Investigating H3K4 Methylation by SET1/MLL Complexes

Ho Yu Alan Au

Wolfson College

University of Oxford

A thesis submitted in fulfilment for the degree of

Doctor of Philosophy

Michaelmas Term 2023

Abstract

H3K4 methylation is a histone modification which often correlates well with active transcription. In mammals, the SET1/MLL complexes are responsible for H3K4 methylation. The conventional view holds that SET1A/B complexes deposit the majority of H3K4me3 at highly-transcribed genes, MLL1/2 complexes deposit H3K4me3 at a subset of lowly-transcribed genes, and MLL3/4 complexes deposit H3K4me1/2 at enhancers. In mammals, H3K4me3 is found at CpG islands - short stretches of DNA with high CpG density that remain unmethylated in an otherwise pervasively methylated genome. The SET1A/B and MLL1/2 complexes all contain a zinc-finger CXXC domain that targets them specifically to CpG islands. Over 70% of all promoters contain a CpG island, and SET1A/B and MLL1/2 complexes were thought to promote transcription by creating a permissive chromatin state through H3K4me3 deposition. However, recent work has shown that SET1A/B complexes can support transcription through an H3K4me3-independent mechanism, and their removal results in surprisingly modest reductions in H3K4me3. Therefore, our understanding of the how SET1/MLL complexes deposit H3K4 methylation and how they regulate transcription is lacking.

Here, I aim to understand how SET1/MLL complexes methylate H3K4 in mouse embryonic stem cells (mESCs) using degrons to rapidly deplete SET1/MLL proteins. Using cell lines in which SET1A/B, MLL1/2, and MLL3/4 can be rapidly depleted on their own, I dissect their contributions to H3K4 methylation. I find that SET1A/B and MLL1/2 have modest contributions to H3K4me3 on their own, but rather synergise to deposit half of all H3K4me3. Using a cell line in which these complexes can be simultanously depleted together, I find that SET1/MLL methyltransferases collectively implement half of all H3K4 methylation in mESCs. I find that the remaining H3K4 methylation is partially dependent on transcription and entirely dependent on shared subunits of SET1/MLL complexes. Surprisingly, I identify a previously uncharacterized isoform of SET1B as a putative candidate for this methyltransferase activity. Taken together, my findings indicate that SET1/MLL complexes collaborate extensively to deposit H3K4 methylation, and that a previously uncharacterized isoform of SET1B may have a major role in H3K4 methylation.

Acknowledgements

First and foremost, I would like to thank Professor Rob Klose for his guidance and support over the past three years - I could not have asked for a better supervisor.

I would like to thank Amy Hughes, whose work with the SET1 complexes provided the impetus for my research. I would also like to thank her for her guidance in the lab during the early days of my PhD. I would like to thank Anna Lastuvkova for generating a large number of the cell lines that were crucial for my research. I would like to thank Jess Kelley for her advice and help with biochemical techniques. I would like to thank Neil Blackledge and Emilia Dimitrova for their continued guidance and support in the lab. I would like to thank Aleks Szczurek for providing stimulating scientific discussions.

I would like to specifically thank Inge de Krijger and Michiel Vermeulen for their assistance in mass spectrometry. I would like to thank Inge for running my IP/MS samples and analysing the data. I would also like to thank Marjorie Fournier for her advice on sample preparation for mass spectrometry.

I would like to thank all members of the Klose lab with whom I have had the pleasure of working, and I thank them for providing a supporting and stimulating scientific environment.

Declaration of Authorship

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

Ho Yu Alan Au

30th November 2023

Contents

List of Figures

List of Tables

List of Common Abbreviations

aa	amino acids (refering to the length of the protein)
ATAC-seq	assay for transposon-accessible chromatin sequencing
BioCap-seq	biotinylated CXXC-affinity-purification sequencing
bp	base pair
cChIP-seq	calibrated chromatin immunoprecipitation sequencing
CGI	CpG island
chr	chromosome
cKO	conditional knockout
CpG	CpG dinucleotide
CTD	C-terminal domain
Cryo-EM	cryoelectron microscopy
dTAG	degradation tag - here, dTAG is used to refer to the small molecule dTAG-13
gRNA	guide RNA
H3K4me1	monomethylated Histone H3 Lysine 4
H3K4me2	dimethylated Histone H3 Lysine 4
H3K4me3	trimethylated Histone H3 Lysine 4
IP	immunoprecipitation
IP/MS	immunoprecipitation followed by mass spectrometry
KO	knockout
lncRNA	long non-coding RNA
mESC	mouse embryonic stem cell
mRNA	messenger RNA
MS	mass spectrometry
NGS	next-generation sequencing
NTD	N-terminal domain
PHD	plant homeodomain
PIC	pre-initation complex
PRC1	polycomb repressive complex 1
PRC2	polycomb repressive complex 2
PRO-seq	precision run-on sequencing

qPCR	quantitative polymerase chain reaction
RNAPII	RNA polymerase II
RRM	RNA recognition motif
RT	room temperature
RT-qPCR	reverse transcription followed by quantitative polymerase chain reaction
SET	Su(var)3-9, Enhancer of Zeste, and Trithorax
TES	transcription end site
TSS	transcription start site
TT-seq	transient transcriptome sequencing
UNT	untreated
WIN	WDR5-interaction motif
WRAD	WDR5, RBBP5, ASH2L, DPY30
WT	wild type

Chapter 1

Introduction

Regulation of gene expression is one of the most fundamental processes in all organisms. Gene expression is a multi-stage process that begins with the transcription of genes by RNA polymerases into messenger RNA (mRNA) that is either translated into proteins or further processed to produce functional RNA molecules such as long non-coding RNAs (lncRNA) and microRNAs (miRNA). Intricate mechanisms have evolved to regulate every step of this process to ensure that the correct genes are expressed to produce the appropriate amount of gene product in the appropriate cell at the appropriate time. Proper regulation of gene expression is required for the development of complex multi-cellular organisms which requires a diverse range of cell types to be generated from the same genome. Accordingly, mis-regulation of gene expression leads to many developmental disorders and is a hallmark of many cancers. The beginning stages of transcription are tightly controlled. Tight control of transcription initiation is particularly important during development because it allows the cell to respond rapidly to signalling pathways and other external cues to ensure that the appropriate gene expression program is established and maintained.

1.1 Transcription by RNA Polymerase II

In eukaryotes, all messenger RNAs (mRNA) are transcribed by RNA polymerase II (RNAPII) to produce proteins and long non-coding RNAs (lncRNA). Transcription by RNAPII is a tightly-coordinated process and can be broadly divided into three stages: initiation, elongation, and termination.

1.1.1 Transcription Initiation

Transcription initiation begins with the formation of a preinitiation complex (PIC) at the promoter. The PIC consists of RNAPII and seven general transcription factors, TFIID, TFIIA, TFIIB, TFIIC, TFIIE, TFIIF, and TFIIH, and is the complex from which RNAPII first begins transcription at the transcription start site (TSS). The TSS is defined as the first nucleotide of a transcript that is synthesised, and the 50 base pairs upstream and downstream of the TSS comprise the ‘core promoter’ (Andersson and Sandelin, 2020, Haberle and Stark, 2018). PIC formation begins with the TATA-binding protein (TBP) subunit of TFIID binding upstream of the TSS. A minority of metazoan promoters contain a TATA-box, a DNA sequence motif specifically recognized by TBP. These promoters typically have a well-defined TSS that results in ‘focused’ initiation from this site (Haberle and Stark, 2018). The majority of metazoan promoters instead have multiple TSSs that are contained within

the core promoter, and supports 'dispersed' initiation in which these TSSs are used with similar frequency (Haberle and Stark, 2018). In TATA-box-containing core promoters, TBP binds directly to the TATA-box; in other core promoters, other subunits of TFIID, TBP-associated factors (TAFs), cooperatively load TBP onto the core promoter. TFIID is further stabilized on the core promoter through interactions with TFIIA, after which TBP bends the DNA in a 90° angle and recruits TFIIB, which also interacts with TFIIB-recognition elements. TFIIF interacts with RNAPII and recruits it to this nascent PIC, following which TFIIE is recruited to further stabilize the PIC and recruit TFIIH. TFIIH contains an ATP-dependent DNA helicase subunit, XPB, which unwinds the promoter DNA to expose the template strand. TFIIH also contains a kinase subunit, CDK7, which phosphorylates serine 5 on the RNAPII C-terminal domain (CTD), a process that is stimulated by the Mediator, a large multi-subunit coactivator that serves to bridge enhancers with promoters (Richter et al., 2022). Once serine 5 of the RNAPII CTD has been phosphorylated, RNAPII begins transcribing. The phosphorylation of serine 5 of the RNAPII CTD also recruits capping enzymes which place an m7G cap on the 5' end of the nascent transcript.

1.1.2 Transcription Elongation

After RNAPII transcribes an initial 25-50 nucleotides, it stops transcribing and is held in a paused state by the DRB sensitivity-inducing factor (DSIF) and negative elongation factor (NELF) (Core and Adelman, 2019). DSIF, which is a heterodimer consisting of SPT4 and SPT5, first binds to RNAPII through SPT5. NELF subsequently recognizes the RNAPII-SPT5 interface and binds to RNAPII, forming a stable RNAPII-DSIF-NELF complex. In this complex, the paused state of RNAPII is stabilised by both DSIF and NELF: SPT5 binds to both the nascent RNA and upstream DNA template near the RNA exit channel and NELF binds to the RNAPII funnel (Vos et al., 2018). This stabilized pausing structure thus restricts RNAPII mobility along the DNA template. Release from promoter-proximal pausing is facilitated by the CDK9 kinase subunit of the positive transcription elongation factor b (pTEF-b), which phosphorylates the C-terminal region (CTR) of SPT5, serine 2 of the RNAPII CTD, and NELF. The phosphorylation of SPT5 converts DSIF into a positive elongation factor, and this also leads to the dissociation of NELF from RNAPII. Dissociation of NELF from RNAPII leads to the recruitment of elongation factors PAF1 and SPT6, which form a stable elongation complex with RNAPII and DSIF. This elongation complex then enters the gene body to begin productive elongation.

As the elongation complex enters into productive elongation, it encounters nucleosomes, which serve as barriers for transcription. The elongation complex recruits a variety of histone chaperones and chromatin remodellers such as CHD1 and FACT which disassemble the nucleosomes to allow passage of RNAPII and reassemble them in its wake. As the elongation complex continues through the gene body, additional elongation factors are recruited to facilitate transcription. For example, during early elongation, the elongation factor Elongin binds to RNAPII to prevent transient pausing (Elmendorf et al., 2001). Additionally, the elongation factor TFIIS recognizes backtracked RNAPII and assists in the cleavage of the transcript to allow for RNAPII to restart productive elongation (Ling et al., 2006). During the later stages of elongation, SCAF4 and SCAF8

proteins bind to the hyperphosphorylated RNAPII CTD and prevent the use of early, alternative polyadenylation sites (polyA sites) (Gregersen et al., 2019). Additionally, the spliceosome co-transcriptionally splices out introns as the transcription elongation complex transcribes through the gene body.

1.1.3 Transcription Termination

As the RNAPII elongation complex reaches the end of the gene, it encounters the polyadenylation signal (PAS), which instructs the termination machinery to cleave the mRNA transcript and polyadenylate the 3' end (Proudfoot, 2016). In mammals, cleavage and polyadenylation is performed by the cleavage and polyadenylation specificity factor (CPSF), cleavage stimulatory factor, and the cleavage factors I and IIm subcomplexes. Once transcript cleavage and polyadenylation has occured, the transcriptional machinery is halted and displaced from DNA. There are two models for how this occurs: the 'torpedo' model, and the allosteric model (Rodríguez-Molina et al., 2022). The 'torpedo' model postulates that after mRNA cleavage at the PAS, RNAPII continues to transcribe past the PAS. The 5'-3' RNA exonuclease XRN2 then degrades this nascent RNA and eventually catches up to the transcribing RNAPII and displaces it off the DNA. The allosteric model posits that transcription of the PAS leads to allosteric changes within RNAPII and dissociation of anti-termination factors and renders it susceptible to termination. These two models are not mutually exclusive and there is evidence that both mechanisms play a role in transcription termination. Recently, it was proposed that transcription of the PAS promotes dephosphorylation of SPT5 by the PNUTS-PP1 phosphatase complex, which slows down RNAPII, allowing XRN2 to displace it off chromatin (Cortazar et al., 2019).

1.2 Regulation of Transcription by Histone Modifications

1.2.1 Histone proteins are extensively modified

The eukaryotic genome is packaged in the nucleus by the wrapping of DNA around histone proteins to form chromatin. The basic unit of chromatin is the nucleosome, which consists of 147bp of DNA wrapped around the histone octamer, which in turn consists of two copies of histones H2A and H2B in dimeric form and histones H3 and H4 arranged in a tetramer. These histone proteins, termed core histone proteins to distinguish them from the H1 and H5 linker histones, have a globular core which serves as a scaffold for the wrapping of DNA, and an unstructured N-terminal tail that can be extensively modified post-translationally. Histones H2A and H2B additionally have an unstructured C-terminal tail that can similarly be modified. These modifications have been extensively catalogued and include methylation, acetylation, ubiquitination, and phosphorylation, among others (Kouzarides, 2007, Millán-Zambrano et al., 2022). Some histone modifications are cell-type specific, such as the addition of a serotonin group to glutamine 5 of histone H3 in serotonergic neurons (Farrelly et al., 2019). In addition to the terminal tails, residues within the globular histone core can also be modified, and many of these residues lie on the lateral surface of the histone octamer or the interfaces between histones. Most histone modifications are reversible, and for most characterized histone modifications, a corresponding

set of enzymes that catalyse their removal have been identified. These enzymes not only regulate the total abundance of a certain histone modification, but may also specifically remove it from certain genomic regions, thus specifying its distribution across the genome.

Histone modifications are often enriched at functional genomic regions and often correlate with transcriptional activity. In particular, histone modifications are often associated with genes and often localised to specific regulatory elements such as promoters and enhancers. Many histone modifications are also specifically found within gene bodies. Broadly, histone modifications are either associated with active transcription or with transcriptional repression. Histone lysine acetylation is almost invariably associated with active transcription, and is generally distributed in a punctate manner at promoters and enhancers. Histone lysine methylation is considerably more variable and can be associated with active transcription or repression depending on the lysine residue that is methylated. For example, H3K9 and H3K27 methylation are both associated with transcriptional repression, and H3K4 and H3K36 methylation are both associated with active transcription. Many histone modifications are associated with specific genomic elements. For example, H3K4me3 is specifically enriched at promoters, H3K36me3 and H2BK120ub1 are broadly distributed across the length of the gene body, and H3K27ac is found at both promoters and enhancers. These distribution patterns suggest that these histone modifications may regulate, or be related to, specific steps of transcription. For example, H3K36me3 has been suggested to support transcription elongation by preventing cryptic transcription initiation at gene bodies, and H3K4me3 has been suggest to support transcription initiation by stabilizing PIC formation at promoters (Sen et al., 2015, Vermeulen et al., 2007, Lauberth et al., 2013).

1.2.2 Histone modifications can affect chromatin structure

One way in which histone modifications can regulate transcription is by altering the local chromatin structure. This is most prominently illustrated by the acetylation and phosphorylation of histone tails, which reduce the basic charge of histones (Millán-Zambrano et al., 2022). Acetylation occurs at lysine residues and phosphorylation occurs at serine, threonine, and tyrosine residues. Neutralization of the basic charge of histones disrupts the electrostatic interactions between histones and DNA, resulting in a less compact chromatin structure and makes the underlying DNA more accessible for DNA binding proteins. Acetylation and phosphorylation of histone tails are therefore associated with active transcription. The less compact chromatin structure also makes the histone tails themselves more accessible. For example, acetylation of histone H3 at various lysine residues has been shown to promote binding of H3K4me3 by the BPTF subunit of the NURF complex (Morrison et al., 2018). Additionally, histone H3 acetylation has been shown to make the histone H3 tail more accessible for the H3K4 methyltransferase MLL1 to methylate H3K4 (Jain et al., 2023).

1.2.3 Many proteins have domains that recognize specific histone modifications

The primary function of histone modifications is their ability to recruit and stabilize the binding of proteins that regulate transcription. Over the years, many specialized protein domains that recognize specific histone

modifications have been identified. For example, the plant homeodomain (PHD) zinc finger has been shown to specifically bind to H3K4me3, and TAF3, a subunit of the general transcription factor TFIID, has been shown to bind H3K4me3 via its PHD (Mellor, 2006, Vermeulen et al., 2007). Being a subunit of TFIID, TAF3 has a key role in transcription initiation, and this was thought to provide an instructive logic for the distribution of H3K4me3 at promoters – when a gene needs to be transcribed, a histone methyltransferase deposits H3K4me3 at its promoter, which then promotes binding of TFIID to initiate transcription. Furthermore, more highly-transcribed genes tend to have more H3K4me3 at the promoter, further highlighting its role in promoting transcription. Many other protein domains have been described to bind specific histone modifications: YEATS domains and bromodomains bind acetylated histone tails; PHDs, chromodomains, and Tudor domains bind methylated histone tails (Hyun et al., 2017). The presence of many such domains in transcriptional regulators provides an instructive logic for the role of histone modifications in regulating transcription, and indeed many transcriptional regulators rely, at least in part, on binding to histone modifications for their functions (Brown et al., 2017, K. Cao et al., 2018, Zhang et al., 2023).

1.3 CpG islands

1.3.1 DNA Methylation

DNA methylation is a stable, heritable chemical modification of nucleotides that is associated with transcriptional repression (Bird, 2002, Schübeler, 2015). In eukaryotes, DNA methyltransferases catalyse the addition of a methyl group to cytosine nucleotides within CpG dinucleotides. DNA methylation functions to repress transcription by recruiting transcriptional repressor complexes which often contain a subunit with a methyl-CpG-binding domain (MBD). Additionally, DNA methylation has been shown to prevent sequence-specific transcription factors from binding to their cognate binding sites. Although DNA methylation is well-conserved throughout eukaryotes, DNA methylation patterns are not uniform across all species. Most prominently, vertebrate genomes are pervasively methylated, with approximately 75% of all CpG dinucleotides being methylated. This is in contrast to fungi, plants, and flies in which DNA methylation is restricted to certain sites. This may be due to the high proportion of transposable elements and viral DNA within vertebrate genomes, and a major biological function of DNA methylation is thought to be the repression of these elements. These methylation patterns are stable and heritable; DNA methylation patterns are established by DNMT3A, DNMT3B, and DNMT3L, and maintained after DNA replication by DNMT1.

1.3.2 CpG islands are closely correlated with sites of transcription initiation

Vertebrate genomes are extensively methylated, with the exception of short, 1-2kb segments of unmethylated DNA referred to as CpG islands (CGIs) (Blackledge and Klose, 2011, Deaton and Bird, 2011). The pervasive methylation of CpG dinucleotides in vertebrate genomes has led to a progressive loss of CpG dinucleotdies throughout evolution due to the deamination of methylcytosine to thymine. As a result, the occurrence

of CpG dinucleotides throughout the genome is much lower than would be expected from the total CG content. CGIs have high CpG density, high CG content, and remain unmethylated. CGIs are strongly associated with sites where transcription initiates, and over 70% of mammalian promoters are associated with a CpG island (Saxonov et al., 2005). In turn, approximately half of all CGIs are found associated with a promoter. CGI promoters include virtually all housekeeping genes and developmentally-regulated genes; non-CGI promoters are typically highly tissue-specific (Haberle and Stark, 2018, Long et al., 2013). CGIs that are not found within promoters are termed orphan CGIs and have also been shown to be sites of transcription initiation, and some have been characterized as alternative gene promoters or as enhancers (Illingworth et al., 2010, Pachano et al., 2021). The close correlation between CGIs and gene promoters has led to the view that CGIs serve as important gene regulatory elements.

1.3.3 The ZF-CXXC domain targets chromatin modifiers to CGIs

CGIs are proposed to play an active role in maintaining a chromatin state that is permissive for transcription (Tazi and Bird, 1990). This appears to be related to the function of a class of proteins that contain a zinc finger CXXC domain (ZF-CXXC). The ZF-CXXC domain specifically targets these proteins to CGIs, where they are proposed to modify chromatin and regulate transcription (Blackledge et al., 2013). Furthermore, these ZF-CXXC-containing proteins have roles in regulating transcription. The first such protein identified was CXXC zinc finger protein 1 (CFP1, also known as CXXC1), which is part of the SET1A/B complexes in mammals (Shin Voo et al., 2000). Subsequent work has shown that CFP1 is the primary determinant of SET1A/B occupancy throughout the genome, and targets the SET1 complexes to CGI promoters through its CXXC domain (Thomson et al., 2010, Brown et al., 2017). The SET1A/B complexes tri-methylate H3K4, and accordingly, H3K4me3 distribution within mammalian genomes is highly correlated with CGIs. Therefore, the finding that the CFP1 subunit of SET1A/B complexes specifically bind CGIs became one of the first indications that CGIs may serve as binding platforms for histone modifiers. All CGI promoters, and indeed all CGIs, contain a peak of H3K4me3, and the boundaries of the underlying CpG island demarcates H3K4me3 enrichment (Hughes et al., 2020, Long et al., 2013). H3K4me3 is found at the promoters of all CGI-containing genes, and levels of H3K4me3 generally correlate with transcription. Subsequent work has shown that two other H3K4 methyltransferases, mixed-lineage leukemia 1 (MLL1) and mixed-lineage leukemia 2 (MLL2) also contain CXXC domains, and their occupancy genome-wide also highly correlates with CGIs (Birke et al., 2002, Bach et al., 2009, Denissov et al., 2014). Given that the SET1A/B and MLL1/2 complexes are thought to be the primary H3K4 tri-methyltransferases in mammals, their association with CGIs through the CXXC domain has led to the view that CGIs are the primary determinant of H3K4me3 deposition. This view was further strengthened by the finding that artificial incorporation of transcriptionally inert CGIs into the genome of mouse embryonic stem cells is sufficient for H3K4me3 deposition (Thomson et al., 2010, Wachter et al., 2014).

On the other hand, a subset of CGI promoters are not transcriptionally active, and CGIs similarly play an active role in maintaining a chromatin state that is repressive for transcription. This was first demonstrated with

the discovery that KDM2B, which contains a CXXC domain, is a component of a variant polycomb repressive complex 1 (vPRC1) (Farcas et al., 2012). Polycomb repressive complexes form an intricate system that primarily repress transcription through the deposition of two histone modifications, mono-ubiquitination of lysine 119 on histone H2A (H2AK119ub1) by PRC1, and tri-methylation of lysine 27 on histone H3 (H3K27me3) by PRC2 (Blackledge and Klose, 2021). Both histone modifications have been shown to be important in repressing transcription and are enriched at the CGI promoters of repressed genes. The polycomb system establishes this repressive chromatin state at CGIs through the targeting of KDM2B via its CXXC domain, and sustains it through positive feedback mechanisms in which both PRC1 and PRC2 recognize and are recruited by H2AK119ub1 and H3K27me3 (Blackledge et al., 2014). At these polycomb-repressed CGI promoters, transcriptional repression is primarily mediated by PRC1, which restricts transcription initiation by limiting PIC formation in an H2AK119ub1-dependent manner (Blackledge et al., 2020, Dobrinić et al., 2021, Szczurek et al., 2023).

Although many CGI-associated proteins have histone modifying activities, recent studies have shown that CGIs do not simply function to influence transcription through chromatin-based mechanisms. KDM2A and KDM2B are CXXC-domain-containing proteins with H3K36 demethylase activity, but it has recently been shown that they primarily constrain transcription from CGIs in a demethylation-independent manner (Turberfield et al., 2019). KDM2A/B both have a long isoform that contains the JmjC domain, the catalytic domain reponsible for demethylase activity, and a short isoform that does not. Conditional knockout of the long isoforms of KDM2A and KDM2B resulted in only modest increases in H3K36me2 at CGI promoters and little change in chromatin accessibiliy or gene expression. In contrast, conditional removal of the CXXC domains of all KDM2A/B isoforms resulted in widespread increases in gene expression, indicating that KDM2A/B mediates gene repression in a chromatin-independent manner. Similarly, recent work has shown that the SET1 complexes can support transcription in a manner that is independent of their H3K4 methyltransferase activity (Hughes et al., 2023). Rapid depletion of SET1A/B results in widespread reduction in transcription but these reductions do not correlate with reductions in H3K4me3. Additionally, only a small N-terminal portion of SET1A and SET1B, which do not include the catalytic SET domain, is able to activate expression of a reporter gene *in vivo*. Therefore, although the histone-modifying activies of CGI-associated proteins are required for transcriptional regulation in some instances, such as for the polycomb repressive complexes, CGIs have more divserse functions than simply to promote changes in histone modifications.

1.4 H3K4 Methylation

1.4.1 Methylation of H3K4 is associated with active transcription

Methylation of lysine 4 on histone H3 (H3K4) is a conserved post-translational modification that is assocaited with active transcription. Lysine methylation of histone tails are exclusively catalysed by proteins containing a SET domain, an acronym for the three founding members of this domain: Su(var)3-9, Enhancer

of Zeste, and Trithorax (Dillon et al., 2005). The SET domain catalyzes the transfer of a methyl group from the donor S-adenosyl methionine (SAM) to the lysine residue. To date, the only non-SET domain histone methyltransferase identified is DOT1L and its homologs, which methylate H3K79 within the core of histone H3. H3K4 can be mono-, di-, or tri-methylated, and these three methylated states vary in their abundance in the nucleus and their distribution across the genome (Barski et al., 2007, Bernstein et al., 2005). Monomethylated H3K4 (H3K4me1) is the most abundant methylated H3K4 species as determined by mass spectrometry, with nearly 30% of all histone H3 molecules being monomethylated at H3K4 in mESCs (Dorighi et al., 2017). Dimethylated H3K4 (H3K4me2) is considerably less abundant, with approximately 1.4% of all histone H3 molecules being dimethylated at H3K4 in mESCs (Dorighi et al., 2017). Trimethylated H3K4 (H3K4me3) is the least abundant of the three methylated states, with approximately 0.35% of all H3 molecules trimethylated at H3K4 in mESCs (Dorighi et al., 2017). Although the exact abundance of each modified H3K4 species can vary considerably depending on the cell type, the approximate ratio of the three methylated H3K4 species remains broadly the same (Leroy et al., 2013). This disparity in abundance is also reflected in their distribution across the genome; H3K4me1 is the most broadly distributed of the three methylated states, followed by H3K4me2 and then H3K4me3 (Bernstein et al., 2005, Barski et al., 2007). H3K4me1 is often characterized as an enhancer-associated modification due to their enrichment at enhancers in flies and mammals. However, H3K4me1 is also a very pervasively deposited modification that is broadly distributed across all actively transcribed regions of the genome, with the exception of actively transcribed promoters which are typically highly-enriched in H3K4me2/3 and are devoid of H3K4me1. H3K4me2 is also broadly distributed across transcribed regions and is also found at enhancers, but unlike H3K4me1 is also found at promoters and often colocalizes with H3K4me3. On the other hand, the distribution of H3K4me3 is much more restrictive than that of H3K4me1/2, being almost exclusively localised to promoters.

