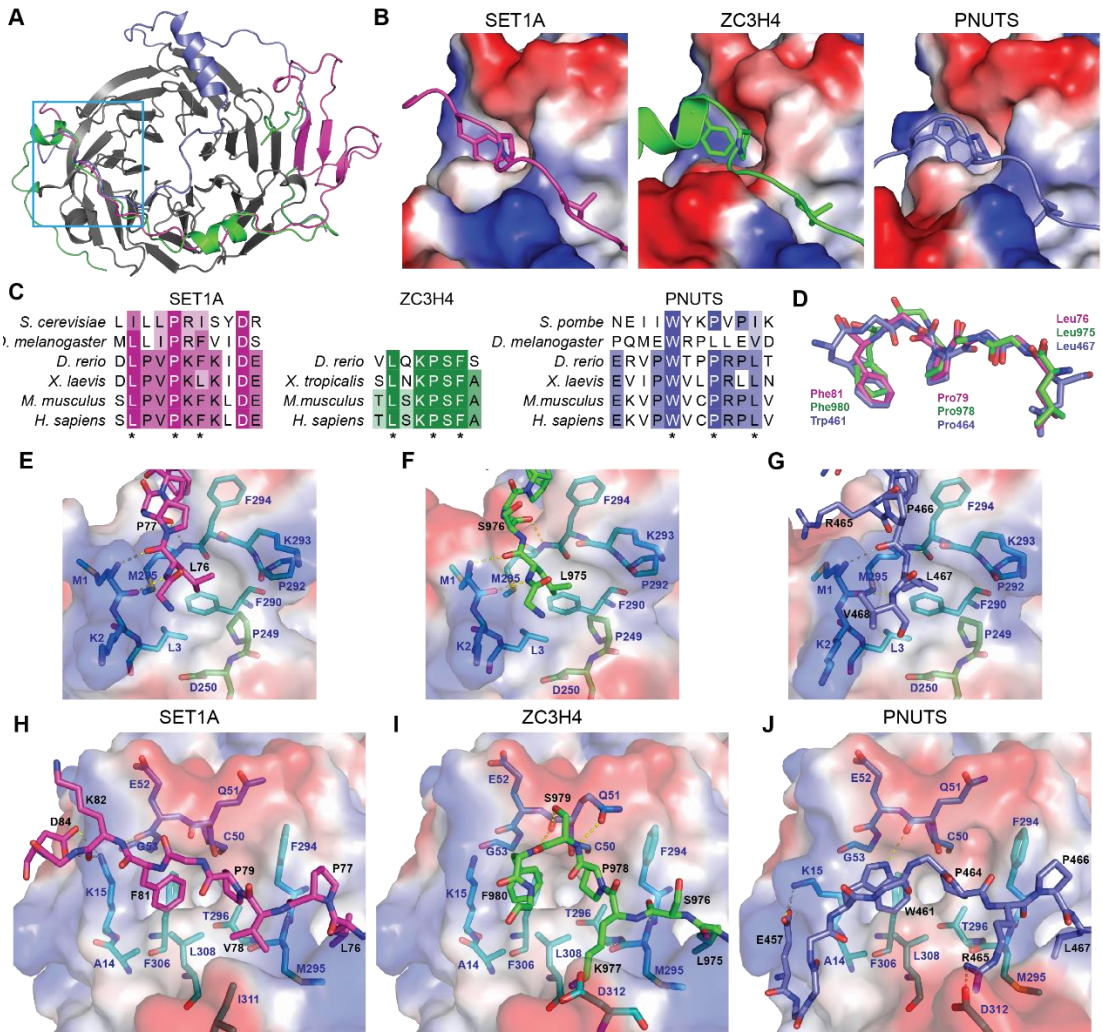


Figure 4.8 SET1A, ZC3H4 and PNUTS share a hydrophobic motif. **A**: Cartoon representation of the overlaid structures of WDR82 in complex with SET1A (magenta), ZC3H4 (green) and PNUTS (slate). For clarity, only one WDR82 molecule is shown. The location of the shared interface is indicated by a blue box. **B**: Close-up views of the hydrophobic motif side chains of SET1A (magenta), ZC3H4 (green), and PNUTS (slate) binding into a groove on the surface of WDR82. **C**: Multiple sequence alignments of the hydrophobic motif sequences of SET1A (magenta), ZC3H4 (green), and PNUTS (slate), shaded by conservation. Asterisks below indicate the motif 'anchor' residues shown in B. **D**: Overlaid stick representation structures of the hydrophobic motifs of SET1A (magenta), ZC3H4 (green), and PNUTS (slate). Relative positions are based on alignment of overall complex structures. For clarity, only the side chains of the 'anchor' residues are shown. **E**: Details of SET1A (magenta) leucine anchor interactions with WDR82. Residues from WDR82 blade 1 are shown in forest, blade 7 in teal. WDR82 residues are labelled in navy, SET1A residues in black. **F**: Details of ZC3H4 (green) leucine anchor interactions with WDR82. Residues from WDR82 blade 1 are shown in forest, blade 7 in teal. WDR82 residues are labelled in navy, ZC3H4 residues in black. **G**: Details of PNUTS (slate) leucine anchor interactions with WDR82. Residues from WDR82 blade 1 are shown in forest, blade 7 in teal. WDR82 residues are labelled in navy, PNUTS residues in black. **H**: Details of SET1A (magenta) proline/hydrophobic anchor interactions with WDR82. Residues from WDR82 blade 7 are shown in teal, blade 6 in marine. WDR82 residues are labelled in navy, SET1A residues in black. For clarity, some SET1A side chains are not shown. **I**: Details of ZC3H4 (green) proline/hydrophobic anchor interactions with WDR82. Residues from WDR82 blade 7 are shown in teal, blade 6 in marine. WDR82 residues are labelled in navy, ZC3H4 residues in black. For clarity, some ZC3H4 side chains

are not shown. **J**: Details of PNUTS (slate) proline/hydrophobic anchor interactions with WDR82. Residues from WDR82 blade 7 are shown in teal, blade 6 in marine. WDR82 residues are labelled in navy, PNUTS residues in black. For clarity, some PNUTS side chains are not shown.

Firstly, each protein has a leucine residue which contacts a hydrophobic patch between blades 6 and 7 of WDR82. Hydrogen bonds between backbone groups serve to stabilise this interaction. For SET1A and ZC3H4, the backbone of this 'anchor' leucine residue (Leu76 and Leu975, respectively) forms a pair of hydrogen bonds with the backbone groups of WDR82 Met1 (Figure 4.8E, F). The peptide backbone of PNUTS runs in the opposite direction to that of SET1A and ZC3H4, meaning these corresponding backbone hydrogen bond donors and acceptors do not line up. Instead, PNUTS utilises the backbone carbonyl of WDR82 Lys293, on the opposite side of the hydrophobic pocket, as a hydrogen bond acceptor to the backbone NH of the Leu467 anchor, whilst the Val468 and Pro466 residues on either side of the Leu467 anchor form hydrogen bonds with the backbone of WDR82 Met1. This alternative pattern of backbone hydrogen bonds serves to position the PNUTS Leu467 anchor side chain into the same hydrophobic pocket on WDR82 as used by SET1A and ZC3H4 (Figure 4.8G).

The path of all three proteins extends across the top of blade 7 of the β propeller, contacting the loop between strands A and B. For SET1A and ZC3H4 this interaction is stabilised by a hydrogen bond between the backbone carbonyl of the residue C-terminal to the leucine anchor (Pro77 and Ser976, respectively) and the backbone NH of WDR82 Met295. Because the backbone of PNUTS runs in the opposite direction to that of SET1A and ZC3H4, it is the backbone carbonyl of Arg465, two residues N-terminal to the anchor leucine, which is positioned as the equivalent hydrogen bond acceptor (Figure 4.8E, F, G).

SET1A, ZC3H4 and PNUTS also bind into a deep hydrophobic pocket in the surface of WDR82 situated between blades 6 and 7 of the β propeller (Figure 4.8A, B). Whilst all three proteins anchor a proline and a large hydrophobic side chain into this pocket, the precise

interactions which stabilise binding of this motif vary, reflecting their different sequence contexts. In the SET1A complex, binding of the Pro79 and Phe81 side chain anchors is stabilised by interactions involving adjacent residues on either side. The side chain of Val78 packs against WDR82 Ile311, whilst residues following the hydrophobic anchors also help to stabilise the motif via backbone hydrogen bonds. In addition, a salt bridge interaction between the side chain of the highly conserved Asp84 and the side chain of WDR82 Lys15 further anchors SET1A (Figure 4.8H).

The structure of WDR82 in the ZC3H4-bound state has a few subtle differences to that of the SET1A complex, which accommodate a slightly different backbone conformation and set of adjacent side chains. Whilst SET1A contacts WDR82 on both sides of the hydrophobic anchor residues, the ZC3H4 backbone forms a short alpha helix and kinks away from the surface of WDR82 after the Phe980 anchor, limiting further interactions with residues that follow the hydrophobic motif (Figure 4.8B). Instead, the conformation of the ZC3H4 backbone allows it to make additional polar contacts within the hydrophobic motif sequence. WDR82 Asp312 projects towards the binding pocket in place of Ile311, forming a salt bridge interaction with the side chain of ZC3H4 Lys977. ZC3H4 Ser979, which lies between the anchor Pro978 and Phe980 residues, forms a pair of hydrogen bonds with WDR82 Gln51, which flips in towards the binding pocket compared to its position in the SET1A-bound structure. The short helix in ZC3H4 also forces the backbone of the hydrophobic anchor residue Phe980 to adopt a subtly different conformation to the equivalent SET1A residue, which positions its backbone NH group to make a hydrogen bond with the WDR82 backbone (Figure 4.8I). Here, despite very similar hydrophobic motifs, SET1A and ZC3H4 make subtly different contacts with WDR82 based on surrounding sequences.

The binding of PNUTS into this hydrophobic pocket again exploits a slightly different set of interactions due to differences in sequence and opposite backbone direction. Like ZC3H4, PNUTS uses a basic residue to make a salt bridge interaction with WDR82 Asp312 (Figure 4.8J). However, unlike SET1A and ZC3H4, in which the anchoring proline and phenylalanine residues are separated by only one amino acid, there are two residues between the Pro464 and Trp461 anchors found in PNUTS. This allows the backbone sufficient conformational flexibility for the tryptophan side chain to insert into the hydrophobic pocket, where it makes a hydrogen bond to the backbone carbonyl of WDR82 Gln51. The WDR82 Lys15 side chain is flipped outwards relative to its confirmation in the SET1A/ZC3H4-bound structures, stabilised by a salt bridge interaction with the side chain of PNUTS Glu457. This slightly expands the hydrophobic pocket, providing sufficient space for the larger tryptophan side chain employed by PNUTS.

This hydrophobic/leucine anchor motif is well conserved within each protein (Figure 4.8C), however differences in surrounding sequences affect how each motif docks onto the surface of WDR82. The conformational changes required in WDR82 to accommodate these differences are minimal and limited to changes in side chain conformation. This demonstrates how intrinsically unstructured regions allow for a greater degree of conformational plasticity than folded domains to mediate binding of different sequences to the same site on a rigid scaffold (Oldfield and Dunker, 2014). It is a remarkable discovery that these three evolutionarily unrelated proteins have converged on the same set of binding interactions, and this overlapping binding interface for all three proteins is likely to be a key determinant of their mutual exclusivity.

4.2.3.2 SET1A and ZC3H4 share a DPR motif

N-terminal to the hydrophobic motif, SET1A and ZC3H4 follow a similar path across the bottom face of WDR82, both binding via a conserved DPR motif into a deep pocket

between blades 2 and 3 of WDR82 (Figure 4.9A, B, C). The DPR motif is stabilised by an internal network of hydrogen bonds between the aspartic acid side chain and the arginine backbone NH and side chain, as well as the restricted backbone conformation of the central proline residue (Figure 4.9D). Whilst the conformation of the DPR motif is very similar in both SET1A and ZC3H4, the interaction predicted with WDR82 is slightly different for each. The pocket in WDR82 is formed by side chains from blade 2 (yellow) and from blade 3 (orange) (Figure 4.9E, F). In the predicted WDR82-SET1A structure, the side chain of SET1A Arg64 makes two hydrogen bonds with the backbone carbonyl of WDR82 Ala223 (Figure 4.9E). In addition, the relative positions of the SET1A Arg62 and WDR82 Phe224 side chains are such that they could form a pi-cation stacking interaction (Figure 4.9G, magenta). However, the structure of the WDR82-ZC3H4 suggests an alternative set of interactions between the DPR motif and WDR82 (Figure 4.9F). The ZC3H4 DPR motif is positioned slightly deeper into the pocket, and the Arg960 side chain forms a more extensive network of hydrogen bonds with the backbone carbonyls of WDR82 Ala223 and Gly206, and the side chain of Asp156. The relative positions of the Arg960 side chain and the aromatic ring of Phe224 are less favourable for a pi-cation stacking interaction (Figure 4.9G, green). These two structures present two potential binding modes for the DPR motif in the same pocket on WDR82. It is unclear whether each is correct for the specific protein, or whether one mode in particular is the true mode of interaction. Experimental structure determination may shed further light on this, however *in vivo* the binding may be a dynamic combination of these modes. Nevertheless, structure prediction reveals a second WDR82 binding motif shared by SET1A and ZC3H4, which is highly conserved in both proteins.

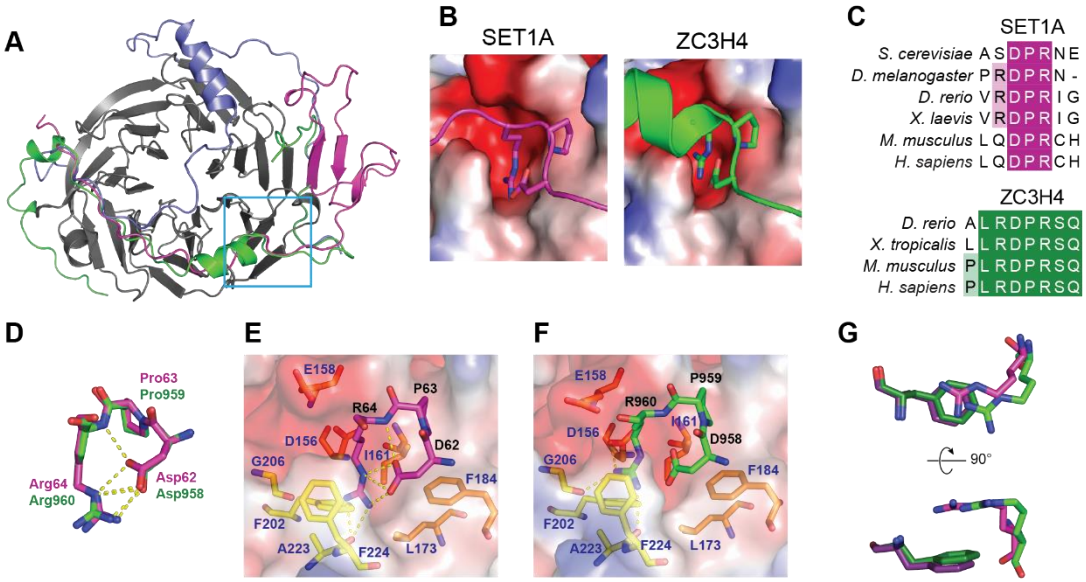


Figure 4.9 SET1A and ZC3H4 share a DPR motif. **A**: Cartoon representation of the overlaid structures of WDR82 in complex with SET1A (magenta), ZC3H4 (green), and PNUTS (slate). For clarity, only one WDR82 molecule is shown. The location of the DPR motif is indicated by a blue box. **B**: Close-up views of the DPR motif side chains of SET1A (magenta) and ZC3H4 (green) binding into a pocket on the surface of WDR82. **C**: Multiple sequence alignments of the DPR motif sequences of SET1A (magenta) and ZC3H4 (green), shaded by conservation. **D**: Aligned stick representation structures of the DPR motifs of SET1A (magenta) and ZC3H4 (green), showing intra-motif polar contacts. **E**: Details of SET1A (magenta) DPR motif interactions with WDR82. Residues from WDR82 blade 2 are shown in yellow, blade 3 in orange. WDR82 residues are labelled in navy, SET1A residues in black. **F**: Details of ZC3H4 (green) DPR motif interactions with WDR82. Residues from WDR82 blade 2 are shown in yellow, blade 3 in orange. WDR82 residues are labelled in navy, ZC3H4 residues in black. **G**: Comparison of the relative positions of WDR82 Phe224 and the DPR motif Arginine side chains in the WDR82-SET1A structure (violet/magenta) and the WDR82-ZC3H4 structure (forest/green).

4.2.3.3 SET1A, ZC3H4 and PNUTS bind the circumference of WDR82

SET1A, ZC3H4, and PNUTS are all predicted to interact with the circumference of WDR82 around blade 4 (Figure 4.10A). The N-terminal β sheet of SET1A extends the sheet of blade 4, whilst ZC3H4 and PNUTS bind in an extended conformation between blades 3 and 4 via a shared 'IPLD' motif (Figure 4.10B). A short helix in the loop between strands C and D of blade 3 is one of the few deviations of WDR82 from an 'ideal' β propeller and adopts slightly different conformations to allow binding via these different modes.

Despite very different interfaces across much of WDR82, ZC3H4 and PNUTS both contain a conserved IPLD motif which is predicted to bind to the circumference of WDR82

between blades 3 and 4 (Figure 4.10C, D). The isoleucine side chain (I946/I520) inserts into a deep pocket formed by side chains from WDR82 blade 4 (red) and blade 3 (orange). In ZC3H4, the adjacent Asn945 side chain also makes a number of hydrogen bonds to WDR82. The proline residue (P947/P521) of the IPLD motif restricts the backbone conformation to position the side chain of the neighbouring leucine residue (L948/L522) against the surface of WDR82. Finally, the motif aspartic acid side chain (D949/D523) makes a salt bridge to the side chain of WDR82 Lys172. The motif is also stabilised throughout by backbone hydrogen bonding to side chains and the backbone of WDR82. Once again, these two unrelated proteins display a remarkable similarity in their mode of binding to WDR82.

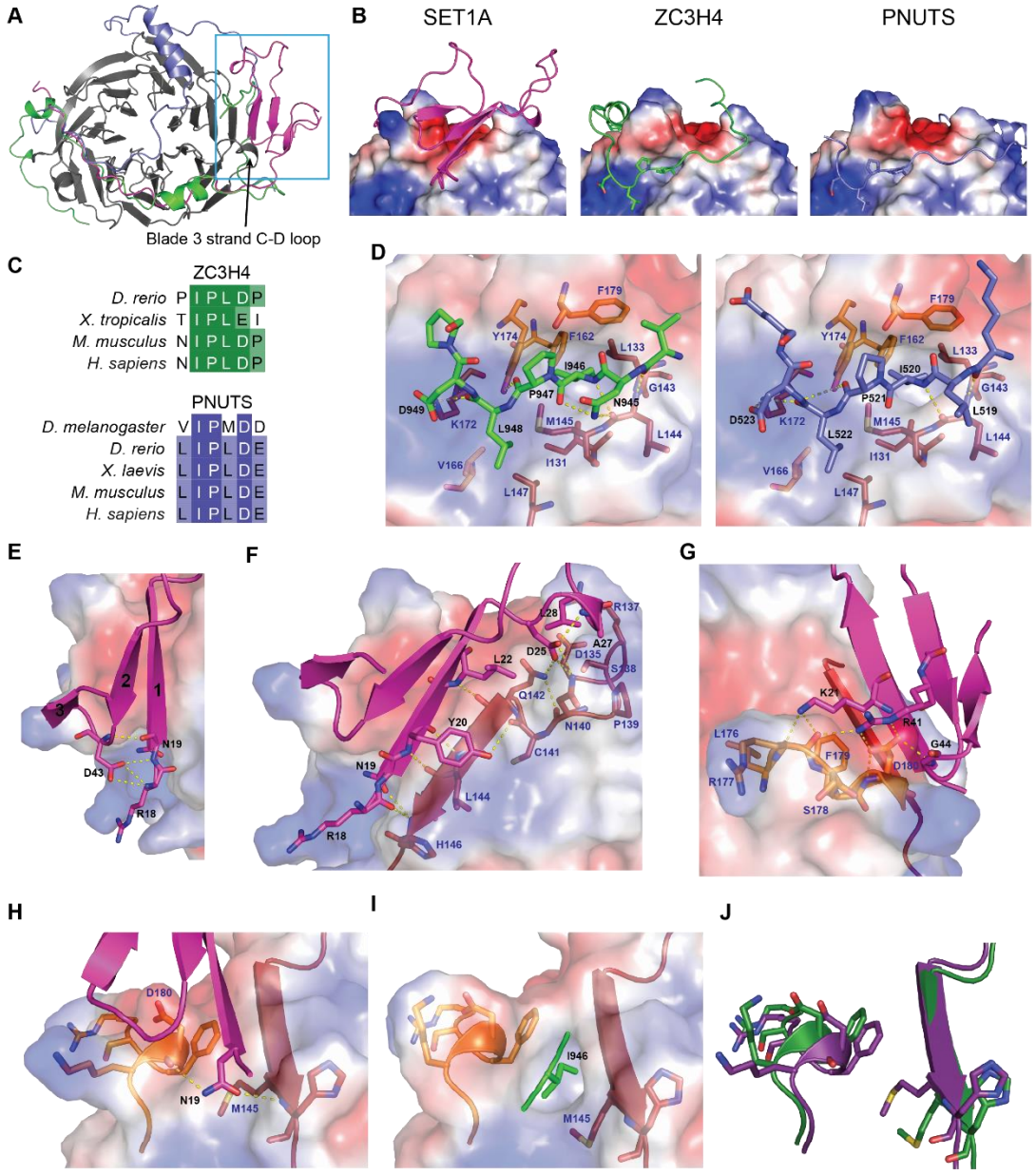


Figure 4.10 SET1A, ZC3H4 and PNUTS contact WDR82 blade 4. **A**: Cartoon representation of the overlaid structures of WDR82 in complex with SET1A (magenta), ZC3H4 (green) and PNUTS (slate). For clarity, only one WDR82 molecule is shown. The location of the overlapping interface is indicated by a blue box. **B**: Close-up views of the SET1A β sheet (magenta), and the ZC3H4 (green) and PNUTS (slate) IPLD motifs on the surface of WDR82. **C**: Multiple sequence alignments of the IPLD motif sequences of ZC3H4 (green) and PNUTS (slate), shaded by conservation. **D**: Details of ZC3H4 (left, green) and PNUTS (right, slate) IPLD motif interactions with WDR82. Residues from WDR82 blade 3 are shown in orange, blade 4 in firebrick. WDR82 residues are labelled in navy, ZC3H4/PNUTS residues in black. **E**: Detail of the SET1A β sheet showing hydrogen bonds made by Asp43. Strands of the sheet are labelled 1 to 3 from N- to C- terminal. **F**: Details of SET1A (magenta) β sheet side chain interactions with blade 4 of WDR82 (firebrick). WDR82 residues are labelled in navy, SET1A residues in black. **G**: Details of SET1A (magenta) β sheet side chain interactions with the WDR82 blade 3 strand C-D loop (orange). WDR82 residues are labelled in navy, SET1A residues in black. **H**: SET1A Asn19 forms a pair of hydrogen bonds which bridge between WDR82 blade 4 (firebrick) and the blade 3 C-D loop (orange). **I**: WDR82 blade 4 (firebrick) and the blade 3 C-D loop (orange) adopt

different conformations to accommodate ZC3H4 Ile946 (green). **J**: Overlay of the structures of WDR82 blade 4 strand D and blade 3 C-D loop from the SET1A-bound (violet) and ZC3H4-bound (forest) structures.

SET1A binds in this region via a very different and more extensive interface, forming a three stranded antiparallel β sheet extension to the β sheet of WDR82 blade 4. The SET1A β sheet itself is stabilised at the base by a network of contacts between Asp43, which lies in the strand 2-3 loop, and the backbone of Arg18 and Asn19 at the base of the first strand (Figure 4.10E). In addition to the β sheet backbone hydrogen bonding, the interaction between SET1A and WDR82 is mediated by an extensive network of side chain interactions. SET1A Asn19 anchors the base of the first β strand by hydrogen bonding to the backbone NH of WDR82 His146, whilst SET1A Tyr20 hydrogen binds to the backbone carbonyl of WDR82 Cys141. The side chain of Leu22 packs against a small hydrophobic patch on the WDR82 surface. The long loop between strands 1 and 2 of the SET1A β sheet forms a short helix which also packs against the surface of WDR82, with the side chain of Asp25 forming a network of hydrogen bonds with WDR82 Asn140 and Gln142. SET1A Leu28 also makes contact with the WDR82 surface in a dip created by WDR82 Asp135 and Arg137, further stabilising the position of the SET1A strand 1-2 loop (Figure 4.10F). The opposite face of the SET1A β sheet packs against the short helix in the strand C-D loop of WDR82 blade 3, stabilised by a network of hydrogen bonds involving the side chains of SET1A Lys21 and Arg41, the side chain of WDR82 Asp180, and various backbone groups from both proteins (Figure 4.10G).

Interestingly, the Asn19 side chain at the base of the SET1A β sheet sits in a small dip between WDR82 blade 4 and the insertion helix in the blade 3 strand C-D loop, in the same position as the anchor isoleucine in the ZC3H4/PNUTS IPLD motifs (Figure 4.10H, I). Subtle conformational differences in WDR82 accommodate this different mode of binding in the same position. Most notably, in the ZC3H4-bound structure, the position of the blade 3 loop C-D helix is shifted away from blade 4, opening up a pocket for the anchor isoleucine

side chain to dock into. The side chain of WDR82 Met145 is also positioned differently, further opening up the cleft (Figure 4.10J). By contrast, this pocket in the SET1A-bound structure is occluded by the side chain of Met145, and the side chain of SET1A Asn19 forms a pair of 'bridging' hydrogen bonds to the backbone of WDR82 His146 and Asp180, which stabilise the closer conformation of the WDR82 blade 3 strand C-D loop helix.

These observations reveal how different proteins can use either similar or very different motifs to occupy shared interfaces on WDR82, which again dictate mutually exclusive binding. Despite a rigid WD40 domain scaffold, subtle plasticity in the conformation of WDR82 may allow it to accommodate these different binding modes.

4.2.3.4 Equivalent mutations in SET1A, ZC3H4 and PNUTS have different effects

Having identified a number of conserved amino acid motifs which seem to mediate specific binding of SET1A, ZC3H4, and PNUTS to WDR82, I next wanted to test whether these predicted interactions are indeed required for binding *in vivo*. I therefore generated SET1A, ZC3H4 and PNUTS expression constructs carrying specific amino acid mutations in the binding motifs I previously identified, and tested their ability to coIP WDR82. Cloning mutations into full-length SET1A and ZC3H4 was technically challenging due to their large size and regions of highly repetitive sequence, so I chose to use smaller fragments that I have previously shown to be sufficient to bind WDR82. Despite their shared features in binding to WDR82, mutations to equivalent motif residues in SET1A, ZC3H4, and PNUTS had different effects.

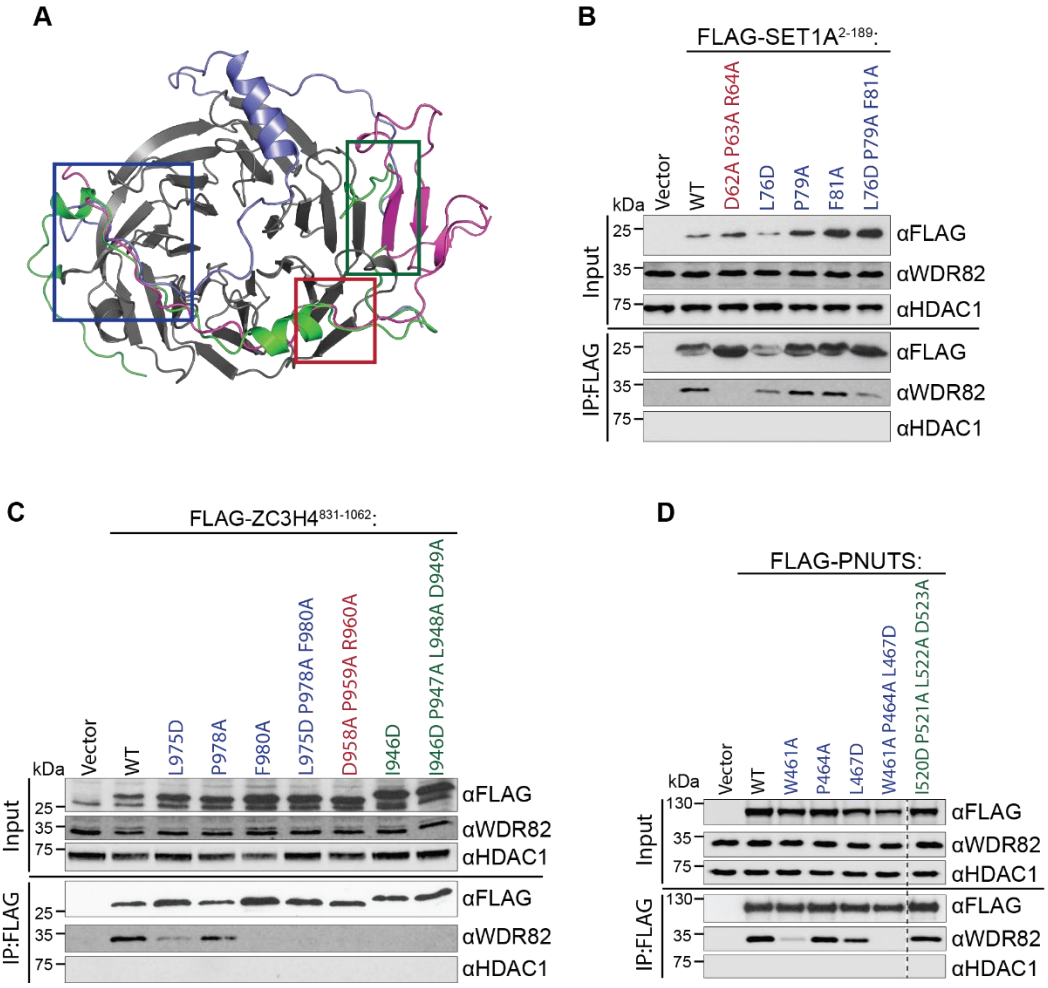


Figure 4.11 Validation of WDR82-binding mutants. **A**: Cartoon representation of the overlaid structures of WDR82 in complex with SET1A (magenta), ZC3H4 (green) and PNUTS (slate). For clarity, only one WDR82 molecule is shown. The locations of mutated interfaces are indicated by coloured boxes. **B:** Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing FLAG-SET1A$^{2-189}$ carrying the indicated mutations, coloured by location as indicated in A. HDAC1 is a loading control for inputs and negative control for IP experiments. **C**: Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing FLAG-ZC3H4$^{831-1062}$ carrying the indicated mutations, coloured by location as indicated in A. HDAC1 is a loading control for inputs and negative control for IP experiments. **D**: Western blot of input and FLAG IP samples from 293T cells transiently transfected with plasmids expressing FLAG-PNUTS carrying the indicated mutations, coloured by location as indicated in A. HDAC1 is a loading control for inputs and negative control for IP experiments. A dashed line indicates where the blot has been cropped for clarity.

Mutation of the conserved DPR motif completely eliminated the interaction of SET1A$^{1-189}$ with WDR82, whereas mutation of Leu76 or of the entire hydrophobic motif (L76D P79A F81A) only reduced the interaction (Figure 4.11B). SET1A makes extensive contacts with WDR82 via its N-terminal β sheet and the DPR motif, so it is perhaps unsurprising that

disruption of the most C-terminal binding motif in the NTD has little effect on overall WDR82 binding.

In contrast, binding of ZC3H4$^{831-1062}$ to WDR82 was much more sensitive to mutations. Mutation of the hydrophobic anchor Phe980, the full hydrophobic motif (L975D P978A F980A), the DPR motif, the Ile946 anchor alone, or the full IPLD motif all eliminated interaction with WDR82 (Figure 4.11C). This may reflect an interaction mode based on multiple relatively low affinity 'hotspots' across an extended interface which is sufficiently destabilised by loss of one interaction that other motifs cannot maintain binding well enough to capture by coIP.

Unlike for ZC3H4, the IPLD motif of PNUTS is dispensable for binding to WDR82, as are a number of other residues I tested (Figure 4.11D). However, mutation of the hydrophobic anchor residue Trp461 greatly reduced binding to WDR82, and mutation of the full hydrophobic motif (W461A P464A L467D) completely abrogated the interaction. The PNUTS interaction with WDR82 is the most extensive of the three proteins (3164 Å^2 buried surface area), making numerous specific contacts. With such an extensive set of interactions single mutations are perhaps less likely to affect binding, however mutation of an extended hotspot such as the hydrophobic motif is sufficient to disrupt the interaction with WDR82.

Together, these results suggest that the interactions predicted by ColabFold are likely to be accurate, as mutations identified from the predicted structures were able to disrupt binding to WDR82.

4.2.4 Towards crystallisation of WDR82-containing complexes

In the absence of experimental structural data, AlphaFold has proved to be a remarkably powerful tool for analysis of the interactions between WDR82 and SET1A,

ZC3H4, and PNUTS. I was able to experimentally validate many of the features observed in the predicted structures, however I also sought to experimentally determine the structures of recombinant WDR82-containing complexes.

Initial attempts to express WDR82 in *E. coli* yielded only insoluble protein in inclusion bodies, so I turned to insect cells as a recombinant expression system more suited to eukaryotic proteins. Using the MultiBac expression system (Bieniossek et al., 2012) and Sf9 cells I was able to produce sufficient quantities of protein for crystallisation. WDR82 was expressed either itself carrying a C-terminal Twin-Strep tag, or untagged and coexpressed with a Twin-Strep-tagged binding partner. Strep affinity purification from cell lysates was followed by tag cleavage using TEV protease and clean up by size exclusion chromatography (Figure A.2). Purified proteins were used to set up small-scale crystallisation trials with commercial screens. Figure 4.12A shows a coomassie-stained SDS-PAGE gel of a selection of the WDR82-SET1A/B complexes used for crystallisation. A summary of all screens and proteins trialled is listed in Table A.2. Initial trials using a complex of WDR82 with SET1A$^{8-189}$ yielded some poor-quality crystals which diffracted to low resolution at best (Figure 4.12B). I used these conditions as a starting point for optimisation experiments to improve crystal quality with this construct, however these failed to improve the crystals (Figure 4.12C). I also tested a variety of changes to the SET1A construct, including removal of the RRM, further truncation of the N-terminal tail, and trimming internal loops, but these all failed to improve crystallisation (Figure 4.12D). Trials with equivalent constructs of SET1B also did not yield high-quality crystals. I also trialled a complex of WDR82 with ZC3H4$^{934-999}$, which did not yield any crystals under conditions tested. Similarly, attempts to crystallise WDR82 alone were also unsuccessful.

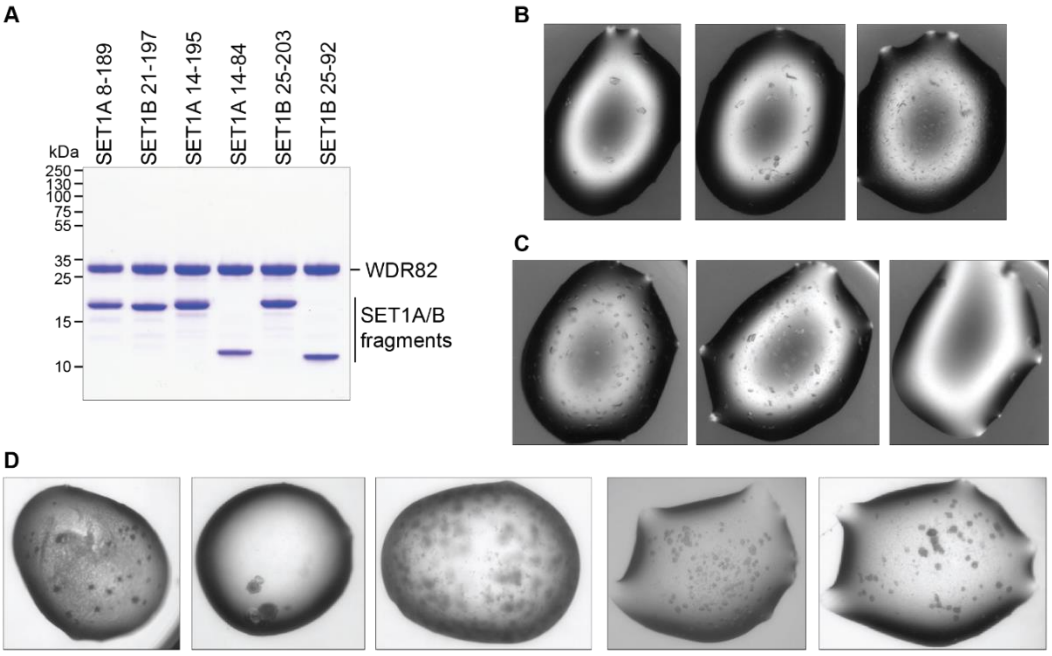


Figure 4.12 Crystallisation of WDR82-containing complexes. **A**: Coomassie-stained SDS-PAGE gel of selected WDR82-SET1A/B complexes used for crystallisation trials. **B**: Example best crystals from initial WDR82-SET1A$^{8-189}$ crystallisation trials. All are from the Morpheus screen (Plate 4), from left to right: drop D6, F6, G6. **C**: Example best crystals from WDR82-SET1A$^{8-189}$ optimisation tests. From left to right: Plate 7 drop B6, Plate 10 drop H11, Plate 10 drop H6. **D**: Example best crystals from further WDR82 complex crystallisation trails. From right to left: Plate 17 drop E10.10, Plate 27 drop B11.00, Plate 27 drop B11.10, Plate 41 drop F1.10, Plate 41 drop F6.00.

These results suggest that WDR82 itself may be generally difficult to crystallise. Further trials under different conditions may be able to grow suitable quality crystals for diffraction, however this was beyond the scope of this project. Given the power of AlphaFold to predict the structures of these relatively small protein complexes, future energies might be better spent in cryo-EM experiments to determine the structure of the larger complexes of which WDR82 is a part and hence understand its role in these wider contexts.

4.3 Summary & discussion

WD40 domains are highly abundant eukaryotic protein domains which act as scaffolds for protein-protein interactions to underpin a wide variety of cellular processes (Schapira et al., 2017; Stirnimann et al., 2010; Xu and Min, 2011). WDR82 is an essential WD40-repeat protein which is a constitutive component of multiple transcription regulatory complexes, and its removal results in wide-ranging changes to transcription (Austenaa et al., 2021,

2015; Bi et al., 2011; Brewer-Jensen et al., 2016; Lee et al., 2010; van Nuland et al., 2013)(Amy Hughes, unpublished data). Whilst previous studies have mapped general regions of SET1A, ZC3H4 and PNUTS that bind WDR82 (Austenaa et al., 2021; Lee and Skalnik, 2008; Lee et al., 2010; Park et al., 2022), a detailed interrogation of these interactions has been lacking. Here I set out to understand the molecular basis of SET1A, ZC3H4, and PNUTS binding to WDR82, in particular to understand how these interactions are mutually exclusive.

Having mapped the WDR82 binding site in SET1A to the 84 amino acid NTD, and using previously mapped regions of ZC3H4 and PNUTS, I took advantage of recent developments in protein structure prediction to interrogate the molecular basis of their interactions with WDR82. The ColabFold implementation of AlphaFold proved to be both easy to use and incredibly powerful, predicting the structure of all three complexes with high confidence within hours. Whilst I was unable to grow crystals and solve the structures experimentally to directly assess the predictions, I was able to experimentally validate the predicted interacting regions. The predicted complex structures revealed that SET1A, ZC3H4 and PNUTS all form extensive contacts with the surface of WDR82, primarily across the bottom face of the β propeller, as well as on the circumference around blade 4. All three proteins are predicted to have several shared interfaces, which dictate that their binding is mutually exclusive. Remarkably, I was able to identify amino acids motifs shared between the WDR82-binding proteins, despite no known evolutionary relationships. The binding modalities of SET1A and ZC3H4 are particularly similar, taking similar paths across the face of WDR82 and employing almost identical hydrophobic and DPR amino acid motifs. The binding of PNUTS to WDR82 is much less similar than SET1A and ZC3H4, with the polypeptide chain running in the opposite direction across the WDR82 surface and covering a larger area. Despite this, PNUTS employs a very similar set of interactions to SET1A and ZC3H4 to anchor a hydrophobic motif in the WDR82 blade 1/7/6 groove. Importantly, I was

able to show that some of the motifs I identified from the predicted structures are indeed required for WDR82 binding. Identification of mutations within each protein that specifically disrupt binding to WDR82 are a powerful tool for further investigation of WDR82 function within each complex. Removal of WDR82 itself allows us to examine its overall role in regulating transcription, however any effects will be a summation of its functions in all contexts. Designing mutations in WDR82 that only disrupt its interaction with one partner are challenging due to the overlapping interfaces of all three proteins. Hence, mutations which disrupt WDR82 incorporation into specific complexes whilst leaving other complexes and WDR82 itself unaffected are a key tool for dissecting the context-specific roles of WDR82.

Until recently, structural studies of WD40 proteins in complexes relied primarily upon crystallography, biasing the literature towards interactors which are amenable to crystallisation, such as small peptide motifs or well-structured domains, rather than larger extended interfaces such as those predicted for SET1A, ZC3H4 and PNUTS. However, the recent proliferation of cryo-EM structures of large multiprotein complexes has begun to reveal the diversity of interactions made by WD40 proteins. For example, extensions of the β sheet of a propeller blade, such as that predicted for SET1A, are rare in solved structures, however structures of the PRC2 complex reveal that SUZ12 forms a 2-stranded antiparallel β sheet extension to blade 7 of the RBBP4 WD40 domain (Figure 4.13A) (Grau et al., 2021). Importantly, structures of these large complexes reveal key roles for WD40 domains as protein-protein interaction scaffolds which bind multiple proteins simultaneously via different interfaces. For example, the structure of the yeast SET1 H3K4 HMT complex reveals how SWD3 acts as a core scaffold component by simultaneously binding the SET1 Win motif via the central pocket on its top face, SWD1 via an extended interface on the bottom face of the β propeller, and SPP1 via the circumference around blades 1 and 2 (Figure 4.13B) (Hsu et al., 2019; Qu et al., 2018). Whilst SWD3 is a core structural

component of the SET1 complex, structures of the PRC2 complex demonstrate how WD40 proteins can also confer additional binding specificities which are important for complex function. The EED component of the PRC2 complex is constitutively anchored into the complex via extensive interactions with other subunits. Importantly, these interactions leave the top face of the EED β propeller exposed. This surface binds more transiently to the H3K27me3 histone modification deposited by PRC2, allosterically activating its catalytic activity and facilitating propagation of repressive chromatin domains (Figure 4.13C) (Chammas et al., 2020; Margueron et al., 2009).

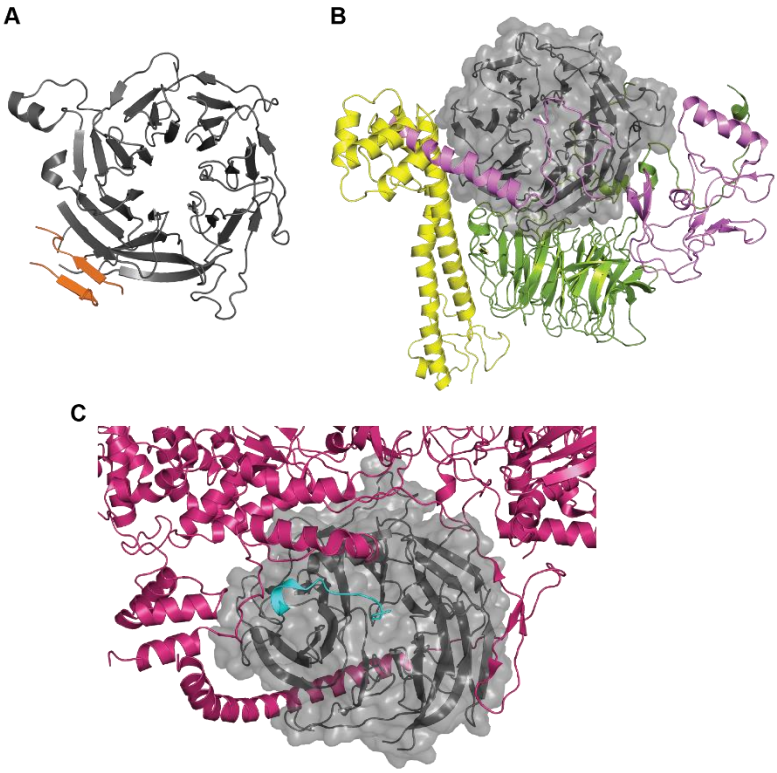


Figure 4.13 WD40 domains are scaffolds for protein-protein interactions. **A**: SUZ12 (orange) forms a two-stranded β sheet extension to blade 7 of the RBBP4 WD40 domain (grey). Structure from 7KSO (Grau et al., 2021). **B**: SWD3 (grey) simultaneously binds SET1 (violet), SWD1 (pea green), and SPP1 (yellow). Modelled from alignment of 6BX3 and 6UGM (Hsu et al., 2019; Qu et al., 2018). **C**: EED (grey) is anchored into the PRC2 complex (warm pink) by extensive contacts which leave its top face free to bind histone H3K27me3 (cyan). Structure from 6WKR (Kasinath et al., 2021).

Whilst SET1A, ZC3H4, and PNUTS are predicted to make extensive contacts across the surface of WDR82, they leave significant interaction surfaces exposed. Notably, they do not

occlude the 'canonical' peptide binding site on the top face of the WDR82 β propeller, raising the possibility that WDR82 simultaneously binds another protein or proteins via this site. Previous work has shown that WDR82 can bind to the heptapeptide repeats of the RNAPII CTD when they are phosphorylated on serine 5 (S5P), providing a direct link between WDR82 and the transcription machinery (Bae et al., 2020; Ebmeier et al., 2017; Lee and Skalnik, 2008; Park et al., 2022). PTM-specific binding by WD40 domains is well characterised in a number of contexts, including Serine/Threonine phosphorylation-specific binding by a number of SCF complexes (Xu and Min, 2011). For example, FBW7 binds Cyclin E only when it is doubly phosphorylated on Thr380 and Ser384 (Hao et al., 2007; Koepp, 2001). Given that this top face of the WDR82 β propeller is left exposed when in complex with SET1A, ZC3H4, and PNUTS, one could envisage a mechanism by which WDR82 links the activity of these complexes to transcription by simultaneously binding phosphorylated RNAPII CTD.

Interestingly, it has been shown that SET1A and ZC3H4 interact with the RNAPII CTD specifically when it is phosphorylated, whereas PNUTS interacts with both phosphorylated and unphosphorylated CTD (Ebmeier et al., 2017; Lee and Skalnik, 2008; Park et al., 2022). This suggests either a difference in specificity for WDR82 when in complex with PNUTS, or additional factors which contribute to PNUTS binding to RNAPII. My structure predictions suggest the top face of WDR82 may be occluded by a long helix of PNUTS, although its relative position was not confidently predicted. This may represent a mechanism by which PNUTS alters the additional binding specificities of WDR82 to provide context-specific functionality. The interaction between WDR82 and S5P CTD is reportedly enhanced by both SET1A and ZC3H4 (Lee and Skalnik, 2008; Park et al., 2022), whilst the specific effects of PNUTS on CTD binding by WDR82 have not been characterised. In the remainder of my thesis, I will examine the molecular function of WDR82 in the context of SET1A, ZC3H4, and PNUTS complexes.

5 Investigating the molecular function of WDR82 in different contexts

5.1 Introduction

The progression of RNAPII through initiation, elongation and termination is accompanied by mRNA processing events such as capping and splicing, and requires numerous regulatory factors. The extended C-terminal domain (CTD) of the largest RNAPII subunit, RBP1, plays a central role in coordinating these activities (Harlen and Churchman, 2017). The YSPTSPS heptapeptide repeats of the RBP1 CTD can be post-translationally modified in a number of ways, most notably by phosphorylation of the non-proline residues (Kim et al., 2009). These different phosphorylation states occur in distinctive patterns across genes and have been implicated in regulating various aspects of transcription and mRNA processing (Cossa et al., 2021; Parua and Fisher, 2020). For example, S5P is associated with transcription initiation and recruits the mRNA capping enzymes, whilst S2P is predominant during elongation and peaks towards the 3' end of genes, where it recruits the 3' end processing machinery (reviewed in Harlen and Churchman, 2017).

Interestingly, it has been shown *in vitro* that WDR82 binds specifically to S5P CTD, and this interaction is enhanced by SET1A and ZC3H4 (Lee and Skalnik, 2008; Park et al., 2022). However, it is not known whether WDR82 mediates CTD binding in the context of the PNUTS complex. Mass spectrometry experiments to identify specific interactors of phosphorylated CTD identified WDR82 as well as the SET1A/B, ZC3H4 and PNUTS complexes, suggesting WDR82 may mediate S5P CTD binding in all three contexts (Ebmeier et al., 2017). This 'adapter' function of WDR82 could serve to specifically localise SET1A/B,

ZC3H4 and PNUTS complexes to initiating RNAPII and allow for its regulation. Interestingly, however, the PNUTS-PP1 complex was also shown to bind unphosphorylated CTD (Ebmeier et al., 2017), suggesting PNUTS may confer a different binding specificity to WDR82 or has alternate mechanisms for binding the RNAPII CTD .

The deposition and removal of RNAPII CTD phosphorylation involves a number of kinases and phosphatases (Cossa et al., 2021; Parua and Fisher, 2020). Notably, the PNUTS-PP1 complex has been shown to dephosphorylate S5P CTD *in vitro* (Carminati et al., 2022; Lee et al., 2010). The PP1 phosphatase has low inherent substrate selectivity and as such must be highly regulated to specify its substrates and limit deleterious aberrant activity (Bollen et al., 2010; Rebelo et al., 2015). The cell achieves this through numerous PP1 interacting proteins (PIPs) such as PNUTS, which are present in sufficiently high concentration to ensure that there is no free PP1 enzyme present in cells (Bollen et al., 2010; Rebelo et al., 2015; Verbinnen et al., 2017). Many PIPs act as general PP1 inhibitors that only release their inhibition in response to specific stimuli, such as their own phosphorylation, or allow binding of only very specific substrates by blocking or extending the substrate-binding surfaces of PP1 (Heroes et al., 2013). PNUTS has been shown to have inhibitory activity towards PP1, however the mechanisms which relieve this inhibition are unknown (Allen et al., 1998; Choy et al., 2014; Kim et al., 2003; Kreivi et al., 1997). WDR82 is not required for PNUTS-PP1 to dephosphorylate RNAPII CTD *in vitro*, however it has been suggested to be required for PNUTS-PP1 regulation *in vivo* (Landsverk et al., 2020).