Both the amount of H3K4 methylation and its distribution at a gene are highly correlated with its level of transcription. H3K4me3 has a well-characterised correlation with transcription as assessed by ChIP-seq: highly-transcribed genes have a larger peak of H3K4me3 at their promoters compared to lowly-transcribed genes (Howe et al., 2017). In mammals, these regions of H3K4me3 enrichment is demarcated by the boundaries of the CpG island underlying the promoter. The promoters of actively transcribed genes invariably have a nucleosome-depleted region at the transcription start site that is flanked by a large peak of H3K4me3 downstream and a smaller peak upstream (Fig. 1.1A). The enrichment of H3K4me2 at the promoter is less reflective of the transcriptional output of the associated gene, but the distribution of H3K4me2 is nonetheless highly correlated with transcription. At highly-transcribed genes, H3K4me3 is highly-enriched at the nucleosomes immediately downstream of the transcription start site. At highly-transcribed genes, the distributions of H3K4me3 and H3K4me2 at promoters are anti-correlated; nucleosomes with the highest H3K4me3 enrichment typically have low H3K4me2 enrichment, and H3K4me2 levels increase at nucleosomes further downstream of the transcription start site as H3K4me3 enrichment decreases (Fig. 1.1A). At highly-transcribed genes, H3K4me2 enrichment is not restricted by the boundaries of the CpG island and often extends into the gene body (Fig. 1.1A). At lowly-transcribed genes, the distribution of H3K4me2 is more similar to that of H3K4me3, with enrich-

ment primarily at the first nucleosomes immediately downstream of the transcription start site (Fig. 1.1B). The height of the H3K4me2 ChIP-seq peak at the promoter does not typically correlate with transcription. Rather, the enrichment of H3K4me2 in gene bodies is more reflective of transcription, with more highly-transcribed genes having more H3K4me2 extending further into the gene body. The distribution of H3K4me1 is anti-correlated with that of H3K4me2 and H3K4me3 at promoters; nucleosomes close to the transcription start site are typically highly-enriched in H3K4me3 and H3K4me2 and devoid of H3K4me1. Additionally, the majority of CpG islands do not have H3K4me1. At highly-transcribed genes, H3K4me1 enrichment flanks the promoter where there is no H3K4me3 and where H3K4me2 levels begin to decline (Fig. 1.1A). Similarly to H3K4me2, highly-transcribed genes tend to have more H3K4me1 in the gene body compared to lowly-transcribed genes. At lowly-transcribed genes, all three methylation states of H3K4 are found at the promoter (Fig. 1.1B).

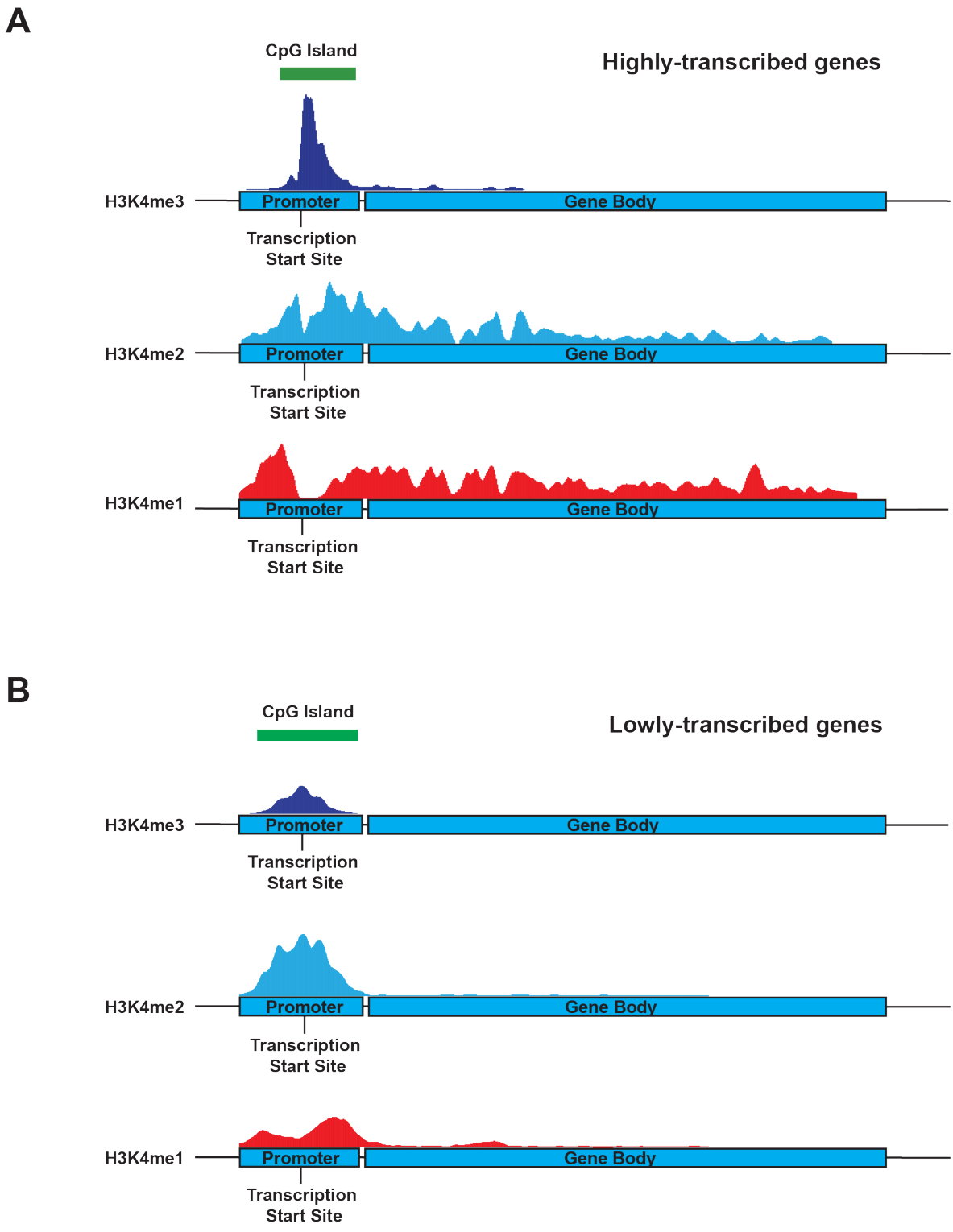

Figure 1.1: Distribution of H3K4 methylation at highly- and lowly-transcribed genes.

Cartoons illustrating the distribution of H3K4 methylation as assessed by ChIP-seq at a representative gene that is **(A)** highly-transcribed or **(B)** lowly-transcribed.

1.4.2 Proposed functions of H3K4 methylation in transcription

H3K4me3

H3K4me3 has primarily been suggested to regulate transcription by recruiting effector proteins that modify histones, remodel chromatin, or are part of the transcriptional machinery. Many domains such as the plant homeodomain (PHD) zinc finger, the tudor domain, and the chromodomain recognize methylated lysine residues on histone proteins, and a subset of these domains specifically bind to H3K4me3 with high affinity (Ruthenburg et al., 2007, Hyun et al., 2017). Accordingly, many proteins that are involved in transcription have one or more of these domains and H3K4me3 is thought to positively regulate transcription by recruiting these proteins or by stabilizing their interactions with chromatin at the promoter.

H3K4me3 interacts with the transcriptional machinery through the TAF3 subunit of the general transcription factor TFIID, which has a PHD finger that binds to H3K4me3 with high affinity (Vermeulen et al., 2007). Subsequent work demonstrated that the interaction between TAF3 and H3K4me3 promoted PIC assembly *in vitro* and enhanced *in vitro* transcription of an acetylated chromatin template (Lauberth et al., 2013). A recent study using *in vitro* reconstituted chromatin templates also reported that the presence of H3K4me3 at the +1 nucleosome of both TATA-containing and TATA-less promoters stimulated transcription (Fisher and Luse, 2023). However, siRNA-mediated knockdown of TAF3 resulted in major changes in gene expression at only a subset of genes that were dependent on p53 (Lauberth et al., 2013). The authors speculate that either the partial depletion of TAF3 by siRNA resulted in sufficient TAF3 to maintain TFIID binding to promoters, or that the TAF3-H3K4me3 interaction is primarily important for rapid transcriptional induction rather than constitutive gene expression.

H3K4me3 has also been shown to recruit other histone modifiers to promoters. For example, ING4 and ING5 are mutually exclusive subunits of the MOZ/MORF H3K23 acetyltransferase complexes and both have PHD fingers which bind H3K4me3 (Lalonde et al., 2013, Champagne et al., 2008). The MOZ/MORF complexes also have a BRPF1 subunit which also has two PHD fingers that have been shown to bind H3K4me3 (Zhang et al., 2023). Accordingly, the distribution of H3K23ac is enriched at active promoters. Another example of an acetyltransferase complex that is recruited by H3K4me3 is the SAGA coactivator complex, which contains histone acetyltransferase activity within its KAT2A/B subunit (Bian et al., 2011). The Sgf29 subunit of SAGA contains tandem tudor domains which recruit SAGA to H3K4me3-enriched promoters where it acetylates H3K9 and H3K14. Other histone modifiers have been shown to bind to H3K4me3. For example, the H3K9 demethylases PHF2 and PHF8 have been reported to bind H3K4me3 (Wen et al., 2010, Feng et al., 2010). Therefore, a major role of H3K4me3 appears to be to establish a chromatin environment that is permissive for transcription by recruiting histone modifying enzymes. Another class of proteins which bind H3K4me3 are chromatin remodellers. One of the first chromatin-acting proteins to be identified as an H3K4me3-binding protein was the BPTF subunit of the NURF chromatin remodelling complex (Wysocka et al., 2006). Another chromatin remodeller that is recruited by H3K4me3 is CHD1, which binds to H3K4me3 through its chromod-

omain (Sims et al., 2005).

Two recent studies have suggested that H3K4me3 does not regulate transcription initiation, but rather regulates the release of RNAPII from promoter-proximal pausing (H. Wang et al., 2023, S. Hu et al., 2023). By depleting shared subunits of the H3K4 methyltransferases in mESCs, these two studies reported that they were able to remove virtually all H3K4 methylation. In both studies, H3K4me3 reductions occurred faster than H3K4me1/2 reductions and resulted in comprehensive reductions in overall transcription. One of the studies was able to demonstrate that these effects were due to H3K4me3 because knockout of the H3K4 demethylases KDM5A and KDM5B, which resulted in slower reductions in H3K4me3, also delayed the onset of transcription defects (H. Wang et al., 2023). Interestingly, both studies reported unperturbed formation of the RNAPII PIC in the absence of H3K4me3, with no reductions in TFIID binding. Accordingly, both studies reported that transcription initiation was not affected by removal of H3K4me3. Rather, removal of H3K4me3 greatly reduced the ability of RNAPII to enter into productive elongation. Both studies implicate the transcription termination complex Integrator in this process, although they differ in their conclusions - with one study claiming that H3K4me3 recruits Integrator to evict paused non-productive elongation complexes to facilitate the entry of elongation-competent RNAPII into the gene body, and the other claiming that loss of H3K4me3 leads to increased Integrator binding at promoters where it prematurely terminates transcription. Therefore, any role of H3K4me3 in regulating pause-release will need to be examined in more detail.

H3K4me2

Compared to H3K4me1 and H3K4me3, H3K4me2 is the least studied of the three H3K4 methylation states. This may be in part due to it lacking a distinct genomic distribution pattern. H3K4me2 is found at promoters, gene bodies, and enhancers, and often co-localizes with H3K4me1 or H3K4me3. In support of this view, protein domains that selectively recognize H3K4me2 are rare; most domains which bind H3K4me2 also bind H3K4me3, and many H3K4me1-binding proteins also bind H3K4me2 (Howe et al., 2017, Hyun et al., 2017). For example, the PHD finger of the Sgf29 subunit of SAGA binds to both H3K4me2 and H3K4me3, although it has a slightly higher affinity for H3K4me3 (Bian et al., 2011). Similarly, H3K9 demethylases such as PHF8 bind to both H3K4me2/3 (Feng et al., 2010). This has led many to consider H3K4me2 to be simply an intermediary state of H3K4 methylation. Interestingly, one process in which H3K4me2 has been specifically implicated is transcriptional memory, in which a recently activated gene undergoes a more rapid re-activation upon exposure to the same activatory signal (Brickner, 2023). This was most prominently characterized in yeast, where H3K4me2 has been shown to have a critical role in the maintenance of transcriptional memory at the *INO1* gene (D'urso et al., 2016, Sump et al., 2022). The *INO1* gene is ordinarily not transcribed but is rapidly induced upon withdrawal of inositol from the growth medium. Prior inositol starvation leads to a more rapid induction of the *INO1* gene during future exposure to inositol starvation. This was termed 'transcriptional memory' and has been reported to rely on the continued association of the RNAPII in a hypophosphorylated form at the *INO1* promoter after the initial inositol starvation, which is dependent on H3K4me2. Whether this mechanism is conserved in higher eukaryotes remains to be determined. In human cells, many genes that

induce more rapidly upon repeated exposure to stimuli, such as genes sensitive to interferon gamma, have been shown to be enriched in H3K4me2 at their promoters. However, whether H3K4me2 plays an instructive role in the re-induction of these genes is unknown (Light et al., 2013).

H3K4me1

H3K4me1 is considered to be a hallmark of active and poised enhancers (Heintzman et al., 2009). However, the functional role of H3K4me1 at enhancers is still subject to debate (Rada-Iglesias et al., 2011). A proteomics screen of H3K4me1-interacting proteins using HeLa cell nuclear extracts demonstrated that subunits of the BAF chromatin remodelling complex bound to *in vitro* reconstituted mononucleosomes containing H3K4me1 with much higher affinity than to mononucleosomes containing H3K4me3 (Local et al., 2018). Additionally, the authors found that BAF complexes remodelled nucleosomes containing H3K4me1 *in vitro* more efficiently than those containing H3K4me3. Another study by the same group found that subunits of the Cohesin complex bound to H3K4me1-containing mononucleosomes with higher affinity than with H3K4me2- or H3K4me3-containing mononucleosomes (Yan et al., 2018). Additionally, by fusing the catalytic SET domain of the H3K4 mono-methyltransferase MLL3 to a catalytically-dead Cas9, the authors were able to demonstrate that targeted generation of H3K4me1 was sufficient to recruit the Cohesin complex to the *Sox2* locus (Yan et al., 2018). However, *in vivo* evidence of a requirement for H3K4me1 for enhancer function genome-wide is lacking.

There is also considerable evidence that H3K4me1 is not required for enhancer function. This mainly comes from studies using point mutations to generate catalytically-dead mutants of MLL3 and MLL4, the primary H3K4 monomethyltransferases in mammals, or catalytically-dead mutants of trithorax-related, the primary H3K4 monomethyltransferases in drosophila (Dorighi et al., 2017, Rickels et al., 2017). These studies have indicated that the methyltransferases themselves are more essential for enhancer function through their catalysis-independent functions.

1.4.3 H3K4 methylation as a consequence of transcription

An increasing body of research suggests that H3K4me3 is, at least in part, a consequence of transcription. Studies perturbing components of the transcription elongation complex alters the distribution of H3K4me3 at gene promoters. For example, knockdown of subunit 9 of the Integrator complex (INTS9) in drosophila results in an accumulation of RNAPII at promoters, which was attributed to a lack of promoter-proximal pause-release (Elrod et al., 2019). This increase in RNAPII at promoters was accompanied by a widespread increase in H3K4me3. On the other hand, knockdown of the histone chaperone FACT in drosophila has been demonstrated to result in a decrease in RNAPII levels at the promoter and an increase in RNAPII levels in the gene body, which the authors attribute to a decrease in promoter-proximal pausing (Tettey et al., 2019). This was accompanied by a reduction in H3K4me3 at the promoter and spreading of H3K4me3 into the gene body. A recent study using human cells demonstrated that depletion of PAF1 results in a genome-wide reduction in H3K4me3 levels, suggesting that elongating RNAPII is required, at least in part, for H3K4me3 deposition (Z.

Wang et al., 2023). The mammalian H3K4 methyltransferases SET1A and SET1B have been shown to interact with RNAPII that is phosphorylated at serine 5 of its CTD through WDR82 (J. H. Lee and Skalnik, 2008). Inhibition of the CDK7 subunit of TFIIH in a human cancer cell line has been shown to reduce H3K4me3 levels downstream of the TSS and reduce the breadth of H3K4me3 enrichment at many promoters, suggesting that H3K4me3 deposition depends, in part, by recruitment of SET1A and SET1B to actively transcribing RNAPII (Ebmeier et al., 2017). Most strikingly, a recent study using the small molecule inhibitor triptolide to inhibit TFIIH in human cells reported that a large proportion of H3K4me3 is lost after 1 hour of triptolide treatment, indicating that active transcription is required for maintenance of H3K4me3 levels genome-wide (Z. Wang et al., 2022).

Despite recent evidence pointing to a role of transcription in directing H3K4me3 deposition, it is clear that transcription is not a requirement for H3K4me3 deposition. Insertion of an artificial CGI-containing sequence into a transcriptionally inert gene desert in mESCs leads to H3K4me3 deposition at the CGI with no corresponding recruitment of RNAPII (Wachter et al., 2014). Additionally, during oogenesis, H3K4me3 accumulates in broad domains at regions which are transcriptionally silent. This accumulation of H3K4me3 appears to be more dependent on the widespread loss of DNA methylation rather than the result of transcriptional activity (Hanna et al., 2018). Therefore, H3K4me3 deposition is likely formed by a combination of transcription-dependent and transcription-independent mechanisms.

1.5 SET1/MLL Complexes Methylate H3K4

1.5.1 SET1/MLL proteins are the major H3K4 methyltransferases

In mammals, H3K4 methylation is deposited by six H3K4 methyltransferases, SET1A, SET1B, MLL1, MLL2, MLL3, and MLL4, which all contain a highly-conserved **S**u(var)3-9, **E**nhancer-of-zeste, **T**ritorax (SET) domain required for catalysis (Dillon et al., 2005). These six methyltransferases can be grouped into three paralogous pairs with high structural similarity: SET1A/B, MLL1/2, and MLL3/4 (Fig. 1.2A, B) (Crump and Milne, 2019). These three pairs contain highly similar proteins that appear to have arisen due to gene duplication during evolution, with each pair being related to a drosophila homolog: SET1A/B are related to drosophila Set1 (dSet1), MLL1/2 are related to Trithorax (Trx), and MLL3/4 are related to Trithorax-related (Trr) (Crump and Milne, 2019). Although these three pairs differ from each other substantially, they share a common architectural feature with the catalytic SET domain situated at the C-terminus of the protein that is followed by a post-SET domain C-terminal to the SET domain that is also required for catalytic activity (Dillon et al., 2005). The remainder of the protein is dedicated towards the methyltransfease-independent activites of these proteins, and includes chromatin-binding domains and interaction domains for subunits that are not required for catalytic activity. Accordingly, the methyltransferase-independent activities of each SET1/MLL protein vary greatly, and is reflected by the variety of different pathologies that result from mutations in the non-catalytic regions of these proteins (Cenik and Shilatifard, 2021, Shilatifard, 2012).

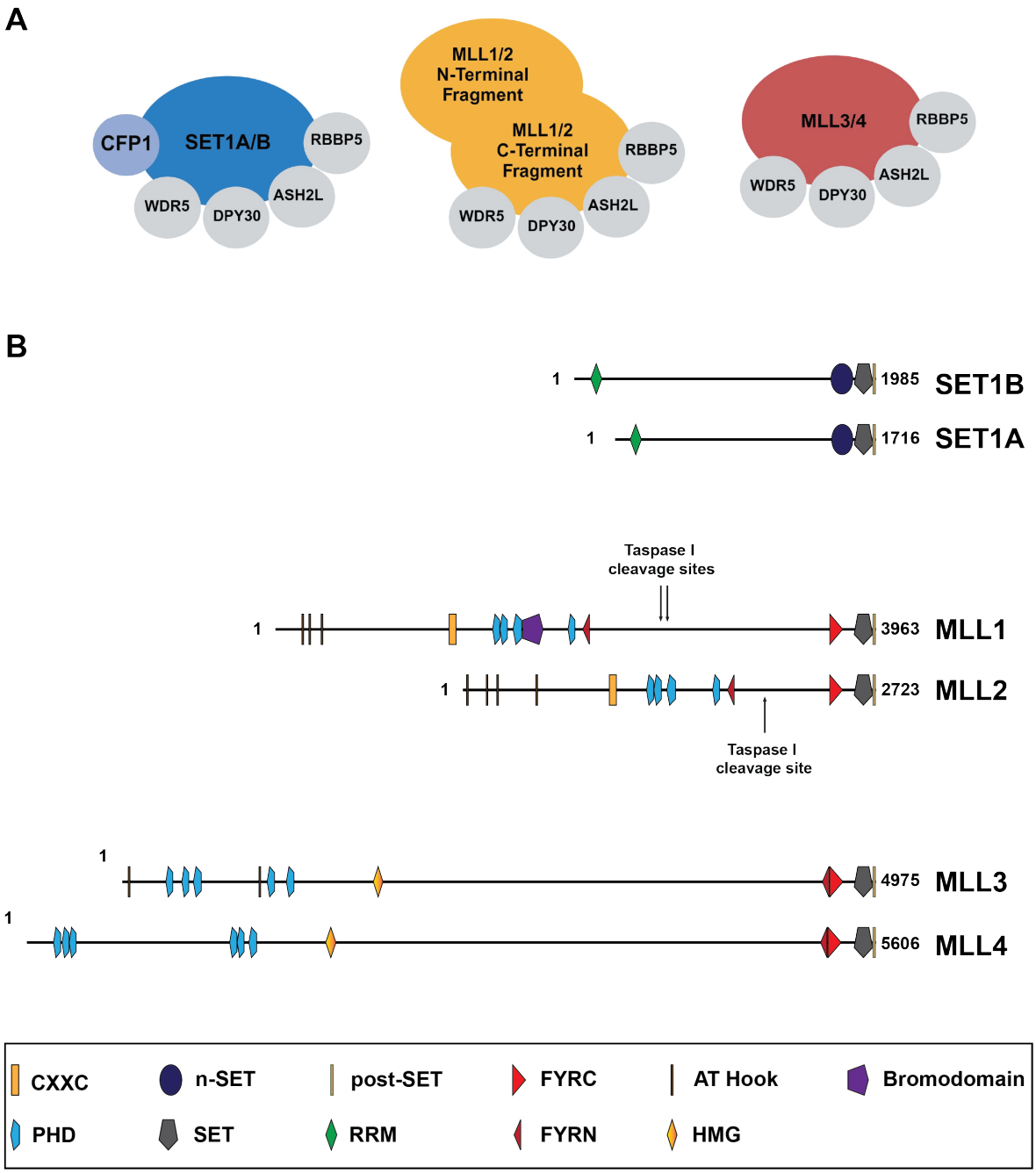

Figure 1.2: The SET1/MLL H3K4 methyltransferases.

A Cartoons illustrating the SET1/MLL complexes. **B** Domain organisation of SET1/MLL proteins. Domain architecture was analysed using SMART.

1.5.2 The WRAD Sub-Complex

A common feature for the SET1/MLL methyltransferases is that their SET domains possess only weak mono-methyltransferase activity on their own and require interaction with four proteins for maximum methyltransferase activity (Shinsky et al., 2015). These four proteins are WD40 repeat protein 5 (WDR5), Retinoblastoma-binding protein 5 (RBBP5), Absent small homeotic-2-like (ASH2L), and Dumpy-30 (DPY-30), and are commonly referred to as the 'WRAD sub-complex' (Fig. 1.2A) (Ernst and Vakoc, 2012). The WRAD sub-complex is conserved and is also found in flies and yeast (Y.-h. Takahashi et al., 2011, Ernst and Vakoc, 2012). Each SET1/MLL protein contains a highly-conserved WDR5-interaction motif (WIN motif) N-terminal to the SET domain that serves as an interaction site for WDR5 (Guarnaccia and Tansey, 2018). For SET1A/B the WIN motif lies within the n-SET domain. ASH2L and RBBP5 exist as a heterodimer within WRAD and also interact with the SET1/MLL proteins, primarily at the SET domain. RBBP5 also interacts extensively with WDR5. DPY30 does not interact with the SET1/MLL proteins, and its association with WRAD is mediated exlusively through interactions with ASH2L. Quantitative proteomics studies have indicated that DPY30 likely exists as a dimer within WRAD (van Nuland et al., 2013). Interestingly, only a small region of each methyltransferase, approximately 400 amino acids encompassing the n-SET (for SET1A/B), SET, and post-SET domains, is required for catalytic activity. Recombinant expression of a C-terminal fragment of MLL1-4 containing the WIN motif, SET domain, and post-SET domain is sufficient to form a complex with WRAD, and this complex is sufficient to generate H3K4me1/2/3 on a nucleosomal substrate *in vitro* (Shinsky et al., 2015, Shinsky and Cosgrove, 2015, Li et al., 2016). For SET1A/B, optimal H3K4 methyltransferase activity towards a nucleosomal substrate requires inclusion of the entire n-SET domain, which facilitates the formation of a complex containing SET1A/B, WRAD, and CFP1 (Kwon et al., 2020).

When recombinantly-expressed SET1/MLL SET domains are incubated with histone H3 peptides as substrate, only weak mono-methylation activity is observed (Shinsky et al., 2015). This is in part due to the presence of a tyrosine at a key position in the active site of the SET1/MLL complexes, which restricts the active site volume and precludes accomodation of methylated lysine residues as a substrate (Y. H. Takahashi et al., 2009, Collins et al., 2005). SET domains typically have either a tyrosine or a phenylalanine at this key position, known as the 'tyr/phe switch', and the presence of phenylalanine results in a larger active site volume and therefore provides intrinsic tri-methyltransferase ability. When the SET1/MLL SET domains are in complex with WRAD, their mono-methylation activity is stimulated (Shinsky et al., 2015). Additionally, the SET domains of SET1A, SET1B, MLL1, and MLL2 attain di- and tri-methyltransferase ability. MLL4 gains di-methyltransferase activity when in complex with WRAD, but MLL3 remains a mono-methyltransferase. Therefore, WRAD has at least two functions: to enable more efficient mono-methyltransferase activity by the SET1/MLL SET domains, and to induce allosteric changes in the SET1/MLL SET domains to allow for di- and tri-methyltransferase activity. Interestingly, it has been shown that WRAD, and more specifically the RBBP5-ASH2L heterdimer, can bind S-adenosyl methionine, suggesting that one mechanism by which WRAD stimulates methyltransferase activity is to promote binding of the methyltransferase complexes to S-adenosyl

methionine (F. Cao et al., 2010, Shinsky et al., 2015). Subsequent structural studies of the MLL1 complex has demonstrated that the MLL1 SET domain naturally adopts an open conformation that does not promote efficient transfer of methyl groups from S-adenosyl methionine to H3K4, and that interactions between the RBBP5-ASH2L heterodimer and the MLL1 SET domain restrict its conformation to produce a catalytically-competent conformation (Li et al., 2016, Rahman et al., 2022). A further role of WRAD is to ensure proper engagement between SET1/MLL proteins and nucleosomes. Interestingly, when the tyrosine at the tyr/phe switch position of MLL1 was mutated to a phenylalanine, robust H3K4 di- and tri-methyltransferase activity was observed on histone H3 peptide substrates, but not reconstituted nucleosomes (Shinsky et al., 2014). Only when the MLL1 mutant was in complex with WRAD could it methylate a nucleosomal substrate. This indicates that WRAD has functions beyond simply altering the active site of the SET domain and has a broader structural role in enabling proper engagment between SET1/MLL proteins and the nucleosome. Interestingly, mono-ubiquitination of lysine 120 on histone H2B (H2BK120ub1) has been shown to stimulate the activity of SET1/MLL complexes *in vitro* and *in vivo* (Kim et al., 2009, Kwon et al., 2020). Recent structural studies have indicated that the molecular basis of this stimulation is mediated through WRAD, and more specifically, by RBBP5. Cryo-electron microscopy studies of the MLL1 complex have shown that RBBP5 forms contacts with the ubiquitin on H2BK120, which 'docks' the complex onto the nucleosome (Fig. 1.3)(Xue et al., 2019, Rahman et al., 2022). Interactions between WRAD and MLL1 serve to orientate the complex on the nucleosome such that it adopts an active conformation with the active site in close proximity to H3K4. Therefore, WRAD functions to position the SET1/MLL complexes on nucleosomes through engagment with H2BK120ub1, and also to induce allosteric changes in the active site of the SET domain to allow for efficient methylation of H3K4.