In Chapter 4 I interrogated the molecular basis of WDR82 binding to SET1A, ZC3H4, and PNUTS and discovered that the 'canonical' top face binding surface of the WDR82 WD40 domain may remain exposed in each of these complexes. This surface could therefore bind another protein, allowing WDR82 to act as a specific adapter module to link each complex to additional proteins such as the RNAPII CTD. I therefore wanted to understand how

incorporation within each complex affects WDR82 binding to the RNAPII CTD, and how this binding behaviour could underpin its function within each complex.

5.2 Results

5.2.1 Investigating CTD binding by WDR82

It has been previously shown that the SET1A/B, ZC3H4, and PNUTS complexes can interact with phosphorylated CTD, however it is not known whether WDR82 mediates CTD binding in all three contexts (Ebmeier et al., 2017). *In vitro* binding assays have shown that WDR82 can bind specifically to S5P CTD and this interaction may be enhanced by SET1A and ZC3H4, however the CTD-binding behaviour of WDR82 in complex with PNUTS has not been examined (Lee and Skalnik, 2008; Park et al., 2022). I therefore wanted to characterise and compare the CTD-binding behaviour of WDR82 in complex with SET1A/B, ZC3H4, and PNUTS. To do this, I performed an *in vitro* pulldown experiment using 4-repeat CTD peptides that were either unphosphorylated or specifically phosphorylated on either serine 2 (S2P) or serine 5 (S5P) of each repeat. Complexes of WDR82 with minimal fragments of SET1A, ZC3H4, and PNUTS that I identified in Chapter 4 were expressed in Sf9 cells and purified via FLAG affinity (Figure 5.1A). Purified proteins were then incubated with CTD peptides that had been coupled to magnetic streptavidin beads. Following washing, bound proteins were eluted by boiling in SDS-PAGE loading dye and analysed by western blot (Figure 5.1B). WDR82 alone bound specifically to S5P peptides, however the signal was only slightly above background, suggesting this interaction may be relatively weak. Consistent with previous reports, WDR82 in complex with SET1A showed higher signal in the S5P pulldown compared to WDR82 alone, suggesting some cooperativity between SET1A and WDR82 that enhances binding to the CTD. The equivalent complex of WDR82 with SET1B showed similar behaviour. Importantly, the SET1A$^{8\text{-}189}$ fragment alone did not bind any of

the peptides at a level detectable above background (indicated by the 'no peptide' control sample), strongly suggesting WDR82 is required for SET1A$^{8-189}$ to associate with the RNAPII CTD. Interestingly, the complexes of WDR82 with the minimal WDR82-binding fragments of ZC3H4 or PNUTS that I identified in Chapter 4 did not bind to any of the CTD peptides under these conditions. This is contrary to previous observations that full-length ZC3H4 enhances WDR82 binding to S5P CTD, suggesting additional regions of ZC3H4 beyond the minimal WDR82-binding fragment I have tested may be required for this behaviour. My size exclusion chromatography data (Figure 3.3) suggests ZC3H4 may form a homo-oligomer, perhaps via coiled-coil regions that were not present in fragment used for these CTD binding assays. Oligomerisation of ZC3H4 could enhance CTD binding by providing multiple binding sites and hence increasing the avidity of the interaction. My data suggest that WDR82 does not bind CTD when in complex with a minimal fragment of PNUTS, however it may be that the affinity was too low or the PNUTS fragment used did not allow for sufficient binding to detect in this assay. Further experiments to quantify the affinity of these interactions and potential oligomerisation of ZC3H4 will be important to fully validate this behaviour. Nevertheless, these results suggest that the individual binding partners of WDR82 do affect its CTD-binding behaviour.

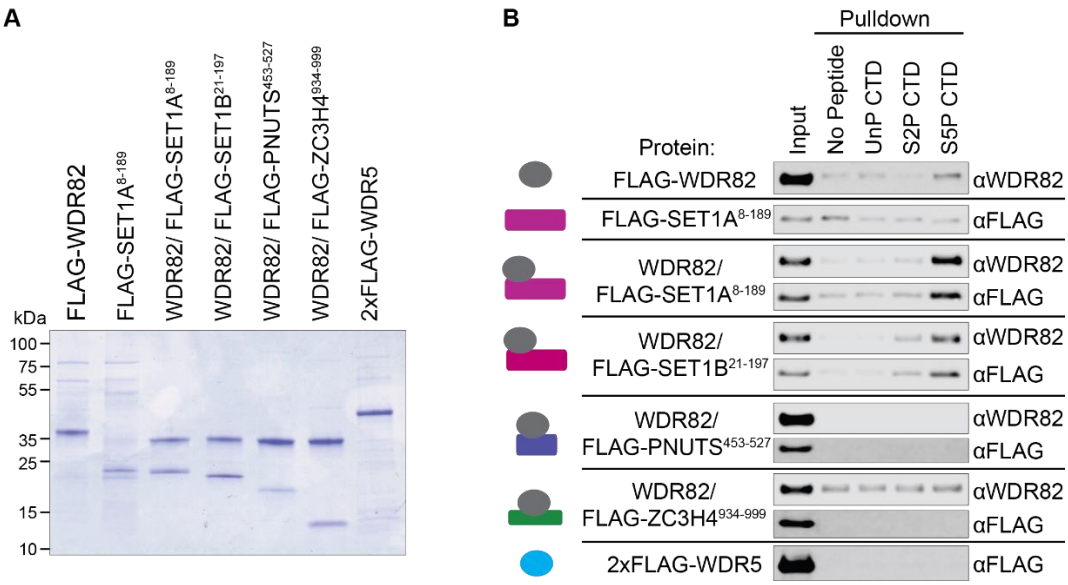


Figure 5.1 *In vitro* CTD binding assay. **A:** Coomassie stained SDS-PAGE gel of recombinant proteins used for CTD binding assays. **B:** Western blot of 2% input and protein pulldown with 4-repeat biotinylated CTD peptides as indicated. This is a representative example of at least two independent replicates for each protein.

To further characterise the CTD-binding behaviour of WDR82 in the context of each complex, I wanted to test whether binding to WDR82 is required for SET1A, ZC3H4, and PNUTS to interact with RNAPII in cells. I therefore made use of the WDR82-binding mutants of each protein that I identified in Chapter 4. Transient expression in 293T cells followed by immunoprecipitation showed that the NTD of SET1A was sufficient to co-IP S5P RNAPII, and mutation of SET1A to disrupt WDR82 binding abrogated the interaction with S5P (Figure 5.2A). This result shows that interaction with WDR82 is required for SET1A binding to S5P CTD.

I was unable to detect any CTD binding *in vitro* with a minimal WDR82-ZC3H4$^{934-999}$ complex (Figure 5.1) and this fragment did not detectably CoIP S5P from 293T cells (Figure 5.2B). The GST tag included to increase the size of this fragment forms dimers, suggesting that dimerisation of this minimal WDR82-ZC3H4 complex is insufficient for CTD binding and additional regions of ZC3H4 may be required for a detectable interaction. Consistent with this possibility, immunoprecipitation of a larger fragment of ZC3H4 (residues 831-1062) did display elevated S5P RNAPII signal over background (Figure 5.2B). I attempted to test the

effect of WDR82 binding mutations on the ability of a ZC3H4$^{831-1062}$ fragment to coIP RNAPII, however I did not detect any S5P CTD pulldown with the WT protein in this experiment (data not shown). In this case, the ZC3H4$^{831-1062}$ fragment used carried only a FLAG tag rather than the FLAG-GST used for the experiment in 2B, suggesting that the GST-induced dimerisation of these fragments may be required for detectable CTD binding. Together, these data suggest that both oligomerisation of ZC3H4 and the presence of additional regions beyond the minimal WDR82-binding fragment may be important for interaction with the CTD. As an alternative strategy to test whether WDR82 is required for ZC3H4 to interact with RNAPII, I aimed to immunoprecipitate ZC3H4 from ESCs in which WDR82 had been removed and to examine whether ZC3H4 could still interact with S5 RNAPII. However, test experiments showed that IP efficiency using an antibody specific to ZC3H4 was too low to detect any RNAPII signal above background levels (Figure 5.2C). Without a clear readout for ZC3H4 interaction with RNAPII CTD, I could not verify whether WDR82 was required for any such interaction. Published data shows that ZC3H4 alone cannot bind to CTD peptides *in vitro*, suggesting any interaction with RNAPII is via WDR82 (Park et al., 2022). Further experiments will be required to test this possibility more robustly *in vivo*.

In contrast, IP of PNUTS demonstrated a strong interaction with S5P RNAPII and, interestingly, mutation of PNUTS to eliminate WDR82 binding had no effect on this interaction, indicating that PNUTS binding to RNAPII is at least partially independent of WDR82. This is could be related to substrate CTD binding by PP1, however the affinity of these enzyme-substrate interactions is generally low to allow for efficient substrate turnover. Alternatively, PNUTS could bind RNAPII via the PAF1 complex, which I identified in Chapter 3 as an interactor of PNUTS. The PAF1 complex makes extensive contacts with the core RNAPII enzyme and so could mediate PNUTS binding to RNAPII independently of the CTD (Vos et al., 2018a).

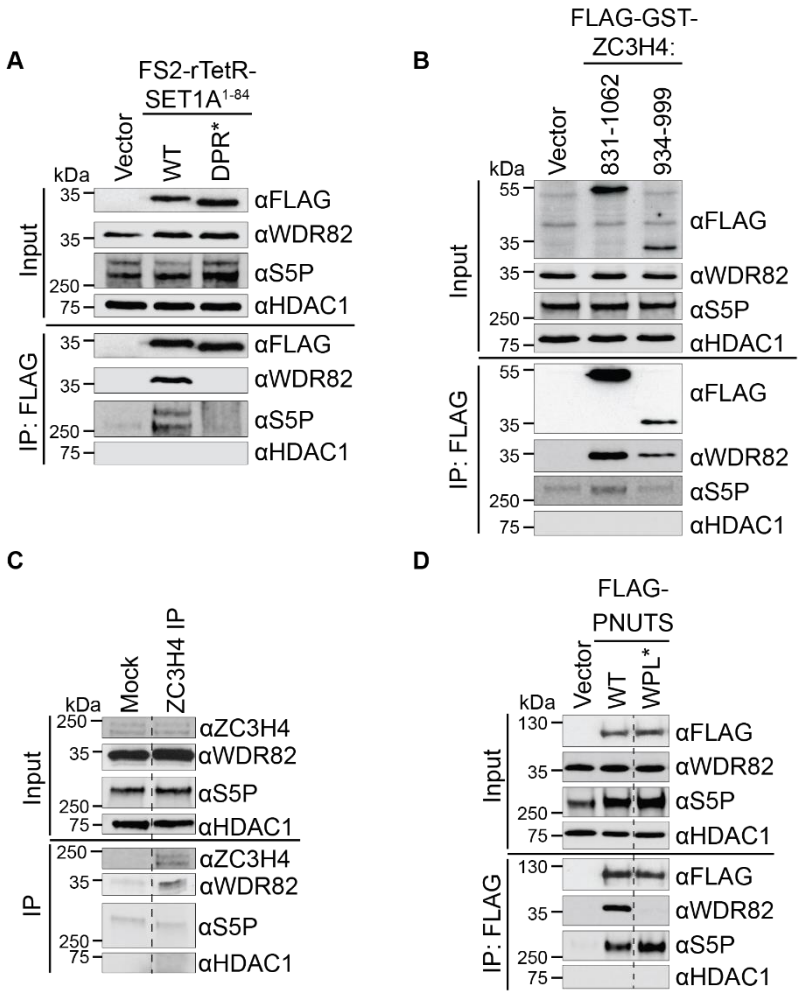


Figure 5.2 Investigating RNAPII binding of WDR82-containing complexes in cells. **A:** Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing either wild-type (WT) or WDR82-binding mutant (DPR*, D62A P63A R64A) SET1A NTD. HDAC1 is a loading control for inputs and negative control for IP experiments. **B:** Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing FLAG-GST-tagged fragments of ZC3H4. HDAC1 is a loading control for inputs and negative control for IP experiments. **C:** Western blot of IP from E14 nuclear extract using antibody specific to ZC3H4. 'Mock' sample was a control IP using anti-FLAG antibody. Dashed line indicates where the blot has been cropped for clarity. **D:** Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing either WT or WDR82-binding mutant (WPL*, W461A P464A L467D) FLAG-tagged PNUTS. HDAC1 is a loading control for inputs and negative control for IP experiments.

Together, these results suggest context-specific functions for WDR82. In the SET1A and possibly ZC3H4 complexes, WDR82 acts as an adapter module to specifically bind S5P CTD, whereas this function was not detectable in PNUTS complex, which has an WDR82-independent mode of interaction with RNAPII.

5.2.2 WDR82 binds PNUTS close to PP1

Having shown that WDR82 in complex with PNUTS does not bind strongly to CTD peptides *in vitro* and PNUTS does not require WDR82 for interaction with RNAPII *in vivo*, I wanted to understand what the role of WDR82 could be in this complex. The PNUTS-PP1 complex has been shown to dephosphorylate the RNAPII CTD and, interestingly, the region of PNUTS which I have shown binds WDR82 is close in linear sequence to the previously identified PP1 binding motif (Figure 5.3A) (Lee et al., 2010). Given the close proximity of the PP1 and WDR82 binding sites on PNUTS, I wanted to understand whether WDR82 could act to regulate PP1 activity.

To investigate this possibility, I used ColabFold to predict the structure of the ternary complex of PP1α, WDR82, and a PNUTS fragment that encompasses the previously identified binding sites for both proteins (Figure 5.3A). The MSAs for all three proteins had excellent coverage (Figure 5.3B) and the structure of the complex was predicted with generally high confidence (Figure 5.3C). The single folded domains of both WDR82 and PP1α were predicted with pLDDT almost entirely above 90 except for the very C-terminal portion of PP1α, which was predicted to be disordered. The PP1 binding motif of PNUTS was predicted with high confidence, as was the WDR82-binding region identified from my previous prediction of the WDR82-PNUTS dimer structure (Figure 5.3C). Interestingly a central region of the PNUTS fragment, which corresponds to the long helix discussed in Section 4.2.2.3, was again predicted with only medium-high confidence. However, unlike the low confidence in its position relative to WDR82 in the dimer prediction, the pAE plot for the trimeric complex (Figure 5.3D) shows high confidence prediction of this region interacting with PP1. The relative positions of WDR82 and PP1 are however less certain, with only moderate PAE scores.

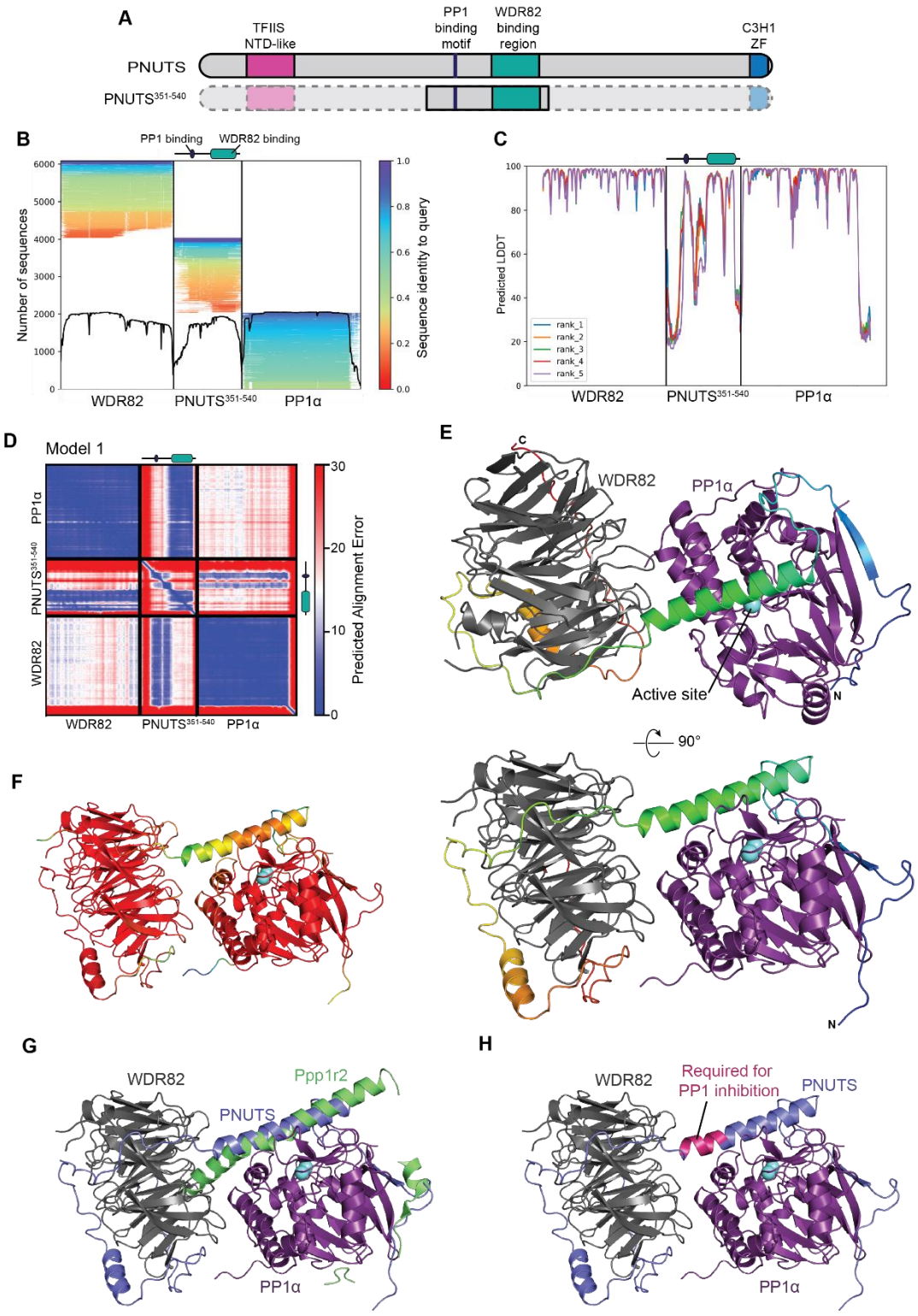


Figure 5.3 Prediction of the WDR82-PNUTS-PP1 complex structure. **A**: Schematic representation of the domain organisation of PNUTS indicating the region used as input for structure prediction (residues 351-540). **B**: Graphical representation of the depth and diversity of the MSA generated by ColabFold structure prediction for WDR82, PNUTS$^{351-540}$, and PP1α. **C**: Graph showing the per-position pLDDT for the 5 predicted structures of the WDR82-PNUTS$^{351-540}$-PP1α complex generated by ColabFold. **D**: Plot of Predicted Alignment Error for the rank 1 WDR82-PNUTS$^{351-540}$-PP1α structure. **E**: Cartoon representation of the rank 1 predicted WDR82-PNUTS$^{351-540}$-PP1α structure generated by ColabFold. For clarity, terminal regions of PNUTS and PP1α predicted with very low

confidence (pLDDT<50) are not shown. WDR82 is in grey, PP1α in purple, and PNUTS$^{351-540}$ is coloured rainbow from N-terminus (blue) to C-terminus (red). PP1α active site metal ions were modelled by alignment of the crystal structure of human PP1α (3V4Y)(O'Connell et al., 2012). **F**: Cartoon representation of the rank 1 predicted WDR82-PNUTS$^{351-540}$-PP1α structure coloured by pLDDT on a spectrum from red (high) to blue (low). **G:** Cartoon representation of the rank 1 predicted WDR82-PNUTS$^{351-540}$-PP1α structure (grey/slate/purple) aligned with the crystal structure of the rat PP1γ-Ppp1r2 complex (green, 2O8A)(Hurley et al., 2007). For clarity, the PP1γ structure is not shown. **H:** Cartoon representation of the rank 1 predicted WDR82-PNUTS$^{351-540}$-PP1α structure with the region of PNUTS required for PP1 inhibition coloured pink.

Examination of the top ranked predicted structure shows how PNUTS forms a predominantly extended structure that wraps around both WDR82 and PNUTS, making extensive contacts with both proteins (Figure 5.3E). The top face of the WDR82 β propeller is orientated towards PP1. Interestingly, the long helix of PNUTS that was predicted in Chapter 4 with only low confidence to bind across the top face of WDR82 is predicted with good confidence to bind across the active site of PP1. This conformation would completely block access of substrates to the active site. Whilst the overall confidence for this region of the structure is lower than for the rest of the structure (pLDDT ~80) (Figure 5.3C, F) it is sufficiently high that the backbone confirmation is likely to be correct and the PAE for its interaction with PP1 is very low, indicating high confidence in the interaction itself. This position of a long helix across the active site of PP1 is highly similar to the previously solved structure of PP1 in complex with Ppp1r2, an inhibitor of PP1 (Figure 5.3G) (Hurley et al., 2007). Previous work has shown that PNUTS can inhibit PP1 activity and residues 445-450, which lie at the C-terminal end of this long helix, closest to the WDR82 binding site (Figure 5.3H), are required for this inhibitory activity (Kim et al., 2003). This suggests that the predicted position of this helix in an inhibitory position across the active site of PP1 is likely to be accurate. The close proximity of WDR82 to this motif raises the possibility that WDR82 could regulate the inhibition of PP1 by PNUTS, however the effect of WDR82 on PP1 inhibition by PNUTS has not previously been tested.

Interestingly, examination of the 5 models predicted by ColabFold reveal two populations in which WDR82 assumes one of two different orientations relative to PP1. The top two ranked structures predict the face of WDR82 to be positioned approximately 90° to the long inhibitory helix (Figure 5.4A), whereas the structures ranked 3rd, 4th, and 5th predict WDR82 to be rotated 180° around the end of the long PNUTS helix and tilted further towards PP1, positioning its top face opposite the PP1 active site (Figure 5.4B). The confidence metrics for these two different conformations are very similar, with the rank 3 structure actually showing subtly higher confidence in the relative positions of WDR82 and PP1, suggesting this may be a more likely conformation (Figure 5.4C). The prediction of these two different structures may reflect a certain degree of conformational flexibility in PNUTS around the C-terminal end of the long inhibitory helix, allowing WDR82 and PP1 to move relative to one another.

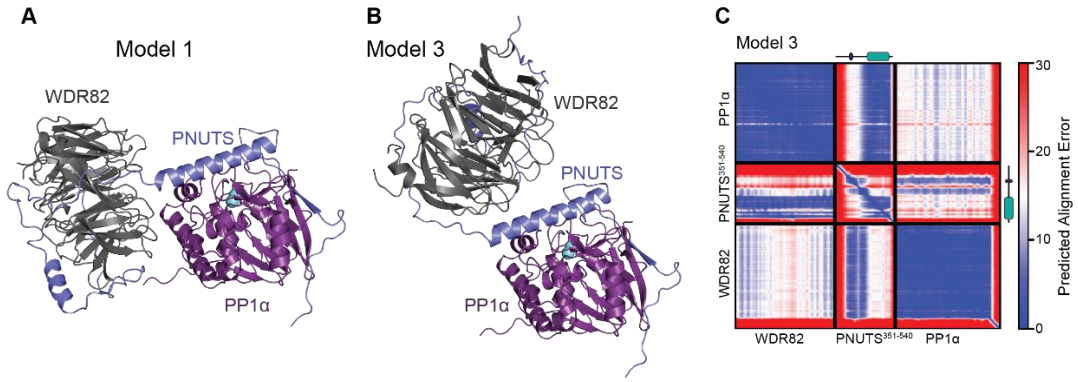


Figure 5.4 Alternative predicted WDR82-PNUTS-PP1 complex structure. **A:** Cartoon representation of the rank 1 predicted WDR82-PNUTS$^{351-540}$-PP1α structure. **B:** Cartoon representation of the rank 3 predicted WDR82-PNUTS$^{351-540}$-PP1α structure. **C:** Plot of Predicted Alignment Error for the rank 3 WDR82-PNUTS$^{351-540}$-PP1α structure.

Given my previous results predicting a potential binding site for this PNUTS helix across the top face of WDR82 (Figure 4.6), one could envisage a mechanism by which PP1 inhibition is relieved by a conformational change which allows for binding of the PNUTS inhibitory helix to WDR82 instead of PP1, opening substrate access to the PP1 active site. Alternatively, WDR82 could contribute to PP1 regulation by binding to substrates such as

the CTD and positioning them close to the PP1 active site. This could act to enhance catalysis by increasing the local substrate concentration or could serve more simply as a substrate specification mechanism.

These results suggest a role for WDR82 beyond simply as a binding adapter that recruits PNUTS-PP1 to RNAPII, but as a substrate presentation or even actively regulatory module. Further experimental investigation of this possibility is an exciting avenue for future research.