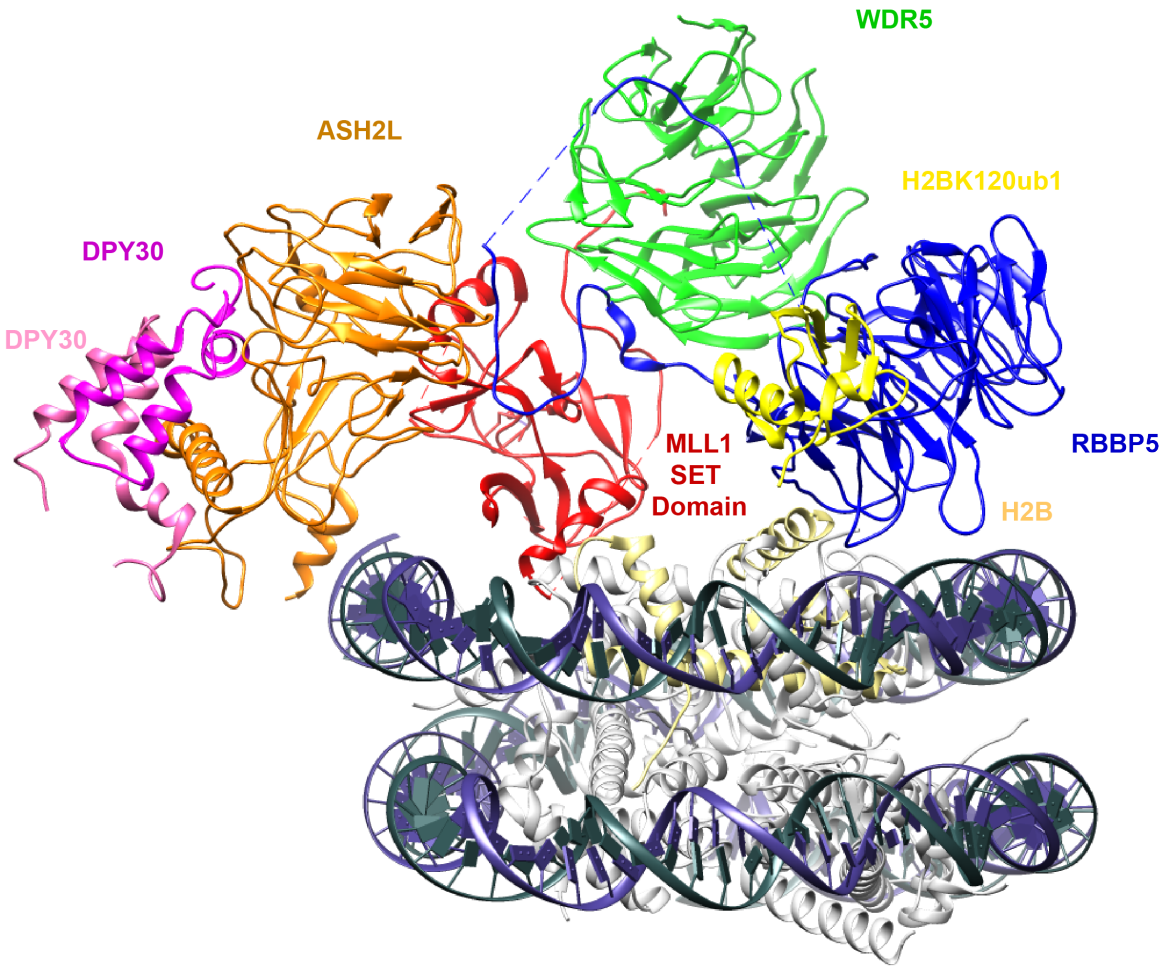

Figure 1.3: Structure of the MLL1 SET domain in complex with WRAD on a nucleosome

Structure of the human MLL1 SET domain in complex with WRAD on a nucleosome containing H2BK120ub1 as determined by cryo-electron microscopy to a resolution of 4.25Å (Rahman et al., 2022, 7UD5). The two copies of DPY30 are in different shades of pink, ASH2L is in orange, MLL1 SET domain is in red, WDR5 is in green, RBBP5 is in navy blue, and the ubiquitin on H2BK120 is in yellow. H2B is in gold, and all other histones are in light grey. DNA is coloured with one strand in dark blue and the other strand in dark green.

1.5.3 SET1A/B Complexes

Roles of SET1A/B during development

SET1A and SET1B are both esssential for mammalian development, during which they carry out non-redundant functions (Kranz and Anastassiadis, 2020). SET1A is required for gastrulation, and SET1A-knockout embryos die at embryonic day 7.5 (E7.5) (Bledau et al., 2014). Furthermore, SET1A is required for embryonic stem cell viability, as no viable ES cells can be derived from SET1A-knockout blastocysts (Bledau et al., 2014). SET1B is required later during development, as SET1B-knockout embryos die at E11.5 (Bledau et al., 2014). In contrast to SET1A, SET1B is not required for ES cell viability. However, SET1B is required for oogenesis, while SET1A is not (Brici et al., 2017). Therefore, despite being highly similar proteins and sharing many of their interactions with other proteins, SET1A and SET1B are not functionally redundant.

SET1A/B domain architecture and complex subunits

SET1A and SET1B are encoded by the *Setd1a* and *Setd1b* genes, respectively, and are the smallest members of the SET1/MLL proteins, with SET1A consisting of 1,716 amino acids and SET1B consisting of 1,985 amino acids. The SET1 proteins are highly similar to each other and are highly similar to the yeast homolog, with the same domain organisation. The domain organisation of SET1A and SET1B are relatively simple. The N-terminus consists of an RNA-binding domain (RRM) and a short N-terminal region that is required for interaction with the WD40-repeat protein 82 (WDR82), and the C-terminus consists of the well conserved SET domain and post-SET domain (Fig. 1.4). The central portion of SET1A and SET1B lack significant sequence conservation and its function is not well understood. Interestingly, the SET1 proteins differ from the MLL proteins in that they contain an n-SET domain N-terminal to the SET domain. The n-SET domain contains an arginine-rich motif, the RXXXRR motif, which has been proposed to have an important role in stimulating methyltransferase activity through interactions with ubiquinated H2B (Worden and Wolberger, 2019, Worden et al., 2020, Bae et al., 2020). Additionally, the n-SET domain contains interaction sites for CFP1, an integral component of SET1 complexes (Fig. 1.4) (Kwon et al., 2020). CFP1 contains a CXXC domain and two PHD fingers, and these domains have been shown to target the SET1 complexes to the CGI promoters of highly-transcribed genes (Brown et al., 2017). Additionally, *in vitro* experiments using SET1 complexes have shown that CFP1 plays a role in the methyltransferase activity of SET1 complexes on nucleosomes containing ubiquinated H2B, a role that is conserved in yeast, where knockout of the homolog Spp1 results in major global reductions in H3K4me3 but not H3K4me1/2 (Kwon et al., 2020). Therefore, CFP1 appears to have a conserved role in SET1 complexes in promoting H3K4 tri-methyltransferase activity, and in mammals has acquired a new role in the targeting of SET1 complexes to CGI chromatin. SET1 proteins also interact with host cell factor 1 (HCF1), biorientation deficient 1 (BOD1), and biorientation deficient like 1 (BOD1L) (Fig. 1.4) (van Nuland et al., 2013). Interestingly, only SET1A interacts with HCF1; SET1B lacks the required HCF1-interacting motif (J. H. Lee et al., 2007, Ciotta et al., 2023). Additionally, BOD1L appears to be a SET1A-specific interactor while BOD1 appears to be a SET1B-specific interactor. BOD1/BOD1L are mammalian homologs

of the yeast Shg1 protein, which is part of the yeast Set1 complex. Interestingly, in both yeast and mESCs, loss of BOD1/BOD1L has been shown to increase global levels of H3K4me2/3, and BOD1/BOD1L have been suggested to restrain the H3K4 methyltransferase activity of SET1 proteins through an as-yet-uncharacterized mechanism (Roguev et al., 2001, Ciotta et al., 2023).

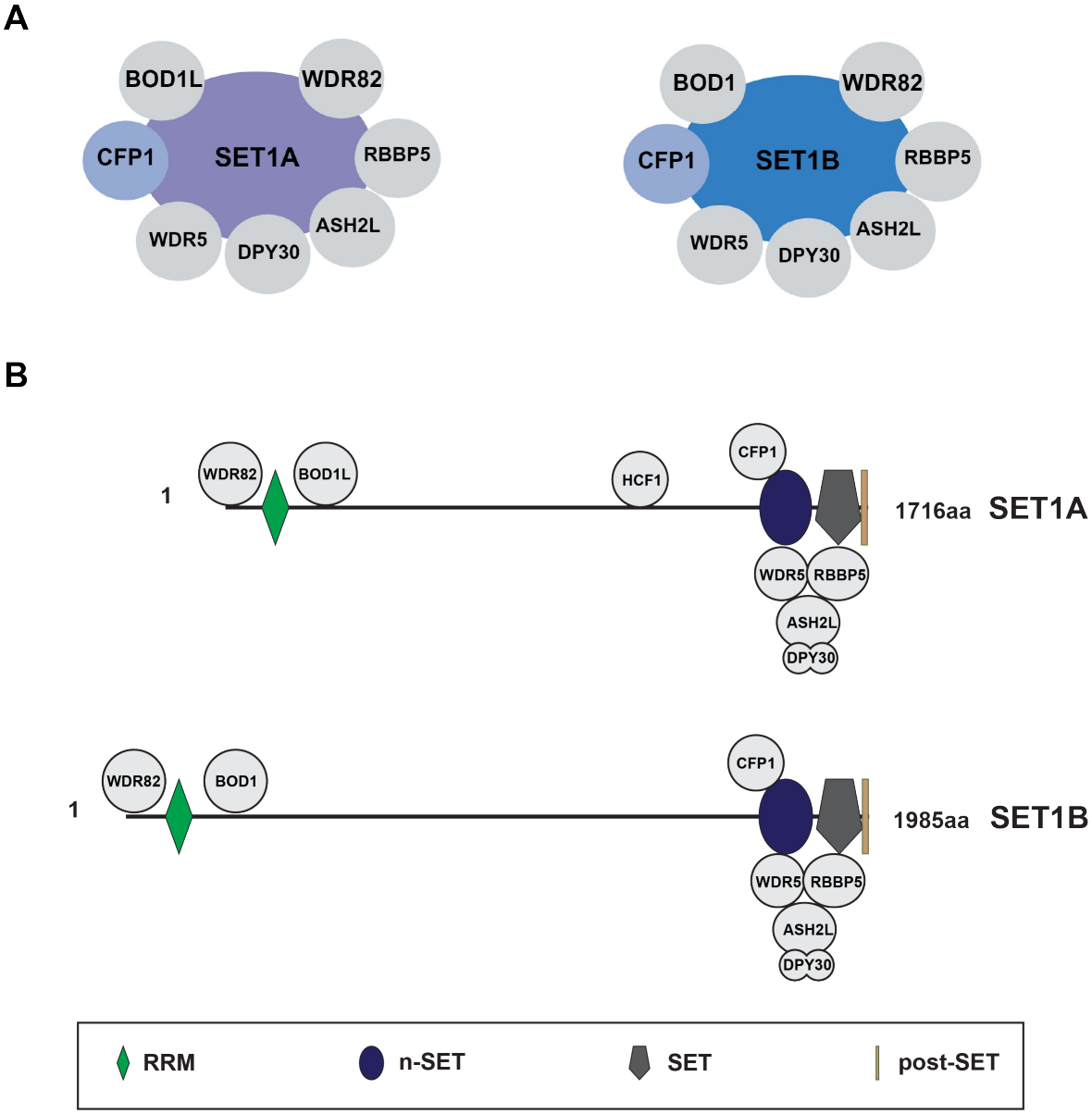

Figure 1.4: SET1A/B domain organisation and interactors.

(A) Cartoons illustrating SET1A/B complexes.
(B) Domain organisation of SET1A/B proteins along with cartoons illustrating binding of known interactors. Domain architecture was analysed using SMART. Cartoons illustrating the known interactors of SET1A/B are also shown at the approximate binding sites on SET1A/B. Cartoon figures of SET1A/B interactors are not to scale.

Interestingly, the interaction between SET1 proteins and WDR82 provides a link between SET1 complexes and the transcriptional machinery. WDR82 is a component of at least four complexes: the SET1 complexes, the PNUTS-PP1 complex, and the ZC3H4 restrictor complex (J. H. Lee and Skalnik, 2008, J.-H. Lee et al., 2010, Estell et al., 2021). All four complexes are involved in transcription: SET1 complexes are H3K4 methyl-

transferases, the PNUTS-PP1 complex is involved in transcription termination at the 3' end, and the ZC3H4 restrictor complex is involved in premature termination at the 5' end of genes, at enhancers, and other non-genic regions. WDR82 has been shown to interact directly with RNAPII that is phosphorylated at serine 5 of its CTD (S5P), but not at serine 2 or unmodified RNAPII (J. H. Lee and Skalnik, 2008). Interestingly, a phosphoproteomics screen revealed that SET1 complexes also interacted with serine 5-phosphorylated RNAPII, suggesting that WDR82 targets SET1 complexes directly to the transcriptional machinery (Ebmeier et al., 2017). Indeed, knockdown of WDR82 in mouse macrophages resulted in a failure to generate *de novo* H3K4me3 peaks in reponse to lipopolysaccharide stimulation (Austenaa et al., 2015).

SET1 complexes localise to CGI promoters of highly-transcribed genes through CFP1

SET1A has been proposed to be the primary H3K4 tri-methyltransferase in mammals, responsible for the majority of H3K4me3 in most cell types. Studies of the mammalian SET1 complexes initially focused on the non-catalytic subunits of the complexes, specifically on CFP1, due to its specific association with CGIs. Early studies using constitutive CFP1-knockout mESCs demonstrated that loss of CFP1 resulted in reductions in H3K4me3 at the promoters of highly-expressed genes, and *de novo* accumulation of H3K4me3 at the promoters of genes induced by the DNA-damaging agent doxorubicin was reliant on CFP1 (Clouaire et al., 2012, Clouaire et al., 2014). Subsequently, it was shown that although all CFP1 binding sites are CGIs, only a subset of CGIs are bound by CFP1. CFP1 contains a CXXC domain that binds CGIs and a PHD finger that binds H3K4me3 (Brown et al., 2017). By generating mutants of the CXXC domain and the PHD domain on their own or in combination, it was demonstrated that CFP1 relies on both H3K4me3-binding and CGI binding in order target the SET1 complexes to CGI promoters of highly-transcribed genes that already contained high levels of H3K4me3 (Brown et al., 2017). Given that highly-transcribed genes have higher levels of H3K4me3 at their promoters, it was therefore assumed that SET1 complexes would be responsible for the majority of H3K4me3 due to their targeting to highly-transcribed genes (Brown et al., 2017). Furthermore, quantitative mass spectrometry experiments indicated that SET1A was the most abundant of the SET1/MLL proteins in HeLa cells (van Nuland et al., 2013). Subsequently, conditional knockout of SET1A and SET1B demonstrated that in mESCs, SET1A is responsible for the majority of H3K4me3 as assessed by immunoblotting, and was required for ES cell viability (Bledau et al., 2014). On the other hand, conditional deletion of the *Setd1b* gene in mESCs resulted in minimal changes in bulk levels of H3K4 methylation, and subsequent studies have demonstrated that SET1B is lowly expressed in many cell types (Bledau et al., 2014, van Nuland et al., 2013). Rather, it was subsequently shown that SET1B plays a larger role in depositing H3K4me3 in the developing oocyte, where it primarily deposits H3K4me3 at highly-transcribed gene promoters (Hanna et al., 2022). However, even in the developing oocyte, SET1B only deposits a minority of the H3K4me3; wheras MLL2 is the primary H3K4 tri-methyltransferase and deposits approximately 80% of all H3K4me3.

SET1 complexes support transcription through an H3K4 methylation-independent mechanism

Recent work has begun to challenge the model that SET1A is the primary H3K4 methyltransferase in mammals. Studies using mES cell lines in which the SET domain of SET1A was constitutively deleted demonstrated that ablating SET1A catalytic activity alone had surprisingly modest effects on H3K4 methylation and no effect on ES cell viability (Sze et al., 2017, K. Cao et al., 2017, Sze et al., 2020). These studies suggested that loss of the catalysis-independent functions of SET1A is the primary reason for loss of ES cell viability. Additionally, the requirement for SET1A for ES cell viability makes it challenging to interpret results from conditional knockout of SET1A due to confounding secondary effects. Recently, the Klose lab has employed the degradation tag (dTAG) degron system, which allows for rapid, near-complete degradation of a tagged protein of interest upon addition of the small bifurcational small molecule dTAG-13 (Nabet et al., 2018). Using an mES cell line in which SET1A and SET1B can be removed rapidly before the onset of cell viability defects, the Klose lab has demonstrated that removing SET1A/B does does not result in major reductions in H3K4me3, despite major transcriptional defects at almost a quarter of all transcribed genes (Hughes et al., 2023). Instead, depletion of SET1A/B resulted in a modest reduction in H3K4me3 across all CGI gene promoters, with highly-transcribed genes exhibiting a more pronounced reduction. Strikingly, reductions in H3K4me3 bore very little correlation with changes in transcription or gene expression, with genes whose expression was greatly reduced after SET1A/B depletion exhibiting similar reductions of H3K4me3 at the promoter as those genes whose expression was unchanged after SET1A/B depletion. This echos previous studies using mESCs in which CFP1 was knocked-out; although removing CFP1 resulted in reductions in H3K4me3 at the promoters of highly-expressed genes, removing CFP1 resulted in reductions in gene expression at primarily moderately- and lowly-transcribed genes (Brown et al., 2017). Indeed, a small N-terminal fragment of SET1A and SET1B was sufficient in induce expression of a reporter gene *in vivo*. Subsequent experiments using transient transcriptome sequencing (TT-seq) to assess ongoing transcription demonstrated that SET1A/B primarily regulated the ability of RNAPII to enter into productive elongation at a subset of moderately- and lowly-transcribed genes by counteracting premature transcription termination by the restrictor complex consisting of WDR82 and ZC3H4 (Hughes et al., 2023). Therefore, at least in mESCs, SET1 complexes can support gene expression through an H3K4 methyltransferase-independent mechanism.

1.5.4 MLL1/2 Complexes

Role of MLL1/2 during development

MLL1 and MLL2 are closely related to drosophila Trithorax (Trx), and are similarly required for the expression of Hox genes (Milne et al., 2002, Glaser et al., 2006). Both MLL1 and MLL2 are required during the early stages of mammalian development, but are not required for the self-renewal of embryonic stem cells (Denissov et al., 2014). Although both MLL1 and MLL2 are required for development, they are required during different stages of development, and MLL1-knockout and MLL2-knockout embryos display different phenotypes. Knockout of MLL1 results in defects in skeletal development and defects in haematopoiesis, resulting in embryonic lethality at E10.5 (Yu et al., 1998). Knockout of MLL2 results in increased apopotosis in embryos and growth retardation leading to embryonic lethality before E11.5 (Glaser et al., 2006). Interestingly, the H3K4 methyltransfease activity of MLL1 appears to dispensible for at least some of its functions during development. Deletion of the SET domain of MLL1 results in the development of viable and fertile adult mice, albeit with defects in skeletal development, indicating that the role of MLL1 during haematopoiesis is independent of its H3K4 methyltransferase activity (Terranova et al., 2006). Subsequent work has demonstrated that deletion of the MLL1 SET domain does not impair the expression of MLL1 target genes in haematopoietic stem cells (Mishra et al., 2014). These findings indicate that MLL1 has both catalysis-dependent and catalysis-independent roles during development. A systematic study of the catalysis-independent roles of MLL2 during development has been lacking, in part due to difficulties in selectively abrogating MLL2 methyltransferase activity: deletion of the MLL2 SET domain has been shown to result in protein instability (Sze et al., 2020).

MLL1/2 domain architecture and interacting partners

MLL1 and MLL2 are encoded by the *Kmt2a* and *Kmt2b* genes, respectively, and like their drosophila counterpart, are both cleaved by Taspase I to form an N-terminal fragment and a C-terminal fragment. These two fragments re-associate after cleavage through interactions between an FY-rich domain on the N-terminal fragment (FYRN domain) and an FY-rich domain on the C-terminal fragment (FYRC domain) (Fig. 1.4) (Hsieh et al., 2003, Takeda et al., 2006). For both MLL1 and MLL2, the N-terminal fragment is substantially larger than the C-terminal fragment: for both MLL1 and MLL2 it creates a 320kDa N-terminal fragment and a 180kDa C-terminal fragment for MLL1 and an 80kDa C-terminal fragment for MLL2. This cleavage separates the SET domain from the CXXC domain: the SET domain is on the C-terminal fragment while the CXXC domain is on the N-terminal fragment (Fig. 1.4). Additionally, the N-terminal fragment contains all the chromatin-interacting domains; in addition to the CXXC domain this includes PHD domains and AT hook domains (Fig. 1.4). This results in a functional separation between the two fragments: the N-terminal fragment is primarily responsible for targeting the proteins to chromatin and the C-terminal fragment is reponsible for H3K4 methylation. MLL2-N exogenously-expressed on its own in MLL2-knockout mouse embryonic stem cells can be detected on chromatin by ChIP, while MLL2-C exogenously-expressed on its own cannot (Denissov et al., 2014). Many leukaemias are driven by a chromsomal translocation that result in the N-terminal frag-

ment of MLL1 being fused to a variety of fusion partners. These fusion proteins are primarily localised to CGI promoters via the CXXC domain present in MLL1-N. Additionally, MLL1/2 associate with menin, an MLL1/2-specific interactor, via their N-terminal fragments (Fig. 1.4). Menin has been shown to recruit MLL1 to Hox genes, and has been shown to be required for their expression (Milne et al., 2004). Additionally, MLL1 interacts with the H3K36me3-reader LEDGF (also known as PSIP1) through its interaction with menin, although this interaction is not observed for MLL2 (Fig. 1.4). The role of menin in the MLL2 complex is less well understood, but it has been suggested that menin preferentially targets both MLL1 and MLL2 complexes to more actively-transcribed genes (Sparbier et al., 2023). Additionally, HCF1 has also been reported to interact to MLL1 and MLL2 through their N-terminal domains.

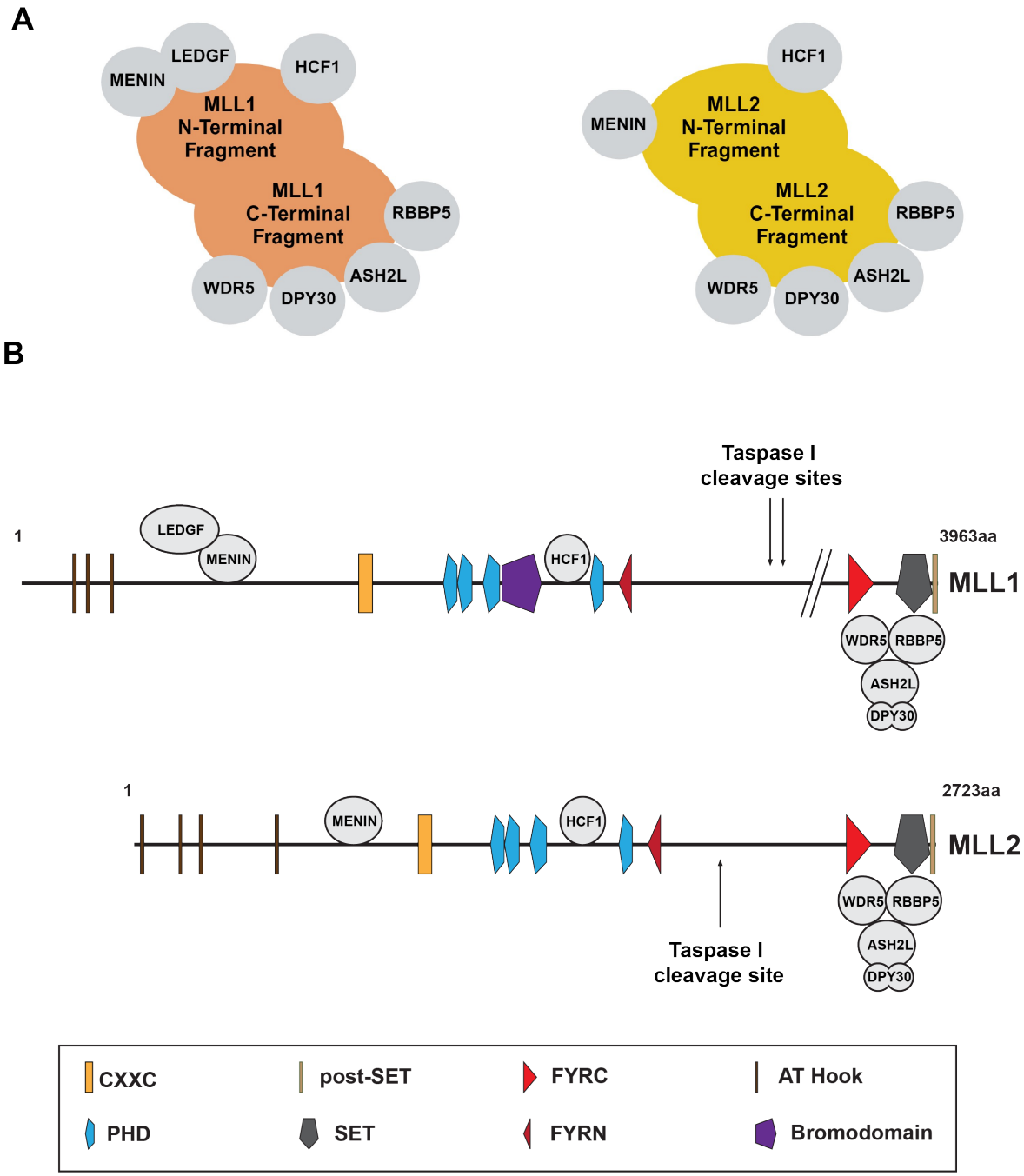

Figure 1.5: MLL1/2 domain organisation and interactors.

(A) Cartoons illustrating the MLL1/2 complexes.
(B) Domain organisation of MLL1/2 proteins along with cartoons illustrating binding of known interactors. Domain architecture was analysed using SMART. Cartoons illustrating the known interactors of MLL1/2 are also shown at the approximate binding sites on MLL1/2. Cartoon figures illustrating MLL1/2-interactors are not to size.

MLL1/2 primarily localize to CGIs via their CXXC domains

Genome-wide analysis of MLL1 and MLL2 occupancy throughout the genome indicates that both MLL1 and MLL2 bind broadly across the genome at CGIs (Denissov et al., 2014, Miyamoto et al., 2020). In contrast to the SET1 complex, MLL1 and MLL2 do not display any preference towards highly-transcribed genes, which suggests that they primarily rely on their CXXC domains for chromatin targeting (Denissov et al., 2014). In

mouse embryonic stem cells, disruption of the MLL2 CXXC domain has been shown to largely abolish MLL2 binding to chromatin (D. Hu et al., 2017). In leukaemia cells, MLL1 fusion proteins in which MLL1-N is fused to a variety of interaction partners invariably localise to CGIs (Milne et al., 2010, Okuda et al., 2014, Kerry et al., 2017). However, it is worth noting that MLL1/2 do not soley rely on their CXXC domains for localisation throughout the genome. Both MLL1 and MLL2 contain PHD fingers in the N-terminal fragment, which have been shown to bind H3K4me3, providing a positive feedback mechanism for sustaining H3K4me3 deposition (Milne et al., 2010, Chang et al., 2010, Vermeulen et al., 2010). Interestingly, multiple studies have reported that MLL1 interacts with the transcription elongation factor RNA Polymerase II-associated Factor 1 (PAF1) via its CXXC domain, an interaction not shared with MLL2, or indeed with any other CXXC-domain containing protein (Milne et al., 2010, Muntean et al., 2010, Miyamoto et al., 2020). Subsequent studies have shown that subtle differences in the amino acid sequence flanking the CXXC domains of MLL1 and MLL2 account for these interactions (Bach et al., 2009).

MLL1/2 primarly contributes to transcription-independent H3K4 methylation

Paradoxically, despite broadly associating with most CGI promoters, MLL1 and MLL2 appear to only be responsible for H3K4 methylation and expression of a subset of genes. Conditional or constitutive knockout of MLL2 results in significant loss of H3K4me3 at a subset of promoters, which typically numbers between 2,000 and 4,000, depending on the study (D. Hu, Garruss, et al., 2013, Denissov et al., 2014, D. Hu et al., 2017, Mas et al., 2018, Douillet et al., 2020). These genes tend to be lowly-expressed and are involved in ESC differentiation and development, such as Hox genes. Additionally, these genes are typically enriched for H3K27me3 and binding of Polycomb proteins. This has led to the view that these promoters represent a 'bivalent' chromatin state bearing both permissive and repressive histone modifications which are primed for the rapid induction of transcription during ES differentiation, upon which the state would resolve into an active chromatin state (D. Hu, Garruss, et al., 2013, Denissov et al., 2014). However, there is little evidence to support this hypothesis at present, and many of these genes retain the ability to be induced upon retinoic-acid differentiation even in the absence of MLL2 and H3K4me3 (Denissov et al., 2014). Interestingly, MLL2 is the more dominant H3K4 methyltransferase in mESCs, and loss of MLL1 has very little effect on H3K4me3 (Denissov et al., 2014). Instead, MLL1 primarily functions redundantly to deposit H3K4me3 with MLL2; removal of both MLL1 and MLL2 leads to further reductions in H3K4me3 in retinoic-acid-differentiated mESCs. Additionally, upon retinoic-acid differentiation, MLL1 deposited H3K4me3 at a subset of promoters in the absence of MLL2. Therefore, the role of MLL1 in depositing H3K4me3 in mESCs are minimal and are limited to targets shared with MLL2.

MLL2 has been particularly associated with depositing H3K4me3 in a transcription-independent manner. Early studies had demonstrated that artificial insertion of a transcriptionally inert CGI into mESCs was sufficient to generate H3K4me3, and it was found that this de novo H3K4me3 generation still occurred when CFP1 was constitutively knocked-out, thus implying that MLL1/2 were primarily responsible for H3K4me3 deposition at transcriptionally-inert CGIs (Wachter et al., 2014). During oogenesis, extensive accumulation

of H3K4me3 occurs, forming broad domains at intergenic regions. Approximately 80% of this transcription-independent H3K4me3 is deposited by MLL2, while H3K4me3 at gene promoters is in part deposited by SET1B (Hanna et al., 2018, Hanna et al., 2022).

Conditional or constitutive knockout of MLL2 in mESCs have typically resulted in modest changes in gene expression or nascent transcription, with only hundreds of genes showing reductions in expression or transcription. Interestingly, these genes tend to be genes which already depend on MLL2 for H3K4me3, and the apparent discrepancy between the thousands of gene that depend on MLL2 for H3K4me3 and the hundreds of gene that depend on MLL2 for gene expression may be that many MLL2 target genes are simply not transcribed to begin with. Interestingly, abolishing MLL2 methyltransferase activity by mutating a key tyrosine residue in the SET domain that does not reduce protein stability has been shown to result in reductions in gene expression in mESCs and during differentiation into primordial germ cells, indicating that in certain instances, methyltransferase activity is required for expression of MLL2 target genes (D. Hu et al., 2017).

1.5.5 MLL3/4 Complexes

Roles of MLL3/4 during development

Both MLL3/4 are required for mammalian development, although they are required at different stages. MLL3 is required for development of the lung, and MLL3-knockout embryos die at birth from asphyxiation (Ashokkumar et al., 2020). MLL4 is required during earlier stages of development, and MLL4-knockout embryos display abnormalities during gastrulation and die at E9.5 (Ashokkumar et al., 2020, J. E. Lee et al., 2013). Furthermore, MLL4 has been shown to be essential for adipogenesis and myogensis (J. E. Lee et al., 2013, J.-E. Lee et al., 2020). Therefore, similar to SET1A/B and MLL1/2, MLL3/4 represent two structurally similar proteins that have distinct roles during development.

Domain Architecture of MLL3/4 and Interacting Partners

MLL3 and MLL4 are encoded by the *Kmt2c* and *Kmt2d* genes, respectively, and are large proteins that are related to the Drosophila H3K4 monomethyltransferase Trithorax-related (Trr) (D. Hu, Gao, et al., 2013, Herz et al., 2012). MLL3 and MLL4 are highly similar proteins and share most of their characterized protein domains. Similar to MLL1 and MLL2, MLL3 and MLL4 have a series of PHD fingers in the N-terminal half of the proteins and the SET domain in the C-terminal half (Fig. 1.6). Interestingly, Drosophila Trr is only homologous to the C-terminal half of MLL3/4; the Drosophila homolog of the N-terminal halves of MLL3/4 is a separate protein called Lost PHD domains of Trithorax-related (LPT) (Herz et al., 2012). Similar to MLL1 and MLL2, MLL3 and MLL4 also have FYRN and FYRC domains, but they are situated close to each other, unlike in MLL1 and MLL2 (Fig. 1.6). Additionally, MLL3 and MLL4 do not undergo cleavage by Taspase I. Whether MLL3/4 bind to H3K4me3 via their PHD fingers *in vivo* is currently not well understood, although *in vitro* studies suggest that they bind to nucleosomes enriched in H3K4me3 and H3 acetylation (Stroynowska-Czerwinska et al., 2023). Initial studies have suggested that a subset of the PHD domains of MLL4 collectively recognize asymmetrically dimethylated H4R3 and that this interaction has some influence on the H3K4 methyltransferase activity of MLL4 *in vivo* (Dhar et al., 2012, Dhar et al., 2018). The requirement for the PHD fingers of MLL3/4 for catalytic activity *in vivo* is inferred by the observation that depletion of LPT in drosophila by RNAi on its own results in reductions in bulk H3K4me1 levels (Herz et al., 2012). Interestingly, constitutive deletion of an N-terminal section of MLL4 resulted in reduced MLL4 occupancy in mESCs along with defects in gene expression comparable to MLL4-knockout cells, indicating that the N-terminal domains of MLL4 have important roles in recruiting MLL4 to chromatin and in its function in transcription (K. Cao et al., 2018).

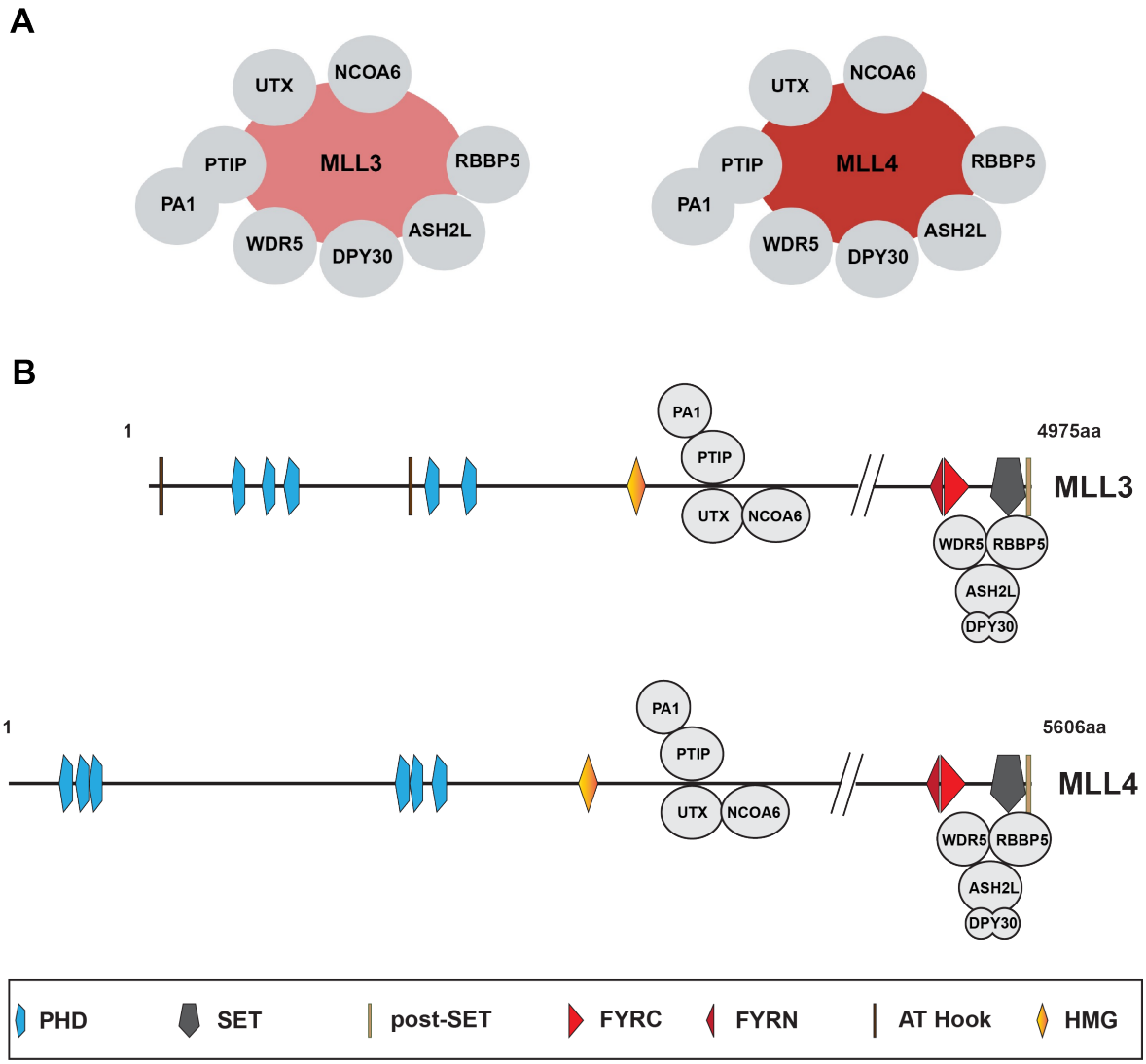

Figure 1.6: MLL3/4 domain organisation and interactors.

(A) Cartoons illustrating MLL3/4 complexes.
(B) Domain organisation of MLL3/4 proteins along with cartoons illustrating binding of known interactors. Domain architecture was analysed using SMART. Cartoons illustrating the known interactors of MLL3/4 are also shown at the approximate binding sites on MLL3/4.

Multiple MLL3/4-specific binding partners have been described, and all of them appear to be shared by MLL3 and MLL4. To date, four primary interacting proteins have been described for MLL3/4 complexes - UTX, NCOA6, PTIP, and PA1 (Fig. 1.6). All four interactors have a role in recruiting MLL3/4 to chromatin or mediating interactions between MLL3/4 and other chromatin-associated proteins such as transcription factors and histone modifiers. UTX has been of particular interest due to its role as an H3K27 demethylase and has been shown to be involved in facilitating MLL3/4 function at enhancers through interacting with the histone acetyltransferase P300 (S. P. Wang et al., 2017). Nuclear co-activator 6 (NCOA6) is a transcriptional co-activator that interacts with transcription factors and nuclear receptors and has been shown to recruit MLL3/4 to target loci (Mahajan and Samuels, 2008). Pax transactivation domain-interacting protein (PTIP) and PTIP-associated protein 1 (PA1) interact with MLL3/4 as a dimer, and has been shown to recruit MLL3/4 to chromatin (Cho et al., 2007, Cho et al., 2009), Schwab et al., 2011). Interestingly, purifications of WDR82 from human

cells and mESCs have identified MLL4, but not MLL3, as an interactor, although this interaction is not well characterized (J.-H. Lee et al., 2010, Jessica Kelley and Rob Klose, unpublished data). Given that WDR82 interacts with the C-terminal domain of RNAPII, it is possible that WDR82 recruits MLL4 to a subset of promoters, although further investigation into this interaction is required.

MLL3/4 primarily function at enhancers to deposit H3K4me1/2

An important difference between MLL3/4 and SET1A/B/MLL1/2 is that they do not contain CXXC domains, nor do they associate with a CXXC-domain containing protein. MLL3/4 themselves do not appear to have any intrinsic ability to target chromatin, and primarily rely on interactions with sequence-specific transcription factors for binding to chromatin. For example, ectopic expression of the adipogenic transciption factor C/EBPβ in pre-adipocytes is sufficient to recruit MLL4 to a subset of enhancers where they deposit H3K4me1 and recruit CBP/P300 (J. E. Lee et al., 2013). Similarly, ectopic expression of lineage-determining transcription factors GATA6 and NEUROD1 in mESCs are sufficient to recruit MLL4 to enhancers (Xie et al., 2023). In MCF-7 cells, knockdown of the pioneer factor FOXA1 by RNAi has been shown to discrupt MLL3 binding (Jozwik et al., 2016). Accordingly, ChIP-seq analyses of MLL3/4 occupancy in a variety of cell types and biological contexts have shown that they primarily localise to distal regions, many of which are annotated enhancers (D. Hu, Gao, et al., 2013, J. E. Lee et al., 2013, C. Wang et al., 2016, Jozwik et al., 2016). In accordance with their dependence on transcription factor recruitment to chromatin, MLL3/4 occupancy as assessed by ChIP-seq varies considerably between cell types and varies between different stages of ES cell differentiation (C. Wang et al., 2016, Xie et al., 2023).

Studies in mESCs have consistently shown, using either mass spectrometry or immunoblotting, that MLL3/4 are responsible for approximately 50% of H3K4me1 (C. Wang et al., 2016, Dorighi et al., 2017). This indicates that although MLL3/4 are responsible for a larger proportion of H3K4me1 compared to SET1A/B and MLL1/2, they are not the only H3K4 mono-methyltransferases. Interestingly, although MLL3/4 are localised to a large number of enhancers in mESCs, they are not responsible for all H3K4me1 at enhancers. Indeed, a recent study reported that H3K4me1 at many MLL3/4-bound enhancers are not dependent on MLL3/4 (Boileau et al., 2023). This indicates that ChIP-seq analysis of MLL3/4 occupancy may not be reflective of their H3K4 methyltransferase activity. Similar to SET1A/B and MLL1/2, MLL3/4 do not contribute equally to H3K4me1 deposition, and MLL4 has been reported to be the major H3K4 monomethyltransferase in most cell types. For example, in mESCs, knockout of MLL3 has little effect on bulk levels of H3K4me1 while knockout of MLL4 or deletion of its SET domain results in a substantial reduction of bulk H3K4me1 levels (K. Cao et al., 2017, K. Cao et al., 2018, Xie et al., 2023). Interestingly, in mouse embryonic fibroblasts, knockdown of MLL4 by shRNA does not result in H3K4me1 reduction unless MLL3 was also knocked out, indicating that MLL3 is able to compensate for loss of MLL4 in more differentiated cell types (D. Hu, Gao, et al., 2013). Additionally, mouse embryos carrying catalytically-dead mutants of MLL3 and MLL4 die at a much earlier stage of development compared to mouse embryos carrying these mutants individually, indicating that MLL3/4 have extensively redundant catalytic roles during development (Xie et al., 2023).

Knockout of MLL3/4 in mESCs leads to a modest reduction in global levels of H3K4me2 as assessed by immunoblotting (J. E. Lee et al., 2013, C. Wang et al., 2016, Xie et al., 2023). ChIP-seq analysis of H3K4me2 levels in MLL3/4-knockout mESCs indicate that these reductions primarily occur at enhancers, and loss of MLL3/4 led to a failure of a subset of enhancers to gain H3K4me2 during adipogenesis (J. E. Lee et al., 2013). On the other hand, knockout of MLL3/4 has very little effect on H3K4me3 in most cell types. Interestingly, MLL4 has been reported to deposit H3K4me3 at broad domains in the mouse cerebella (Dhar et al., 2018). This is in accordance with *in vitro* studies which demonstrates that MLL4 is capable of generating H3K4me3 on a nucleosomal substrate, although its H3K4 tri-methyltransferase activity is much weaker than its H3K4 di- and mono-methyltransferase activity (Kwon et al., 2020). This suggests that MLL4 may have a greater role in depositing H3K4me3 in certain biological contexts.

The role of catalysis in MLL3/4 function

The role of the H3K4 mono-methyltransferase activity of MLL3/4 in their functions at enhancers remains debated. The role of MLL3/4 H3K4 methyltransferase activity in enhancer function has been well studied due to the identification of a key tyrosine residue in the SET domains of MLL3 (Y4792) and MLL4 (Y5477) that is essential for methyltransfease activity but not required for protein stability or chromatin association (Dorighi et al., 2017). By mutating these tyrosine residues, mES cells in which MLL3/4 are catalytically dead have been generated. Bulk levels of H3K4me1 in these cell lines are approximately 50% of wild type cells and are comparable to MLL3/4-knockout cells (Dorighi et al., 2017). Using this mES cell line in which MLL3/4 are rendered catalytically-dead, it has been demonstrated that the methyltransferase activity of MLL3/4 are partially responsible for H3K27ac deposition at enhancers, but largely dispensable for the loading of RNAPII at enhancers, eRNA transcription, and gene expression. Interestingly, consititutive knockout of the MLL3/4 proteins themselves resulted in major loss of H3K27ac, RNAPII binding, and eRNA transcription from MLL3/4-bound enhancers, with major reductions in gene expression, indicating that non-catalytic activites of MLL3/4 play a larger role in enhancer activation in mESCs. A subsequent study has also shown that the methyltransferase activity of MLL3/4 is dispensible for differentiation of mESCs to all three germ layers (Xie et al., 2023).

Subsequent studies have shed some light on the catalysis-independent roles of MLL3/4 at enhancers, and suggest that the interaction between MLL3/4 and UTX is crucial for P300 recruitment and subsequent enhancer activation (S. P. Wang et al., 2017). However, in both mESCs and drosophila, loss of MLL3/4 catalytic activity results in modest reductions in H3K27ac, indicating that H3K4me1 has at least some role in recruiting CBP/P300 to enhancers or stabilizing their interaction with enhancers (Rickels et al., 2017, Dorighi et al., 2017). On the other hand, some studies suggest that MLL3/4-dependent H3K4me1 have important roles in enhancer activation. Studies have shown that the catalytic activity of MLL3/4 is required for the maintenance of a subset of enhancer-promoter interactions in mESCs and that MLL3/4-dependent H3K4me1 is required for binding of the BAF complex to enhancers (Yan et al., 2018, Local et al., 2018). Given that simultaneous mutations of the SET domains of MLL3 and MLL4 leads to early embryonic lethality, MLL3/4-dependent H3K4me1 clearly have a prominent role in regulating transcription or some other process during development,

although further investigations are required (Xie et al., 2023).

1.6 Aims of thesis

Despite the importance of H3K4 methylation in the regulation of transcription, how SET1/MLL complexes shape the deposition of H3K4 methylation across the genome in mammalian cells is not well understood. Previous work from the Klose lab has demonstrated that the fundamental paradigm of SET1 complexes being responsible for the majority of H3K4me3 in mESCs is not accurate, and that previous analyses using perturbation methods that do not deplete SET1A/B proteins with sufficient rapidity are confounded by secondary effects. This illustrates a need for a systematic investigation of all SET1/MLL complexes using rapid degradation approaches in order to functionally dissect their roles in H3K4 methylation.

In this thesis, I aim to investigate how SET1/MLL complexes methylate H3K4 across the genome in mouse embryonic stem cells. Using degron-mediated depletion of SET1/MLL proteins, either on their own or in combination with each other, I investigate their contributions to H3K4 methylation by quantitatively mapping all three H3K4 methylation states. I first focus on studying the four CpG-island-associated H3K4 methyltransferases, SET1A/B and MLL1/2, and show that they synergise to deposit H3K4me3 at CpG islands, but do not account for all H3K4me3 at CpG island promoters (Chapter 3). I next re-evaluate our understanding of the contribution of MLL3/4 to higher H3K4 methylation states and investigate whether they account for the remaining H3K4me3 (Chapter 4). Having found that all six SET1/MLL methyltransfeases only account for approximately half of all H3K4 methylation, I find that the remaining H3K4 methylation is dependent on active transcription and requires the WRAD subcomplex (Chapters 4 and 5). I then use biochemical techniques to characterize this novel, WRAD-containing H3K4 methyltransferase (Chapter 5). Finally, I use mass spectrometry to identify a putative alternative SET1B isoform which may be responsible for a large proportion of all H3K4 methylation (Chapter 5).

The findings presented in this thesis demonstrate that the six characterized SET1/MLL complexes collaborate extensively to deposit H3K4 methylation, but they do not account for all H3K4 methylation in mESCs. This work provides evidence of a novel methyltransferase that requires both the WRAD subcomplex and transcription for H3K4 methylation, and which may be a previously uncharacterized isoform of SET1B. The work presented in this thesis highlights the complexity of the H3K4 methylation system and provides a basis on which to further explore how SET1/MLL complexes influence transcription and how it relates to H3K4 methylation.

Chapter 2

Materials and Methods

2.1 Cell Culture Methods

2.1.1 Mouse embryonic stem cell culture

Mouse embryonic stem cells were cultured in Dulbeco's Minimum Essential Medium supplemented with 10% fetal calf serum, 1x penicillin/streptomycin, 2mM L-glutamine, 10ng/ml leukaemia inhibitory factor (LIF), 1x non-essential amino acids, and 0.5mM beta-mercaptoethanol. Cells were cultured in 37°C, with 5% CO_2.

2.1.2 HEK293T cell culture

HEK293T cells were cultured in Dulbeco's Minimum Essential Medium supplemented with 10% fetal calf serum, 1x pen/strep, L-glutamine, non-essential amino acids, and b-mercaptoethanol. Cells were cultured in 37°C, with 5% CO_2.

2.1.3 Drosphila cell culture

SG4 drosophila cells were cultured in Schneider's Drosophila Medium supplemented with 1x pen/strep and 10% heat-inactivated fetal calf serum. SG4 cells were cultured at 25°C.

2.2 Generation of stable transgenic cell lines

To generate stable transgenic mES cell lines, mESCs were transiently transfected with a construct expressing the guide RNA (gRNA) and the Cas9 endonuclease, and a targeting construct containing the transgenes flanked by 500bp-1000bp homology arms to allow for homology-directed repair following CRISPR-Cas9 targeting. Additionally, the gRNA construct contained the puromycin resistance gene to allow for selection of transfected cells. 2µg of targeting construct and 500ng of gRNA construct were transfected into mESCs using Lipofectamine 3000 according to the manufacturer's instructions. The following day, transfected cells were plated across a range of seeding densities and subjected to a 48 hour purmoycin selection using 1µg/ml of puromycin. Colonies derived from individual cells were allowed to form for the next 5-10 days before they

were picked and transferred to a 96-well plate. Clones were genotyped by PCR and homozygotes were expanded for further analysis.

2.3 Calibrated native ChIP-seq (cnChIP-seq)

All calibrated native ChIP-seq experiments were performed in biological triplicate.

2.3.1 Preparation of native chromatin

All histone modification ChIP-seq experiments were performed using calibrated native ChIP-seq. 50 million mESCs and 20 million SG4 cells were washed once with PBS and mixed in PBS. mESC-SG4 pellets were pelleted by centrifugation for 5 minutes at 1,500 rpm and resuspended in 1ml of ice-cold RSB buffer (10mM Tris pH8.0, 10mM NaCl, 3mM $MgCl_2$). Nuclei were then released by the addition of 28ml of ice-cold RSB buffer supplemented with 0.1% NP-40 and mixed by inversion for 10 times. Nuclei were pelleted by centrifugation for 5 minutes at 1,500g at 4°C. Nuclei were washed by resuspending in 5ml ice-cold RSB buffer supplemented with 0.25M sucrose and 3mM calcium chloride and pelleted by centrifugation for 5 minutes at 1,500g at 4°C. Washed nuclei were then resuspended in 1ml ice-cold RSB buffer supplemented with 0.25M sucrose, 3mM calcium chloride, and 1X protease inhibitor cocktail (Roche). 200U of micrococcal nuclease (MNase) (Fermentas) was then added to digest chromatin by incubation at 37°C for 5 minutes, with gentle agitation very minute. MNase digestion was stopped by the addition of 1µl of 1M EDTA. Insoluble nuclear material was pelleted by centrifugation for 5 minutes at 5,000rpm at 4°C and the supernatant (S1) containing MNase-digested chromatin was collected. The nuclear pellet was then resuspended in 300µl of nucleosome release buffer (10mM Tris pH 7.5, 10mM NaCl, 0.2nM EDTA) and incubated for 1 hour at 4°C with gentle rotation (16rpm). Insoluble nuclear material was then pelleted by centrifugation for 5 minutes at 5,000rpm at 4°C and the supernatant (S2) containing MNase-digested chromatin was collected. S1 and S2 supernatants were mixed, aliquoted, frozen on dry ice, and stored at -80°C. To confirm mono-nucleosomal digestion of chromatin, a small aliquot of chromatin (40µl) was purified using a column-based DNA purification kit (QIAGEN QIAquick PCR Purification Kit) and analysed by gel electrophoresis.

2.3.2 Chromatin immunoprecipitation

For each ChIP reaction, 200µl of S1+S2 chromatin was diluted 10-fold in Native ChIP Incubation Buffer (70mM NaCl, 10mM Tris pH 7.5, 2mM $MgCl_2$). Diluted chromatin was centrifuged for 5 minutes at 3,000rpm at 4°C to pellet any insoluble aggregates. 1ml of supernatant was taken for ChIP, the remaining 1ml of supernatant was taken as the input. For each input sample, 100µl of diluted native chromatin was purified using a column-based DNA purification kit (Zymo Research ChIP clean and concentrator) and eluted in 12µl of elution buffer. For each ChIP reaction, the appropriate amount of antibody was added to 1ml of diluted native chromatin and incubated overnight (16 hours) at 4°C with rotation (16rpm). Protein A agarose beads

(Repligen) were used to capture immuno-chromatin complexes. Beads were first blocked in Native ChIP Incubation Buffer supplemented with 1mg/ml yeast tRNA and 1mg/ml BSA overnight at 4°C with rotation (16rpm). Blocked beads were then washed twice with 1ml Native ChIP Incubation Buffer by centrifugation at 1,000g for 3 minutes at 4°C, then resuspended with Native ChIP Incubation Buffer. For each ChIP reaction, 20µl of packed, blocked beads were added and incubated for 1 hour at 4°C with rotation (16rpm). Beads were then washed 4 times by centrifugation at 1,000g for 1 minute at 4°C and resuspension with 1ml Native ChIP Wash Buffer (20mM Tris pH 7.5, 2mM EDTA, 125mM NaCl, 0.1% Triton), followed by one wash with TE buffer. Samples were then eluted using 100ul of ChIP Elution Buffer (0.1M $NaHCO_3$, 1% SDS) by incubating at room temperature (24°C) with shaking (1,200rpm) for 30 minutes, followed by centrifugation at 1,000g for 1 minute at room temperature to pellet beads. Supernatant was taken as the eluate and purified using a column-based DNA purification kit (Zymo Research ChIP clean and concentrator) and eluted in 12µl of elution buffer. For H3K4me3 ChIP experiments, ChIP reactions were set up in duplicate and purified on the same column, followed by elution with 24µl of elution buffer. ChIP samples were then analysed by qPCR to assess the reproducibility and quality of the experiment and then sequenced.