5.2.3 WDR82 affects CTD phosphorylation in vivo

From the predicted structure of the WDR82-PNUTS-PP1α complex, I hypothesised that WDR82 may have a role in regulating PP1 activity and hence CTD dephosphorylation. If WDR82 were to aid PNUTS regulation of PP1 activity, one might expect its removal to affect the balance of cellular CTD phosphorylation. To test this possibility, I examined the levels of total (NTD), S2P, and S5P RNAPII, as well as SET1A, ZC3H4, and PNUTS, in nuclear protein extracts following rapid removal of WDR82 using the dTAG system (Figure 5.5A). Loss of WDR82 was associated with increased CTD phosphorylation after just two hours of dTAG-13 treatment, which would be consistent with a disruption of PNUTS-PP1 phosphatase activity (Figure 5.5B). However, by 24hrs of WDR82 removal CTD phosphorylation returned to untreated levels, whilst PNUTS and total ZC3H4 protein levels were increased. This suggests a mechanism by which increased levels of PNUTS and/or ZC3H4 could compensate for loss of WDR82 to rebalance RNAPII phosphorylation. Interestingly, ZC3H4 runs as two close but distinct bands by SDS-PAGE and the increase in total ZC3H4 signal was primarily due to increased intensity in the upper of these two bands. There are no known variant isoforms of ZC3H4, suggesting this upper band may be a post-translationally modified form of the protein. Consistent with this possibility, a number of residues in ZC3H4 have been found to be phosphorylated (Huttlin et al., 2010; Villén et al., 2007). The function and

regulation of these modifications are unknown, however given their increase following loss of WDR82, it is possible that the WDR82-PNUTS-PP1 complex could contribute to ZC3H4 dephosphorylation.

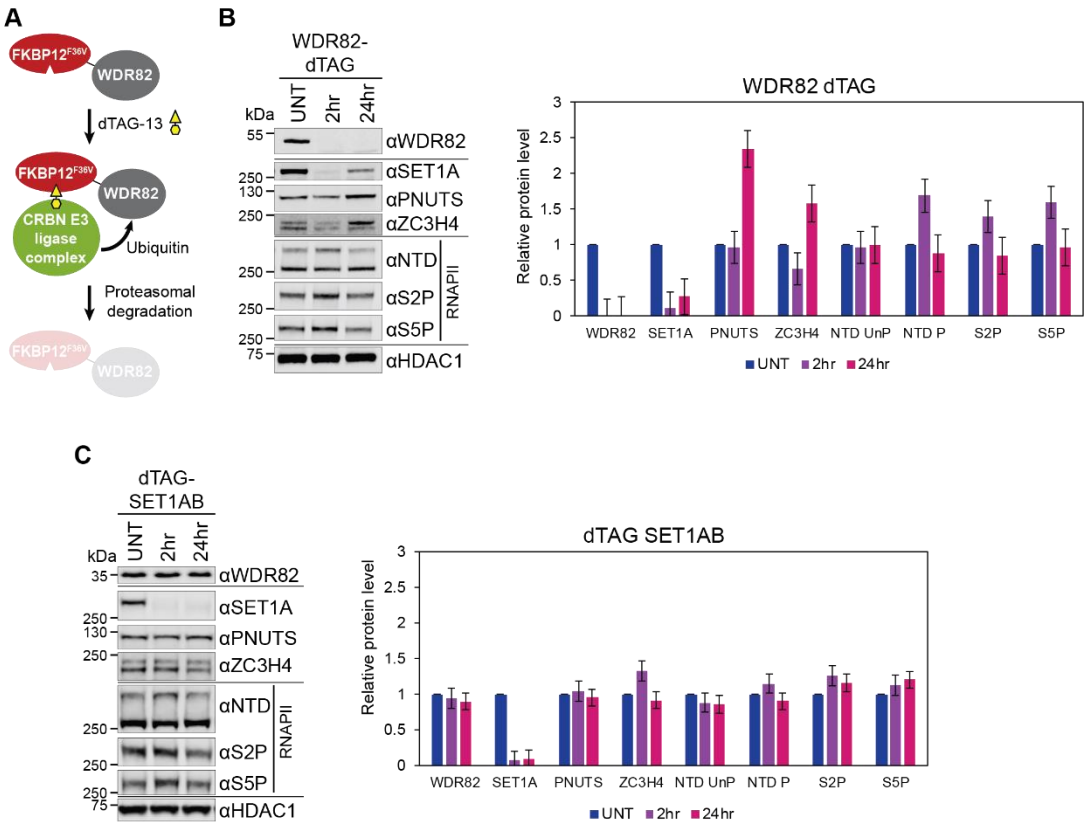


Figure 5.5 Examination of the effects of WDR82 depletion on RNAPII phosphorylation. **A**: A schematic of the dTAG system applied to WDR82. Treatment of cells expressing FKBP12^{F36V}(dTAG)-WDR82 with dTAG-13 brings the CRBN E3 ligase complex into close proximity to WDR82, resulting in its polyubiquitination and proteasomal degradation. Figure adapted from original by Amy Hughes. **B:** Representative western blot of nuclear extract from WDR82-dTAG cells treated as indicated, and quantification of western blot signal from at least three replicates, normalised to HDAC1 signal with untreated (UNT) level set as 1. Quantification of ZC3H4 signal was taken as the total across both bands. 'NTD UnP' is the lower band detected with anti-RNAPII NTD antibody and 'NTD P' is the upper band, representing all phosphorylated RNAPII. Error bars represent standard error. **C:** as for B, with dTAG-SET1AB cells.

Removal of WDR82 is also associated with almost complete degradation of SET1A, consistent with previous observations that SET1A is unstable without WDR82. To confirm that the effects on ZC3H4, PNUTS, and RNAPII following WDR82 removal are not due to concurrent loss of SET1A/B, I also examined bulk protein levels following rapid removal of

SET1A and SET1B (Figure 5.5C). The changes in RNAPII phosphorylation following SET1AB removal are minimal, whilst PNUTS is unaffected and ZC3H4 is increased slightly at the 2hr time point. This results suggests the effects on RNAPII phosphorylation seen following WDR82 removal are not due to loss of SET1A/B.

To further dissect the contribution of each WDR82-containing complex to the effects on RNAPII phosphorylation seen following WDR82 removal, I also examined the effects of PNUTS and ZC3H4 removal on bulk nuclear protein levels. Interestingly, ZC3H4 removal (Figure 5.6A) was associated with rapid increases in the levels of both WDR82 and PNUTS, suggesting that either ZC3H4 regulates WDR82 and PNUTS levels, or PNUTS expression is increased to compensate for the loss of ZC3H4 activity. RNAPII phosphorylation was unaffected by ZC3H4 loss, perhaps due to a rapid compensatory effect of increased PNUTS levels.

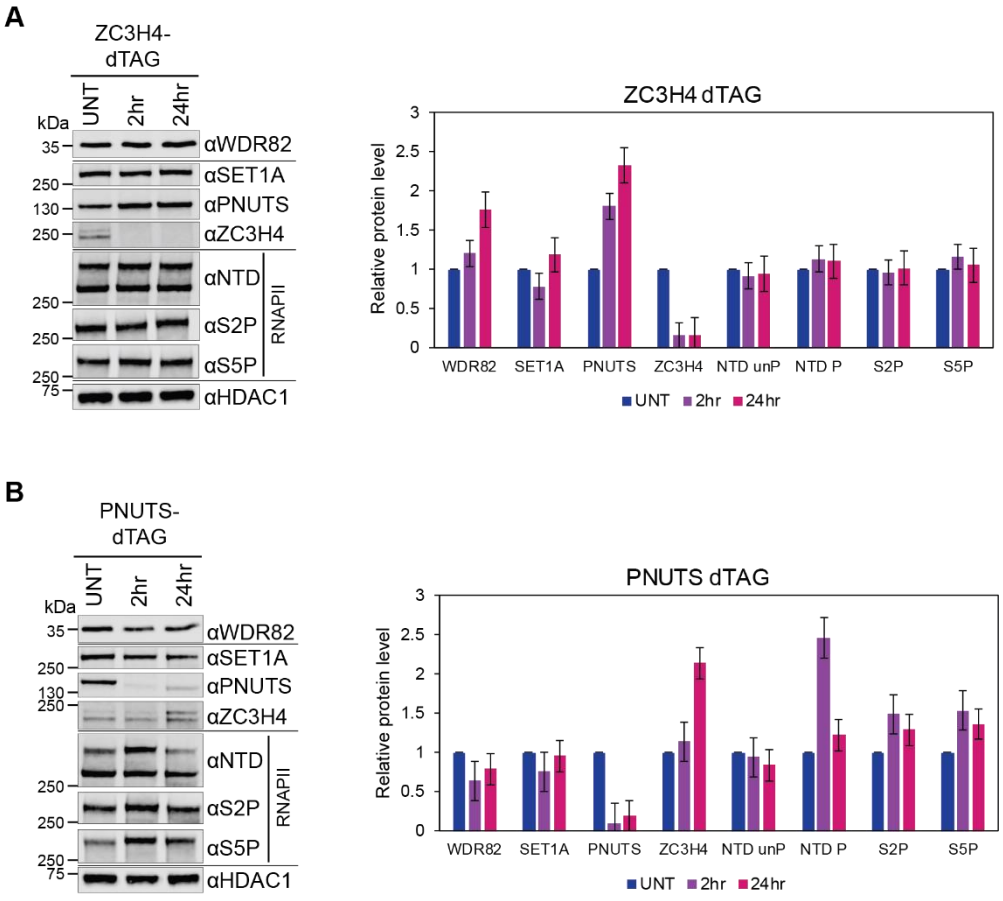
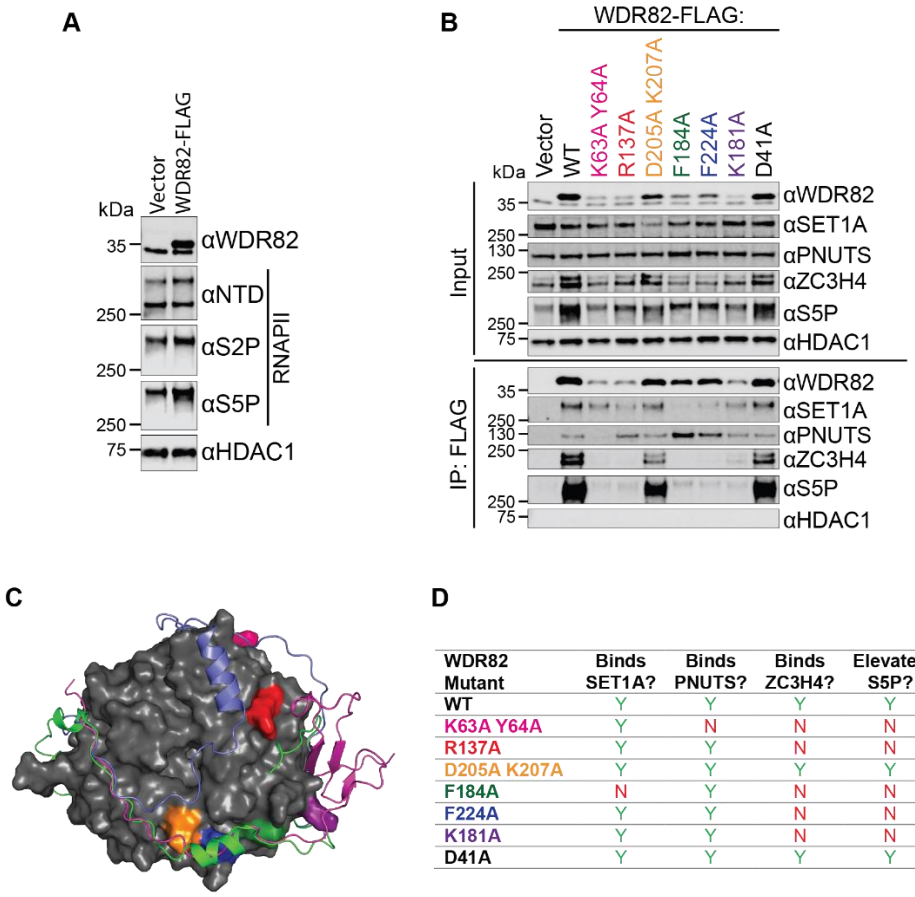


Figure 5.6 Examination of the effects of ZC3H4 and PNUTS depletion on RNAPII phosphorylation. **A:** Representative western blot of nuclear extract from ZC3H4-dTAG cells treated as indicated, and quantification of western blot signal from at least three replicates, normalised to HDAC1 signal with untreated (UNT) level set as 1. Quantification of ZC3H4 signal was taken as the total across both bands. 'NTD UnP' is the lower band detected with anti-RNAPII NTD antibody and 'NTD P' is the upper band, representing all phosphorylated RNAPII. Error bars represent standard error. **B:** as for A, with PNUTS-dTAG cells.

PNUTS removal most closely copied the effects on RNAPII phosphorylation seen following loss of WDR82 (Figure 5.6B). RNAPII phosphorylation was significantly increased by 2hrs of PNUTS removal and to a greater extent than for WDR82 removal. This supports a role for the WDR82-PNUTS-PP1 complex in dephosphorylation of RNAPII CTD. Phosphorylation was recovered somewhat by 24hrs, but not all the way to untreated levels, suggesting the cell cannot completely compensate for loss of PNUTS. Levels of PNUTS recovered slightly by the 24hr time point, which may represent selection bias towards cells which maintain some level of PNUTS expression. Interestingly, similarly to WDR82 removal, total ZC3H4 levels were increased following 24hrs of PNUTS removal and this was again due to an increase in the upper, post-translationally modified, species. The mechanism and function of this effect is unclear, however there does seem to be a relationship between PNUTS and ZC3H4 protein levels and possibly some ability for these two complexes to compensate for each other to regulate RNAPII phosphorylation.

Interestingly, I found that WDR82 overexpression in 293T cells is also associated with increased phosphorylation of RNAPII (Figure 5.7A). In order to understand which WDR82-containing complex may underpin this effect, I generated mutations in WDR82 to disrupt binding to specific complexes and examined their effects on RNAPII phosphorylation (Figure 5.7B). Due to the overlapping binding sites of SET1A, ZC3H4, and PNUTS on WDR82, design of mutations which disrupted binding of only one partner was challenging, and I was only able to define mutations which disrupted binding to SET1A and ZC3H4, PNUTS and ZC3H4, or ZC3H4 alone (Figure 5.7D). None of the mutants I tested eliminated binding to only SET1A or PNUTS (data not shown), and a number of the mutant proteins expressed at lower

level that the WT, suggesting they may destabilise WDR82 to some extent. Interestingly, the hyperphosphorylation of RNAPII is reliant on the ability of WDR82 to bind ZC3H4, as overexpression of mutations which eliminate ZC3H4 binding did not affect phosphorylation (Figure 5.7D). This result suggests the WDR82-ZC3H4 complex may promote RNAPII phosphorylation, however the mechanism for this effect is unclear. Furthermore, this observation is counter to the effects seen when ZC3H4 is depleted, where I observed no change in RNAPII phosphorylation. Overexpression of WDR82 may upset the balance of its different activities so these results should not be over-interpreted. However, this observation suggests the WDR82-ZC3H4 complex may regulate RNAPII phosphorylation and further characterisation of WDR82-ZC3H4 effects on RNAPII phosphorylation in a more native system would be informative.



WDR82 Mutant	Binds SET1A?	Binds PNUTS?	Binds ZC3H4?	Elevated S5P?
WT	Y	Y	Y	Y
K63A Y64A	Y	N	N	N
R137A	Y	Y	N	N
D205A K207A	Y	Y	Y	Y
F184A	N	Y	N	N
F224A	Y	Y	N	N
K181A	Y	Y	N	N
D41A	Y	Y	Y	Y

Figure 5.7 WDR82 overexpression is associated with increased RNAPII phosphorylation. **A:** Western blot of nuclear extract from 293T cells transfected with either empty vector or a plasmid expressing

FLAG-tagged WDR82. HDAC1 is presented as a loading control. **B:** Western blot of input nuclear extract and FLAG IP samples from 293T cells transiently transfected with plasmids expressing WT or mutant FLAG-tagged WDR82. Mutants are coloured as in C. HDAC1 is a loading control for inputs and negative control for IP experiments. **C:** Cartoon representation of the overlaid structures of WDR82 in complex with SET1A (magenta), ZC3H4 (green) and PNUTS (slate). Mutated WDR82 residues are coloured as in B. D41 is not visible from this viewpoint. **D:** Table summarising effects of WDR82 mutations on binding to SET1A, PNUTS, and ZC3H4, and on RNAPII S5P levels. Mutants are coloured as in C.

Overall, these results suggest an intimate relationship between the WDR82-containing complexes, particularly PNUTS and ZC3H4, and RNAPII phosphorylation. However, the nuclear extraction procedure used to generate these samples does not strip very stably bound factors, such as engaged polymerase, from chromatin. These results therefore might not fully reflect the effects on the chromatin-bound protein fractions that may be more relevant for transcription regulation. Further examination of specifically chromatin-based effects, for example by ChIP-seq, will be important to further characterise the relationships between these complexes, RNAPII phosphorylation, and transcription itself.

5.3 Summary and discussion

The SET1A, ZC3H4, and PNUTS complexes have previously been shown to associate with phosphorylated RNAPII CTD, however the role of WDR82 in this interaction has not been investigated in all these contexts. I therefore sought to understand and directly compare the CTD-binding behaviour of WDR82 when in complex with SET1A, ZC3H4, and PNUTS, and characterise how this behaviour could underpin the role of WDR82 in each complex.

I have shown that the WDR82-SET1A/B complex binds specifically to S5P CTD peptides *in vitro*, and the SET1A NTD interacts with S5P RNAPII in a WDR82-dependent manner *in vivo*. These results suggest that the function of WDR82 within the SET1 complexes is as an adapter module which links the activities of the SET1 complexes to RNAPII specifically when it carries S5P CTD, which is predominantly associated with initiating transcription. This

function has previously been proposed to recruit the SET1A/B H3K4 HMT activity to sites of active transcription, allowing for H3K4me3 deposition which in turn promotes transcriptional activity (Franks et al., 2017; Lee and Skalnik, 2008). However, recent work from the Klose lab has shown that SET1A catalytic activity is dispensable for gene activation by SET1A, while tethering the NTD of SET1A to the promoter is sufficient to activate reporter gene expression (Hughes et al., 2022). Importantly, interaction with WDR82 was required for this activity. This result suggested that WDR82-SET1A/B binding to RNAPII CTD may function to directly regulate RNAPII, rather than simply to recruit the SET1 complexes to sites of active transcription. The mechanism by which the WDR82-SET1A NTD complex could promote transcription is unclear, but may be via stabilisation of RNAPII binding at promoters.

Furthermore, we showed that SET1A/B act to support transcription by counteracting premature transcription termination by ZC3H4. We hypothesised that this behaviour may be mediated by the shared subunit WDR82. My work in Chapter 3 indicates that levels of WDR82 are sufficiently high, with a significant pool of monomeric protein, that competition for binding to WDR82 is unlikely to be a factor in this functional competition. Whilst I have shown that the WDR82-SET1A complex binds CTD both *in vitro* and *in vivo*, I could not conclusively characterise the ability of the WDR82-ZC3H4 complex to bind CTD. However, others have identified S5P-specific CTD binding for WDR82 in complex with full-length ZC3H4 (Park et al., 2022). This suggests my results may be due to technical limitations, such as low affinity under the condition tested or insufficiency of the minimal WDR82-binding fragment of ZC3H4 used. The role of WDR82 within ZC3H4 complexes requires further examination, however it may be as an adapter for CTD binding, similar to in SET1A/B complexes. In support of this possibility, the similarities in the molecular basis of SET1A and ZC3H4 binding to WDR82 (Chapter 4) would suggest they have similar effects on the additional binding behaviours of WDR82. The RNAPII CTD is sufficiently large that both the

SET1A/B and ZC3H4 complexes could bind simultaneously and it is therefore unlikely that direct binding competition for the CTD underpins their functional competition. However, WDR82-mediated binding to the RNAPII CTD would allow for colocalisation of the two complexes and hence their activities, allowing them to come into direct competition, with outputs determined by integration of the relative levels and activities of each complex.

The role of WDR82 within the PNUTS complex may be subtly different. The PNUTS-WDR82 complex I tested did not bind any CTD peptides *in vitro* and interaction with WDR82 was not required for PNUTS to bind RNAPII in cells. Furthermore, prediction of the WDR82-PNUTS-PP1α complex structure showed that WDR82 is positioned close to the PP1 active site. I therefore hypothesise that rather than binding RNAPII CTD to stably anchor the complex to RNAPII, WDR82 may act with PNUTS to regulate PP1 activity. A minimal PNUTS-PP1 complex has previously been shown to dephosphorylate the RNAPII CTD *in vitro*, however the PNUTS construct used in the experiments lacked the inhibitory helix, and WDR82 was not present (Choy et al., 2014). Larger fragments of PNUTS which include the inhibitory helix I observed in the predicted WDR82-PNUTS-PP1α complex structure have been shown to strongly inhibit PP1 activity *in vitro*. However, PNUTS complexes purified from HEK293T cells were able to dephosphorylate RNAPII CTD *in vitro*, suggesting the inhibitory activity of PNUTS is not absolute (Kim et al., 2003; Lee et al., 2010). The role of WDR82 in regulating PP1 activity has not been directly examined, however it has been suggested to be required for proper PNUTS-PP1 activity *in vivo* (Landsverk et al., 2020). I hypothesise that WDR82 allows for release of PP1 inhibition through conformation changes in PNUTS which alter binding of the inhibitory helix to allow access to the PP1 active site. The possible stimulus for this conformation change is unclear, but it may be PNUTS phosphorylation or binding to the CTD. Alternatively, WDR82 binding to substrates may act to 'present' them to the PP1 active site or it could simply enhance catalysis by increasing the local concentration of substrate. Further experiments will be necessary to assess these

hypotheses. For example, *in vitro* phosphatase assays would be useful to determine whether WDR82 affects the activity or specificity of the PNUTS-PP1 complex.

Rapid removal of WDR82 was associated with an increase in CTD phosphorylation that is most closely recapitulated by removal of PNUTS. This is consistent with the function of PNUTS-PP1 as a CTD phosphatase complex and suggests that WDR82 is required for normal phosphatase activity. Removal of WDR82 or PNUTS is also associated with extensive changes to transcription, however it is unclear whether these are due to disruption of RNAPII phosphorylation or if changes in transcription result in altered phosphorylation profiles (Austenaa et al., 2015; Ciurciu et al., 2013; Cortazar et al., 2019)(Amy Hughes, unpublished data). Detailed investigation of the effects of WDR82 or PNUTS removal on transcription and on chromatin-bound RNAPII, for example using TT-seq and ChIP-seq, respectively, will be important to understand the extent of WDR82-PNUTS-PP1-mediated regulation of RNAPII phosphorylation and how this could affect transcription itself.

Interestingly, I found that WDR82 overexpression in 293T cells was associated with increases in RNAPII phosphorylation which were dependent upon WDR82 binding to ZC3H4. This result was surprising, as WDR82 removal from ESCs was also associated with increased RNAPII phosphorylation. The mechanism for this effect is unclear and could be due to a number of factors. Firstly, my analysis of protein levels following dTAG treatments suggested that the levels of PNUTS and ZC3H4 are closely related, with each increased when the other was removed. However, when WDR82 was removed both proteins increased. This could represent a mutual cross-regulation which relies on WDR82. Either removal or overexpression of WDR82 may decouple this dependence. Alternatively, the effects I observed by western blot of nuclear extracts could represent a change in the distribution of RNAPII between the chromatin-bound and free soluble fractions. ZC3H4 has been shown to promote premature transcription termination and an increase in this

activity could result in more RNAPII being evicted from chromatin shortly following initiation, increasing the soluble pool detected by bulk western blot (Austenaa et al., 2021; Estell et al., 2021). It would be interesting to further examine the effects of ZC3H4 removal on chromatin-bound RNAPII by ChIP-seq and to dissect the contribution of WDR82 to ZC3H4 distribution and function.

In conclusion, my data suggest that WDR82 takes on different roles in different contexts. In the SET1A/B and ZC3H4 complexes, WDR82 seems to act as a binding adapter to S5P RNAPII CTD, allowing for colocalisation of opposing activities. In contrast, the PNUTS-PP1 complex interacts with WDR82 in a distinct manner to its incorporation into the SET1A and ZC3H4 complexes (Chapter 4) and WDR82 may take on a different function in this complex to regulate PP1 activity rather than providing a stable interaction with the RNAPII CTD. It will be interesting to further interrogate the effects of WDR82 on PP1 catalytic activity both *in vitro* and *in vivo*. Importantly, the specific mutations I identified in Chapter 4 which disrupt SET1A, ZC3H4, and PNUTS binding to WDR82 will be crucial in interrogating its role in these different complexes.

6 Conclusions & future directions

WDR82 is a small (35.1kDa) WD40-repeat protein that is essential for cell viability and organism development (Bi et al., 2011; Cheng et al., 2004; Lee and Skalnik, 2005; Miller et al., 2001; Soares and Buratowski, 2012). Loss of WDR82 from cells is associated with widespread effects on transcription including changes in gene body transcription, termination, and upstream antisense transcription (Austenaa et al., 2015)(Amy Hughes, unpublished data). Previous work has identified WDR82 as component of multiple complexes which regulate transcription, namely the SET1 histone methyltransferase complexes, the ZC3H4 complex, and the PNUTS-PP1 phosphatase complex (Austenaa et al., 2021; Brewer-Jensen et al., 2016; Lee et al., 2010; Lee and Skalnik, 2005; Miller et al., 2001; van Nuland et al., 2013). However, the molecular basis of WDR82 incorporation into these complexes and its function in each context was not well characterised. Furthermore, it has been proposed that WDR82 acts as an adaptor module which could link each complex to transcription by binding S5P CTD (Ebmeier et al., 2017; Lee and Skalnik, 2008; Park et al., 2022). Here I have biochemically defined the repertoire of WDR82-containing complexes in mESCs and have shown that cellular WDR82 concentration is not limiting (Chapter 3). In addition, I have used highly accurate protein structure predictions to determine the molecular basis of WDR82 incorporation into the SET1A, ZC3H4, and PNUTS complexes and have validated these interactions by *in vivo* binding experiments (Chapter 4). Remarkably I discovered that, despite no known evolutionary relationships, SET1A and ZC3H4 bind WDR82 via very similar interfaces, whereas PNUTS employs a distinct binding mode which nevertheless shares some features with SET1A and ZC3H4. Finally, I sought to understand the molecular function of WDR82 in each context. I propose that WDR82 functions within the SET1A and ZC3H4 complexes as an adapter module that binds S5P CTD to link these

complexes and their activities to initiating RNAPII. In contrast, I hypothesise that WDR82 may function in the PNUTS complex to help regulate PP1 activity. (Chapter 5).