2.4 SNAP-ChIP

To determine the specificity of H3K4me1 and H3K4me2 antibodies, barcoded synthetic modified nucleosomes carrying H3K4me0/1/2/3 from the SNAP-ChIP K-MetStat Panel (EpiCypher) was used as a spike-in for native ChIP. Briefly, native chromatin was prepared as described above, but 2µl of SNAP-ChIP nucleosomes were added to purified nuclei immediately before micrococcal nuclease digestion. ChIP was performed as described above with H3K4me1 or H3K4me2 antibodies, and the specificity was assessed by qPCR using primers specifically annealing to barcoded DNA from H3K4me0/1/2/3-modified SNAP-ChIP nucleosomes.

Primer name	Forward	Reverse
SNAP-ChIP H3K4me0	5'-CGTATCGCGCGCATAATA-3'	5'-CGCGTAACGACGTACC-3'
SNAP-ChIP H3K4me1	5'-CGTATCGCGCGCATAATA-3'	5'-CGTTAACGCGTTTCGT-3'
SNAP-ChIP H3K4me2	5'-CGTATCGCGCGCATAATA-3'	5'-CGTACGTCGTGTCGAA-3'
SNAP-ChIP H3K4me3	5'-CGTATCGCGCGCATAATA-3'	5'-GTTCGCGACACCGTTC-3'

Table 2.1: Primers used for SNAP-ChIP qPCR

2.5 Calibrated double-crosslinked ChIP-seq

All calibrated double-crosslinked ChIP-seq experiments were performed in biological triplicate.

2.5.1 Double crosslinking cells

mESCs were washed once in PBS and counted. 50 million mESCs were resuspended in 10ml PBS and crosslinked with 2mM disuccinimidyl glutarate (DSG) for 45 minutes at 25°C with gentle agitation. Cells were then immediately crosslinked for 15 minutes with 1% formaldehyde 25°C with gentle agitation. Cells were then quenched with 1.5ml 1M glycine (125µM final concentration) for 5 minutes at 25°C with gentle agitation. Crosslinked cells were pelleted by centrifugation at 4°C at 1,000g for 5 minutes, frozen on dry-ice, and stored at -80°C before use. HEK293T cells in which RPB1 was tagged with 3X T7 tags were used for calibration. 50 million T7-RPB1 HEK293T cells were double-crosslinked as described for mESCs. After double-crosslinked cells were pelleted by centrifugation, cells were resuspended in ice-cold PBS and aliquoted, then frozen on dry-ice and stored at -80°C before use.

2.5.2 Preparation of double-crosslinked chromatin

50 million crosslinked mESCs were resuspended in 9ml ice-cold lysis buffer (50mM HEPES pH 7.9, 140mM NaCl, 1mM EDTA, 10% glycerol, 0.5% NP-40, 0.25% Triton X-100). 2 million crosslinked T7-RPB1 HEK293T cells were resuspended in 1 ml ice-cold lysis buffer 1 and added to resuspended mESCs. Cells were incubated on a roller at 4°C for 10 minutes, then pelleted by centrifugation at 1,000g for 5 minutes at 4°C. Nuclei were then resuspended in ice-cold lysis buffer 2 (10mM Tris-HCl pH 8, 200mM NaCl, 1mM EDTA, 0.5mM EGTA) and incubated on a roller at 4°C for 5 minutes. Nuclei were then pelleted by centrifugation at 1,000g for 5 minutes at 4°C. Nuclei were then resuspended in 1ml ice-cold sonication buffer (10mM Tris-HCl pH 8, 100mM NaCl, 1mM EDTA, 0.5mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine) and transferred to sonication tubes with 200µl of sonication beads. Nuclei were then sonicated using a Bioruptor Pico for 30 cycles, with each cycle consisting of 30 seconds on, 30 seconds off. Sonicated chromatin was then transferred to a new 1.5ml microcentrifuge tube and 100µl of ice-cold sonication buffer supplemented with 10% Triton was added and mixed by inversion. Insoluble nuclear material was pelleted by centrifugation at 20,000g for 10 minutes at 4°C. The supernatant containing sonicated, double-crosslinked chromatin was then taken for ChIP or frozen on dry-ice and stored at -80°C before use.

2.5.3 Chromatin immunoprecipitation

100µl of chromatin was used for each ChIP reaction. 2 ChIP reactions were performed for each sample in order to obtain sufficient DNA for next-generation sequencing library preparation. Chromatin was first pre-cleared before ChIP as follows. Protein A agarose beads were blocked for at least 30 minutes at 4°C with rotation (16rpm) with ChIP dilution buffer (1% Triton X-100, 1mM EDTA, 20mM Tris-HCl pH 8.0, 150mM NaCl) supplemented with 0.2mg/ml BSA and 50µg/ml yeast tRNA. For each pair of ChIP reactions, 250µl of sonicated chromatin was added to 2.25ml ChIP dilution buffer containing 1x protease inhibitor cocktail (Roche). 40µl of blocked beads were added to 2.5ml of diluted chromatin and rotated (16rpm) at 4°for 30 minutes. Beads were collected by centrifugation for 4 minutes at 1,000g at 4°C. Supernatant conataining

pre-cleared chromatin was taken for ChIP: 1ml of chromatin was used for each ChIP reaction, and 100µl of the remainder was taken as the input. The appropriate amount of antibody was added to each 1ml aliquot of pre-cleared chromatin and rotated at 4°C overnight (16 hours). Input sample was stored overnight at -80°C. After overnight incubation with antibody, 20ul of packed, blocked, pre-washed protein A agarose beads were added to each 1ml ChIP sample and rotated (16rpm) at 4°C for 1 hour. Beads were then washed once in low-salt buffer (0.1% SDS, 1% Triton X-100, 2mM EdTA, 20mM Tris-HCl pH 8, 150mM NaCl), high-salt buffer (0.1% SDS, 1% Triton X-100, 2mM EdTA, 20mM Tris-HCl pH 8, 500mM NaCl), LiCl immune complex wash buffer (250mM LiCl, 0.1% Np40, 1% sodium deoxycholate, 1mM EDTA, 10mM Tris-HCl pH 8.0) and twice with TE buffer. For each wash, beads were collected by centrifugation for 1 minute at 1,000g at 4°C and resuspended in 1ml ice-cold wash buffer, then rotated (16rpm) for 4 minutes at 4°C. Samples were eluted by resuspending beads in 100µl of elution buffer (1% SDS, 0.1 M $NaHCO_3$) and incubated at room temperature (24°C) for 30 minutes with shaking at 1,200rpm. Beads were then collected by centrifugation at room temperature at 1,000g for 1 minute. 100µl eluates were transferred to new 1.5 microcentrifuge tubes and de-crosslinked by addition of 4µl 5M NaCl, and RNAs were removed by addition of 2µl RNase, and incubated at 65°overnight (16 hours). Input samples were similarly treated. After overnight de-crosslinking and RNase treatment, samples were treated with 1µl of proteinase K for 1 hour at 45°C. Samples were then purified using Zymo ChIP clean and concentrator kit and eluted in 15µl of elution buffer. Duplicate ChIP samples were pooled after purification.

2.6 Next Generation Sequencing

All sequencing experiments were performed in biological triplicates. Sequencing data from individual replicates were analysed independently and replicates which correlated well were merged. All heatmaps and metaplots presented in this thesis were produced from genome coverage tracks of three merged biological replicates.

2.6.1 Sequencing library preparation

All sequencing libraries were prepared using NEBNext Ultra II Library Preparation Kit as per the manufacturer's protocol with no size selection of DNA fragments. For native ChIP-seq experiments, 5ng of DNA was used as input material and amplified using 7 cycles during the indexing step. For double-crosslinked ChIP-seq experiments, 10ng of DNA was used as input material and amplified using 8 cycles during the indexing step. For double-crosslinked ChIP-seq experiments where ChIP DNA concentration was too low to be measured, the entire volume of ChIP DNA was used as input DNA and amplified for 11 cycles during the indexing step. For double-crosslinked ChIP-seq experiments, prior to library preparation, DNA was sonicated again to reduce the average size of the DNA fragments to avoid amplification bias against larger fragments. DNA was sonicated using a Diagenode PicoRuptor sonicator for 20 cycles with 30 seconds ON and 30 seconds OFF in each cycle, with a brief spin down after 10 cycles. For all library preparations, the final purification step using AMPure beads was repeated to remove primer dimers.

2.6.2 Quality control and quantification of libraries

All sequencing libraries were analysed using the DNA high-sensitivity assay on a Bioanalyzer instrument (Agilent) to confirm correct size distribution of DNA fragments in each library and to obtain the average fragment size for quantification. All libraries were quantified using a Qubit fluorometric assay to obtain the concentration of each library and the molarity was calculated using 660g/mol as the molar concentration of DNA and size-corrected using the average size obtained by the Bioanalyzer instrument. Libraries were diluted to 6nM and pooled for sequencing.

2.6.3 Sequencing

Next generation sequencing was performed on an Illumina NextSeq 500 instrument according to the manufacturer's instructions. Next generation sequencing was performed using NextSeq 500/550 V2 High Output sequencing kits. Sequencing was performed using paired-end sequencing with 75 cycles.

2.7 Bioinformatic analysis of next-generation sequencing data

2.7.1 Sequencing data alignment

ChIP-seq reads were aligned to the genome using bowtie2, with the '–no-mixed' and '–no-discordant' options. Mouse reads were aligned to the mm10 genome assembly, drosophila reads were aligned to the dm6 genome assembly, and human reads were aligned to the hg19 genome assembly. Only reads which mapped uniquely to the genome were retained. Mapped reads were sorted using 'Sambamba sort' and duplicates were removed using 'Sambamba markup'. Mapped, filtered reads were indexed using 'samtools index' and the number of mapped reads were counted using 'samtools idxstats'.

2.7.2 Spike-in normalisation

To facilitate quantitative comparisons between ChIP-seq samples, spike-in normalisation was employed. Spike-in normalization was performed by randomly downsampling the reads that mapped to the mm10 genome to reflect the number of mapped spike-in reads in that sample. All ChIP-seq experiments for the same factor or histone modification (e.g. H3K4me3) were downsampled together. First, the number of reads mapping to the mm10 genome or the spike-in genome were counted using samtools. A first normalisation factor was calculated using the following formula:

$$Normalisation\ Factor = \frac{1}{Number\ of\ SpikeIn\ reads\ in\ ChIP}$$

For double-crosslinking ChIP-seq, the normalisation factor was further multiplied by the ratio between the number of spike-in reads in the input sample and the number of mouse reads in the input sample to correct for variations during the mixing of HEK cells with mESCs:

$$\textit{Input Corrected Normalisation Factor}$$

$$= \textit{Normalisation Factor} \times \frac{\textit{Number of SpikeIn reads in input}}{\textit{Number of mouse reads in input}}$$

For each replicate of each ChIP experiment, the largest normalization factor was used to calculate a first downsampling factor for all samples within that replicate:

$$\textit{Downsampling Factor 1} = \frac{\textit{Normalisation Factor}}{\textit{Largest Normalisation Factor of Replicate}}$$

The number of mm10 reads after downsampling using downsampling factor 1 was calculated, and the untreated sample with the lowest readcount was identified. A readcount adjustment factor was calculated for all untreated samples such that the number of reads after downsampling would be the same in all untreated samples:

$$\textit{Readcount Adjustment Factor} = \frac{\textit{Lowest Untreated Readcount}}{\textit{Readcount After Downsampling}}$$

This readcount adjustment factor was then applied across all samples of the replicate to obtain a second downsampling factor:

$$\textit{Downsampling Factor 2} = \textit{Downsampling Factor 1} \times \textit{Readcount Adjustment Factor}$$

Downsampling Factor 2 was then used to randomly downsample the mapped mm10 reads using Sambamba.

2.7.3 Visualisation of sequencing data

The downsampled BAM file of each replicate was used to generate a BigWig file for visualisation and for heatmap and metaplot analyses. Replicates which correlated well were merged using Sambamba for further analyses. Merged BAM files were used to generate BigWig files for visualisation and for heatmap and metaplot analyses. All heatmaps and metaplots presented here were generated using merged BigWig files.

To generate BigWig files for visualisation, the pileup function of MACS2 was used to generate a bedgraph file. WigtoBigWig was then used to generate a BigWig file from the bedgraph file.

Heatmaps and metaplots were generated using the ComputeMatrix, plotHeatmap, and plotProfile tools from Deeptools.

2.7.4 Differential enrichment analysis

To perform differential enrichment analysis, the number of mapped mm10 and spike-in reads of each sample before downsampling were first counted in an appropriate interval set using multiBamSummary from deeptools. For the H3K4me3 differential enrichment analysis presented in this thesis, mm10 reads were counted in an interval set consisting of all non-redundant mm10 promoters (TSS $\pm$ 2kb) which overlapped with an H3K4me3 peak (see 'functional annotation'). Dm6 reads were counted in an interval set consisting of all refGene-annotated dm6 promoters (TSS $\pm$ 2kb). Differential enrichment analysis was performed using DESeq2 using a custom R script which incorporates spike-in normalization by calculating DESeq2 size factors using mm10 and dm6 read counts. A change was considered to be significant if the significant threshold of $P_{adj} < 0.05$ was met. Fold change cut-offs were assigned for each analysis depending on the aim of the experiment (see main text).

2.7.5 Peak calling

Peak calling was performed using MACS2 on BAM files of individual ChIP-seq replicates before downsampling, with the appropriate input BAM file as the control file. 'BAMPE' was specified for all peak calling. For H3K4me3 peaks, the 'broad' peak calling option was specified and a minimum false discovery rate cut-off of 0.05 ('–broad-cutoff 0.05') was applied. For T7-MLL3/4 and RBBP5 ChIP-seq, the 'narrow' peak calling option was specified and a minimum false discovery rate cut-off of 0.05 ('-q 0.01') was applied. HOMER mergePeaks was used to merge peaks from each replicate, with 'd -given' specified to ensure only peaks which overlap between the three replicates are merged and taken forward for further analysis.

2.7.6 Functional annotation of genomic regions of interest

A custom-built, non-redundant gene set of 20,633 genes from the mm10 genome was used as the basis for functional annotation of the genome. This non-redundant gene set was obtained by filtering out very short genes with poor mappability and transcripts which are highly similar from the refGene mm10 gene set (Rose et al., 2016). For analysis of all promoters, the TSS $\pm$ 2kb from this non-redundant gene set was used. For analysis of all gene bodies, the region encompassing the TSS to the TES of these 20,633 genes was used.

To sort all promoters and gene bodies according to transcription, the 20,633 non-redundant genes were annotated with RPKM values derived from transient transcriptome sequencing (TT-seq) data from untreated SET1A/B-dTAG cells (Hughes et al., 2023). To analyse H3K4me3 and H3K4me2 levels at promoters according to transcription, a list of H3K4me3-associated or H3K4me2-associated promoters was first generated. Bedtools intersect was used to identify promoters from the non-redundant promoter set (n=20,633) which overlapped with a H3K4me2 or H3K4me3 peak. H3K4me2 and H3K4me3 peaks were generated from SET1A/B-dTAG H3K4me2/3 ChIP-seq experiments presented in this thesis. H3K4me2-associated and H3K4me3-associated promoters were then sorted according to RPKM values from TT-seq data from untreated SET1A/B-dTAG cells, then divided into quartiles.

For differential enrichment analysis of H3K4me3 levels at promoters, the set of non-redundant promoters which overlapped with an H3K4me3 peak was used for analysis.

'Active enhancers' are defined in this study using a previously generated interval set in which an ATAC-seq peak and an H3K27ac ChIP-seq peak overlap (n=12,006)(Fursova et al., 2021). This set of active enhancers are further divided into 'intergenic' (n=4,156) and 'intragenic' (n=7,850) dependending on whether they occur within the gene body of an annotated gene.

CpG islands are defined using a previously generated BioCap-seq peak set (n=27,047)(Long et al., 2013).

2.8 Extraction of nuclear proteins

To extract nuclear proteins, the size of the cells pellet was first estimated. Cells were resuspended in 10 volumes of Buffer A (10mM HEPES pH 7.9, 1.5mM $MgCl_2$, 10mM KCl, 0.5mM DTT, 0.5mM PMSF and 1x protease inhibitor cocktail) and incubated on ice for 10 minutes to allow cells to swell. Cells were collected by centrifugation at 500g for 5 minutes at 4°C and resuspended in 3 volumes of Buffer A supplemented with 0.1% NP-40, then incubated on ice for 10 minutes. Released nuclei were then collected by centrifugation at 1,500g for 5 minutes at 4°C and resuspended in 1 volume of Buffer C (5mM HEPES pH 7.9, 26% glycerol, 1.5mM $MgCl_2$, 0.2mM EDTA, 0.5mM DTT, 1x protease inhibitor cocktail). The volume of nuclei suspension was then accurately estimated and 5M NaCl was added to a final concentration of 400mM. Nuclei suspensions were then rotated (16rpm) at 4°C for 1 hour to extract nuclear proteins. Nuclei were then centrifuged at 13,000rpm for 20 minutes at 4°C to pellet insoluble material. The soluble fraction was then taken as nuclear extract and quantified using the Bradford Assay.

2.9 Acid extraction of histones

Unless stated otherwise, all buffers were ice-cold and centrifugation was performed at 4°C. Cells were washed once in 1ml RSB buffer and pelleted by centrifugation at 1,500rpm for 5 minutes. Cells were resuspended in 1ml RSB buffer supplemented with 0.5% NP-40 and incubated on ice for 10 minutes. Cells were collected by centrifugation at 500g for 5 minutes and resuspended in 0.5ml 5mM $MgCl_2$. 0.5ml 0.8M HCl was added and cells were incubated on ice for 20 minutes, then centrifuged for 20 minutes at 13,000rpm. Supernatant was then transferred to a new tube and trichloroacetic acid was added to a final concentration of 25%, then incubated on ice for 30 minutes. Histones were pelleted by centrifugation at 13,000rpm for 15 minutes, then washed twice with 1ml ice-cold acetone by vortexing to resuspend and centrifugation at 13,000rpm for 15 minutes. After the second acetone wash, acetone was aspirated off and the histone pellet was left to air-dry at room temperature until acetone had evaporated. The histone pellet was resuspended in 80µl 1X SDS loading buffer and boiled for 90°C for 5 minutes and stored at -80°until use. Prior to immunoblotting, to ensure equal loading of histones across samples, histone extracts were quantified by resolving on a 15% SDS-PAGE gel, which was then stained with Coomassie Blue. Coomassie-Blue-stained histone H3 bands were then quantified

using a Li-Cor FC Odessey imager and used for determining loading volumes.

2.10 Immunoblotting

Extracts were resolved by SDS-PAGE using either Tris-Glycine gels or Tris-Acetate gels. Tris-Acetate gels were generally used for immunoblotting for proteins larger than 200kDa and Tris-Glycine gels were generally used for immunoblotting for proteins smaller than 200kDa. The percentage of acrylamide/bis-acrylamide incorporated into each Tris-Glycine gel was adjusted according to the size of the target protein. For immunoblotting most proteins, proteins were transferred onto nitrocellulose membranes by semi-dry transfer using the TransBlot Turbo system (Biorad). For SET1B, MLL1-N, MLL2-N, MLL3, and MLL4 western blots, proteins were transferred onto nitrocellulose membranes by wet transfer using modified Towbin's buffer (25mM Tris, 192mM glycine, 20% ethanol, 0.01% SDS) for 22 hours at 4°C. Membranes were blocked for 1 hour at room temperature with 5% milk-PBST with gentle agitation and then incubated with primary antibody, diluted in 5% milk-PBST, overnight with gentle agitation at 4°C. Membranes were then washed 3X with PBST, 10 minutes each wash, at room temperature with gentle agitation, then incubated in secondary antibody, diluted in 5% milk-PBST, at room temperature for 1 hour with gentle agitation. Membranes were either incubated in a HRP-conjugated secondary antibody for enhanced chemiluminescence or a secondary antibody conjugated with a fluorescent dye for imaging on a Li-Cor FC Odessey imager.

2.11 Co-immunoprecipitation

For each immunoprecipitation reaction, 600µg of nuclear extract was used. 600µg of nuclear extract was first topped-up to 550µl with BC150 buffer (150mM KCl, 10% glycerol, 50mM HEPES pH7.9, 0.5mM EDTA, 0.5mM DTT) supplemented with 1x Protease Inhibitor Cocktail. Nuclear extracts were centrifuged for 5 minutes at 4°C at 13,000rpm to pellet any precipitates. Supernatant was transferred to a fresh Protein Lobind tube. 50µl was taken as the input sample and stored at -80°C. The appropriate amount of antibody was added to the remainder and 25U of Benzonase nuclease (Sigma) was added. Samples were rotated (16rpm) at 4°C overnight (16 hours). Protein A agarose beads were washed twice in 1ml BC150 buffer by centrifugation at 1,000g at 4°C for 4 minutes and blocked by resuspending in BC150 supplemented with 1% fish skin gelatin and 0.2mg/ml BSA, then rotated (16rpm) at 4°overnight (16 hours). Beads were then washed twice in BC150 and brought back to the original volume with BC150. After overnight incubation with primary antibody, 50µl of 50% slurry of blocked beads was added to each IP sample. Samples were then rotated at 4°C for 4 hours, then washed 4 times with BC150 supplemented with 0.02% NP-40. For each wash, beads were collected by centrifugation at 1,000 for 1 minute at 4°C and then resuspended in 1ml of ice-cold BC150+0.02% NP-40, then rotated (16 hours) at 4°C for 10 minutes. After the final wash, beads were resuspended in 30µl of 2x SDS loading buffer and boiled for 5 minutes at 95°C. Beads were collected by centrifugation at 1,000g for 4 minutes at room temperature and the supernatant was taken as the immunoprecipitate.

2.12 Size-exclusion chromatography

Size-exclusion chromatography of nuclear extracts was performed on a Superose 6 Increase 10/300 GL column. The column was previously calibrated using Thyroglobulin (669kDa), Ferritin (440kDa), Aldolase (158kDa), Conalbumin (75kDa), and Ovalbumin (43kDa). Nuclear extracts were prepared as described above. Nuclear extracts were digested with Benzonase nuclease (Sigma) at 250U/mg, then dialysed overnight into BC200 (200mM KCl, 10% glycerol, 50mM HEPES pH7.9, 0.5mM EDTA, 0.5mM DTT). Dialysed nuclear extracts were centrifuged at 13,000rpm for 10 minutes at 4°C to pellet precipitates. Supernatant was re-quantified and 1.7 mg of dialysed extract was loaded onto the column. 250µl fractions were collected in a 96-well deep-well plate and precipitated with trichloroacetic acid. Samples were resuspended in Lammeli Buffer and 2µl 1M Tris pH 8 was added to neutralize all samples. Fractions were then analysed by immunoblotting.

2.13 WDR5-FLAG immunoprecipitation followed by mass spectrometry

For mass spectometry experiments, cells were harvested by scrapping in ice-cold PBS supplemented with 1x protease inhibitor cocktails instead of by trypsinisation. Cells were pelleted by centrifugation at 4°at 1000g for 5 minutes. Cell pellets were stored at -80°until use. Nuclear extracts were prepared as described as above and immediately used for FLAG immunoprecipitation.

2.13.1 FLAG immunoprecipitation

For WDR5-FLAG IP/MS, 1.4mg of nuclear extract was used per IP. 1.4mg of nuclear extract was diluted in nuclear extract buffer C without DTT or NaCl to give a final NaCl concentration of 150mM. All samples were topped up to the same volume with nuclear extract buffer C containing 150mM NaCl. 1.4µl of Benzonanse nuclease (Sigma) was added to each sample and incubated for 30 minutes at 4°C with rotation at 16rpm. Anti-FLAG M2 affinity gel (Sigma, A2220) was washed 3 times in BC150 buffer without glycerol. For each wash, beads were centrifuged at 1000g for 3 minutes at 4°C and resuspended with 1ml BC150 without glycerol. After the final wash, beads were resuspended in the original volume with BC150 without 10% glycerol. Nuclear extracts were centrifuged for 10 minutes at 13,000rpm at 4°C and the supernatant was transferred to a new protein lo-bind tube. 25ul of washed FLAG M2 agarose beads (50µl slurry) was added to precleared nuclear extracts and incubated at 4°C for 4 hours with rotation at 16rpm. Beads were then washed 3 times with BC150 without glycerol but supplemented with 0.02% NP-40. For each wash, beads were incubated in washing buffer at 4°C with 16rpm rotation. Beads were then washed for a further 3 times as above with BC150 without glycerol or NP-40. Proteins were then eluted under acidic conditions using glycine as follows. 50µl of 0.1M glycine-HCl, pH 3.5 was added to beads. Beads were incubated for 10 minutes with gentle shaking at 900rpm at room temperature. Beads were pelleted by centrifugation for 3 minutes at 1000g and the supernatant containing the

eluate was immediated neutralized by transferring to fresh tubes containing 5µl of 0.5M Tris-HCL, pH 7.4 with 1.5M NaCl. The elution step was repeated for a total of two elutions and the eluates were pooled. Eluates were stored at -20°C until use.

2.13.2 In-solution trypsin digestion of protein samples for mass spectrometry

FLAG IP eluates were denatured by adding freshly made 8M urea dissolved in 100mM ammonium bicarbonate buffer pH 7.8 to a final concentration of 4M urea. Samples were incubated for 10 minutes at room temperature with shaking at 650rpm. Cysteines were then reduced with 10mM TCEP (Tris(2-carboxyethyl)phosphine) and incubated for 30 minutes at room temperature without shaking. Cysteines were then alkylated with 50mM CIAM (2-chloroacetamide) freshly prepared in 8M urea/100mM ammonium bicarbonate buffer solution and incubated in the dark for 30 minutes at room temperature without shaking. Samples were then pre-digested with LysC at a ratio of 1µg of LysC for 100ug of protein sample and incubated at 37°C for 2 hours with shaking at 650rpm. Samples were then diluted using 100mM ammonium bicarbonate buffer to a final concentration of 2M urea. $CaCl_2$ was then added to a final concentration of 2mM. Trypsin was added at a ratio of 1µg of trypsin for 40µg of protein sample. Samples were digested with trypsin for 19 hours at 37°C with shaking at 800rpm. Trypsin digestion was stopped with 5% formic acid. Trypsin-digested samples were centrifuged for 30 minutes at 12,700rpm at 20°C. Supernatant was collected and desalted onto C18 columns.

2.13.3 Desalting peptides using C18 resin

C18 stage-tips were prepared by inserting C18 resin into a 10µl tip using a needle. C18 resin was activated with 60µl of 100% acetonitrile as such: 60µl of 100% acetonitrile was added to the resin and centrifuged for 4 minutes at 4000rpm at room temperature until liquid passes through the resin. This step was repeated. Activated resin was then washed with 60µl of 0.1% trifluoroacetic acid by centrifugation at 4000rpm at room temperature for 4 minutes until liquid goes through the resin. Resin was then centrifuged again for 2 minutes at 8000rpm at room temperature. Samples were loaded onto the C18 resin, 80µl at a time until entire sample volume has been loaded. Resin was centrifuged at 4000rpm for 4 minutes at room temperature until liquid has passed through the resin. After entire sample volume has passed through resin, resin was centrifuged one final time at 8000rpm for 2 minutes at room temperature. Resin was then washed twice with 60µl of 100% acetonitrile as described above. Peptides were then eluted twice with 60µl of 50% acetonitrile/0.1% trifluoroacetic acid into a new tube. For each elution, resin was centrifuged at 4000rpm at room temperature until all liquid passes through. A final spin for 2 minutes at 8000rpm at room temperature was then performed. Peptides were dried in a SpeedVac overnight for 20 hours. Dried, desalted peptides were stored at -20°until injection into mass spectrometer.