An important aspect of understanding the function of WDR82 is a detailed picture of the molecular identity of each complex. The composition of the SET1 complexes has been well characterised in many cell types and I was able to identify the majority of previously defined core complex components as interactors of SET1A and WDR82. In contrast, the ZC3H4 complex has only recently emerged as an important regulator of transcription and its biochemical characterisation has been limited. I have shown that the constitutive core complex is likely to consist of only ZC3H4 and WDR82, however I also uncovered many interactions with RNA-binding proteins. This suggests the function of ZC3H4 may be closely linked with binding to RNA and further characterisation of RNA-mediated interactions may shed light on the mechanisms by which ZC3H4 promotes premature transcription termination.

Despite no known evolutionary relationship between the two proteins, I found that ZC3H4 and SET1A interact with WDR82 in a remarkably similar way. Furthermore, published data suggests WDR82 has similar CTD-binding behaviour in both complexes (Lee and Skalnik, 2008; Park et al., 2022). I have shown that the WDR82-SET1A complex binds specifically to S5P CTD, however I was unable to conclusively verify the CTD-binding behaviour of the ZC3H4-WDR82 complex. Optimisation of these experiments to obtain a clear readout will be important to understand the function of WDR82 in the ZC3H4 complex. We have recently shown that SET1A supports transcription genome-wide by opposing ZC3H4-mediated premature transcription termination and have proposed that this may be mediated by WDR82 (Hughes et al., 2022). I have shown that there is sufficient monomeric WDR82 present in cells that competition between SET1A and ZC3H4 is unlikely to be a result of competition for binding to WDR82. Instead, WDR82 binding to S5P RNAPII

CTD may colocalise the SET1 and ZC3H4 complexes, allowing for their functional competition. It will be important to examine whether WDR82 is required for SET1A and ZC3H4 activity genome-wide. Knockout of SET1A or ZC3H4 followed by rescue with the WDR82-binding mutants I have identified here would allow for *in vivo* dissection of each complex individually to ask whether interaction with WDR82 is required for their opposing activities.

The distribution of PNUTS across genes has been shown to closely reflect that of RNAPII, however the interactions which underpin this localisation are not well understood (Cortazar et al., 2019; Verheyen et al., 2015)(Amy Hughes, unpublished data). My results suggest that WDR82 in complex with PNUTS does not bind the RNAPII CTD and I have shown that PNUTS can interact with RNAPII independently of WDR82. Therefore, there must exist an additional interaction or interactions which allow PNUTS to associate with RNAPII. Interestingly, I identified a stable interaction between the PNUTS-PP1 complex and the elongation factor PAF1 that is independent of RNAPII. The PAF1 complex forms extensive contacts with the core RNAPII enzyme and hence could mediate PNUTS interaction with RNAPII (Vos et al., 2018a). Interestingly, previous research suggests PNUTS and PAF1 have opposing effects on RNAPII elongation rate. PAF1 has been shown to promote elongation, whilst PNUTS has been proposed to restrict elongation speed (Cortazar et al., 2019; Hou et al., 2019; Liu et al., 2022). It will be important to examine the functional relationship between these two complexes further, as it could underpin both PNUTS association with RNAPII and transcriptional regulation throughout gene bodies. This could be addressed with an integrated experimental approach combining biochemical characterisation of the physical association between PNUTS, PAF1, and RNAPII with genomics experiments to examine, for example, the effects of PNUTS removal on transcription and PAF1 distribution.

In addition, I have shown that PNUTS binds WDR82 in a manner distinct from SET1A and ZC3H4, and have hypothesised that WDR82 functions in the PNUTS complex to regulate PP1 activity. Further exploration of this hypothesis *in vitro* will be key to understanding the role of WDR82 in the PNUTS complex. Previous experiments examining the catalytic activity of the PNUTS-PP1 complex towards various substrates *in vitro* have used different PNUTS fragments and did not include WDR82 (Kim et al., 2003; Kreivi et al., 1997; Landsverk et al., 2020; Lee et al., 2010; Wu et al., 2018). These assays therefore came to differing conclusions about the extent to which PNUTS inhibits or promotes PP1 activity. *In vitro* phosphatase assays using PNUTS constructs guided by my predicted structures will allow for a more refined assessment of PP1 regulation and the role of WDR82 in this. RNAPII phosphorylation *in vivo* is clearly affected by loss of WDR82 or PNUTS, however the relationship between this effect and changes in transcription is unclear. Furthermore, the PNUTS-PP1 complex has also been shown to dephosphorylate Spt5 and this activity is important for control of RNAPII elongation speed (Cortazar et al., 2019). The extensive changes to transcription and RNAPII phosphorylation observed following loss of PNUTS are therefore difficult to interpret as they may represent disrupted dephosphorylation of multiple substrates as well as potential phosphatase-independent activities that affect transcription. Detailed characterisation of PP1 regulation by the PNUTS complex *in vitro* will be an important process to define components or interactions that can be disrupted to examine specific functional relationships *in vivo*.

Appendices

Table A.1 Sequences of synthesised cDNAs

cDNA	Sequence
SET1A	ACGTAGTCTAGACGCCATGGCAATGAATAATTTTGTTTAACTTTAAGAAGGAGATATACATC GCGGCCGCATGCATCATCACCATCATCATAGCAGCGGTGTTGATCTGGGCACCGAAAATCTG TATTTTCAGAGCATGGATCAAGAAGGCGGAGGCGACGGTCAGAAAGCACCGAGCTTTCAGT GGCGTAACTATAAACTGATTGTTGATCCGGCTCTGGATCCGGCACTGCGTCGTCCGAGCCAG AAAGTTTATCGTTATGATGGTGTTCATTTCAGCGTGAGCGATAGCAAATATACACCGGTTGA AGATCTGCAGGATCCGCGTTGTCATGTTCGTAGCAAAGCACGTGATTTTAGCCTGCCGGTTC CGAAATTCAAACTGGATGAATTTTACATTGGTCAGATCCCGCTGAAAGAAGTTACCTTTGCAC GTCTGAATGATAATGTGCGTGAAACCTTTCTGAAAGACATGTGCCGTAAATATGGTGAAGTT GAAGAAGTGGAAATTCTGCTGCATCCGCGTACACGTAAACATCTGGGTTTAGCACGTGTTCT GTTTACCAGCACACGTGGTGCAAAAGAAACCGTTAAAAATCTGCATCTGACCAGCGTTATGG GCAACATTATTCATGCACAGCTGGATATTAAAGGTCAGCAGCGTATGAAATATTACGAGCTG ATTGTGAATGGTAGCTATACACCGCAGACCGTTCCGACCGGTGGTAAAGCACTGAGCGAAA AATTTCAAGGTAGCGGTGCAGCAGCAGAAACCACCGAAGCACGTCGTCGTAGCAGCAGCGA TACCGCAGCATATCCGGCAGGCACCACCGTTGGTGGCACCCCTGGTAATGGCACCCCGTGTA GCCAGGATACCAATTTTAGCAGCAGTCGTCAGGATACCCCGAGCAGCTTTGGTCAGTTTACT CCGCAGAGCAGCCAGGGTACACCGTATACCAGCCGTGGTAGCACCCCGTATTCACAGGATA GCGCATATAGCAGCTCAACCACCAGTACCAGCTTTAAACCGCGTCGTAGCGAAAATAGCTAT CAGGATAGCTTTAGCCGTCGTCATTTTAGCACCAGCAGCGCACCGGCAACCACCGCAACCGC AACCAGTGCCACCGCAGCGGCAACCGCAGCAAGCAGCAGCTCATCAAGCAGTAGCAGTAGC TCAAGCAGTTCAAGCTCAAGTAGCAGCGCAAGCCAGTTTCGTGGTAGCGATTCAAGCTATCC GGCTTATTATGAAAGCTGGAATCGTTATCAGCGTCATACCAGCTATCCGCCTCGTCGTGCAAC CCGTGAAGATCCGAGCGGTGCCAGCTTTGCAGAAAATACCGCAGAACGTTTTCCGCCTAGCT ATACCAGTTATCTGGCACCGGAACCGAATCGTAGCACCGATCAGGATTATCGTCCGCCTGCA AGCGAAGCACCGCCTCCGGAACCGCCTGAACCTGGTGGTGGTGGCGGTGGTTCAGGTGGCG GAGGCGGAGGTGGTGGCGGAGGTGGCGGTGGCGCACCTAGTCCGGAACGTGAAGAAGCA CGTACCCCTCCGCGTCCGGCAAGTCCGGCACGTAGTGGTAGCCCTGCACCGGAAACAACCAA TGAAAGCGTTCCGTTTGCACAGCATAGCAGCCTGGATAGCCGTATTGAAATGCTGCTGAAAG AGCAGCGTAGCAAATTTAGCTTTCTGGCAAGTGATACCGAAGAAGAAGAGGAAAATTCAAG CGCAGGTCCGGGTGCACGTGATGCCGGTGCCGAAGTTCCGAGTGGTGCAGGTCATGGTCCG TGTACACCTCCACCGGCTCCGGCAAATTTTGAAGATGTTGCACCGACCGGTAGTGGTGAACC GGGTGCAGCACGTGAAAGCCCGAAAGCAAATGGTCAGAATCAGGCAAGCCCGTGTAGTAG CGGTGAAGATATGGAAATTAGTGATGATGATCGTGGTGGTAGTCCGCCACCGGCACCGACA CCACCTCAGCAGCCTCCACCTCCTCCTCCGCCTCCACCACCGCCACCTCCGCCATATCTGGCAA GCCTGCCGCTGGGTTATCCTCCGCATCAGCCTGCATATCTGCTGCCTCCACGTCCGGATGGTC CGCCTCCTCCGGAATATCCTCCTCCTCCTCCACCTCCTCCGCCACATATTTATGATTTTGTTAAT AGCCTGGAACTGATGGATCGTCTGGGTGCACAGTGGGGTGGTATGCCGATGAGCTTTCAAA TGCAGACCCAGATGCTGACCCGTCTGCATCAGCTGCGTCAAGGTAAAGGTCTGACCGCAGCC TCAGCAGGTCCGCCAGGTGGTGCATTTGGTGAAGCATTTCTGCCGTTTCCACCGCCTCAAGA AGCAGCATACGGTCTGCCGTATGCACTGTATACCCAGGGTCAAGAAGGTCGCGGTAGTTATA GCCGTGAAGCATATCATCTGCCGCTGCCGATGGCAGCAGAACCGCTGCCGTCAAGCAGCGTT AGTGGCGAAGAAGCCCGTCTGCCGCATCGTGAAGAGGCAGAAATTGCAGAAAGCAAAGTTC TGCCGAGCGCAGGTACAGTTGGTCGTGTTCTGGCAACCCTGGTTCAAGAAATGAAAAGCATT ATGCAGCGTGATCTGAATCGCAAAATGGTTGAAAATGTTGCGTTTGGTGCCTTTGATCAGTG

	GTGGGAATCAAAAGAAGAAAAAGCAAAACCGTTTCAGAACGCAGCAAAACAGCAGGCCAA AGAAGAGGATAAAGAAAAAATGAAGCTGAAAGAACCGGGTATGCTGAGCCTGGTTGATTG GGCAAAAAGCGGTGGTATTACCGGTATTGAAGCCTTTGCCTTTGGTAGCGGTCTGCGTGGTG CCCTGCGTCTGCCGAGTTTTAAAGTGAAACGTAAAGAACCGTCCGAAATTAGCGAAGCCAGC GAAGAAAAACGTCCGCGTCCTAGCACACCAGCCGAAGAGGATGAGGATGATCCGGAACGC GAAAAAGAAGCAGGCGAACCGGGTCGTCCGGGTACAAAACCGCCTAAACGTGATGAAGAA CGTGGTAAAACCCAGGGTAAACATCGTAAAGGTTTTACACTGGATAGCGAAGGTGAGGAAG CAAGCCAAGAAAGCAGCAGTGAAAAAGATGAAGATGATGATGACGAGGATGAAGAGGACG AAGAACAAGAAGAAGCAGTTGACGCAACCAAAAAAGAGGCCGAAGCAAGCGACGGCGAA GATGAGGATAGCGATTCTAGCAGCCAGTGTAGCCTGTATGCAGATAGTGATGGTGAAAATG GTAGTACCAGCGATAGCGAAAGCGGTAGCAGTTCTAGCAGTAGTAGCTCATCATCAAGTTCT AGTTCTTCAAGCAGCTCAGAAAGCTCTAGCGAAGAGGAAGAACAGAGCGCAGTTATTCCGA GCGCAAGTCCTCCGCGTGAAGTGCCGGAACCTCTGCCAGCACCGGATGAAAAACCGGAAAC CGATGGTCTGGTTGATAGTCCGGTTATGCCGCTGTCAGAAAAAGAAACACTGCCGACACAGC CAGCCGGTCCGGCAGAAGAACCACCACCGAGCGTTCCTCAGCCTCCTGCAGAACCTCCGGCA GGTCCTCCTGATGCAGCACCGCGTCTGGATGAACGTCCGAGCAGCCCGATTCCGCTGTTACC GCCACCGAAAAAACGCCGTAAAACCGTGAGCTTTAGCGCAGCAGAAGAAGCGCCTGTTCCT GAACCGAGCACAGCCGCACCGCTGCAGGCGAAAAGCAGCGGTCCGGTGAGCCGTAAAGTT CCGCGTGTTGTTGAACGTACCATTCGTAATCTGCCTCTGGATCATGCAAGCCTGGTTAAAAGC TGGCCTGAAGAGGTTGCCCGTGGTGGTCGTAATCGTGCAGGCGGTCGTGTGCGTAGCACTG AGGAAGAAGAAGCCACCGAAAGCGGCACCGAAGTTGATCTGGCAGTTCTGGCCGATCTGGC ACTGACACCGGCACGTCGCGGTCTGGCTACCCTGCCGACAGGTGATGATAGTGAAGCAACC GAAACCTCAGATGAAGCAGAACGTCCGTCACCTCTGCTGAGCCATATTCTGCTGGAACATAA CTATGCACTGGCAATTAAACCGCCTCCTACAACTCCGGCACCTCGTCCGCTGGAACCTGCGCC TGCATTAGCAGCACTGTTTAGCAGCCCTGCAGATGAAGTTCTGGAAGCACCGGAAGTTGTTG TTGCAGAAGCCGAAGAACCTAAACAGCAGCTGCAGCAACAGCATCCGGAACAAGAGGGTG AAGAAGAGGAAGAGGACGAGGAAGAAGAGTCAGAAAGCAGCGAAAGCTCAAGCTCTAGCT CATCAGATGAAGAAGGTGCAATTCGTCGTCGTTCACTGCGTAGTCATACCCGTCGTCGTCGC CCTCCGTTGCCTCCTCCTCCTCCTCCACCGCCTAGCTTTGAACCGCGTAGCGAATTTGAGCAG ATGACCATTCTGTATGATATTTGGAATAGTGGCCTGGACCTGGAAGATATGAGCTATCTGCG CCTGACCTATGAACGTCTGCTGCAACAGACCAGCGGTGCCGATTGGCTGAACGATACCCATT GGGTTCAGCATACAATTACCAATCTGAGCACCCCGAAACGCAAACGTCGTCCGCAGGATGGT CCTCGTGAACATCAGACCGGTAGCGCACGTTCAGAAGGTTATTATCCGATTAGCAAGAAAGA AAAAGATAAATATCTGGATGTGTGCCCTGTGAGCGCACGTCAGCTGGAAGGTGGTGATACA CAGGGTACAAATCGTGTGCTGAGCGAACGTCGTAGTGAACAGCGTCGTCTGCTGAGTGCAA TTGGCACCAGCGCAATTATGGATAGTGATCTGCTGAAACTGAACCAGCTGAAATTCCGTAAG AAAAAACTGCGTTTTGGTCGTAGCCGTATCCATGAATGGGGTCTGTTTGCAATGGAACCGAT TGCAGCCGATGAAATGGTTATTGAATATGTGGGTCAGAACATTCGTCAGATGGTTGCCGATA TGCGTGAAAAACGTTATGTGCAAGAAGGTATTGGTAGCAGCTACCTGTTTCGTGTTGATCAT GATACCATTATCGATGCCACCAAATGTGGTAATCTGGCACGTTTTATCAATCATTGTTGTACC CCGAATTGCTACGCCAAAGTTATTACCATTGAAAGCCAGAAAAAAATCGTGATCTATTCCAA GCAGCCGATTGGTGTGGATGAAGAAATTACCTATGATTACAAATTCCCGCTTGAGGACAACA AAATTCCGTGTCTGTGTGGTACAGAAAGCTGTCGTGGTAGTCTGAACTAATAATGTACAAGA TCCGAATTCAAGCTTAGCAAC
SET1B$^{1-209}$	ATGGAAAACTCTCACCCTCACCACCACCATCAGCAGCCTCCACCTCAACCTGGACCTAGCGGC GAGAGAAGAAACCACCACTGGCGGAGCTACAAGCTGATGATCGACCCCGCTCTGAAGAAGG GCCACCACAAGCTGTACAGATACGACGGCCAGCACTTCAGCCTGGCCATGAGCAGCAACAG ACCCGTGGAAATCGTCGAGGACCCTAGAGTCGTCGGCATCTGGACCAAGAACAAAGAACTG GAACTGAGCGTGCCCAAGTTCAAGATCGACGAGTTCTACGTGGGCCCCGTGCCTCCTAAGCA

	AGTGACATTCGCCAAGCTGAACGACAACGTGCGCGAGAACTTCCTGAGAGACATGTGCAAG AAGTACGGCGAGGTCGAGGAAGTGGAAATCCTGTACAACCCCAAGACCAAGAAGCACCTGG GAATCGCCAAGGTGGTGTTCGCCACAGTCAGAGGCGCCAAAGAAGCCGTGCAGCATCTGCA CAGCACAAGCGTGATGGGCAACATCATCCACGTCGAGCTGGACACCAAGGGCGAGACAAGA ATGAGATTCTACGAGCTGCTGGTCACCGGCAGATACACCCCTCAGACACTGCCTGTGGGAGA GCTGGATGCTATC

Table A.2 Summary of crystallisation trials. All samples included full length WDR82 and the additional fragment listed in column 3. All reagents were from Molecular Dimensions. Details of the Morpheus optimisation screen ('Morpheus opt') are given in Figure A.1. Ratios given are protein: precipitant.

Plate	Protein code	Fragment	Conc (mg/ml)	Screen	Temperature	Ratio 1	Ratio 2	Notes
1	W82S	SET1A 8-189	6	PGA	18	1:1	-	
2	W82S	SET1A 8-189	6	JCSG+	18	1:1	-	
3	W82S	SET1A 8-189	6	MIDAS	18	1:1	-	
4	W82S	SET1A 8-189	6	Morpheus	18	1:1	-	
5	W82S	SET1A 8-189	6	ProPlex	18	1:1	-	
6	W82S	SET1A 8-189	8	Morpheus opt	4	1:1	-	Microseed
7	W82S	SET1A 8-189	8	Morpheus opt	18	1:1	-	Microseed
8	W82S	SET1A 8-189	8	Morpheus opt	4	1:1	-	Silver bullets additive screen
9	W82S	SET1A 8-189	8	Morpheus opt	18	1:1	-	Silver bullets additive screen
10	W82S	SET1A 8-189	8	Morpheus	18	1:1	-	Microseed
11	W82S	SET1A 8-189	8	Morpheus 2	18	1:1	-	Microseed
12	W82S B4	SET1B 21-197	12	PACT	4	1:1	2:1	
13	W82S B4	SET1B 21-197	12	ProPlex	4	1:1	2:1	
14	W82S B4	SET1B 21-197	12	JCSG+	4	1:1	2:1	
15	W82S B4	SET1B 21-197	12	Morpheus	4	1:1	2:1	
16	W82S B4	SET1B 21-197	12	MIDAS	4	1:1	2:1	
17	W82S B4	SET1B 21-197	12	PACT	RT	1:1	2:1	
18	W82S B4	SET1B 21-197	12	MIDAS	RT	1:1	2:1	
19	W82S B4	SET1B 21-197	12	JCSG+	RT	1:1	2:1	
20	W82S B4	SET1B 21-197	12	ProPlex	RT	1:1	2:1	
21	W82S B4	SET1B 21-197	12	Morpheus	RT	1:1	2:1	
22	W82S A10	SET1A 14-195	5.3	MIDAS	RT	1:1	2:1	

23	W82S A10	SET1A 14-195	5.3	ProPlex	RT	1:1	2:1	
24	W82S A10	SET1A 14-195	5.3	Morpheus	RT	1:1	2:1	
25	W82S A10	SET1A 14-195	5.3	JCSG+	RT	1:1	2:1	
26	W82S A10	SET1A 14-195	5.3	PACT	RT	1:1	2:1	
27	W82S A11	SET1A 14-84	12.6	ProPlex	RT	1:1	2:1	
28	W82S A11	SET1A 14-84	12.6	Morpheus	RT	1:1	2:1	
29	W82S A11	SET1A 14-84	12.6	MIDAS	RT	1:1	2:1	
30	W82S A11	SET1A 14-84	12.6	JCSG+	RT	1:1	2:1	
31	W82S A11	SET1A 14-84	12.6	PACT	RT	1:1	2:1	
32	W82S B9	SET1B 25-203	7.1	ProPlex	RT	1:1	2:1	
33	W82S B9	SET1B 25-203	7.1	Morpheus	RT	1:1	2:1	
34	W82S B9	SET1B 25-203	7.1	MIDAS	RT	1:1	2:1	
35	W82S B9	SET1B 25-203	7.1	PACT	RT	1:1	2:1	
36	W82S B9	SET1B 25-203	7.1	JCSG+	RT	1:1	2:1	
37	W82S B10	SET1B 25-92	6.3	MIDAS	RT	1:1	2:1	
38	W82S B10	SET1B 25-92	6.3	Morpheus	RT	1:1	2:1	
39	W82S B10	SET1B 25-92	6.3	JCSG+	RT	1:1	2:1	
40	W82S B10	SET1B 25-92	6.3	PGA	RT	1:1	2:1	
41	W82S A10	SET1A 14-195	10.1	Proples	RT	1:1	-	Microseed
42	W82S A10	SET1A 14-195	10.1	Morpheus	RT	1:1	-	Microseed
43	W82S A10	SET1A 14-195	10.1	Index	RT	1:1	-	Microseed
44	W82S A10	SET1A 14-195	10.1	MIDAS+	RT	1:1	-	Microseed
45	W82S A11	SET1A 14-84	7.5	Morpheus	RT	1:1	-	Microseed
46	W82S A11	SET1A 14-84	7.5	Proplex	RT	1:1	-	Microseed
47	WDR82		8	PGA	RT	1:1	2:1	
48	WDR82		8	JCSG+	RT	1:1	2:1	

49	WDR82		8	PACT	RT	1:1	2:1	
50	WDR82		8	Morpheus	RT	1:1	2:1	
51	WDR82		8	MIDAS	RT	1:1	2:1	
52	W82S A10	SET1A 14-195	5	Proplex	RT	1:1	2:1	
53	W82S A10	SET1A 14-195	5	Morpheus	4	1:1	2:1	Microseed
54	W82S A10	SET1A 14-195	5	Proplex	4	1:1	2:1	Microseed
55	WDR82		4	JCSG+	RT	1:1	2:1	
56	WDR82		4	Morpheus	RT	1:1	2:1	
57	WDR82		4	MIDAS	RT	1:1	2:1	
58	WDR82		4	PACT	RT	1:1	2:1	
59	WDR82		4	Proplex	RT	1:1	2:1	
60	W82S A18	SET1A 14-88	6.5	JCSG+	RT	1:1	-	
61	W82S A18	SET1A 14-88	6.5	PACT	RT	1:1	-	
62	W82S A18	SET1A 14-88	6.5	Morpheus	RT	1:1	-	
63	W82S A18	SET1A 14-88	6.5	Proplex	RT	1:1	-	
64	W82S A17	SET1A 14-84 Δ29-32	6.8/5.6	JCSG+	RT	1:1	1:1	
65	W82S A17	SET1A 14-84 Δ29-32	6.8/5.6	Morpheus	RT	1:1	1:1	
66	W82S A17	SET1A 14-84 Δ29-32	6.8/5.6	Proplex	RT	1:1	1:1	
67	W82S A19	SET1A 14-88 Δ29-32	6.2/4.8	Proplex	RT	1:1	1:1	
68	W82S A19	SET1A 14-88 Δ29-32	6.2/4.8	Morpheus	RT	1:1	1:1	
69	W82S A19	SET1A 14-88 Δ29-32	6.2/4.8	JCSG+	RT	1:1	1:1	
70	W82 Z8	ZC3H4 934-999	5.6	Morpheus	RT	1:1	2:1	
71	W82 Z8	ZC3H4 934-999	5.6	Proplex	RT	1:1	2:1	
72	W82 Z8	ZC3H4 934-999	5.6	JCSG+	RT	1:1	-	
73	W82 Z8	ZC3H4 934-999	5.6	MIDAS	RT	1:1	-	
73	W82 Z8	ZC3H4 934-999	5.6	PACT	RT	1:1	-	

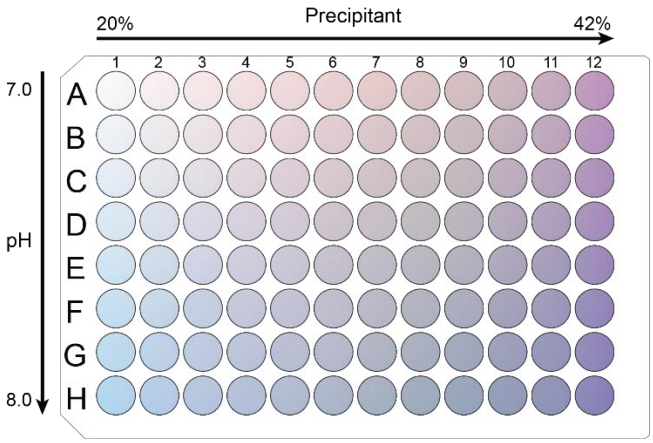


Figure A.1 Schematic of Morpheus optimisation crystallisation screen. Screen was prepared using 0.1M Morpheus buffer mix 2 prepared according to manufacturer's protocol to give the pH indicated, and precipitant mix 2 used at percentage v/v indicated.