2.13.4 Mass spectrometry

Liquid chromatography tandem mass spectrometry and the subsequent data analysis was performed by Inge de Krijger. Dried desalted peptides were resuspended in 10µl of 5% formic acid 5% acetonitrile before

subjecting to liquid chromatography tandem mass spectrometry (LC-MS/MS). Raw data was analysed using the MaxQuant computational platform and quantitative analyses were performed using the Perseus computational platform (Tyanova, Temu, Sinitcyn, et al., 2016, Tyanova, Temu, and Cox, 2016). A negative control IP was performed using a cell line in which no protein was FLAG-tagged. To identify proteins interacting with WDR5, comparative analysis was performed using Perseus to determine proteins enriched in the FLAG-IP condition relative to the negative control. Significance was calculated using an ANOVA test. Proteins were considered valid WDR5 interactors if they met a significance threshold of $p < 0.05$ and a t-test difference > 3.

2.14 Antibodies used

Antibody	Manufacturer	Reference	Dilution for Western Blotting	Amount for ChIP	Amount for Co-IP
SET1A	Klose lab	-	1:500	-	-
SET1B	Cell Signaling Technology	D1U5D	1:500	-	-
MLL1-C	Active Motif	61926	1:500	-	-
MLL1-C	Cell Signaling Technology	D6G9N	1:500	-	-
MLL1-N	Cell Signaling Technology	D2M7U	1:500	-	-
MLL2-C	Cell Signaling Technology	E6A8V	1:2000	10µl	-
MLL2-N	Cell Signaling Technology	E3M1V	1:500	-	-
MLL3	Sigma Aldrich	ABE-1850	1:500	-	-
MLL4	Abcam	ab213721	1:500	-	-
WDR5	Cell Signaling Technology	D9E1I	1:1000	-	-
WDR5	Bethyl	A302-429A	-	-	5ug (5µl)
RBBP5	Cell Signaling Technology	D3I6P	1:1000	10µl	15µl
ASH2L	Cell Signaling Technology	D93F6	1:1000	-	-
T7	Cell Signaling Technology	D9E1X	1:500	6µl	-
H3	Cell Signaling Technology	96C10	1:1000	-	-
H3K4me1	Cell Signaling Technology	D1A9	1:1000	5µl	-
H3K4me2	Thermofisher	710796	1:1000	2µg (4µl)	-
H3K4me3	Klose lab	-	1:2000	1.5µl	-
SUZ12	Cell Signaling Technology	D39F6	1:1000	-	-
TBP	Abcam	-	1:500	-	-

Table 2.2: List of antibodies used in this study

2.15 Reverse transcription followed by quantitative polymerase chain reaction (RT-qPCR)

2.15.1 RNA extraction

Cells were resuspended in 1ml of TRIzol reagent and incubated for 5 minutes at room temperature. 0.2ml of chloroform was added to each sample and samples were shaken vigorously for 15 seconds, then incubated for 2 minutes at room temperature. Samples were centrifuged at 12,000g for 15 minutes at 4°C, and the upper colourless aqueous phase containing RNA was transferred to a fresh tube. 0.5ml isopropanol was added to the colourless aqueous phase and incubated for 10 minutes at room temperature to precipitate RNA. RNA was pelleted by centrifugation at 12,000g for 10 minutes at 4°C, then washed twice with 1ml of ice-cold 70% ethanol by centrifugation at 7,600g for 5 minutes at 4°C. RNA pellet was air-dried at room temperature for 10 minutes, then resuspended in 30µl of nuclease-free water, then quantified using Nanodrop. RNA samples were then directly subjected to DNase treatment to remove contaminating genomic DNA. 5µg RNA was used for DNase treatment using the Turbo DNA Free kit (Life Technologies, AM1907) according to the manufacturer's instructions.

2.15.2 Reverse transcription

200ng of DNase-treated RNA was used for reverse transcription using the ImProm-II™ Reverse Transcription System (Promega) according to the manufacturer's instructions using random hexamer primers for first-strand cDNA synthesis. A negative control reaction in which no reverse transcriptase was added was used to assess the amount of genomic DNA contamination during qPCR. cDNA was diluted in water before qPCR amplification.

2.15.3 qPCR

qPCR was performed by mixing 5µl of DNA sample with 0.75µl of 10µM forward primer, 0.75µl of 10µM reverse primer, 7.5µl of Bioline SensiMix SYBR No-ROX mastermix, and 1.75µl of H_2O, and cycling as follows: 95°C for 10 minutes, 40 cycles of 95°C for 10 seconds, 60°C for 45 seconds, 72°C for 15 seconds, followed by a melting step from 65°to 95°.

Primer name	**Forward**	**Reverse**
Oct4 intronic	5'-TCTGTGTCCTTATTCTGCTGCT-3'	5'-GCAGAGGGTGACTCAGAAGG-3'
Hspg2 intronic	5'-GATCCCATCAATGACGGCGA-3'	5'-GTGCTCCTGGATCCTATGCC-3'
Gapdh intronic	5'-TGTAGTGAGCCCCAGGCTAT-3'	5'-ATAGCTGATGGCTGCAGGTT-3'

Table 2.3: List of primers used for RT-qPCR

2.16 Cell lines used

Cell Line	Genotype	Source
dTAG-SET1A/B	*FKBP12^{F36V} -Setd1A; FKBP12^{F36V}-Setd1B*	Amy Hughes; Hughes et al., 2023
MLL1/2-dTAG	*Kmt2a-FKBP12^{F36V}; Kmt2b-FKBP12^{F36V}*	Amy Hughes, Anna Lastuvkova, this study
SET1A/B/MLL1/2-dTAG	*FKBP12^{F36V}-Setd1A; FKBP12F36V-Setd1B; Kmt2a-FKBP12F36V; Kmt2b-FKBP12F36V*	Anna Lastuvkova, Amy Huhges, this study
MLL3/4-dTAG	*3XT7-2XStrepII-FKBP12^{F36V}-Kmt2C; 3XT7-2XStrepII-FKBP12^{F36V}-eGFP-Kmt2d*	This study
SET1A/B/ MLL1/2/3/4-dTAG	*FKBP12^{F36V}-Setd1A; FKBP12^{F36V}-Setd1B; Kmt2a-FKBP12^{F36V}; Kmt2b-FKBP12^{F36V}; 3XT7-2XStrepII-FKBP12^{F36V}-Kmt2C; 3XT7-2XStrepII-FKBP12^{F36V}-eGFP-Kmt2d*	This study
WDR5-dTAG	*Wdr5-FKBP12^{F36V}-2XStrepII-3XT7*	Anna Lastuvkova, this study
RBBP5-dTAG	*RBBP5-FKBP12^{F36V}-2XStrepII-3XT7*	This study
SET1A/B/MLL1/2/ 3/4/WDR5-dTAG	*FKBP12^{F36V}-Setd1A; FKBP12^{F36V}-Setd1B; Kmt2a-FKBP12^{F36V}; Kmt2b-FKBP12^{F36V}; 3XT7-2XStrepII-FKBP12^{F36V}-Kmt2C; 3XT7-2XStrepII-FKBP12^{F36V}-eGFP-Kmt2d; Wdr5-FKBP12^{F36V}-2XStrepII-3XT7*	Anna Lastuvkova, this study
SET1A/B/MLL1/2/ 3/4/RBBP5-dTAG	*FKBP12^{F36V}-Setd1A; FKBP12^{F36V}-Setd1B; Kmt2a-FKBP12^{F36V}; Kmt2b-FKBP12^{F36V}; 3XT7-2XStrepII-FKBP12^{F36V}-Kmt2C; 3XT7-2XStrepII-FKBP12^{F36V}-eGFP-Kmt2d; Rbbp5-FKBP12^{F36V}-2XStrepII-3XT7*	This study
SET1A/B/MLL1/2/ 3/4-dTAG-WDR5-FLAG	*FKBP12^{F36V}-Setd1A; FKBP12^{F36V}-Setd1B; Kmt2a-FKBP12^{F36V}; Kmt2b-FKBP12^{F36V}; 3XT7-2XStrepII-FKBP12^{F36V}-Kmt2C; 3XT7-2XStrepII-FKBP12^{F36V}-eGFP-Kmt2d; Wdr5-3XFLAG*	Anna Lastuvkova, this study

Table 2.4: List of cell lines used in this study

Target	gRNA sequence	PAM
SET1A	TAAATGAGCAAAGATGGACC	AGG
SET1B	TTCAGGTTGGGTTAACGGCA	TGG
MLL1-C	CCAGTGTCACTGCCACAAGC	AGG
MLL2-C	TGTCCAACAACCCCGGGGAC	AGG
MLL3	CCGACGACATCCTAGTCACC	AGG
MLL4	GTCCATCCCTCTCTCCGACT	GGG
WDR5	GAAGAGTGACTGCTAAGTCC	TGG
RBBP5	CAGAACTGCTGTGAAGACCG	TGG

Table 2.5: gRNA sequences for the transgenic cell lines used

2.17 Publicly-available NGS datasets used

Dataset	Cell line	Source	Accession
H3K27ac ChIP-seq	Bap1fl/fl, untreated	Fursova et al., 2021	GSE161996
ATAC-seq	Bap1fl/fl, untreated	Fursova et al., 2021	GSE161996
BioCap-seq	wild-type mouse embryonic stem cell V6.5	Long et al., 2013	GSE43512
CpG density	N/A	King and Klose, 2017	N/A
TT-seq	dTAG-SET1A/B, untreated	Hughes et al., 2023	GSE199805
RNAPII ChIP-seq	dTAG-SET1A/B, untreated	Hughes et al., 2023	GSE199805
T7-SET1A ChIP-seq	T7-dTAG-SET1A, untreated	Hughes et al., 2023	GSE199805
T7-Med14 ChIP-seq	MED14-T7 Med13/13l$^{fl/fl}$, untreated	Dimitrova et al., 2022	GSE185930
CAGE-seq	OS25 mouse embryonic stem cells	Forrest et al., 2014	DRA000991

Table 2.6: List of publicly-available NGS datasets used

Chapter 3

CpG island-associated SET1/MLL complexes collaborate to deposit H3K4 methylation

Introduction

H3K4 methylation in the mammalian genome is deposited by SET1/MLL complexes, with SET1A/B and MLL1/2 complexes thought to be the primary methyltransferases responsible for depositing H3K4me3. (Shilatifard, 2012, Crump and Milne, 2019, Hughes et al., 2020). All four of theses complexes contain a CXXC domain: CFP1, a SET1-complex-specific subunit, contains a CXXC domain, and the N-terminal fragments of MLL1 and MLL2 each contain a CXXC domain. Accordingly, the close correlation between H3K4me3 and CpG islands has been presumed to be a result of the CXXC domains of SET1A/B/MLL1/2-complexes directing them to CpG islands (Hughes et al., 2020). Additionally, the enrichment of H3K4me3 at highly-transcribed genes is thought to be a result of high SET1A occupancy at these promoters, which is determined in part by the preference of CFP1 to bind CpG islands with high H3K4me3 levels through its PHD finger and also through direct interactions between SET1 complexes and the RNA polymerase II pre-initiation complex (Brown et al., 2017, Ebmeier et al., 2017). These observations have led to the view that SET1 complexes, like their yeast and drosophila counterparts, are the primary contributors of H3K4me3 in mammalian cells (Shilatifard, 2012). On the other hand, MLL1/2 are primarily responsible for maintaining H3K4me3 levels at lowly-transcribed genes (Denissov et al., 2014). Indeed, MLL1/2 bind broadly across all CpG islands and its occupancy does not appear to be correlated with levels of transcription of the associated gene (Denissov et al., 2014). Additionally, MLL2 have been reported to deposit the majority of transcription-independent H3K4me3 in oocytes (Hanna et al., 2018). These observations have led to a model in which SET1A/B deposit H3K4me3 in a transcription-dependent manner at highly-transcribed genes while MLL1/2 deposit H3K4me3 in a transcription-independent manner at lowly-transcribed genes (Cenik and Shilatifard, 2021).

The view that SET1A is the primary contributor of H3K4me3 in mouse embryonic stem cells (mESCs) has been increasingly challenged. Early studies which came to this conclusion did so by analysing bulk H3K4 methylation levels after depleting SET1A protein by conditional knockout or RNA interference (Bledau et al.,

2014, Fang et al., 2016). This presents a problem because SET1A is essential for mESC maintenance and self-renewal, and the use of slow-acting methods to deplete SET1A means that the primary effects of SET1A removal on H3K4 methylation cannot be distinguished from those caused by secondary effects stemming from declining cell viability. Recently, using degron technology to rapidly deplete SET1A/B in mESCs, the Klose lab has found that despite causing widespread and profound transcriptional defects, removing SET1A/B has a surprisingly modest effect on H3K4me3 levels genome wide (Hughes et al., 2023). At 2 hours after SET1A/B depletion, a pervasive but modest reduction in H3K4me3 levels was observed, followed by a recovery in H3K4me3 levels after 24 hours of SET1A/B depletion (Hughes et al., 2023). These findings indicate that SET1A/B depletion does not account for the majority of H3K4me3 and suggests that other methyltransferases must also contribute centrally to H3K4me3.

Despite extensive study of SET1/MLL complexes, the exact roles of SET1A/B and MLL1/2 in depositing H3K4 methylation remain unclear, since acute depletion of SET1A/B and constitutive deletion of MLL1/2 result in relatively modest reductions in H3K4me3 levels at most genes. Additionally, the role of SET1A/B and MLL1/2 complexes in depositing H3K4me1/2 have not been extensively characterized with high resolution at the genome-wide level. In this chapter, I first seek to build on our previous findings and systematically examine how SET1A/B shape H3K4 methylation. I then investigate the role of MLL1/2 in H3K4 methylation using a degron system to enable rapid depletion of MLL1/2.

3.1 SET1A/B contribute modestly to H3K4 methylation

3.1.1 SET1A/B primarily deposits H3K4me3 at highly-transcribed genes

The Klose lab has previously generated a SET1A/B-dTAG cell line in which SET1A and SET1B can be rapidly depleted after 2 hours of treatment with the dTAG-13 compound (hereafter referred to as 'dTAG') (Fig. 3.1A)(Hughes et al., 2023). Using this cell line, Klose lab has shown that removing SET1A/B results in a modest but widespread decrease in H3K4me3 levels across the genome after 2 hours of dTAG treatment. Surprisingly, H3K4me3 levels recovered to levels comparable to those in untreated cells after 24 hours of dTAG treatment, suggesting an uncharacterised compensation mechanism, presumably via other methyltransferases. I first wanted to characterize these findings in more detail by performing a more extensive dTAG treatment time-course. I decided to add an 8-hour treatment to further characterize acute changes and a 48-hour treatment to further characterize the recovery in H3K4me3. In accordance with previous findings, analysis of bulk H3K4me3 levels did not reveal any major changes during this time-course, with a reduction to approximately 80% after 2 hours of dTAG treatment and a recovery to approximately 120% after 48 hours (Fig. 3.1B). I next performed calibrated ChIP-seq (cChIP-seq) to quantitatively investigate changes in H3K4me3 levels genome-wide. As expected, H3K4me3 levels at promoters decreased modestly after 2 hours of dTAG treatment (Fig. 3.1C). Interestingly, reductions in H3K4me3 levels were already less pronounced after 8 hours of dTAG treatment, and H3K4me3 levels continued to recover progressively after 24 hours of dTAG treatment. After 48 hours of SET1A/B removal, H3K4me3 levels had exceeded those in untreated cells. Therefore, removing SET1A/B leads to an immediate but modest reduction of H3K4me3 followed by a rapid but sustained recovery of H3K4me3 levels that lasts for multiple days.

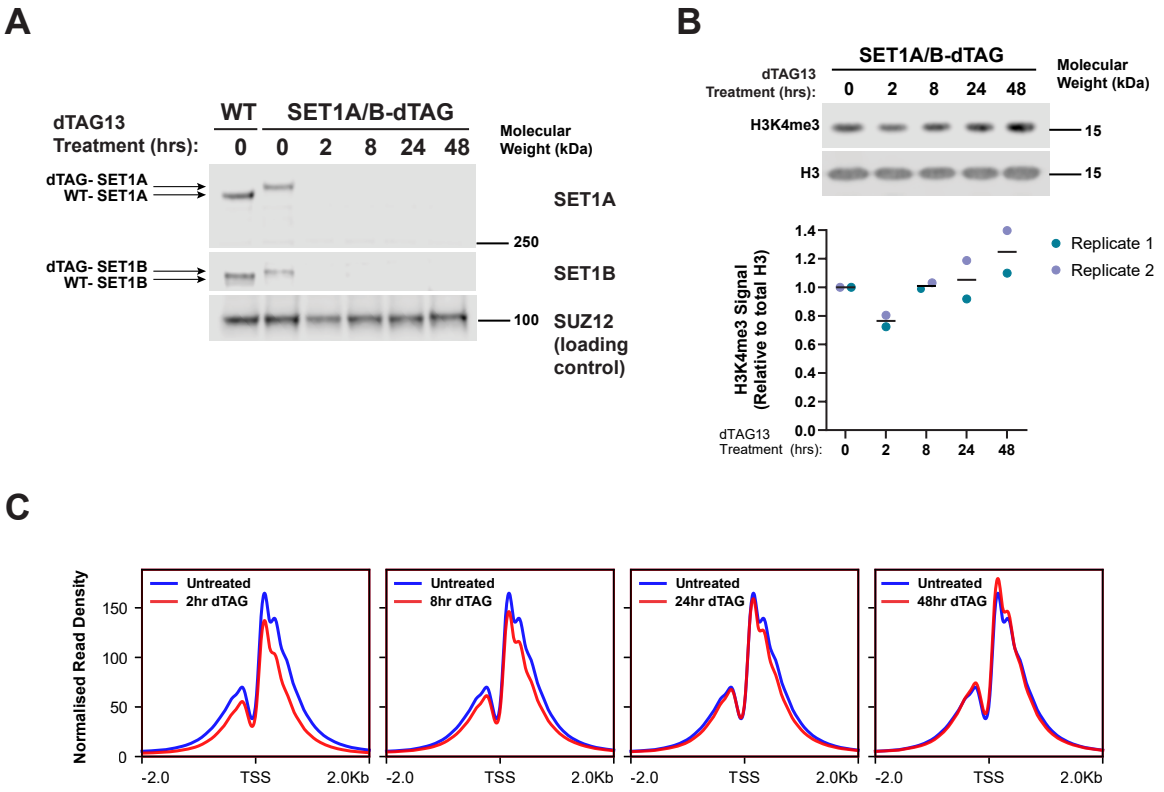

Figure 3.1: SET1A/B depletion results in modest reductions in H3K4me3.

(A) Immunoblotting of SET1A and SET1B from nuclear extracts of SET1A/B-dTAG cells following dTAG13 treatment. SUZ12 was used as a loading control.
(B) Immunoblotting of H3K4me3 from histone extracts of SET1A/B-dTAG cells following dTAG13 treatment. Histone H3 was used as a loading control. Quantitative analysis of bulk H3K4me3 levels are shown below. H3K4me3 signal was normalised to total H3 signal and shown as relative to the untreated sample. Individual values from 2 biological replicates are plotted, with a line showing the mean of the two values.
(C) Metaplots showing H3K4me3 cChIP-seq signal at all promoters (n=20,633) after SET1A/B-depletion.

ChIP-seq experiments have shown that SET1A/B occupancy tends to favour highly-transcribed genes (Brown et al., 2017). This would suggest that H3K4me3 levels at highly-transcribed genes should be more sensitive to SET1A/B removal. To test this hypothesis, I divided all promoters containing an H3K4me3 peak into quartiles according to levels of transcription using a previously generated transient transcriptome sequencing (TT-seq) dataset in untreated SET1A/B-dTAG cells, with genes in the 1st quartile being the most highly-transcribed (Hughes et al. 2023). Analysis of H3K4me3 levels in promoters across these quartiles revealed that while H3K4me3 in the promoters of highly-transcribed genes are somewhat more sensitive to SET1A/B removal, the general trend across all quartiles is that of a moderate decrease after 2 hours of SET1A/B removal (Fig. 3. 2A). Interestingly, I noted that the lowly-transcribed genes of the 4th quartile experienced a slightly more pronounced recovery after 48 hours of SET1A/B-depletion (Fig. 3. 2B, C).

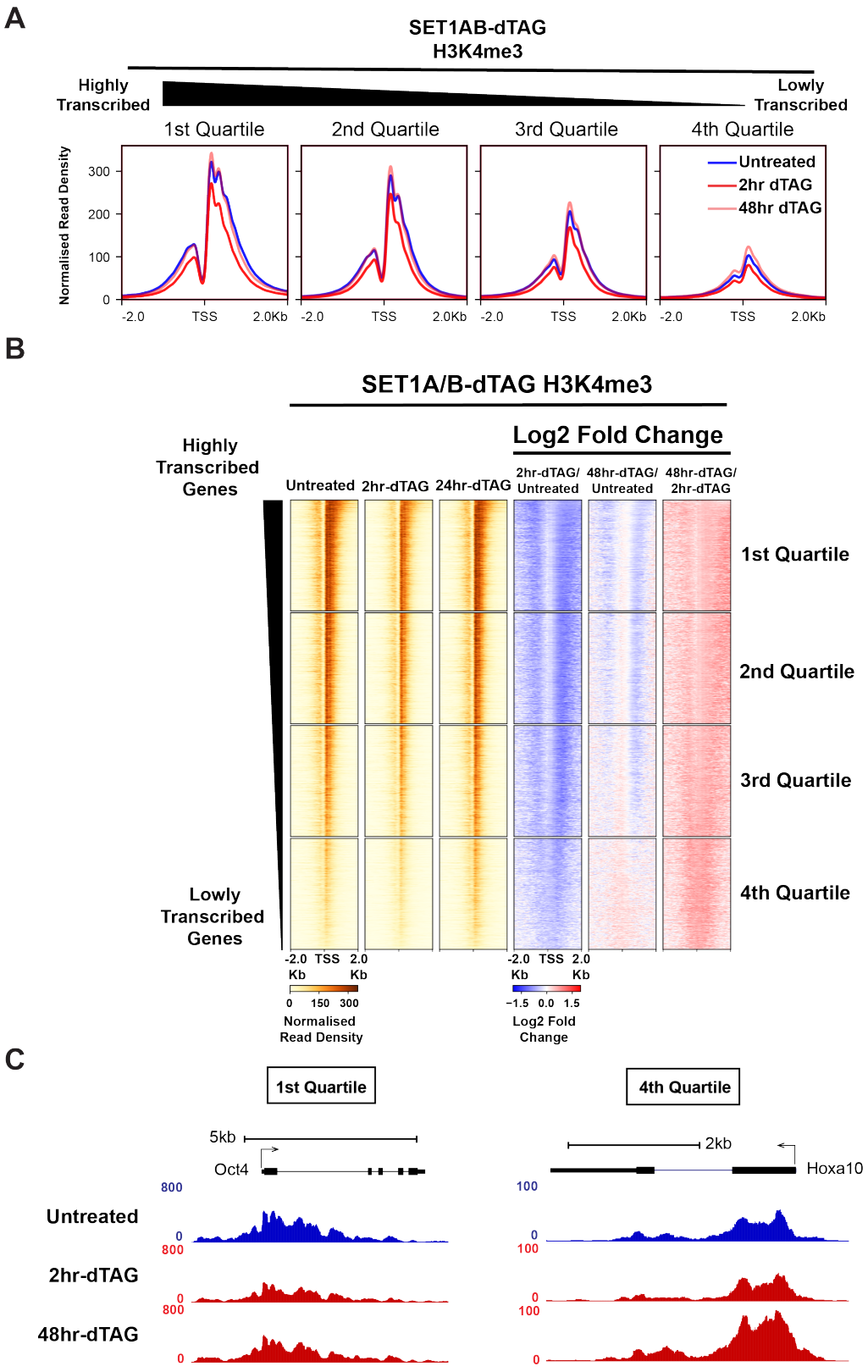

Figure 3.2: Highly-transcribed genes are slightly more sensitive to SET1A/B-depletion.

(A) Metaplots showing H3K4me3 cChIP-seq signal over H3K4me3-associated promoters (n=14,724) divided into quartiles based on levels of transcription in untreated cells.
(B) Heatmaps showing H3K4me3 cChIP-seq signal over promoters divided into quartiles as shown in (A). Also shown are heatmaps of the $\log_2$ fold change between dTAG-treated and untreated H3K4me3 cChIP-seq signal and between 48hr-dTAG-treated and 2hr-dTAG-treated H3K4me3 cChIP-seq signal. All heatmaps are sorted in descending order based on transcription levels in untreated cells.
(C) Genome coverage tracks of H3K4me3 at a representative gene of the 1st quartile (*Oct4*) and a representative gene of the 4th quartile (*Hoxa10*) after SET1A/B-depletion.

Although the general trend across promoters is a moderate decrease in H3K4me3 levels after 2 hours of SET1A/B removal, I noted that a small subset of genes appeared to be more sensitive to SET1A/B removal than others. Differential enrichment analysis revealed that H3K4me3 levels at 706 promoters were reduced by more than 2 fold after 2 hours of SET1A/B-depletion (Fig. 3. 3A). However, by 48 hours of dTAG treatment, only 30 promoters continued to exhibit a 2-fold reduction in H3K4me3 levels (Fig. 3. 3A, B). This indicates that very few genes are entirely-dependent on SET1A/B for H3K4me3 deposition and that virtually all promoters that exhibit H3K4me3 reductions upon SET1A/B depletion experience some H3K4me3 recovery (Fig. 3.3C). Interestingly, I found that these 706 promoters were not enriched for highly-transcribed genes, but were fairly well distributed across the four transcription quartiles (Fig. 3. 3D). Accordingly, levels of total RNA Polymerase II were broadly similar at the 706 SET1A/B-dependent promoters and SET1A/B-independent promoters (Fig. 3.3E). Surprisingly, SET1A/B-dependent genes had markedly lower levels of SET1A (as assessed by T7 cChIP-seq in a T7-dTAG-SET1A cell line, Hughes et al., 2023). Similarly, MLL2 levels (as assessed by cChIP-seq in untreated cells - see Chapter 5) were also markedly lower at SET1A/B-dependent promoters (Fig. 3.3G). Additionally, these promoters had lower CpG density and therefore, lower levels of non-methylated CpG dinucleotides as assessed by BioCap, a technique which specifically identifies non-methylated DNA (Blackledge et al., 2012) (Fig. 3.3H, I). Taken together, these observations demonstrate that there is a small number of promoters with low CpG-density at which SET1A/B are the primary contributors of H3K4me3, likely through a binding modality that is independent of the CXXC domain of CFP1. The recovery of H3K4me3 at these genes also suggests that there is a compensatory H3K4 methyltransferase acting at these genes independently of CGI-binding.

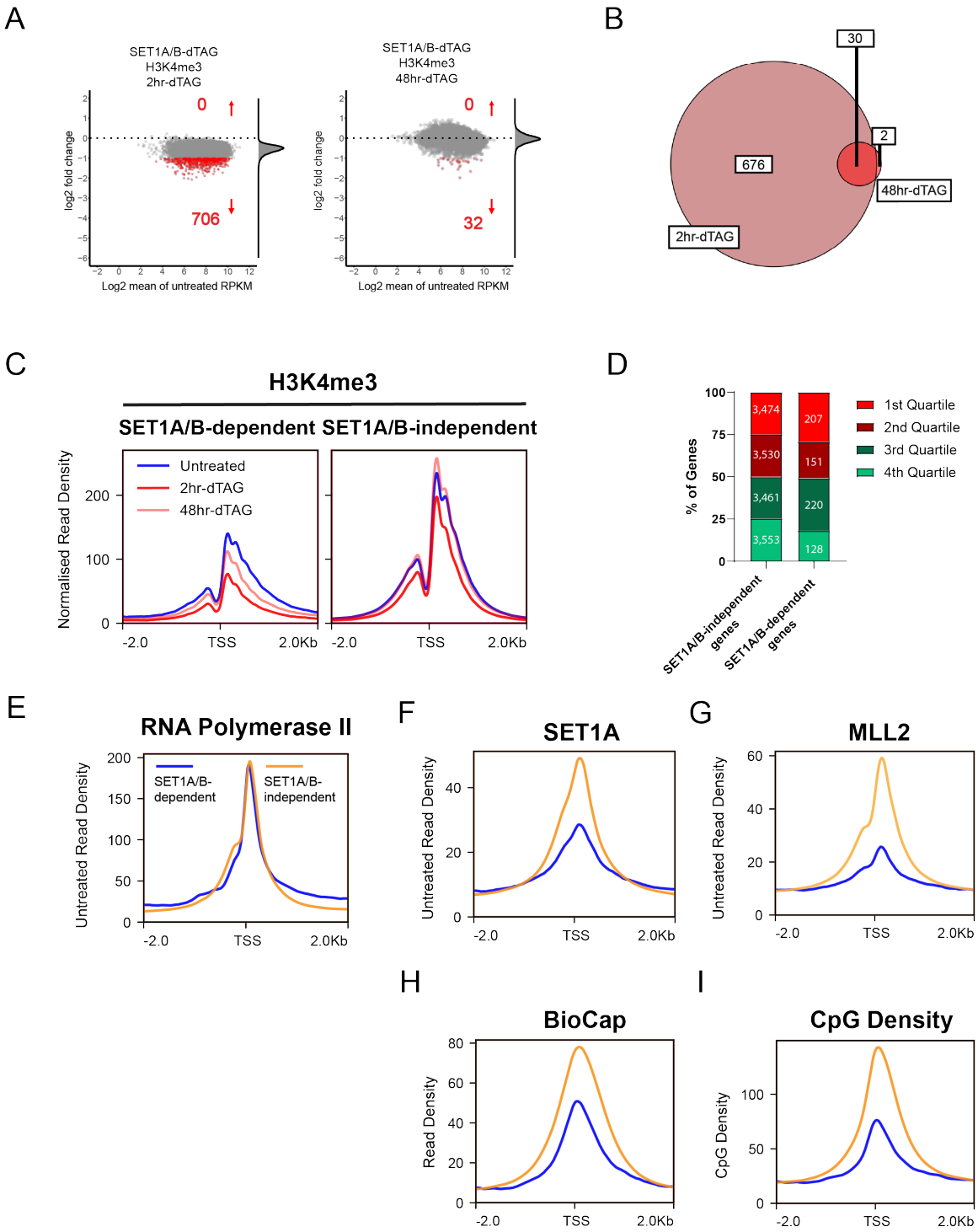

Figure 3.3: Only a small subset of genes are highly-sensitive to SET1AB-depletion.

(A) MA plots showing differential enrichment of H3K4me3 across all H3K4me3-associated promoters (n=14,724) between 2hr-dTAG-treated (left) or 48hr-dTAG-treated (right) and untreated cells. Promoters demonstrating at least 2-fold change in H3K4me3 levels are highlighted in red and the number of these promoters is indicated in red.
(B) Venn diagram showing overlap between promoters demonstrating at least 2-fold reduction in H3K4me3 after 2 hours of dTAG treatment and after 48 hours of dTAG treatment.
(C) Metaplots showing H3K4me3 cChIP-seq signal at SET1A/B-dependent (n=706) and -independent (n=14,018) promoters after 2 hours and 48 hours of SET1A/B-depletion.
(D) Bar graph showing distribution of the 706 SET1A/B-sensitive genes across the four promoter quartiles divided by transcription levels. The number of genes from each quartile are shown.
(E - I) Metaplots showing cChIP-seq signal of (E) total RNA Polymerase II, (F) SET1A, (G) MLL2 in untreated mESCs, (H) BioCap signal and (I) CpG density in wild-type mESCs at SET1A/B-dependent and SET1A/B-independent promoters.

3.1.2 Validation of H3K4-methylation antibodies

Having substantiated and expanded our previous finding that SET1A/B do not contribute greatly to H3K4me3 on their own, I next asked how SET1A/B shaped H3K4me1/2 across the genome. However, I first needed to have confidence that the H3K4me1 and H3K4me2 antibodies would accurately capture H3K4me1/2 distribution by ChIP-seq. A major technical problem that has affected the reproducibility of investigations into H3K4 methylation has been the specificity of the antibodies against the three methylated states in ChIP experiments (Shah et al., 2018).

Antibody specificity has traditionally been assessed using microarray-based methods, but these do not necessarily recapitulate their performance in ChIP experiments. Epicypher has recently developed a technique termed SNAP-ChIP to assess the specificity of antibodies in histone modification ChIP experiments (Shah et al., 2018). This technique uses synthetic, post-translationally-modified nucleosomes with barcoded DNA as a spike-in control during the ChIP experiment (Fig. 3.4A). Pull-down of target and non-specific epitopes can then be assessed quantitatively by qPCR using primers designed to anneal to the barcodes. Using this method, the Klose lab has recently shown that our in-house-generated H3K4me3 antibody exhibits high specificity for H3K4me3 with little cross-reactivity against H3K4me1/2 (Amy Hughes, unpublished data). However, the specificity of the commercially-available H3K4me1 and H3K4me2 antibodies I would use for subsequent ChIP-seq experiments had yet to be characterized.

I therefore used the SNAP-ChIP technique to characterize the specificity of our commercially available H3K4me1 and H3K4me2 antibodies. I found that the H3K4me2 antibody cross-reacted the most with H3K4me1, but still maintained a several-hundredfold enrichment of H3K4me2 signal over H3K4me1 signal at all four of the amounts of antibody tested (Fig. 3.4B). I found that the H3K4me1 antibody cross-reacted the most with H3K4me2 and unmodified H3K4, but also maintained a several-hundredfold enrichment of H3K4me1 signal over these two epitopes (Fig. 3. 4B). Therefore, these H3K4me1 and H3K4me2 antibodies used for ChIP-seq experiments described hereafter have the required specificity to interrogate H3K4me1/2 throughout the genome.

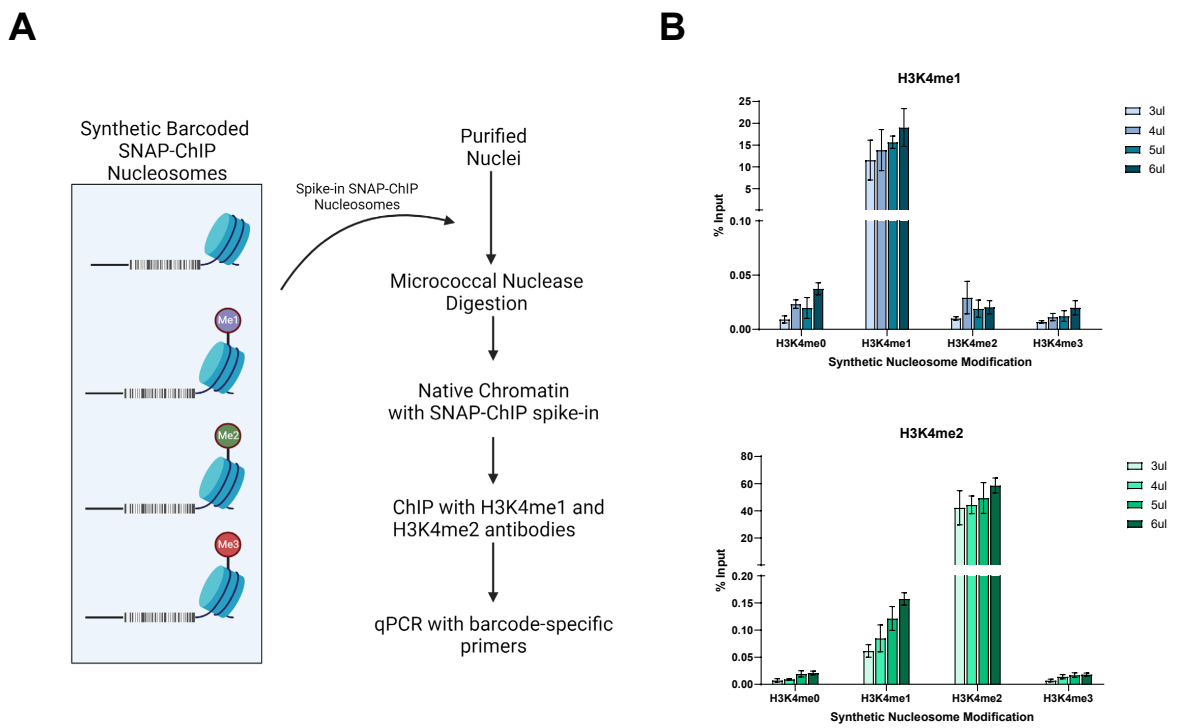

Figure 3.4: H3K4me1 and H3K4me2 antibodies are highly specific.

(A) Schematic illustrating the SNAP-ChIP experimental workflow.
(B) ChIP-qPCR analysis of H3K4me1 and H3K4me2 SNAP-ChIP experiments. Mean values with standard error of the mean from 3 replicates are shown.

3.1.3 H3K4me1/2 broadly increase upon SET1A/B-depletion

Having confirmed that our H3K4me1 and H3K4me2 antibodies were specific for their targets, I next wanted to know how SET1A/B contribute to H3K4me1/2. Surprisingly, bulk analysis of H3K4me2 levels by immunoblotting revealed that H3K4me2 levels increased after SET1A/B depletion (Fig. 3.5A). This increase is modest at the early timepoints of 2 hours and 8hrs, and primarily manifests from 24 hours of SET1A/B removal onwards. To understand whether this increase in H3K4me2 was uniform across the genome, I performed cChIP-seq for H3K4me2. Analysis of H3K4me2 levels across all promoters broadly recapitulates changes in bulk H3K4me2 levels seen in the western blots: a modest increase after 2 hours and 8 hours of dTAG treatment followed by a more pronounced increase after 24 hours and 48 hours of dTAG treatment (Fig. 3. 5B). Interestingly, I noted that this increase in H3K4me2 after 2 hours of dTAG treatment was unique to promoters; H3K4me2 levels across gene bodies and enhancers in fact showed a slight decrease after 2 hours of dTAG treatment (Fig. 3. 5C, D).

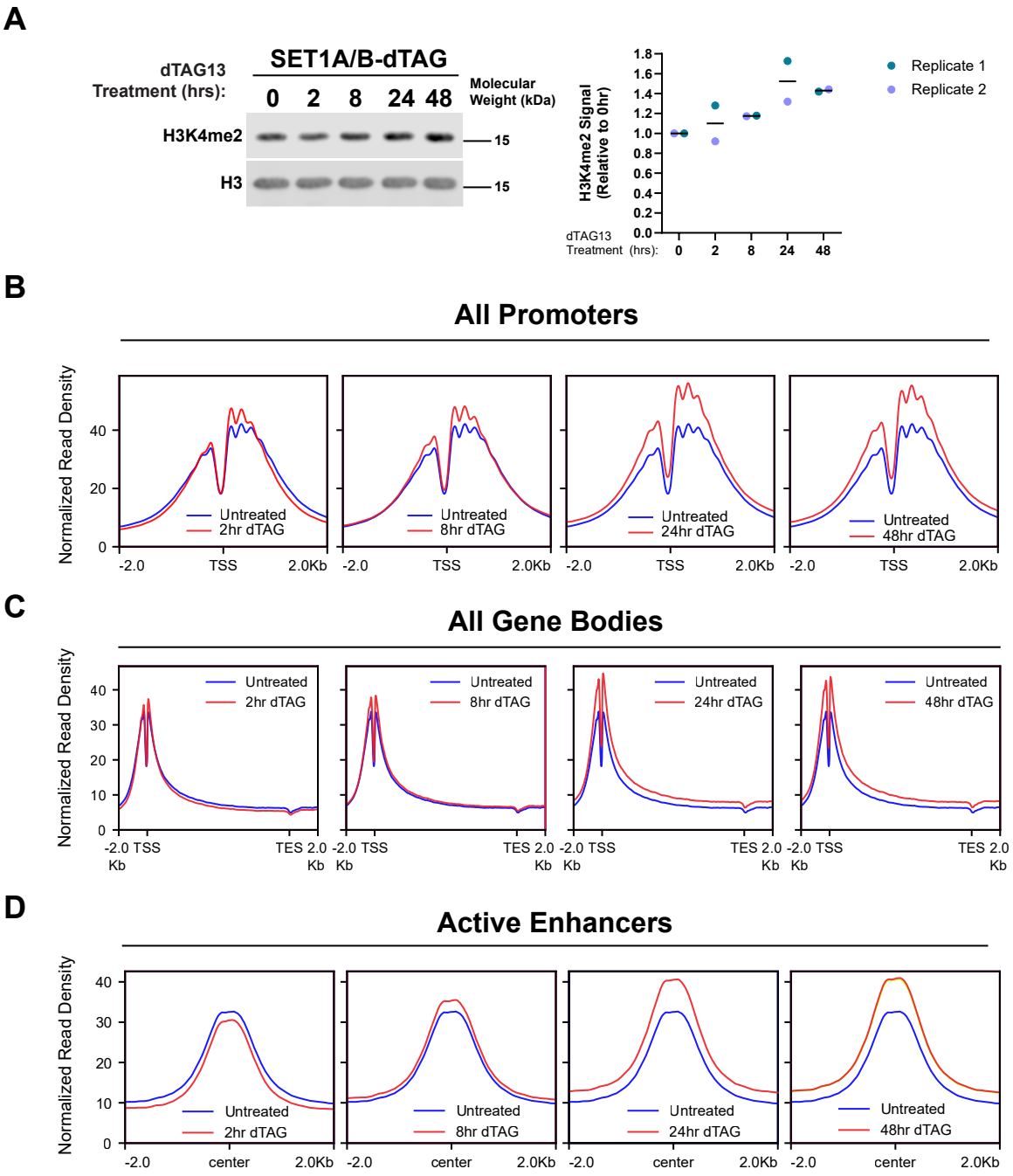

Figure 3.5: H3K4me2 increases after SET1A/B depletion.

(A) Western blot analysis of bulk H3K4me2 levels after SET1A/B-depletion. Histone H3 was used as a loading control. Quantification of H3K4me2 signal is shown on the right - H3K4me2 signal was normalised to total H3 signal and shown as relative to the untreated condition. Individual values from two biological replicates are shown along with a line showing the mean of the two values.
(B - D) Metaplots showing H3K4me2 cChIP-seq signal over **(B)** all promoters (n=20,633), **(C)** all gene bodies (n=20,633), and **(D)** all active enhancers (n=12,006) after SET1A/B-depletion.

I therefore focused on promoters for a more detailed analysis on the effect of removing SET1A/B on H3K4me2. Analysis of H3K4me2 levels across promoters divided into quartiles based on transcription levels revealed that H3K4me2 increased rapidly at the promoters of highly transcribed genes upon SET1A/B depletion (Fig. 3. 6A-C). At these genes (1st and 2nd quartiles), this increase was particularly pronounced just downstream of the transcription start site at the +1 and +2 nucleosomes, where H3K4me2 enrichment is ordinarily low and H3K4me3 enrichment is ordinarily high. In more lowly transcribed genes (3rd and 4th quartiles), this region is less enriched for H3K4me3 and more enriched for H3K4me2. At these more lowly-transcribed genes, H3K4me2 levels do not increase appreciably after 2 hours of SET1A/B depletion, but only increase after prolonged SET1A/B removal. Given that H3K4me3 at highly-transcribed genes are more sensitive to SET1A/B removal, this suggests that the increase in H3K4me2 levels seen after 2 hours of SET1A/B depletion may be a result of either demethylation of H3K4me3 or accumulation of H3K4me2 due to lack of conversion to H3K4me3 by SET1A/B.

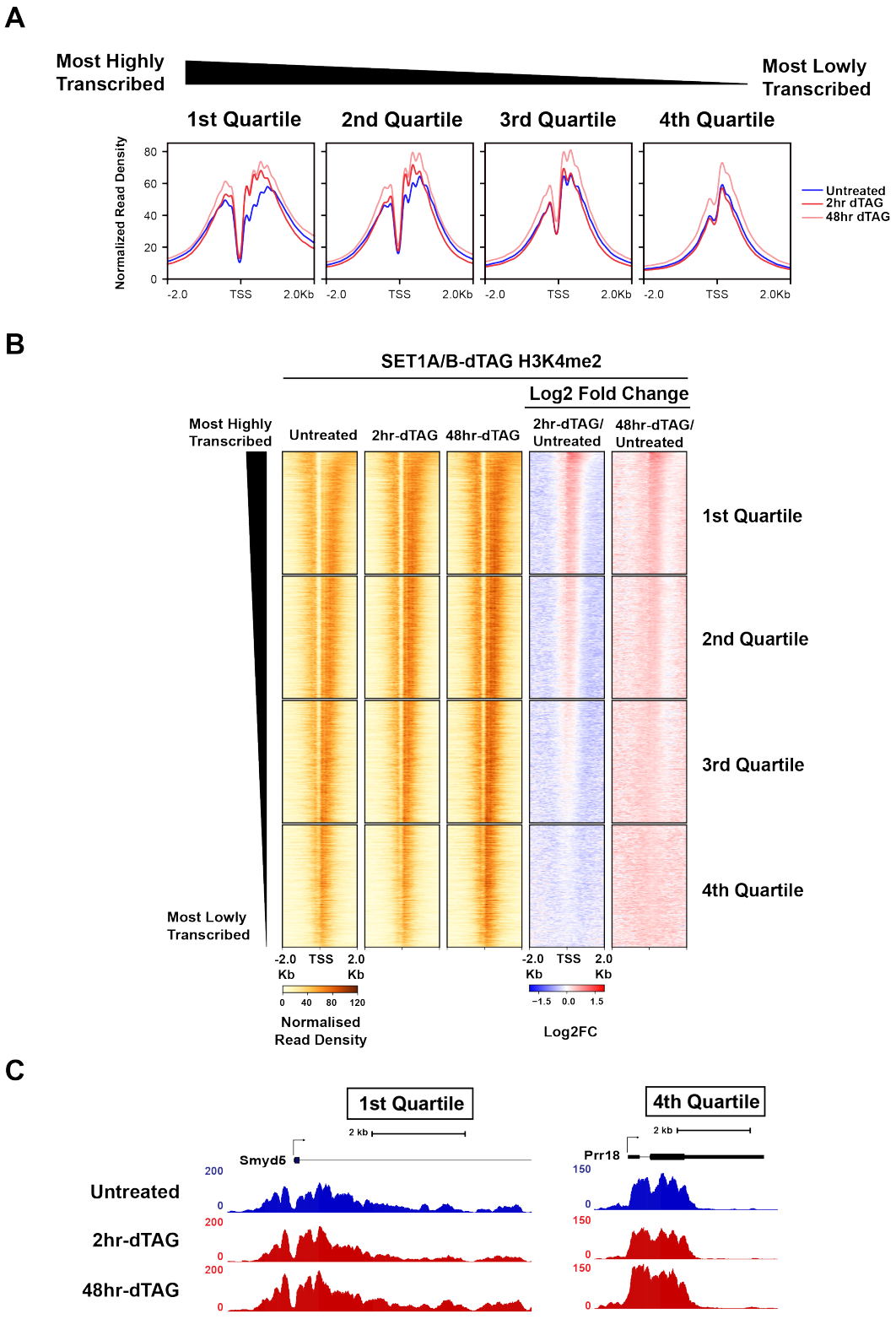

Figure 3.6: H3K4me2 levels increase at promoters of highly transcribed genes upon acute SET1A/B-depletion.

(A) Metaplots showing H3K4me2 ChIP-seq signal across H3K4me2-associated promoters (n=15,379) in untreated cells and cells treated with dTAG for 2 hours or 48 hours. Promoters were divided into quartiles according to levels of transcription of the associated gene in untreated SET1AB-dTAG cells as assessed by TT-seq.
(B) Heatmaps showing the $\log_2$ fold change of H3K4me2 ChIP-seq signal between cells treated with dTAG for 2 hours or 48 hours and untreated cells across promoters of the four quartiles shown in (A). Heatmaps are sorted by transcription levels of the associated gene in untreated cells.
(C) genome coverage tracks of H3K4me2 ChIP-seq experiments across one representative gene from the 1st quartile, *Smyd5*, and one representative gene from the 4th quartile, *Prr18*.

In contrast to H3K4me3, H3K4me1 is highly enriched within gene bodies and enhancers. Given the broad increase in H3K4me2 within these regions, I wanted to examine how depleting SET1A/B would affect H3K4me1. Western blot analysis indicated that bulk H3K4me1 levels also increased after SET1A/B depletion, and cChIP-seq analysis showed that similarly to H3K4me2, H3K4me1 levels increased across gene bodies and at enhancers, but only after 48 hours of dTAG treatment (Fig. 3.7). Notably, this increase in H3K4me1 occurs at regions of the genome where H3K4me3 levels are ordinarily low or non-existent. Additionally, this increase occurs after prolonged SET1A/B-depletion, when H3K4me3 levels have also increased beyond levels in untreated cells. Furthermore, this increase occurs uniformly across the genome irrespective of transcription levels of the associated gene. These observations indicate that this increase in H3K4me1 is unlikely to have resulted from demethylation of H3K4me3 or accumulation of H3K4me1 due to lack of conversion by SET1A/B to higher methylation states. Rather, these observations suggest that in the absence of SET1A/B, there is an increased H3K4-mono- and di- methyltransferase activity within the cell that may be explained by compensation from other H3K4 methyltransferases in the absence of SET1A/B. Taken together, these observations show that SET1A/B are not major contributors to H3K4 methylation on their own, and their most prominent role seems to be the conversion of H3K4me2 to H3K4me3 at highly transcribed genes.

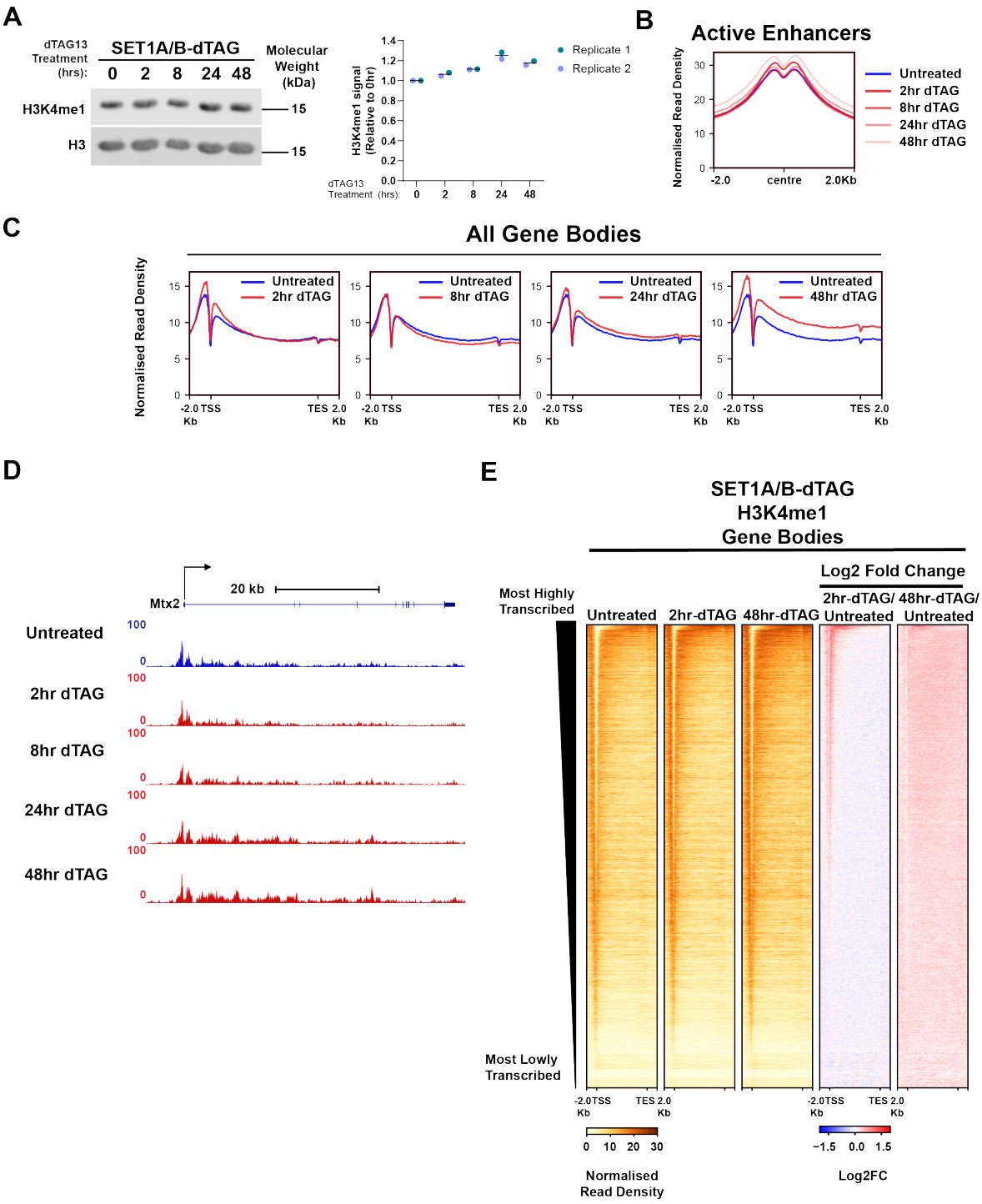

Figure 3.7: H3K4me1 broadly increases upon SET1A/B-depletion.

(A) Western blot analysis of bulk H3K4me1 levels after SET1A/B-depletion. Histone H3 is used as a loading control. Quantification of H3K4me1 signal is shown on the right - H3K4me1 signal is normalised to total H3 signal and shown as relative to the untreated condition. Individual values from two biological replicates are shown, with a line showing the mean of the two values.
(B) Metaplots showing H3K4me1 ChIP-seq signal across all active enhancers (n=12,006) after SET1A/B-depletion.
(C) Metaplots showing H3K4me1 ChIP-seq signal across all gene bodies (n=20,633) after SET1A/B-depletion.
(D) Genome coverage tracks of H3K4me1 at an example gene, *Mtx2*, after SET1A/B-depletion.
(E) H3K4me1 cChIP-seq signal over all gene bodies as shown in (B) after 2 hours and 48 hours of dTAG treatment. Also shown are heatmaps of the $\log_2$ ratio of H3K4me1 cChIP-seq signal between the dTAG-treated and untreated conditions. All heatmaps are sorted in descending order by transcription levels in untreated cells.

3.2 MLL1/2 deposit H3K4me2/3 broadly across the genome

3.2.1 A degron system to rapidly deplete MLL1 and MLL2

The finding that SET1A/B contribute minimally to H3K4 methylation on their own suggests that other H3K4 methyltransferases have an underappreciated role in depositing H3K4 methylation. Given that MLL1 and MLL2 are H3K4 methyltransferases that bind CpG islands broadly throughout the genome, I wanted to investigate the extent of their contributions to H3K4 methylation.