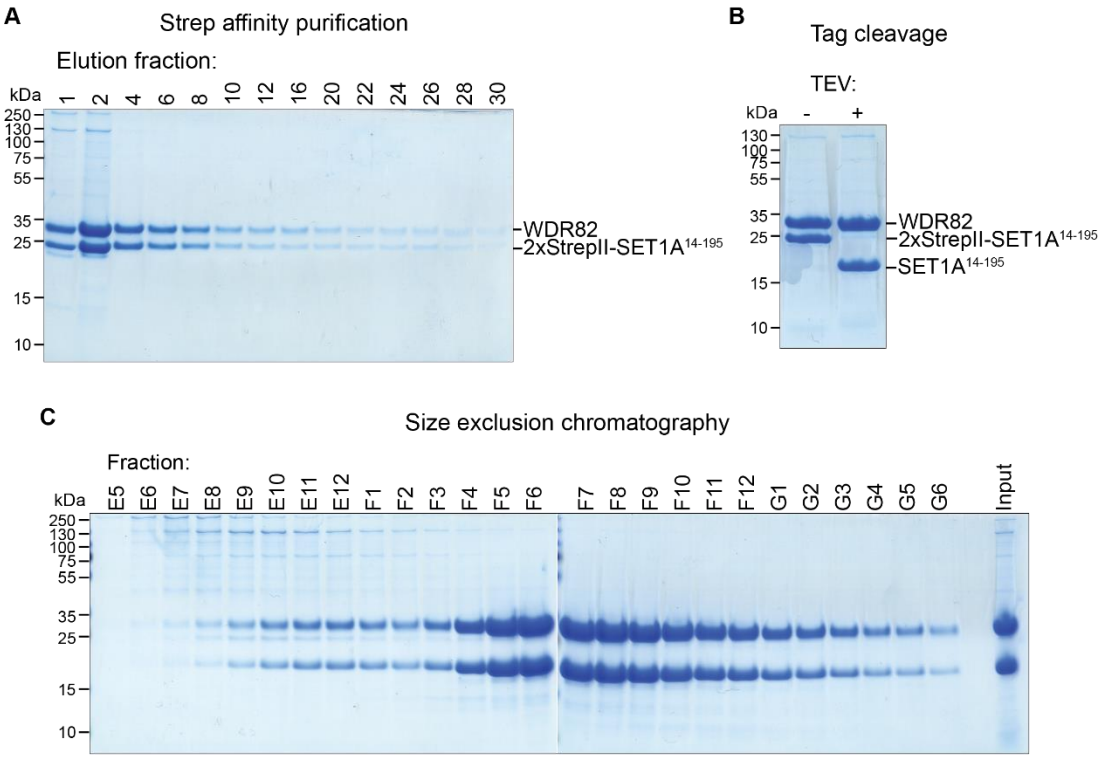


Figure A.2 Example of protein purification for crystallography. **A:** Coomassie-stained SDS-PAGE gel of Strep affinity purification elution fractions from purification of WDR82-SET1A$^{14-195}$. **B:** Coomassie-stained SDS-PAGE gel showing samples of Strep affinity-purified WDR82-SET1A$^{14-195}$ complex before and after tag cleavage using TEV protease. **C:** Coomassie-stained SDS-PAGE gel showing size exclusion chromatography fractions from purification of WDR82-SET1A$^{14-195}$ complex. Fractions F4 to F12 were pooled.

Bibliography

Adelman K, Henriques T. 2018. Transcriptional speed bumps revealed in high resolution. *Nature*. doi:10.1038/d41586-018-05971-8

Adelman K, Lis JT. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet*. doi:10.1038/nrg3293

Adelman K, Wei W, Ardehali MB, Werner J, Zhu B, Reinberg D, Lis JT. 2006. Drosophila Paf1 Modulates Chromatin Structure at Actively Transcribed Genes. *Mol Cell Biol* **26**:250–260. doi:10.1128/mcb.26.1.250-260.2006

Allen BL, Taatjes DJ. 2015. The Mediator complex: A central integrator of transcription. *Nat Rev Mol Cell Biol*. doi:10.1038/nrm3951

Allen PB, Kwon YG, Nairn AC, Greengard P. 1998. Isolation and characterization of PNUTS, a putative protein phosphatase 1 nuclear targeting subunit. *J Biol Chem* **273**:4089–4095. doi:10.1074/jbc.273.7.4089

Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA. 2013. Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **499**:360–363. doi:10.1038/nature12349

AlphaFold. n.d. https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb

Andersen PK, Lykke-Andersen S, Jensen TH. 2012. Promoter-proximal polyadenylation sites reduce transcription activity. *Genes Dev* **26**:2169–2179. doi:10.1101/gad.189126.112

Ardehali MB, Yao J, Adelman K, Fuda NJ, Petesch SJ, Webb WW, Lis JT. 2009. Spt6 enhances the elongation rate of RNA polymerase II in vivo. *EMBO J* **28**:1067–1077. doi:10.1038/emboj.2009.56

Aslanidis C, de Jong PJ. 1990. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res* **18**:6069–6074. doi:10.1093/nar/18.20.6069

Austenaa LMI, Barozzi I, Simonatto M, Masella S, Della Chiara G, Ghisletti S, Curina A, de Wit E, Bouwman BAM, de Pretis S, Piccolo V, Termanini A, Prosperini E, Pelizzola M, de Laat W, Natoli G. 2015. Transcription of Mammalian cis-Regulatory Elements Is Restrained by Actively Enforced Early Termination. *Mol Cell* **60**:460–474. doi:10.1016/j.molcel.2015.09.018

Austenaa LMI, Piccolo V, Russo M, Prosperini E, Polletti S, Polizzese D, Ghisletti S, Barozzi I, Diaferia GR, Natoli G. 2021. A first exon termination checkpoint preferentially suppresses extragenic transcription. *Nat Struct Mol Biol* **28**:337–346. doi:10.1038/s41594-021-00572-y

Bae HJ, Dubarry M, Jeon J, Soares LM, Dargemont C, Kim J, Geli V, Buratowski S. 2020. The

Set1 N-terminal domain and Swd2 interact with RNA polymerase II CTD to recruit COMPASS. *Nat Commun* **11**:1–10. doi:10.1038/s41467-020-16082-2

Bannister AJ, Kouzarides T. 2011. Regulation of chromatin by histone modifications. *Cell Res*. doi:10.1038/cr.2011.22

Barman P, Reddy D, Bhaumik SR. 2019. Mechanisms of antisense transcription initiation with implications in gene expression, genomic integrity and disease pathogenesis. *Non-coding RNA*. doi:10.3390/ncrna5010011

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**:823–837. doi:10.1016/j.cell.2007.05.009

Bartkowiak B, Liu P, Phatnani HP, Fuda NJ, Cooper JJ, Price DH, Adelman K, Lis JT, Greenleaf AL. 2010. CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. *Genes Dev* **24**:2303–2316. doi:10.1101/gad.1968210

Beckedorff F, Blumenthal E, DaSilva LF, Aoi Y, Cingaram PR, Yue J, Zhang A, Dokaneheifard S, Valencia MG, Gaidosh G, Shilatifard A, Shiekhattar R. 2020. The Human Integrator Complex Facilitates Transcriptional Elongation by Endonucleolytic Cleavage of Nascent Transcripts. *Cell Rep* **32**:107917. doi:10.1016/j.celrep.2020.107917

Belotserkovskaya R, Saunders A, Lis JT, Reinberg D. 2004. Transcription through chromatin: Understanding a complex FACT. *Biochim Biophys Acta - Gene Struct Expr*. doi:10.1016/j.bbaexp.2003.09.017

Benjamin B, Sanchez AM, Garg A, Schwer B, Shuman S. 2021. Structure-function analysis of fission yeast cleavage and polyadenylation factor (CPF) subunit Ppn1 and its interactions with Dis2 and Swd22. *PLoS Genet* **17**. doi:10.1371/journal.pgen.1009452

Bentley DL. 2014. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet*. doi:10.1038/nrg3662

Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L, Dreyfuss G. 2012. U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150**:53–64. doi:10.1016/j.cell.2012.05.029

Berger I, Craig A. 2010. MultiBac Expression System User Manual. doi:10.13140/2.1.2563.6645

Bernecky C, Herzog F, Baumeister W, Plitzko JM, Cramer P. 2016. Structure of transcribing mammalian RNA polymerase II. *Nature* **529**:551–554. doi:10.1038/nature16482

Bernecky C, Plitzko JM, Cramer P. 2017. Structure of a transcribing RNA polymerase II-DSIF complex reveals a multidentate DNA-RNA clamp. *Nat Struct Mol Biol* **24**:809–815. doi:10.1038/nsmb.3465

Bi Y, Lv Z, Wang Y, Hai T, Huo R, Zhou Z, Zhou Q, Sha J. 2011. WDR82, a Key Epigenetics-Related Factor, Plays a Crucial Role in Normal Early Embryonic Development in Mice. *Biol Reprod* **84**:756–764. doi:10.1095/biolreprod.110.084343

Bieniossek C, Imasaki T, Takagi Y, Berger I. 2012. MultiBac: Expanding the research toolbox

for multiprotein complexes. *Trends Biochem Sci*. doi:10.1016/j.tibs.2011.10.005

Blazek D, Kohoutek J, Bartholomeeusen K, Johansen E, Hulinkova P, Luo Z, Cimermancic P, Ule J, Peterlin BM. 2011. The cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev* **25**:2158–2172. doi:10.1101/gad.16962311

Bledau AS, Schmidt K, Neumann K, Hill U, Ciotta G, Gupta A, Torres DC, Fu J, Kranz A, Stewart AF, Anastassiadis K. 2014. The H3K4 methyltransferase Setd1a is first required at the epiblast stage, whereas Setd1b becomes essential after gastrulation. *Dev* **141**:1022–1035. doi:10.1242/dev.098152

Bollen M, Peti W, Ragusa MJ, Beullens M. 2010. The extended PP1 toolkit: Designed to create specificity. *Trends Biochem Sci*. doi:10.1016/j.tibs.2010.03.002

Booth GT, Parua PK, Sansó M, Fisher RP, Lis JT. 2018. Cdk9 regulates a promoter-proximal checkpoint to modulate RNA polymerase II elongation rate in fission yeast. *Nat Commun* **9**:1–10. doi:10.1038/s41467-018-03006-4

Bösken CA, Farnung L, Hintermair C, Schachter MM, Vogel-Bachmayr K, Blazek D, Anand K, Fisher RP, Eick D, Geyer M. 2014. The structure and substrate specificity of human Cdk12/Cyclin K. *Nat Commun* **5**:1–14. doi:10.1038/ncomms4505

Brewer-Jensen P, Wilson CB, Abernethy J, Mollison L, Card S, Searles LL. 2016. Suppressor of sable [Su(s)] and Wdr82 down-regulate RNA from heat-shock-inducible repetitive elements by a Mechanism that involves transcription termination. *RNA* **22**:139–154. doi:10.1261/rna.048819.114

Brown DA, Di Cerbo V, Feldmann A, Ahn J, Ito S, Blackledge NP, Nakayama M, McClellan M, Dimitrova E, Turberfield AH, Long HK, King HW, Kriaucionis S, Schermelleh L, Kutateladze TG, Koseki H, Klose RJ. 2017. The SET1 Complex Selects Actively Transcribed Target Genes via Multivalent Interaction with CpG Island Chromatin. *Cell Rep* **20**:2313–2327. doi:10.1016/j.celrep.2017.08.030

Buccitelli C, Selbach M. 2020. mRNAs, proteins and the emerging principles of gene expression control. *Nat Rev Genet*. doi:10.1038/s41576-020-0258-4

Buratowski S. 2009. Progression through the RNA Polymerase II CTD Cycle. *Mol Cell*. doi:10.1016/j.molcel.2009.10.019

Buratowski S. 2003. The CTD code. *Nat Struct Biol*. doi:10.1038/nsb0903-679

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

Carminati M, Manav MC, Bellini D, Passmore LA. 2022. A direct interaction between CPF and Pol II links RNA 3'-end processing to transcription. *bioRxiv* 2022.07.28.501803. doi:10.1101/2022.07.28.501803

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest ARR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A,

Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, Bono B, Gatta G Della, Bernardo D Di, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SPT, Kruger A, Kummerfeld SK, Kurochkin I V., Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Babu MM, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schönbach C, Sekiguchi K, Semple CAM, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y. 2005. Molecular biology: The transcriptional landscape of the mammalian genome. *Science (80-)* **309**:1559–1563. doi:10.1126/science.1112014

Casañal A, Kumar A, Hill CH, Easter AD, Emsley P, Degliesposti G, Gordiyenko Y, Santhanam B, Wolf J, Wiederhold K, Dornan GL, Skehel M, Robinson C V., Passmore LA. 2017. Architecture of eukaryotic mRNA 3'-end processing machinery. *Science (80-)* **358**:1056–1059. doi:10.1126/science.aao6535

Cermakova K, Demeulemeester J, Lux V, Nedomova M, Goldman SR, Smith EA, Srb P, Hexnerova R, Fabry M, Madlikova M, Horejsi M, De Rijck J, Debyser Z, Adelman K, Courtney Hodges H, Veverka V. 2021. A ubiquitous disordered protein interaction module orchestrates transcription elongation. *Science (80-)* **374**:1113–1121. doi:10.1126/science.abe2913

Chammas P, Mocavini I, Di Croce L. 2020. Engaging chromatin: PRC2 structure meets function. *Br J Cancer*. doi:10.1038/s41416-019-0615-2

Chao SH, Price DH. 2001. Flavopiridol Inactivates P-TEFb and Blocks Most RNA Polymerase II Transcription in Vivo. *J Biol Chem* **276**:31793–31799. doi:10.1074/jbc.M102306200

Chen X, Qi Y, Wu Z, Wang X, Li J, Zhao D, Hou H, Li Y, Yu Z, Liu W, Wang M, Ren Y, Li Z, Yang H, Xu Y. 2021. Structural insights into preinitiation complex assembly on core promoters. *Science (80-)* **372**. doi:10.1126/science.aba8490

Cheng B, Price DH. 2007. Properties of RNA polymerase II elongation complexes before and after the P-TEFb-mediated transition into productive elongation. *J Biol Chem* **282**:21901–21912. doi:10.1074/jbc.M702936200

Cheng H, He X, Moore C. 2004. The Essential WD Repeat Protein Swd2 Has Dual Functions

in RNA Polymerase II Transcription Termination and Lysine 4 Methylation of Histone H3. *Mol Cell Biol* **24**:2932–2943. doi:10.1128/mcb.24.7.2932-2943.2004

Cheung ACM, Cramer P. 2011. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature* **471**:249–253. doi:10.1038/nature09785

Chiu AC, Suzuki HI, Wu X, Mahat DB, Kriz AJ, Sharp PA. 2018. Transcriptional Pause Sites Delineate Stable Nucleosome-Associated Premature Polyadenylation Suppressed by U1 snRNP. *Mol Cell* **69**:648-663.e7. doi:10.1016/j.molcel.2018.01.006

Cho NH, Cheveralls KC, Brunner AD, Kim K, Michaelis AC, Raghavan P, Kobayashi H, Savy L, Li JY, Canaj H, Kim JYS, Stewart EM, Gnann C, McCarthy F, Cabrera JP, Brunetti RM, Chhun BB, Dingle G, Hein MY, Huang B, Mehta SB, Weissman JS, Gómez-Sjöberg R, Itzhak DN, Royer LA, Mann M, Leonetti MD. 2022. OpenCell: Endogenous tagging for the cartography of human cellular organization. *Science (80-)* **375**. doi:10.1126/science.abi6983

Choudhury R, Singh S, Arumugam S, Roguev A, Stewart AF. 2019. The Set1 complex is dimeric and acts with Jhd2 demethylation to convey symmetrical H3K4 trimethylation. *Genes Dev* **33**:550–564. doi:10.1101/gad.322222.118

Choy MS, Hieke M, Kumar GS, Lewis GR, Gonzalez-DeWhitt KR, Kessler RP, Stein BJ, Hessenberger M, Nairn AC, Peti W, Page R. 2014. Understanding the antagonism of retinoblastoma protein dephosphorylation by PNUTS provides insights into the PP1 regulatory code. *Proc Natl Acad Sci U S A* **111**:4097–4102. doi:10.1073/pnas.1317395111

Ciurciu A, Duncalf L, Jonchere V, Lansdale N, Vasieva O, Glenday P, Rudenko A, Vissi E, Cobbe N, Alphey L, Bennett D. 2013. PNUTS/PP1 Regulates RNAPII-Mediated Gene Expression and Is Necessary for Developmental Growth. *PLoS Genet* **9**:e1003885. doi:10.1371/journal.pgen.1003885

Clapier CR, Cairns BR. 2009. The biology of chromatin remodeling complexes. *Annu Rev Biochem*. doi:10.1146/annurev.biochem.77.062706.153223

Clouaire T, Webb S, Bird A. 2014. Cfp1 is required for gene expression-dependent H3K4 trimethylation and H3K9 acetylation in embryonic stem cells. *Genome Biol* **15**:451. doi:10.1186/s13059-014-0451-x

Clouaire T, Webb S, Skene P, Illingworth R, Kerr A, Andrews R, Lee JH, Skalnik D, Bird A. 2012. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes Dev* **26**:1714–1728. doi:10.1101/gad.194209.112

ColabFold. n.d. https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

Connelly S, Manley JL. 1988. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev* **2**:440–452. doi:10.1101/gad.2.4.440

Corden JL. 1990. Tails of RNA polymerase II. *Trends Biochem Sci*. doi:10.1016/0968-

0004(90)90236-5

Corden JL, Cadena DL, Ahearn JM, Dahmus ME. 1985. A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II. *Proc Natl Acad Sci U S A* **82**:7934–7938. doi:10.1073/pnas.82.23.7934

Core L, Adelman K. 2019. Promoter-proximal pausing of RNA polymerase II: A nexus of gene regulation. *Genes Dev*. doi:10.1101/gad.325142.119

Cortazar MA, Sheridan RM, Erickson B, Fong N, Glover-Cutter K, Brannan K, Bentley DL. 2019. Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism. *Mol Cell* **76**:896-908.e4. doi:10.1016/j.molcel.2019.09.031

Cossa G, Parua PK, Eilers M, Fisher RP. 2021. Protein phosphatases in the RNAPII transcription cycle: Erasers, sculptors, gatekeepers, and potential drug targets. *Genes Dev*. doi:10.1101/GAD.348315.121

Cossa G, Roeschert I, Prinz F, Baluapuri A, Silveira Vidal R, Schülein-Völk C, Chang YC, Ade CP, Mastrobuoni G, Girard C, Wortmann L, Walz S, Lührmann R, Kempa S, Kuster B, Wolf E, Mumberg D, Eilers M. 2020. Localized Inhibition of Protein Phosphatase 1 by NUAK1 Promotes Spliceosome Activity and Reveals a MYC-Sensitive Feedback Control of Transcription. *Mol Cell* **77**:1322-1339.e11. doi:10.1016/j.molcel.2020.01.008

Cramer P. 2019. Organization and regulation of gene transcription. *Nature*. doi:10.1038/s41586-019-1517-4

Czudnochowski N, Bösken CA, Geyer M. 2012. Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. *Nat Commun* **3**:1–12. doi:10.1038/ncomms1846

Danko CG, Hah N, Luo X, Martins AL, Core L, Lis JT, Siepel A, Kraus WL. 2013. Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Mol Cell* **50**:212–222. doi:10.1016/j.molcel.2013.02.015

de Santa F, Barozzi I, Mietton F, Ghisletti S, Polletti S, Tusi BK, Muller H, Ragoussis J, Wei CL, Natoli G. 2010. A large fraction of extragenic RNA Pol II transcription sites overlap enhancers. *PLoS Biol* **8**:e1000384. doi:10.1371/journal.pbio.1000384

Dehé PM, Dichtl B, Schaft D, Roguev A, Pamblanco M, Lebrun R, Rodríguez-Gil A, Mkandawire M, Landsberg K, Shevchenko Anna, Shevchenko Andrej, Rosaleny LE, Tordera V, Chávez S, Stewart AF, Géli V. 2006. Protein interactions within the Set1 complex and their roles in the regulation of histone 3 lysine 4 methylation. *J Biol Chem* **281**:35404–35412. doi:10.1074/jbc.M603099200

Dharmarajan V, Lee JH, Patel A, Skalnik DG, Cosgrove MS. 2012. Structural basis for WDR5 interaction (Win) motif recognition in human SET1 family histone methyltransferases. *J Biol Chem* **287**:27275–27289. doi:10.1074/jbc.M112.364125

Ding L, Paszkowski-Rogacz M, Winzi M, Chakraborty D, Theis M, Singh S, Ciotta G, Poser I, Roguev A, Chu WK, Choudhary C, Mann M, Stewart AF, Krogan N, Buchholz F. 2015. Systems Analyses Reveal Shared and Diverse Attributes of Oct4 Regulation in Pluripotent Cells. *Cell Syst* **1**:141–151. doi:10.1016/j.cels.2015.08.002

Dingar D, Tu WB, Resetca D, Lourenco C, Tamachi A, De Melo J, Houlahan KE, Kalkat M, Chan PK, Boutros PC, Raught B, Penn LZ. 2018. MYC dephosphorylation by the PP1/PNUTS phosphatase complex regulates chromatin binding and protein stability. *Nat Commun* **9**. doi:10.1038/s41467-018-05660-0

Dujardin G, Lafaille C, Petrillo E, Buggiano V, Gómez Acuña LI, Fiszbein A, Godoy Herz MA, Nieto Moreno N, Muñoz MJ, Alló M, Schor IE, Kornblihtt AR. 2013. Transcriptional elongation and alternative splicing. *Biochim Biophys Acta - Gene Regul Mech*. doi:10.1016/j.bbagrm.2012.08.005

Eaton JD, Francis L, Davidson L, West S. 2020. A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes. *Genes Dev* **34**:132–145. doi:10.1101/gad.332833.119

Eaton JD, West S. 2020. Termination of Transcription by RNA Polymerase II: BOOM! *Trends Genet*. doi:10.1016/j.tig.2020.05.008

Ebmeier CC, Erickson B, Allen BL, Allen MA, Kim H, Fong N, Jacobsen JR, Liang K, Shilatifard A, Dowell RD, Old WM, Bentley DL, Taatjes DJ. 2017. Human TFIIH Kinase CDK7 Regulates Transcription-Associated Chromatin Modifications. *Cell Rep* **20**:1173–1186. doi:10.1016/j.celrep.2017.07.021

Elrod ND, Henriques T, Huang KL, Tatomer DC, Wilusz JE, Wagner EJ, Adelman K. 2019. The Integrator Complex Attenuates Promoter-Proximal Transcription at Protein-Coding Genes. *Mol Cell* **76**:738-752.e7. doi:10.1016/j.molcel.2019.10.034

Erickson B, Sheridan RM, Cortazar M, Bentley DL. 2018. Dynamic turnover of paused pol II complexes at human promoters. *Genes Dev* **32**:1215–1225. doi:10.1101/gad.316810.118

Estell C, Davidson L, Steketee PC, Monier A, West S. 2021. Zc3h4 restricts non-coding transcription in human cells. *Elife* **10**. doi:10.7554/ELIFE.67305

Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, Žídek A, Bates R, Blackwell S, Yim J, Ronneberger O, Bodenstein S, Zielinski M, Bridgland A, Potapenko A, Cowie A, Tunyasuvunakool K, Jain R, Clancy E, Kohli P, Jumper J, Hassabis D. 2022. Protein complex prediction with AlphaFold-Multimer. *bioRxiv* 2021.10.04.463034. doi:10.1101/2021.10.04.463034

Florens L, Carozza MJ, Swanson SK, Fournier M, Coleman MK, Workman JL, Washburn MP. 2006. Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **40**:303–311. doi:10.1016/j.ymeth.2006.07.028

Fournier M, Bourriquen G, Lamaze FC, Côté MC, Fournier É, Joly-Beauparlant C, Caron V, Gobeil S, Droit A, Bilodeau S. 2016. FOXA and master transcription factors recruit Mediator and Cohesin to the core transcriptional regulatory circuitry of cancer cells. *Sci Rep* **6**:1–11. doi:10.1038/srep34962

Francette AM, Tripplehorn SA, Arndt KM. 2021. The Paf1 Complex: A Keystone of Nuclear Regulation Operating at the Interface of Transcription and Chromatin. *J Mol Biol*. doi:10.1016/j.jmb.2021.166979

Franks TM, McCloskey A, Shokirev M, Benner C, Rathore A, Hetzer MW. 2017. Nup98 recruits the Wdr82-Set1A/COMPASS complex to promoters to regulate H3K4 trimethylation in hematopoietic progenitor cells. *Genes Dev* **31**:2222–2234. doi:10.1101/gad.306753.117

Fraser NW, Sehgal PB, Darnell JE. 1978. DRB-induced premature termination of late adenovirus transcription. *Nature* **272**:590–593. doi:10.1038/272590a0

Frietze S, Farnham PJ. 2011. Transcription factor effector domains. *Subcell Biochem* **52**:261–277. doi:10.1007/978-90-481-9069-0_12

Fuchs G, Voichek Y, Benjamin S, Gilad S, Amit I, Oren M. 2014. 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol* **15**:R69. doi:10.1186/gb-2014-15-5-r69

Fujinaga K, Irwin D, Huang Y, Taube R, Kurosu T, Peterlin BM. 2004. Dynamics of Human Immunodeficiency Virus Transcription: P-TEFb Phosphorylates RD and Dissociates Negative Effectors from the Transactivation Response Element. *Mol Cell Biol* **24**:787–795. doi:10.1128/mcb.24.2.787-795.2004

Giardina C, Perez-Riba M, Lis JT. 1992. Promoter melting and TFIID complexes on Drosophila genes in vivo. *Genes Dev* **6**:2190–2200. doi:10.1101/gad.6.11.2190

Gilmour DS, Lis JT. 1986. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. *Mol Cell Biol* **6**:3984–3989. doi:10.1128/mcb.6.11.3984-3989.1986

Glover-Cutter K, Larochelle S, Erickson B, Zhang C, Shokat K, Fisher RP, Bentley DL. 2009. TFIIH-Associated Cdk7 Kinase Functions in Phosphorylation of C-Terminal Domain Ser7 Residues, Promoter-Proximal Pausing, and Termination by RNA Polymerase II. *Mol Cell Biol* **29**:5455–5464. doi:10.1128/mcb.00637-09

Grau D, Zhang Y, Lee CH, Valencia-Sánchez M, Zhang J, Wang M, Holder M, Svetlov V, Tan D, Nudler E, Reinberg D, Walz T, Armache KJ. 2021. Structures of monomeric and dimeric PRC2:EZH1 reveal flexible modules involved in chromatin compaction. *Nat Commun* **12**:1–12. doi:10.1038/s41467-020-20775-z

Gregersen LH, Mitter R, Ugalde AP, Nojima T, Proudfoot NJ, Agami R, Stewart A, Svejstrup JQ. 2019. SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell* **177**:1797-1813.e18. doi:10.1016/j.cell.2019.04.038

Greifenberg AK, Hönig D, Pilarova K, Düster R, Bartholomeeusen K, Bösken CA, Anand K, Blazek D, Geyer M. 2016. Structural and Functional Analysis of the Cdk13/Cyclin K Complex. *Cell Rep* **14**:320–331. doi:10.1016/j.celrep.2015.12.025

Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA. 2007. A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells. *Cell* **130**:77–88. doi:10.1016/j.cell.2007.05.042

Guo S, Yamaguchi Y, Schilbach S, Wada T, Lee J, Goddard A, French D, Handa H, Rosenthal A. 2000. A regulator of transcriptional elongation controls vertebrate neuronal development. *Nature* **408**:366–369. doi:10.1038/35042590

Haberle V, Stark A. 2018. Eukaryotic core promoters and the functional basis of transcription initiation. *Nat Rev Mol Cell Biol*. doi:10.1038/s41580-018-0028-8

Haddad JF, Yang Y, Takahashi Y hei, Joshi M, Chaudhary N, Woodfin AR, Benyoucef A, Yeung S, Brunzelle JS, Skiniotis G, Brand M, Shilatifard A, Couture JF. 2018. Structural Analysis of the Ash2L/Dpy-30 Complex Reveals a Heterogeneity in H3K4 Methylation. *Structure* **26**:1594-1603.e4. doi:10.1016/j.str.2018.08.004

Hao B, Oehlmann S, Sowa ME, Harper JW, Pavletich NP. 2007. Structure of a Fbw7-Skp1-Cyclin E Complex: Multisite-Phosphorylated Substrate Recognition by SCF Ubiquitin Ligases. *Mol Cell* **26**:131–143. doi:10.1016/j.molcel.2007.02.022

Harlen KM, Churchman LS. 2017. The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nat Rev Mol Cell Biol*. doi:10.1038/nrm.2017.10

Hartzog GA, Fu J. 2013. The Spt4-Spt5 complex: A multi-faceted regulator of transcription elongation. *Biochim Biophys Acta - Gene Regul Mech*. doi:10.1016/j.bbagrm.2012.08.007

Hartzog GA, Wada T, Handa H, Winston F. 1998. Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in Saccharomyces cerevisiae. *Genes Dev* **12**:357–369. doi:10.1101/gad.12.3.357

He C, Liu N, Xie D, Liu Y, Xiao Y, Li F. 2019. Structural basis for histone H3K4me3 recognition by the N-terminal domain of the PHD finger protein Spp1. *Biochem J* **476**:1957–1973. doi:10.1042/BCJ20190091

He Y, Yan C, Fang J, Inouye C, Tjian R, Ivanov I, Nogales E. 2016. Near-atomic resolution visualization of human transcription promoter opening. *Nature* **533**:359–365. doi:10.1038/nature17970

Henriques T, Scruggs BS, Inouye MO, Muse GW, Williams LH, Burkholder AB, Lavender CA, Fargo DC, Adelman K. 2018. Widespread transcriptional pausing and elongation control at enhancers. *Genes Dev* **32**:26–41. doi:10.1101/gad.309351.117

Heroes E, Lesage B, Görnemann J, Beullens M, Van Meervelt L, Bollen M. 2013. The PP1 binding code: A molecular-lego strategy that governs specificity. *FEBS J*. doi:10.1111/j.1742-4658.2012.08547.x

Hieb AR, Halsey WA, Betterton MD, Perkins TT, Kugel JF, Goodrich JA. 2007. TFIIA Changes the Conformation of the DNA in TBP/TATA Complexes and Increases their Kinetic Stability. *J Mol Biol* **372**:619–632. doi:10.1016/j.jmb.2007.06.061

Higgs MR, Reynolds JJ, Winczura A, Blackford AN, Borel V, Miller ES, Zlatanou A, Nieminuszczy J, Ryan EL, Davies NJ, Stankovic T, Boulton SJ, Niedzwiedz W, Stewart GS. 2015. BOD1L Is Required to Suppress Deleterious Resection of Stressed Replication Forks. *Mol Cell* **59**:462–477. doi:10.1016/j.molcel.2015.06.007

Higgs MR, Sato K, Reynolds JJ, Begum S, Bayley R, Goula A, Vernet A, Paquin KL, Skalnik DG, Kobayashi W, Takata M, Howlett NG, Kurumizaka H, Kimura H, Stewart GS. 2018. Histone Methylation by SETD1A Protects Nascent DNA through the Nucleosome Chaperone Activity of FANCD2. *Mol Cell* **71**:25-41.e6.

doi:10.1016/j.molcel.2018.05.018

Horikoshi M, Hai T, Lin YS, Green MR, Roeder RG. 1988. Transcription factor ATF interacts with the TATA factor to facilitate establishment of a preinitiation complex. *Cell* **54**:1033–1042. doi:10.1016/0092-8674(88)90118-3

Hou L, Wang Y, Liu Y, Zhang N, Shamovsky I, Nudler E, Tian B, Dynlacht BD. 2019. Paf1C regulates RNA polymerase II progression by modulating elongation rate. *Proc Natl Acad Sci U S A* **116**:14583–14592. doi:10.1073/pnas.1904324116

Howe FS, Fischl H, Murray SC, Mellor J. 2017. Is H3K4me3 instructive for transcription activation? *BioEssays* **39**:1–12. doi:10.1002/bies.201600095

Hsu PL, Shi H, Leonen C, Kang J, Chatterjee C, Zheng N. 2019. Structural Basis of H2B Ubiquitination-Dependent H3K4 Methylation by COMPASS. *Mol Cell* **76**:712-723.e4. doi:10.1016/j.molcel.2019.10.013

Hughes AL, Kelley JR, Klose RJ. 2020a. Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation. *Biochim Biophys Acta - Gene Regul Mech*. doi:10.1016/j.bbagrm.2020.194567

Hughes AL, Kelley JR, Klose RJ. 2020b. Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation. *Biochim Biophys Acta - Gene Regul Mech* **1863**:194567. doi:10.1016/j.bbagrm.2020.194567

Hughes AL, Szczurek AT, Kelley JR, Lastuvkova A, Turberfield AH, Dimitrova E, Blackledge NP, Klose RJ. 2022. A CpG island-encoded mechanism protects genes from premature transcription termination. *bioRxiv* 2022.03.24.485638. doi:10.1101/2022.03.24.485638

Hughes CS, Moggridge S, Müller T, Sorensen PH, Morin GB, Krijgsveld J. 2019. Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat Protoc* **14**:68–85. doi:10.1038/s41596-018-0082-x

Hurley TD, Yang J, Zhang L, Goodwin KD, Zou Q, Cortese M, Dunker AK, DePaoli-Roach AA. 2007. Structural basis for regulation of protein phosphatase 1 by inhibitor-2. *J Biol Chem* **282**:28874–28883. doi:10.1074/jbc.M703472200

Huttlin EL, Jedrychowski MP, Elias JE, Goswami T, Rad R, Beausoleil SA, Villén J, Haas W, Sowa ME, Gygi SP. 2010. A tissue-specific atlas of mouse protein phosphorylation and expression. *Cell* **143**:1174–1189. doi:10.1016/j.cell.2010.12.001

Hyun K, Jeon J, Park K, Kim J. 2017. Writing, erasing and reading histone lysine methylations. *Exp Mol Med*. doi:10.1038/emm.2017.11

Imbalzano AN, Zaret KS, Kingston RE. 1994. Transcription factor (TF) IIB and TFIIA can independently increase the affinity of the TATA-binding protein for DNA. *J Biol Chem* **269**:8280–8286. doi:10.1016/s0021-9258(17)37190-9

Ivanov D, Kwak YT, Guo J, Gaynor RB. 2000. Domains in the SPT5 Protein That Modulate Its Transcriptional Regulatory Properties. *Mol Cell Biol* **20**:2970–2983. doi:10.1128/mcb.20.9.2970-2983.2000

Jagiello I, Beullens M, Stalmans W, Bollen M. 1995. Subunit structure and regulation of protein phosphatase-1 in rat liver nuclei. *J Biol Chem* **270**:17257–17263. doi:10.1074/jbc.270.29.17257

Jain BP, Pandey S. 2018. WD40 Repeat Proteins: Signalling Scaffold with Diverse Functions. *Protein J*. doi:10.1007/s10930-018-9785-7

Jasnovidova O, Stefl R. 2013. The CTD code of RNA polymerase II: A structural view. *Wiley Interdiscip Rev RNA*. doi:10.1002/wrna.1138

Jerebtsova M, Klotchenko SA, Artamonova TO, Ammosova T, Washington K, Egorov V V., Shaldzhyan AA, Sergeeva M V., Zatulovskiy EA, Temkina OA, Petukhov MG, Vasin A V., Khodorkovskii MA, Orlov YN, Nekhai S. 2011. Mass spectrometry and biochemical analysis of RNA polymerase II: Targeting by protein phosphatase-1. *Mol Cell Biochem* **347**:79–87. doi:10.1007/s11010-010-0614-3

Jonkers I, Kwak H, Lis JT. 2014. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* **2014**. doi:10.7554/eLife.02407

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583–589. doi:10.1038/s41586-021-03819-2

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. doi:10.1021/bi401715v

Kadonaga JT. 2012. Perspectives on the RNA polymerase II core promoter. *Wiley Interdiscip Rev Dev Biol*. doi:10.1002/wdev.21

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L, Dreyfuss G. 2010. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468**:664–668. doi:10.1038/nature09479

Kamieniarz-Gdula K, Gdula MR, Panser K, Nojima T, Monks J, Wiśniewski JR, Riepsaame J, Brockdorff N, Pauli A, Proudfoot NJ. 2019. Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Mol Cell* **74**:158-172.e9. doi:10.1016/j.molcel.2019.01.027

Kamieniarz-Gdula K, Proudfoot NJ. 2019. Transcriptional Control by Premature Termination: A Forgotten Mechanism. *Trends Genet*. doi:10.1016/j.tig.2019.05.005

Kaplan CD, Laprade L, Winston F. 2003. Transcription elongation factors repress transcription initiation from cryptic sites. *Science (80-)* **301**:1096–1099. doi:10.1126/science.1087374

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G,

Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (80-)* **316**:1484–1488. doi:10.1126/science.1138341

Kasinath V, Beck C, Sauer P, Poepsel S, Kosmatka J, Faini M, Toso D, Aebersold R, Nogales E. 2021. JARID2 and AEBP2 regulate PRC2 in the presence of H2AK119ub1 and other histone modifications. *Science (80-)* **371**. doi:10.1126/science.abc3393

Katayama S, Tomaru Y, Kasukawa T, Waki K, Nakanishi M, Nakamura M, Nishida H, Yap CC, Suzuki M, Kawai J, Suzuki H, Carninci P, Hayashizaki Y, Wells C, Frith M, Ravasi T, Pang KC, Hallinan J, Mattick J, Hume DA, Lipovich L, Batalov S, Engström PG, Mizuno Y, Faghihi MA, Sandelin A, Chalk AM, Mottagui-Tabar S, Liang Z, Lenhard B, Wahlestedt C. 2005. Molecular biology: Antisense transcription in the mammalian transcriptome. *Science (80-)* **309**:1564–1566. doi:10.1126/science.1112009

Kettenberger H, Armache KJ, Cramer P. 2003. Architecture of the RNA polymerase II-TFIIS complex and implications for mRNA cleavage. *Cell* **114**:347–357. doi:10.1016/S0092-8674(03)00598-1

Kieft R, Zhang Y, Marand AP, Moran JD, Bridger R, Wells L, Schmitz RJ, Sabatini R. 2020. Identification of a novel base J binding protein complex involved in RNA polymerase II transcription termination in trypanosomes. *PLoS Genet* **16**:e1008390. doi:10.1371/journal.pgen.1008390

Kim J, Kim JA, McGinty RK, Nguyen UTT, Muir TW, Allis CD, Roeder RG. 2013. The n-SET Domain of Set1 Regulates H2B Ubiquitylation-Dependent H3K4 Methylation. *Mol Cell* **49**:1121–1133. doi:10.1016/j.molcel.2013.01.034

Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, Greenblatt JF, Buratowski S. 2004. The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* **432**:517–522. doi:10.1038/nature03041

Kim M, Suh H, Cho EJ, Buratowski S. 2009. Phosphorylation of the yeast Rpb1 C-terminal domain at serines 2,5, and 7. *J Biol Chem* **284**:26421–26426. doi:10.1074/jbc.M109.028993

Kim TK, Ebright RH, Reinberg D. 2000. Mechanism of ATP-dependent promoter melting by transcription factor IIH. *Science (80-)* **288**:1418–1421. doi:10.1126/science.288.5470.1418

Kim TK, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G, Greenberg ME. 2010. Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465**:182–187. doi:10.1038/nature09033

Kim Young Mi, Watanabe T, Allen PB, Kim Young Myoung, Lee SJ, Greengard P, Nairn AC, Kwon YG. 2003. PNUTS, a protein Phosphatase 1 (PP1) NUclear Targeting Subunit: Characterization of its PP1 and RNA-binding domains and regulation by phosphorylation. *J Biol Chem* **278**:13819–13828. doi:10.1074/jbc.M209621200

Knezetic JA, Luse DS. 1986. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell* **45**:95–104. doi:10.1016/0092-8674(86)90541-6

Koch F, Fenouil R, Gut M, Cauchy P, Albert TK, Zacarias-Cabeza J, Spicuglia S, De La Chapelle AL, Heidemann M, Hintermair C, Eick D, Gut I, Ferrier P, Andrau JC. 2011. Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. *Nat Struct Mol Biol* **18**:956–963. doi:10.1038/nsmb.2085

Koepp DM. 2001. Phosphorylation-Dependent Ubiquitination of Cyclin E by the SCFFbw7 Ubiquitin Ligase. *Science (80-)* **294**:173–177. doi:10.1126/science.1065203

Kornberg RD, Lorch Y. 1999. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*. doi:10.1016/S0092-8674(00)81958-3

Kostrewa D, Zeller ME, Armache KJ, Seizl M, Leike K, Thomm M, Cramer P. 2009. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* **462**:323–330. doi:10.1038/nature08548

Krebs AR, Imanci D, Hoerner L, Gaidatzis D, Burger L, Schübeler D. 2017. Genome-wide Single-Molecule Footprinting Reveals High RNA Polymerase II Turnover at Paused Promoters. *Mol Cell* **67**:411-422.e4. doi:10.1016/j.molcel.2017.06.027

Kreivi JP, Trinkle-Mulcahy L, Lyon CE, Morrice NA, Cohen P, Lamond AI. 1997. Purification and characterisation of p99, a nuclear modulator of protein phosphatase 1 activity. *FEBS Lett* **420**:57–62. doi:10.1016/S0014-5793(97)01485-3

Krissinel E, Henrick K. 2007. Inference of Macromolecular Assemblies from Crystalline State. *J Mol Biol* **372**:774–797. doi:10.1016/j.jmb.2007.05.022

Kwak H, Fuda NJ, Core LJ, Lis JT. 2013. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science (80-)* **339**:950–953. doi:10.1126/science.1229386

Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR, Weirauch MT. 2018. The Human Transcription Factors. *Cell*. doi:10.1016/j.cell.2018.01.029

Landsverk HB, Sandquist LE, Bay LTE, Steurer B, Campsteijn C, Landsverk OJB, Marteijn JA, Petermann E, Trinkle-Mulcahy L, Syljuåsen RG. 2020. WDR82/PNUTS-PP1 Prevents Transcription-Replication Conflicts by Promoting RNA Polymerase II Degradation on Chromatin. *Cell Rep* **33**:108469. doi:10.1016/j.celrep.2020.108469

Landsverk HB, Sandquist LE, Sridhara SC, Rødland GE, Sabino JC, De Almeida SF, Grallert B, Trinkle-Mulcahy L, Syljuasen RG. 2019. Regulation of ATR activity via the RNA polymerase II associated factors CDC73 and PNUTS-PP1. *Nucleic Acids Res* **47**:1797–1813. doi:10.1093/nar/gky1233

Larochelle S, Amat R, Glover-Cutter K, Sansó M, Zhang C, Allen JJ, Shokat KM, Bentley DL, Fisher RP. 2012. Cyclin-dependent kinase control of the initiation-to-elongation switch of RNA polymerase II. *Nat Struct Mol Biol* **19**:1108–1115. doi:10.1038/nsmb.2399

Lee J-H, Skalnik DG. 2008. Wdr82 Is a C-Terminal Domain-Binding Protein That Recruits the Setd1A Histone H3-Lys4 Methyltransferase Complex to Transcription Start Sites of Transcribed Human Genes. *Mol Cell Biol* **28**:609–618. doi:10.1128/mcb.01356-07

Lee JH, Skalnik DG. 2005. CpG-binding protein (CXXC finger protein 1) is a component of the

mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J Biol Chem* **280**:41725–41731. doi:10.1074/jbc.M508312200

Lee JH, Tate CM, You JS, Skalnik DG. 2007. Identification and characterization of the human Set1B histone H3-Lys 4 methyltransferase complex. *J Biol Chem* **282**:13419–13428. doi:10.1074/jbc.M609809200

Lee JH, Voo KS, Skalnik DG. 2001. Identification and Characterization of the DNA Binding Domain of CpG-binding Protein. *J Biol Chem* **276**:44669–44676. doi:10.1074/jbc.M107179200

Lee JH, You J, Dobrota E, Skalnik DG. 2010. Identification and characterization of a novel human PP1 phosphatase complex. *J Biol Chem* **285**:24466–24476. doi:10.1074/jbc.M110.109801

Lee SJ, Lee JK, Maeng YS, Kim YM, Kwon YG. 2009. Langerhans cell protein 1 (LCP1) binds to PNUTS in the nucleus: Implications for this complex in transcriptional regulation. *Exp Mol Med* **41**:189–200. doi:10.3858/emm.2009.41.3.022

Li W, You B, Hoque M, Zheng D, Luo W, Ji Z, Park JY, Gunderson SI, Kalsotra A, Manley JL, Tian B. 2015. Systematic Profiling of Poly(A)+ Transcripts Modulated by Core 3' End Processing and Splicing Factors Reveals Regulatory Rules of Alternative Cleavage and Polyadenylation. *PLoS Genet* **11**:e1005166. doi:10.1371/journal.pgen.1005166

Liu P, Kenney JM, Stiller JW, Greenleaf AL. 2010. Genetic organization, length conservation, and evolution of RNA polymerase II carboxyl-terminal domain. *Mol Biol Evol* **27**:2628–2641. doi:10.1093/molbev/msq151

Liu X, Bushnell DA, Wang D, Calero G, Kornbergt RD. 2010. Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. *Science (80-)* **327**:206–209. doi:10.1126/science.1182015

Liu Z, Wu A, Wu Z, Wang T, Pan Y, Li B, Zhang X, Yu M. 2022. TOX4 facilitates promoter-proximal pausing and C-terminal domain dephosphorylation of RNA polymerase II in human cells. *Commun Biol* **5**:1–15. doi:10.1038/s42003-022-03214-1

Lopez T, Dalton K, Frydman J. 2015. The Mechanism and Function of Group II Chaperonins. *J Mol Biol*. doi:10.1016/j.jmb.2015.04.013

Lorch Y, LaPointe JW, Kornberg RD. 1987. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell* **49**:203–210. doi:10.1016/0092-8674(87)90561-7

Louder RK, He Y, López-Blanco JR, Fang J, Chacón P, Nogales E. 2016. Structure of promoter-bound TFIID and model of human pre-initiation complex assembly. *Nature* **531**:604–609. doi:10.1038/nature17394

Lu H, Flores O, Weinmann R, Reinberg D. 1991. The nonphosphorylated form of RNA polymerase II preferentially associates with the preinitiation complex. *Proc Natl Acad Sci U S A* **88**:10004–10008. doi:10.1073/pnas.88.22.10004

Lu X, Zhu X, Li Y, Liu M, Yu B, Wang Yu, Rao M, Yang H, Zhou K, Wang Yao, Chen Y, Chen M,

Zhuang S, Chen LF, Liu R, Chen R. 2016. Multiple P-TEFbs cooperatively regulate the release of promoter-proximally paused RNA polymerase II. *Nucleic Acids Res* **44**:6853–6867. doi:10.1093/nar/gkw571

Lykke-Andersen S, Žumer K, Molska EŠ, Rouvière JO, Wu G, Demel C, Schwalb B, Schmid M, Cramer P, Jensen TH. 2021. Integrator is a genome-wide attenuator of non-productive transcription. *Mol Cell* **81**:514-529.e6. doi:10.1016/j.molcel.2020.12.014

Ma J, An K, Zhou JB, Wu NS, Wang Y, Ye ZQ, Wu YD. 2019. WDSPdb: An updated resource for WD40 proteins. *Bioinformatics* **35**:4824–4826. doi:10.1093/bioinformatics/btz460

Margueron R, Justin N, Ohno K, Sharpe ML, Son J, Drury WJ, Voigt P, Martin SR, Taylor WR, De Marco V, Pirrotta V, Reinberg D, Gamblin SJ. 2009. Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature* **461**:762–767. doi:10.1038/nature08398

Marshall NF, Peng J, Xie Z, Price DH. 1996. Control of RNA polymerase II elongation potential by a novel carboxyl- terminal domain kinase. *J Biol Chem* **271**:27176–27183. doi:10.1074/jbc.271.43.27176

Marshall NF, Price DH. 1995. Purification of P-TEFb, a transcription factor required for the transition into productive elongation. *J Biol Chem* **270**:12335–12338. doi:10.1074/jbc.270.21.12335

Mehta V, Chamousset D, Law J, Ooi S, Campuzano D, Nguyen V, Boisvert F-M, Moorhead GB, Trinkle-Mulcahy L. 2022. Subcellular distribution of PP1 isoforms in holoenzyme complexes. *bioRxiv* 2022.09.09.507380. doi:10.1101/2022.09.09.507380

Meinhart A, Kamenski T, Hoeppner S, Baumli S, Cramer P. 2005. A structural perspective of CTD function. *Genes Dev*. doi:10.1101/gad.1318105

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, Lander ES, Bernstein BE. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**:553–560. doi:10.1038/nature06008

Miller T, Krogan NJ, Dover J, Erdjument-Bromage H, Tempst P, Johnston M, Greenblatt JF, Shilatifard A. 2001. COMPASS: A complex of proteins associated with a trithorax-related SET domain protein. *Proc Natl Acad Sci U S A* **98**:12902–12907. doi:10.1073/pnas.231473398

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nat Methods* **19**:679–682. doi:10.1038/s41592-022-01488-1

Missra A, Gilmour DS. 2010. Interactions between DSIF (DRB sensitivity inducing factor), NELF (negative elongation factor), and the Drosophila RNA polymerase II transcription elongation complex. *Proc Natl Acad Sci U S A* **107**:11301–11306. doi:10.1073/pnas.1000681107

Morgan MAJ, Shilatifard A. 2020. Reevaluating the roles of histone-modifying enzymes and their associated chromatin modifications in transcriptional regulation. *Nat Genet*

52:1271–1281. doi:10.1038/s41588-020-00736-4

Murton BL, Chin WL, Ponting CP, Itzhaki LS. 2010. Characterising the Binding Specificities of the Subunits Associated with the KMT2/Set1 Histone Lysine Methyltransferase. *J Mol Biol* **398**:481–488. doi:10.1016/j.jmb.2010.03.036

Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. 2007. RNA polymerase is poised for activation across the genome. *Nat Genet* **39**:1507–1511. doi:10.1038/ng.2007.21

Ng HH, Robert F, Young RA, Struhl K. 2003. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell* **11**:709–719. doi:10.1016/S1097-2765(03)00092-3

Notredame C, Higgins DG, Heringa J. 2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**:205–217. doi:10.1006/jmbi.2000.4042

Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, Andersson R, Hoof I, Schein A, Andersen PR, Andersen PK, Preker P, Valen E, Zhao X, Pelechano V, Steinmetz LM, Sandelin A, Jensen TH. 2013. Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol* **20**:923–928. doi:10.1038/nsmb.2640

O'Connell N, Nichols SR, Heroes E, Beullens M, Bollen M, Peti W, Page R. 2012. The molecular basis for substrate specificity of the nuclear NIPP1:PP1 holoenzyme. *Structure* **20**:1746–1756. doi:10.1016/j.str.2012.08.003

Odho Z, Southall SM, Wilson JR. 2010. Characterization of a novel WDR5-binding site that recruits RbBP5 through a conserved motif to enhance methylation of histone H3 lysine 4 by mixed lineage leukemia protein-1. *J Biol Chem* **285**:32967–32976. doi:10.1074/jbc.M110.159921

Oh JM, Di C, Venters CC, Guo J, Arai C, So BR, Pinto AM, Zhang Z, Wan L, Younis I, Dreyfuss G. 2017. U1 snRNP telescripting regulates a size-function-stratified human genome. *Nat Struct Mol Biol* **24**:993–999. doi:10.1038/nsmb.3473

Oldfield CJ, Dunker AK. 2014. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* **83**:553–584. doi:10.1146/annurev-biochem-072711-164947

Palangat M, Renner DB, Price DH, Landick R. 2005. A negative elongation factor for human RNA polymerase II inhibits the anti-arrest transcript-cleavage factor TFIIS. *Proc Natl Acad Sci U S A* **102**:15036–15041. doi:10.1073/pnas.0409405102

Park K, Zhong J, Jang JS, Kim Jihyun, Kim HJ, Lee JH, Kim Jaehoon. 2022. ZWC complex-mediated SPT5 phosphorylation suppresses divergent antisense RNA transcription at active gene promoters. *Nucleic Acids Res* **50**:3835–3851. doi:10.1093/nar/gkac193

Parua PK, Booth GT, Sansó M, Benjamin B, Tanny JC, Lis JT, Fisher RP. 2018. A Cdk9-PP1 switch regulates the elongation-termination transition of RNA polymerase II. *Nature* **558**:460–464. doi:10.1038/s41586-018-0214-z

Parua PK, Fisher RP. 2020. Dissecting the Pol II transcription cycle and derailing cancer with

CDK inhibitors. *Nat Chem Biol*. doi:10.1038/s41589-020-0563-4

Parua PK, Kalan S, Benjamin B, Sansó M, Fisher RP. 2020. Distinct Cdk9-phosphatase switches act at the beginning and end of elongation by RNA polymerase II. *Nat Commun* **11**:1–13. doi:10.1038/s41467-020-18173-6

Patel AB, Greber BJ, Nogales E. 2020. Recent insights into the structure of TFIID, its assembly, and its binding to core promoter. *Curr Opin Struct Biol*. doi:10.1016/j.sbi.2019.10.001

Patel AB, Louder RK, Greber BJ, Grünberg S, Luo J, Fang J, Liu Y, Ranish J, Hahn S, Nogales E. 2018. Structure of human TFIID and mechanism of TBP loading onto promoter DNA. *Science (80-)* **362**. doi:10.1126/science.aau8872

PDBe < PISA < EMBL-EBI. n.d. https://www.ebi.ac.uk/pdbe/pisa/pistart.html

Pei Y, Shuman S. 2002. Interactions between fission yeast mRNA capping enzymes and elongation factor Spt5. *J Biol Chem* **277**:19639–19648. doi:10.1074/jbc.M200015200

Porrua O, Libri D. 2015. Transcription termination and the control of the transcriptome: Why, where and how to stop. *Nat Rev Mol Cell Biol*. doi:10.1038/nrm3943

Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH, Jensen TH. 2008. RNA exosome depletion reveals transcription upstream of active human promoters. *Science (80-)* **322**:1851–1854. doi:10.1126/science.1164096

Proudfoot NJ. 2016. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science (80-)*. doi:10.1126/science.aad9926

Proudfoot NJ. 2011. Ending the message: Poly(A) signals then and now. *Genes Dev*. doi:10.1101/gad.17268411

Qiu Y, Gilmour DS. 2017. Identification of regions in the Spt5 subunit of DRB sensitivity-inducing factor (DSIF) that are involved in promoter-proximal pausing. *J Biol Chem* **292**:5555–5570. doi:10.1074/jbc.M116.760751

Qu Q, Takahashi Y hei, Yang Y, Hu H, Zhang Y, Brunzelle JS, Couture JF, Shilatifard A, Skiniotis G. 2018. Structure and Conformational Dynamics of a COMPASS Histone H3K4 Methyltransferase Complex. *Cell* **174**:1117-1126.e12. doi:10.1016/j.cell.2018.07.020

Quevedo M, Meert L, Dekker MR, Dekkers DHW, Brandsma JH, van den Berg DLC, Ozgür Z, IJcken WFJ va., Demmers J, Fornerod M, Poot RA. 2019. Mediator complex interaction partners organize the transcriptional network that defines neural stem cells. *Nat Commun* **10**:1–15. doi:10.1038/s41467-019-10502-8

Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, Sharp PA, Young RA. 2010. C-Myc regulates transcriptional pause release. *Cell* **141**:432–445. doi:10.1016/j.cell.2010.03.030

Rasmussen EB, Lis JT. 1993. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci U S A* **90**:7923–7927.

doi:10.1073/pnas.90.17.7923

Rebelo S, Santos M, Martins F, da Cruz e Silva EF, da Cruz e Silva OAB. 2015. Protein phosphatase 1 is a key player in nuclear events. *Cell Signal*. doi:10.1016/j.cellsig.2015.08.007

Richard P, Manley JL. 2009. Transcription termination by nuclear RNA polymerases. *Genes Dev*. doi:10.1101/gad.1792809

Robinson PJ, Trnka MJ, Bushnell DA, Davis RE, Mattei PJ, Burlingame AL, Kornberg RD. 2016. Structure of a Complete Mediator-RNA Polymerase II Pre-Initiation Complex. *Cell* **166**:1411-1422.e16. doi:10.1016/j.cell.2016.08.050

Roeder RG, Rutter WJ. 1969. Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature* **224**:234–237. doi:10.1038/224234a0

Rothbart SB, Strahl BD. 2014. Interpreting the language of histone and DNA modifications. *Biochim Biophys Acta - Gene Regul Mech*. doi:10.1016/j.bbagrm.2014.03.001

Rougvie AE, Lis JT. 1988. The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* **54**:795–804. doi:10.1016/S0092-8674(88)91087-2

Schapira M, Tyers M, Torrent M, Arrowsmith CH. 2017. WD40 repeat domain proteins: A novel target class? *Nat Rev Drug Discov*. doi:10.1038/nrd.2017.179

Schier AC, Taatjes DJ. 2020. Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev*. doi:10.1101/gad.335679.119

Schmid M, Jensen TH. 2018. Controlling nuclear RNA levels. *Nat Rev Genet*. doi:10.1038/s41576-018-0013-2

Schulze WM, Stein F, Rettel M, Nanao M, Cusack S. 2018. Structural analysis of human ARS2 as a platform for co-transcriptional RNA sorting. *Nat Commun* **9**:1–15. doi:10.1038/s41467-018-04142-7

Sentenac A. 1985. Eukaryotic RNA polymerase. *Crit Rev Biochem Mol Biol* **18**:31–90. doi:10.3109/10409238509082539

Shao W, Zeitlinger J. 2017. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat Genet* **49**:1045–1051. doi:10.1038/ng.3867

Shetty A, Kallgren SP, Demel C, Maier KC, Spatt D, Alver BH, Cramer P, Park PJ, Winston F. 2017. Spt5 Plays Vital Roles in the Control of Sense and Antisense Transcription Elongation. *Mol Cell* **66**:77-88.e5. doi:10.1016/j.molcel.2017.02.023

Shi X, Finkelstein A, Wolf AJ, Wade PA, Burton ZF, Jaehning JA. 1996. Paf1p, an RNA polymerase II-associated factor in Saccharomyces cerevisiae, may have both positive and negative roles in transcription. *Mol Cell Biol* **16**:669–676. doi:10.1128/mcb.16.2.669

Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR, Frank J, Manley JL. 2009. Molecular Architecture of the Human Pre-mRNA 3' Processing Complex. *Mol Cell*

33:365–376. doi:10.1016/j.molcel.2008.12.028

Shinsky SA, Monteith KE, Viggiano S, Cosgrove MS. 2015. Biochemical reconstitution and phylogenetic comparison of human SET1 family core complexes involved in histone methylation. *J Biol Chem* **290**:6361–6375. doi:10.1074/jbc.M114.627646

Shlyueva D, Stampfel G, Stark A. 2014. Transcriptional enhancers: From properties to genome-wide predictions. *Nat Rev Genet*. doi:10.1038/nrg3682

Skaar JR, D'Angiolella V, Pagan JK, Pagano M. 2009. SnapShot: F Box Proteins II. *Cell* **137**:5–6. doi:10.1016/j.cell.2009.05.040

Skaar JR, Pagan JK, Pagano M. 2013. Mechanisms and function of substrate recruitment by F-box proteins. *Nat Rev Mol Cell Biol* **14**:369–81. doi:10.1038/nrm3582

Soares LM, Buratowski S. 2012. Yeast Swd2 is essential because of antagonism between Set1 histone methyltransferase complex and APT (associated with Pta1) termination factor. *J Biol Chem* **287**:15219–15231. doi:10.1074/jbc.M112.341412

Soutourina J. 2018. Transcription regulation by the Mediator complex. *Nat Rev Mol Cell Biol*. doi:10.1038/nrm.2017.115

Spitz F, Furlong EEM. 2012. Transcription factors: From enhancer binding to developmental control. *Nat Rev Genet*. doi:10.1038/nrg3207

Srivastava R, Ahn SH. 2015. Modifications of RNA polymerase II CTD: Connections to the histone code and cellular function. *Biotechnol Adv*. doi:10.1016/j.biotechadv.2015.07.008

Steurer B, Janssens RC, Geverts B, Geijer ME, Wienholz F, Theil AF, Chang J, Dealy S, Pothof J, Van Cappellen WA, Houtsmuller AB, Marteijn JA. 2018. Live-cell analysis of endogenous GFP-RPB1 uncovers rapid turnover of initiating and promoter-paused RNA Polymerase II. *Proc Natl Acad Sci U S A* **115**:E4368–E4376. doi:10.1073/pnas.1717920115

Stirnimann CU, Petsalaki E, Russell RB, Müller CW. 2010. WD40 proteins propel cellular networks. *Trends Biochem Sci* **35**:565–574. doi:10.1016/j.tibs.2010.04.003

Su J, Miao X, Archambault D, Mager J, Cui W. 2021. ZC3H4-a novel Cys-Cys-Cys-His-type zinc finger protein-is essential for early embryogenesis in mice. *Biol Reprod* **104**:325–335. doi:10.1093/biolre/ioaa215

Sun Y, Zhang Y, Hamilton K, Manley JL, Shi Y, Walz T, Tong L. 2018. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc Natl Acad Sci U S A* **115**:E1419–E1428. doi:10.1073/pnas.1718723115

Syrovatkina V, Alegre KO, Dey R, Huang XY. 2016. Regulation, Signaling, and Physiological Functions of G-Proteins. *J Mol Biol*. doi:10.1016/j.jmb.2016.08.002

Sze CC, Cao K, Collings CK, Marshall SA, Rendleman EJ, Ozark PA, Chen FX, Morgan MA, Wang L, Shilatifard A. 2017. Histone H3K4 methylation-dependent and -independent functions of set1A/COMPASS in embryonic stem cell self-renewal and differentiation. *Genes Dev* **31**:1732–1737. doi:10.1101/gad.303768.117

Sze CC, Ozark PA, Cao K, Ugarenko M, Das S, Wang L, Marshall SA, Rendleman EJ, Ryan CA, Zha D, Douillet D, Chen FX, Shilatifard A. 2020. Coordinated regulation of cellular identity–associated H3K4me3 breadth by the COMPASS family. *Sci Adv* **6**:eaaz4764. doi:10.1126/sciadv.aaz4764

Tatomer DC, Elrod ND, Liang D, Xiao MS, Jiang JZ, Jonathan M, Huang KL, Wagner EJ, Cherry S, Wilusz JE. 2019. The Integrator complex cleaves nascent mRNAs to attenuate transcription. *Genes Dev* **33**:1525–1538. doi:10.1101/gad.330167.119

Ter Haar E, Harrison SC, Kirchhausen T. 2000. Peptide-in-groove interactions link target proteins to the β-propeller of clathrin. *Proc Natl Acad Sci U S A* **97**:1096–1100. doi:10.1073/pnas.97.3.1096

Thomson JP, Skene PJ, Selfridge J, Clouaire T, Guy J, Webb S, Kerr ARW, Deaton A, Andrews R, James KD, Turner DJ, Illingworth R, Bird A. 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* **464**:1082–6. doi:10.1038/nature08924

Tirode F, Busso D, Coin F, Egly JM. 1999. Reconstitution of the transcription factor TFIIH: Assignment of functions for the three enzymatic subunits, XPB, XPD, and cdk7. *Mol Cell* **3**:87–95. doi:10.1016/S1097-2765(00)80177-X

Tropberger P, Schneider R. 2010. Going global: Novel histone modifications in the globular domain of H3. *Epigenetics* **5**:112–117. doi:10.4161/epi.5.2.11075

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, Velankar S, Kleywegt GJ, Bateman A, Evans R, Pritzel A, Figurnov M, Ronneberger O, Bates R, Kohl SAA, Potapenko A, Ballard AJ, Romera-Paredes B, Nikolov S, Jain R, Clancy E, Reiman D, Petersen S, Senior AW, Kavukcuoglu K, Birney E, Kohli P, Jumper J, Hassabis D. 2021. Highly accurate protein structure prediction for the human proteome. *Nature* **596**:590–596. doi:10.1038/s41586-021-03828-1

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. 2016. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods*. doi:10.1038/nmeth.3901

Van Dyke MW, Roeder RG, Sawadogo M. 1988. Physical analysis of transcription preinitiation complex assembly on a class II gene promoter. *Science (80-)* **241**:1335–1338. doi:10.1126/science.3413495

van Nuland R, Smits AH, Pallaki P, Jansen PWTC, Vermeulen M, Timmers HTM. 2013. Quantitative Dissection and Stoichiometry Determination of the Human SET1/MLL Histone Methyltransferase Complexes. *Mol Cell Biol* **33**:2067–2077. doi:10.1128/mcb.01742-12

Van Oss SB, Shirra MK, Bataille AR, Wier AD, Yen K, Vinayachandran V, Byeon IJL, Cucinotta CE, Héroux A, Jeon J, Kim J, VanDemark AP, Pugh BF, Arndt KM. 2016. The Histone Modification Domain of Paf1 Complex Subunit Rtf1 Directly Stimulates H2B Ubiquitylation through an Interaction with Rad6. *Mol Cell* **64**:815–825. doi:10.1016/j.molcel.2016.10.008

Vanoosthuyse V, Legros P, van der Sar SJA, Yvert G, Toda K, Le Bihan T, Watanabe Y, Hardwick K, Bernard P. 2014. CPF-Associated Phosphatase Activity Opposes

Condensin-Mediated Chromosome Condensation. *PLoS Genet* **10**. doi:10.1371/journal.pgen.1004415

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, Zídek A, Green T, Tunyasuvunakool K, Petersen S, Jumper J, Clancy E, Green R, Vora A, Lutfi M, Figurnov M, Cowie A, Hobbs N, Kohli P, Kleywegt G, Birney E, Hassabis D, Velankar S. 2022. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res* **50**:D439–D444. doi:10.1093/nar/gkab1061

Verbinnen I, Ferreira M, Bollen M. 2017. Biogenesis and activity regulation of protein phosphatase 1. *Biochem Soc Trans*. doi:10.1042/BST20160154

Verheyen T, Görnemann J, Verbinnen I, Boens S, Beullens M, Van Eynde A, Bollen M. 2015. Genome-wide promoter binding profiling of protein phosphatase-1 and its major nuclear targeting subunits. *Nucleic Acids Res* **43**:5771–5784. doi:10.1093/nar/gkv500

Villén J, Beausoleil SA, Gerber SA, Gygi SP. 2007. Large-scale phosphorylation analysis of mouse liver. *Proc Natl Acad Sci U S A* **104**:1488–1493. doi:10.1073/pnas.0609836104

Vlaming H, Mimoso CA, Field AR, Martin BJE, Adelman K. 2022. Screening thousands of transcribed coding and non-coding regions reveals sequence determinants of RNA polymerase II elongation potential. *Nat Struct Mol Biol* **29**:613–620. doi:10.1038/s41594-022-00785-9

Voo KS, Carlone DL, Jacobsen BM, Flodin A, Skalnik DG. 2000. Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *MolCell Biol* **20**:2108–2121. doi:10.1128/MCB.20.6.2108-2121.2000

Vos SM, Farnung L, Boehning M, Wigge C, Linden A, Urlaub H, Cramer P. 2018a. Structure of activated transcription complex Pol II–DSIF–PAF–SPT6. *Nature* **560**:607–612. doi:10.1038/s41586-018-0440-4

Vos SM, Farnung L, Urlaub H, Cramer P. 2018b. Structure of paused transcription complex Pol II–DSIF–NELF. *Nature* **560**:601–606. doi:10.1038/s41586-018-0442-2

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

Wall MA, Coleman DE, Lee E, Iñiguez-Lluhi JA, Posner BA, Gilman AG, Sprang SR. 1995. The structure of the G protein heterotrimer Giα1β1γ2. *Cell* **83**:1047–1058. doi:10.1016/0092-8674(95)90220-1

Wang L, Collings CK, Zhao Z, Cozzolino KA, Ma Q, Liang K, Marshall SA, Sze CC, Hashizume R, Savas JN, Shilatifard A. 2017. A cytoplasmic COMPASS is necessary for cell survival and triple-negative breast cancer pathogenesis by regulating metabolism. *Genes Dev* **31**:2056–2066. doi:10.1101/gad.306092.117

Wang Z, Song A, Xu H, Hu S, Tao B, Peng L, Wang J, Li J, Yu J, Wang L, Li Z, Chen X, Wang M, Chi Y, Wu J, Xu Y, Zheng H, Chen FX. 2022. Coordinated regulation of RNA polymerase

II pausing and elongation progression by PAF1. *Sci Adv* **8**. doi:10.1126/sciadv.abm5504

Washington K, Ammosova T, Beullens M, Jerebtsova M, Kumar A, Bollen M, Nekhai S. 2002. Protein phosphatase-1 dephosphorylates the C-terminal domain of RNa polymerase-II. *J Biol Chem* **277**:40442–40448. doi:10.1074/jbc.M205687200

Weber CM, Ramachandran S, Henikoff S. 2014. Nucleosomes are context-specific, H2A.Z-Modulated barriers to RNA polymerase. *Mol Cell* **53**:819–830. doi:10.1016/j.molcel.2014.02.014

Wei Y, Redel C, Ahlner A, Lemak A, Johansson-Åkhe I, Houliston S, Kenney TMG, Tamachi A, Morad V, Duan S, Andrews DW, Wallner B, Sunnerhagen M, Arrowsmith CH, Penn LZ. 2022. The MYC oncoprotein directly interacts with its chromatin cofactor PNUTS to recruit PP1 phosphatase. *Nucleic Acids Res* **50**:3505–3522. doi:10.1093/nar/gkac138

Wen Y, Shatkin AJ. 1999. Transcription elongation factor hSPT5 stimulates mRNA capping. *Genes Dev* **13**:1774–1779. doi:10.1101/gad.13.14.1774

West S, Gromak N, Proudfoot NJ. 2004. Human 5′ → 3′ exonuclease Xm2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* **432**:522–525. doi:10.1038/nature03035

Wong KH, Jin Y, Struhl K. 2014. TFIIH Phosphorylation of the Pol II CTD Stimulates Mediator Dissociation from the Preinitiation Complex and Promoter Escape. *Mol Cell* **54**:601–612. doi:10.1016/j.molcel.2014.03.024

Worden EJ, Zhang X, Wolberger C. 2020. Structural basis for COMPASS recognition of an H2B-ubiquitinated nucleosome. *Elife* **9**. doi:10.7554/eLife.53199

Workman JL, Kingston RE. 1998. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu Rev Biochem*. doi:10.1146/annurev.biochem.67.1.545

Wu CH, Yamaguchi Y, Benjamin LR, Horvat-Gordon M, Washinsky J, Enerly E, Larsson J, Lambertsson A, Handa H, Gilmour D. 2003. NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes Dev* **17**:1402–1414. doi:10.1101/gad.1091403

Wu D, De Wever V, Derua R, Winkler C, Beullens M, Van Eynde A, Bollen M. 2018. A substrate-trapping strategy for protein phosphatase PP1 holoenzymes using hypoactive subunit fusions. *J Biol Chem* **293**:15152–15162. doi:10.1074/jbc.RA118.004132

Wu M, Wang PF, Lee JS, Martin-Brown S, Florens L, Washburn M, Shilatifard A. 2008. Molecular Regulation of H3K4 Trimethylation by Wdr82, a Component of Human Set1/COMPASS. *Mol Cell Biol* **28**:7337–7344. doi:10.1128/mcb.00976-08

Wysocka J, Myers MP, Laherty CD, Eisenman RN, Herr W. 2003. Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1. *Genes Dev* **17**:896–911. doi:10.1101/gad.252103

Xing Z, Lin A, Li C, Liang K, Wang S, Liu Y, Park PK, Qin L, Wei Y, Hawke DH, Hung MC, Lin C, Yang L. 2014. LncRNA directs cooperative epigenetic regulation downstream of

chemokine signals. *Cell* **159**:1110–1125. doi:10.1016/j.cell.2014.10.013

Xu C, Bian C, Lam R, Dong A, Min J. 2011. The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. *Nat Commun* **2**:227. doi:10.1038/ncomms1237

Xu C, Min J. 2011. Structure and function of WD40 domain proteins. *Protein Cell*. doi:10.1007/s13238-011-1018-1

Yamada T, Yamaguchi Y, Inukai N, Okamoto S, Mura T, Handa H. 2006. P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation. *Mol Cell* **21**:227–237. doi:10.1016/j.molcel.2005.11.024

Yamaguchi Y, Takagi T, Wada T, Yano K, Furuya A, Sugimoto S, Hasegawa J, Handa H. 1999. NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation. *Cell* **97**:41–51. doi:10.1016/S0092-8674(00)80713-8

Yang Y, Joshi M, Takahashi YH, Ning Z, Qu Q, Brunzelle JS, Skiniotis G, Figeys D, Shilatifard A, Couture JF. 2020. A non-canonical monovalent zinc finger stabilizes the integration of Cfp1 into the H3K4 methyltransferase complex COMPASS. *Nucleic Acids Res* **48**:421–431. doi:10.1093/nar/gkz1037

Zatreanu D, Han Z, Mitter R, Tumini E, Williams H, Gregersen L, Dirac-Svejstrup AB, Roma S, Stewart A, Aguilera A, Svejstrup JQ. 2019. Elongation Factor TFIIS Prevents Transcription Stress and R-Loop Accumulation to Maintain Genome Stability. *Mol Cell* **76**:57-69.e9. doi:10.1016/j.molcel.2019.07.037

Zeitlinger J, Stark A, Kellis M, Hong JW, Nechaev S, Adelman K, Levine M, Young RA. 2007. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **39**:1512–1516. doi:10.1038/ng.2007.26

Zhang H, Li M, Gao Y, Jia C, Pan X, Cao P, Zhao X, Zhang J, Chang W. 2014. Structural implications of Dpy30 oligomerization for MLL/SET1 COMPASS H3K4 trimethylation. *Protein Cell* **6**:147–151. doi:10.1007/s13238-014-0127-z

Zhang P, Chaturvedi CP, Tremblay V, Cramet M, Brunzelle JS, Skiniotis G, Brand M, Shilatifard A, Couture JF. 2015. A phosphorylation switch on RbBP5 regulates histone H3 Lys4 methylation. *Genes Dev* **29**:123–128. doi:10.1101/gad.254870.114

Zhang P, Lee H, Brunzelle JS, Couture JF. 2012. The plasticity of WDR5 peptide-binding cleft enables the binding of the SET1 family of histone methyltransferases. *Nucleic Acids Res* **40**:4237–4246. doi:10.1093/nar/gkr1235

Zhang Y, Sun Y, Shi Y, Walz T, Tong L. 2020. Structural Insights into the Human Pre-mRNA 3'-End Processing Machinery. *Mol Cell* **77**:800-809.e6. doi:10.1016/j.molcel.2019.11.005

Zhou Q, Li T, Price DH. 2012. RNA polymerase II elongation control. *Annu Rev Biochem* **81**:119–143. doi:10.1146/annurev-biochem-052610-095910