To understand the role of MLL1/2 in depositing H3K4 methylation in mESCs, I generated a cell line in which MLL1/2 can be rapidly depleted using degron technology (Nabet et al., 2018). MLL1 and MLL2 are both produced as single proteins which are then cleaved by Taspase I to form an N-terminal fragment, containing the CXXC domain, and a C-terminal fragment, containing the catalytic SET domain (Hsieh et al., 2003, Takeda et al., 2006). After cleavage, the N-terminal fragment and the C-terminal fragment form a heterodimer. Devising a strategy with which to tag these proteins with $FKBP12^{F36V}$ has proven challenging, given that tagging the proteins at the C-terminus risks perturbing the SET domain, and tagging the N-terminus of the proteins may not result in degradation of the cleaved C-terminal fragments. The Klose lab had previously designed a tagging strategy in which endogenous MLL1 and MLL2 were internally tagged to fuse $FKBP12^{F36V}$ to the N-terminus of the C-terminal fragment (Amy Hughes and Anna Lastuvkova) (Fig. 3.8A). It was envisaged that upon the addition of dTAG, the cleaved C-terminal fragments, which contain the SET domain required for H3K4 methyltransferase activity, would be degraded by targeted proteolysis before dimerization with the N-terminal fragments. Using this strategy, the Klose lab had generated an MLL2-dTAG cell line, in which I further targeted MLL1 with the $FKBP12^{F36V}$ tag to generate the MLL1/2-dTAG cell line. Characterization of MLL1 and MLL2 proteins by immunoblotting of nuclear extracts demonstrate that the C-terminal fragments of both MLL1 and MLL2 are fully depleted after 2 hours of dTAG treatment, comparable to the degradation of SET1A and SET1B in the SET1A/B-dTAG cell line (Fig. 3.8B). Interestingly, immunoblotting using antibodies against the N-terminal fragments of MLL1 and MLL2 reveal that both N-terminal fragments are also depleted after dTAG treatment, albeit at a slower rate (Fig. 3.8B). This is in line with previous studies indicating that both fragments are unstable without dimerization (Yokoyama et al., 2011). I also noted that the full-length MLL2 was readily observable by immunoblotting using the antibody against the N-terminal fragment (MLL2-N), in contrast to full-length MLL1, which was not. Full-length MLL2 was also depleted after dTAG treatment, indicating that in addition to degrading the C-terminal fragments, the $FKBP12^{F36V}$ tag also resulted in the uncleaved MLL2 protein being degrading after dTAG treatment. Therefore, the MLL1/2-dTAG cell line generated here allows for the near-complete depletion of the entire MLL1/2 proteins.

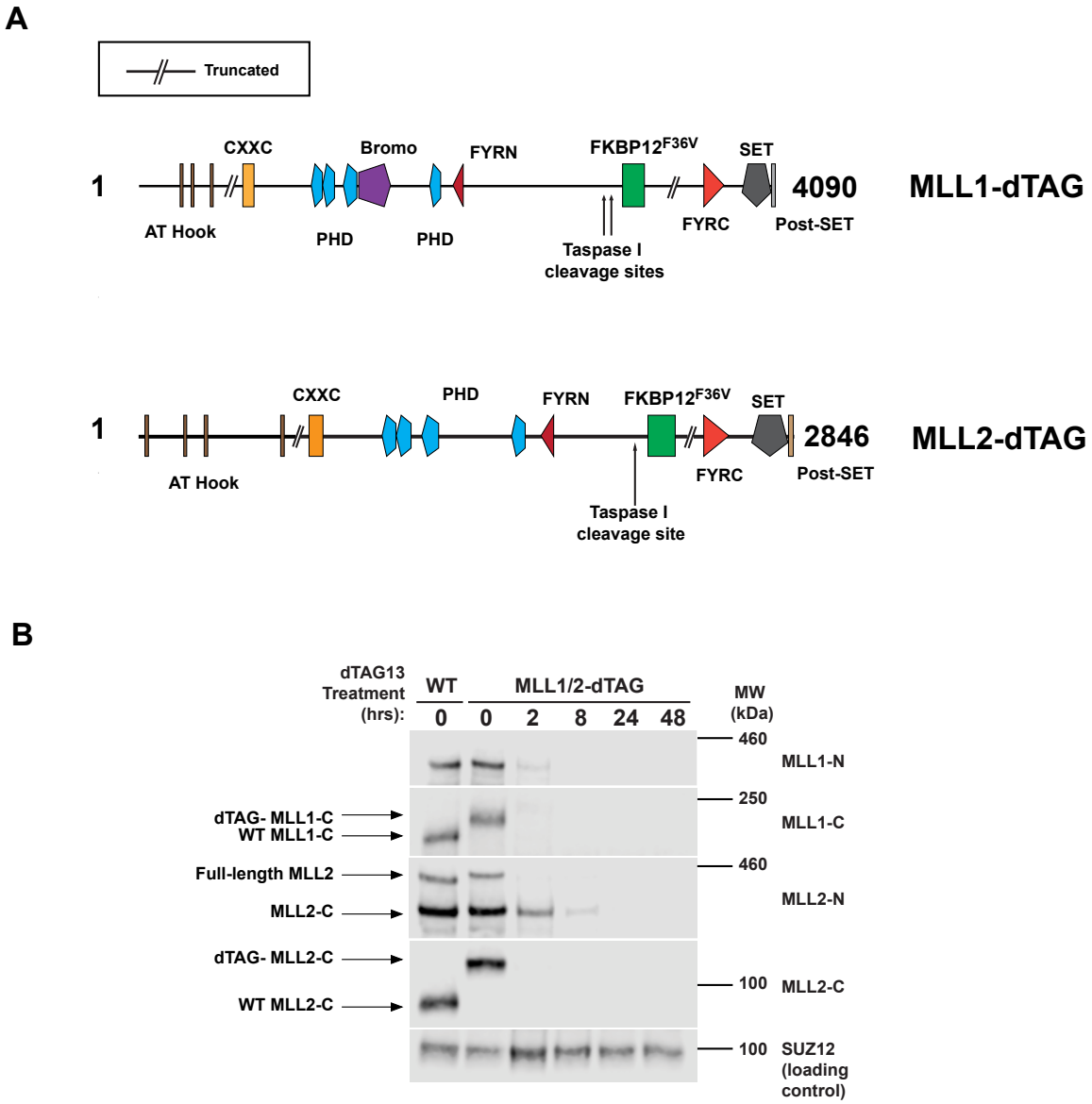

Figure 3.8: Generating an MLL1/2-dTAG cell line.

(A) Schematic illustrating the internal tagging strategy for fusing the degron tag to MLL1 and MLL2. Protein domain architecture was analysed using SMART.
(B) Western blots demonstrating robust degradation of MLL1 and MLL2 after dTAG treatment. SUZ12 was used as a loading control. MLL1-N: MLL1 N-terminal fragment, MLL1-C: MLL1 C-terminal fragment, MLL2-N: MLL2 N-terminal fragment, MLL2-C: MLL2 C-terminal fragment.

3.2.2 MLL1/2 deposit H3K4me3 at all CpG-island promoters but are only essential at a subset of lowly-transcribed genes

I first sought to investigate whether MLL1/2 have a greater role in depositing H3K4me3 than previously reported. In line with previous reports, I observed a mild decrease in bulk levels of H3K4me3 from 2 hours of dTAG treatment onwards (Fig. 3.9A). I next sought to investigate how changes in H3K4me3 occurred throughout the genome after MLL1/2 depletion by performing cChIP-seq. In contrast to SET1A/B-depletion, MLL1/2-depletion resulted in a larger reduction in H3K4me3 levels across the genome. Metaplot analysis demonstrated that H3K4me3 levels across promoters decreased after 2 hours of dTAG treatment and reached their lowest after 8 hours of dTAG treatment, followed by a recovery thereafter (Fig. 3.9B). I noted that this pattern was not uniform across the genome, with H3K4me3 at the promoters of lowly-transcribed genes being much more severely reduced after MLL1/2-depletion compared to highly-transcribed genes. Additionally, H3K4me3 levels did not recover at lowly-transcribed genes, with H3K4me3 levels reaching their lowest after 8 hours of dTAG treatment and remaining at those levels at subsequent timepoints. These findings reinforce the notion that MLL1/2 are responsible for depositing H3K4me3 at lowly-transcribed genes. However, I noted that H3K4me3 across all promoters were reduced after 8 hours of MLL1/2-depletion, including at highly-transcribed genes (Fig. 3.9B-C). This demonstrates that MLL1/2 also have a role in depositing H3K4me3 at highly-transcribed genes which may be compensated for over time by other methyltransferases in the absence of MLL1/2. Additionally, I noted that across all promoters, H3K4me3 upstream of the transcription start site was more sensitive to MLL1/2-depletion compared to H3K4me3 downstream of the transcription start site, an effect not seen after SET1A/B-depletion (Fig. 3.9C0. Moreover, this reduction was fairly uniform across promoters of all four quartiles. Taken together, these findings indicate that MLL1/2 deposit H3K4me3 broadly across all promoters, but only H3K4me3 peaks upstream of the transcription start site and H3K4me3 peaks at lowly-transcribed genes are primarily dependent on MLL1/2.

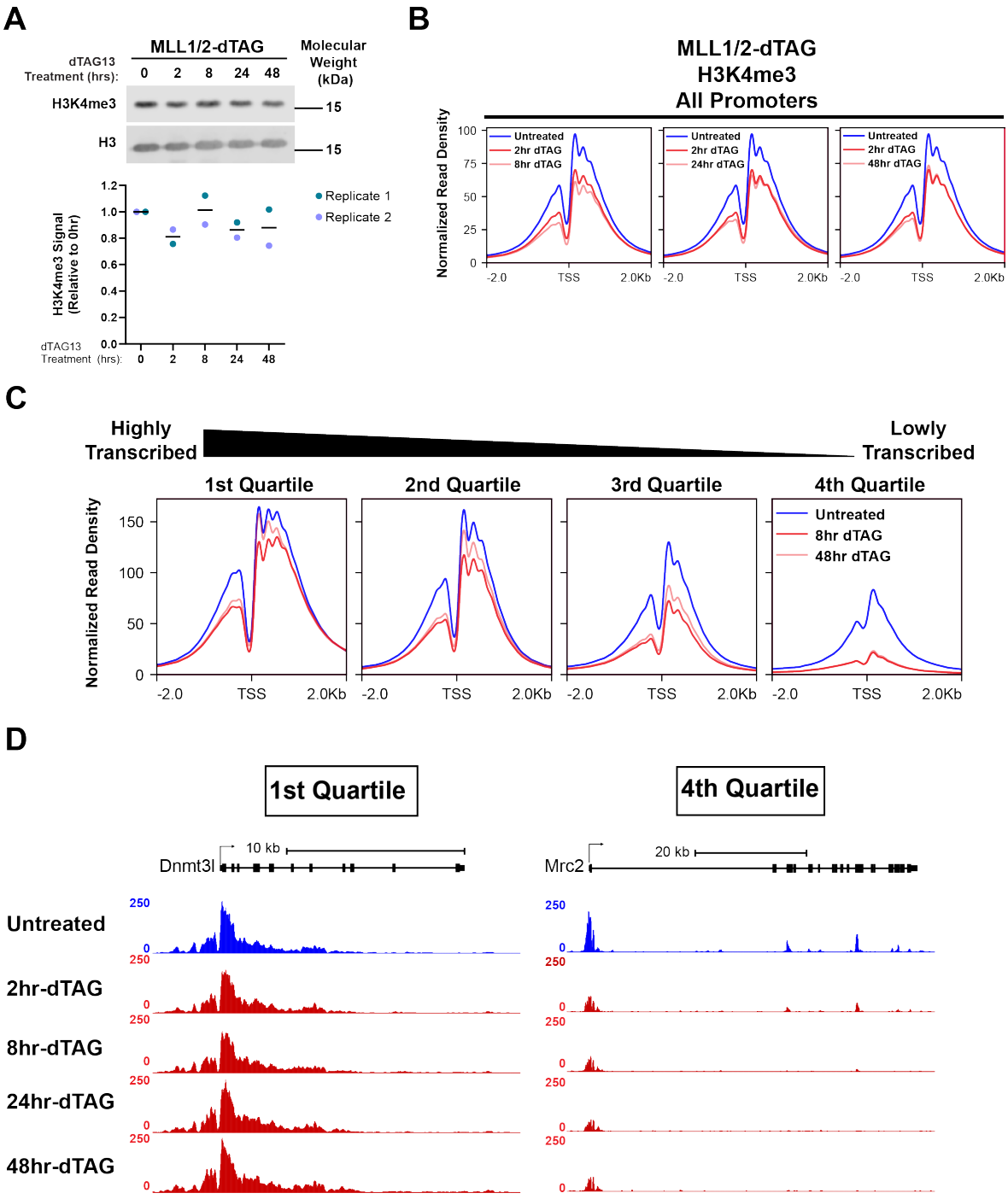

Figure 3.9: MLL1/2 deposit H3K4me3 broadly across the genome.

(A) Western blot analysis of bulk H3K4me3 levels after MLL1/2-depletion. Histone H3 is used as a loading control. Quantification of H3K4me3 signal is shown below - H3K4me3 signal is normalised to total H3 signal and shown as relative to the untreated condition. Individual values from 2 biological replicates are plotted with a line showing the mean of the two values.
(B) Metaplots showing H3K4me3 levels at promoters following MLL1/2-depletion (n=20,633).
(C) Metaplots showing H3K4me3 at H3K4me3-associated promoters (n=14,724) divided into quartiles based on transcription levels.
(D) Genome coverage tracks showing H3K4me3 at a representative gene from the 1st quartile (*Dnmt3l*) and the 4th quartile (*Mrc2*).

I next sought to identify and characterize the promoters where H3K4me3 levels were most sensitive to MLL1/2-depletion. Therefore, I defined MLL1/2-dependent promoters as those where H3K4me3 was reduced by more than 2-fold after 2 hours of dTAG treatment and continued to be reduced by more than 2-fold after 48hrs of treatment. Differential enrichment analysis revealed that after 2 hours of dTAG treatment, H3K4me3 was reduced by more than 2-fold at 2,792 promoters (Fig. 3.10A). Of these 2,792 promoters, H3K4me3 remained reduced by more than 2-fold at 2,693 promoters, indicating that H3K4me3 at these promoters is highly dependent on MLL1/2 (Fig. 3.10B). MLL1/2-dependent promoters were lowly-transcribed, with lower levels of RNA Polymerase II occupancy and lower levels of H3K4me3 (Fig. 3.10C). Surprisingly, MLL1/2-dependent promoters also had lower levels of MLL2 occupancy. This suggests that dependency of a promoter on MLL1/2 for H3K4me3 levels may not rely on high levels of MLL1/2 occupancy, but may result from lack of binding by other H3K4 methyltransferases.

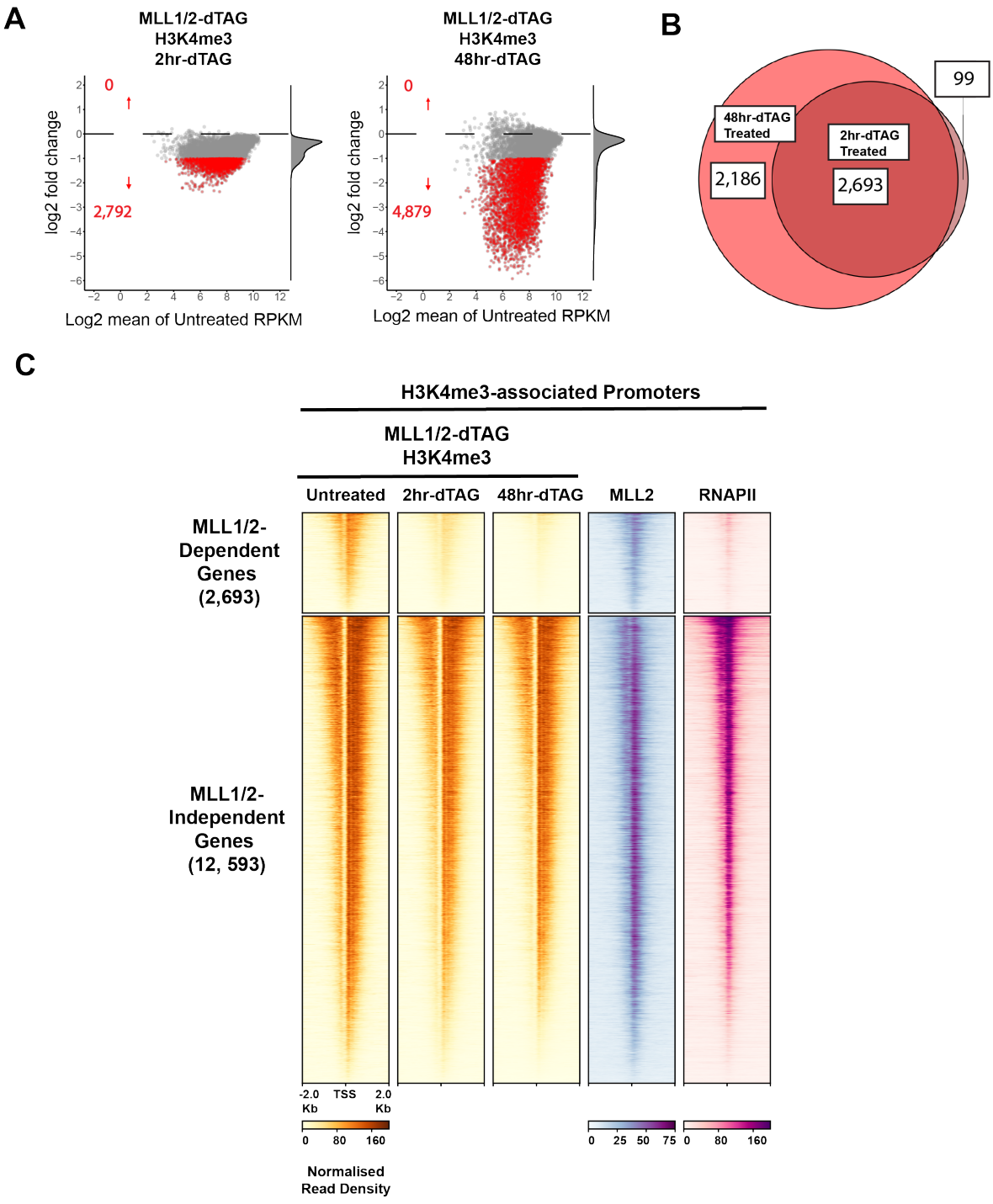

Figure 3.10: A subset of promoters are strongly-dependent on MLL1/2 for H3K4me3.

(A) MA plots showing log2 fold changes in H3K4me3 at H3K4me3-associated promoters (n=14,724) following 2 hours or 48 hours of MLL1/2-depletion. Promoters showing more than 2-fold reduction in H3K4me3 after dTAG treatment are shown in red.
(B) Venn diagram showing overlap between promoters showing more than 2-fold reduction in H3K4me3 after 2 hours and 48 hours of MLL1/2-depletion. Promoters with more than 2-fold reduction in H3K4me3 after both 2 hours and 48 hours of MLL1/2-depletion are classified as MLL1/2-dependent and all other H3K4me3-associated promoters are classified as MLL1/2-independent.
(C) Heatmaps showing H3K4me3 levels at MLL1/2-dependent and MLL1/2-independent promoters following 2 hours or 48 hours of MLL1/2-depletion, along with MLL2 and total RNA polymerase II ChIP-seq signal in untreated cells. Heatmaps are sorted in descending order by H3K4me3 levels.

3.2.3 MLL1/2 are major contributors of H3K4me2 deposition at lowly-transcribed genes and at enhancers

Given the major role MLL1/2 have in depositing H3K4me3, I wanted to understand if they also contribute to depositing H3K4me2. Surprisingly, bulk levels of H3K4me2 decreased significantly after MLL1/2 depletion, to approximately 60% (Fig. 3.11A). However, this reduction was slower, with major reductions manifesting primarily after 24 hours and 48 hours of dTAG treatment. Accordingly, cChIP-seq analysis demonstrates that across all promoters, H3K4me2 levels were mildly reduced after 2 hours of dTAG treatment, reaching their lowest point after 8 hours of dTAG treatment and with no recovery thereafter (Fig. 3.11B). I noted that sites of H3K4me3 and H3K4me2 reductions were highly correlated. Indeed, analysis of H3K4me2 levels across promoters of different levels of transcription indicated that H3K4me2 levels decreased the most at promoters of lowly-transcribed genes, similar to the effect observed for H3K4me3 (Fig. 3.11B). The greater overall reduction in H3K4me2 levels was not due more genes being sensitive to MLL1/2 removal but rather due to the comparatively greater enrichment of H3K4me2 at lowly-transcribed genes. Indeed, at the 2,693 promoters which were most dependent on MLL1/2 for H3K4me3, levels of H3K4me2 were also largely depleted (Fig. 3.11C). This indicates that promoters of a subset of lowly-transcribed genes require MLL1/2 for both H3K4me2 and H3K4me3. Additionally, removal of MLL1/2 also resulted in reductions in H3K4me2 upstream of the transcription start site at highly-transcribed genes, although H3K4me2 levels downstream of the transcription start site were only mildly reduced (Fig. 3.11D). This is similar to the effect removing MLL1/2 has on H3K4me3, where H3K4me3 upstream of the transcription start site of highly-transcribed genes is more severely reduced. Interestingly, I found that H3K4me2 levels were also substantially reduced at active enhancers after MLL1/2-depletion (Fig. 3.11E). This indicates that MLL1/2 broadly deposits H3K4me2/3 at lowly-transcribed genes and regulatory elements with little active transcription.

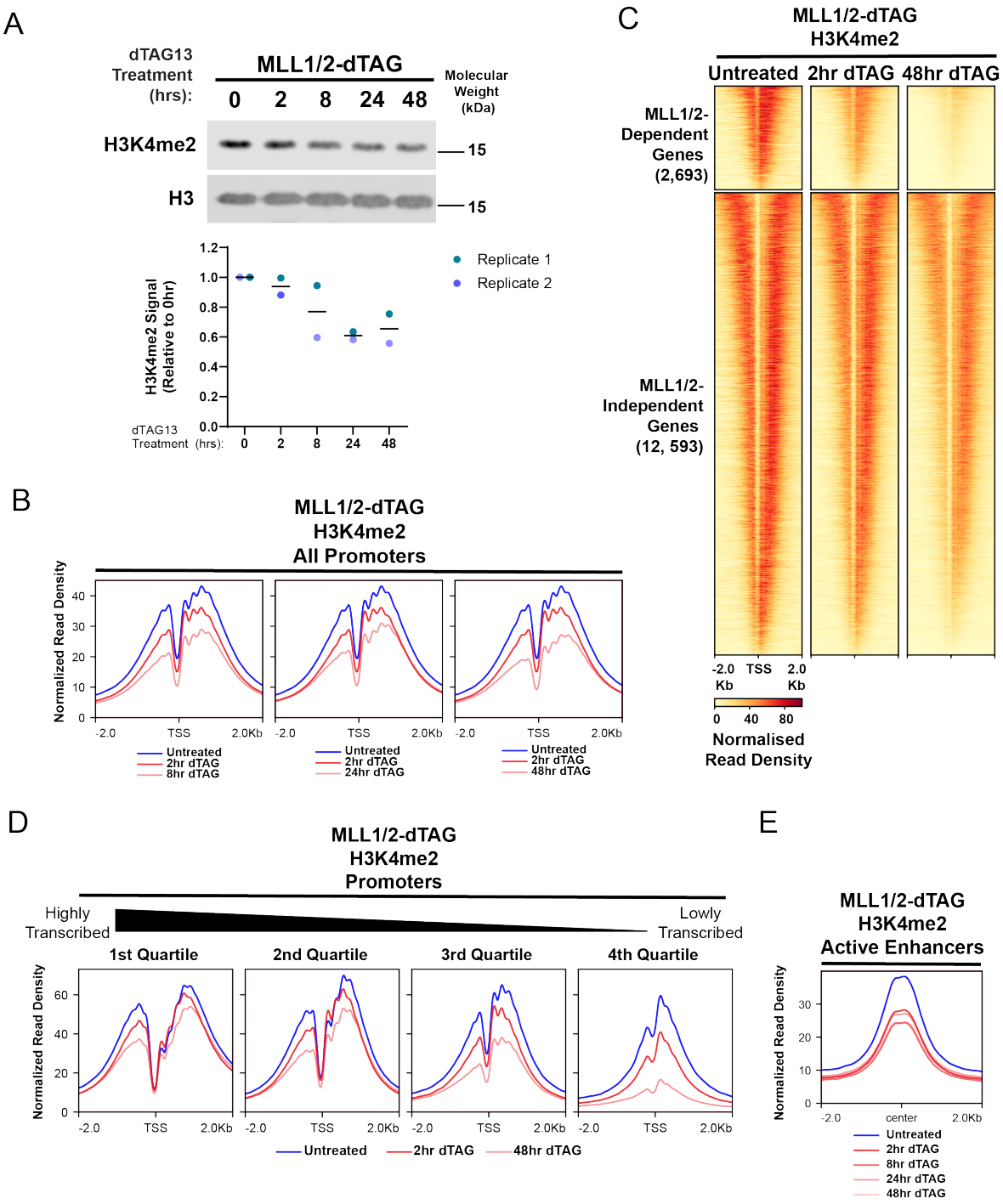

Figure 3.11: Removing MLL1/2 leads to substantial reductions in H3K4me2.

(A) Western blots showing bulk H3K4me2 levels after MLL1/2-depletion. Histone H3 is used as a loading control. Quantification of H3K4me2 signal is shown below - H3K4me2 signal is normalised to total H3 levels and shown as relative to the untreated condition. Individual values from 2 biological replicates are shown, with a line showing the mean of the two values.
(B) Metaplots showing H3K4me2 cChIP-seq signal at all promoters (n=20,633) and H3K4me2-containing promoters (n=15,379) divided into quartiles based on transcription levels.
(C) Heatmaps showing H3K4me2 levels at promoters that are dependent on MLL1/2 for H3K4me3 deposition. Heatmaps are sorted by H3K4me2 levels in untreated cells.
(D) Heatmaps showing H3K4me2 levels after MLL1/2-depletion for 2 hours or 48 hours at H3K4me2-associated promoters divided into quartiles based on transcription.
(E) Metaplot showing H3K4me2 levels at active enhancers (n=12,006) after MLL1/2-depletion.

3.2.4 MLL1/2 have minor contributions to H3K4me1

I next wanted to understand if MLL1/2 also contribute to H3K4me1. Interestingly, cChIP-seq revealed that H3K4me1 moderately increased around transcription start sites after 2 hours of MLL1/2-depletion (Fig. 3.12A). This increase was fairly uniform across all promoters irrespective of transcription levels of the associated genes (Fig. 3.12B). This effect was also seen after SET1A/B-depletion, albeit largely restricted to highly transcribed genes (Fig. 3.7E). Given that MLL1/2 deposits H3K4me3 broadly across all promoters, and highly-transcribed genes tend exhibit more pronounced reductions in H3K4me3 after SET1A/B-depletion, this rapid increase in H3K4me1 around transcription start sites could be a result of demethylation of H3K4me2/3 to H3K4me1 upon SET1A/B or MLL1/2-depletion. Alternatively, this increase could also be the result of accumulation of H3K4me1 due to the lack of conversion by SET1A/B or MLL1/2 to H3K4me3, although the rapidity of the increase suggests that demethylation of H3K4me3 is more likely to be the primary cause. Additionally, H3K4me1 levels also increased moderately at enhancers, where removal of MLL1/2 leads to reductions in H3K4me2 (Fig. 3.12A), further substantiating the hypothesis that H3K4me1 increases at these sites due to demethylation of H3K4me2/3 upon MLL1/2-depletion. In contrast, H3K4me1 decreased broadly across gene bodies upon MLL1/2-depletion (Fig. 3.12A, B), although this was not evident until after 8 hours of dTAG treatment. Interestingly, this reduction was more severe at the bodies of genes which depend on MLL1/2 for H3K4me2/3 (Fig. 3.12B, C), although substantial levels of H3K4me1 still remain. This indicates that MLL1/2 have minor contributions to H3K4me1 at lowly-transcribed genes which depend on them for H3K4me2/3, but otherwise have do not contribute to H3K4me1 deposition across the genome to any great extent. Taken together, these findings indicate that MLL1/2 are primarily H3K4 di- and tri-methyltransferases, with their methyltransferase activities most prominent at lowly-transcribed genes.

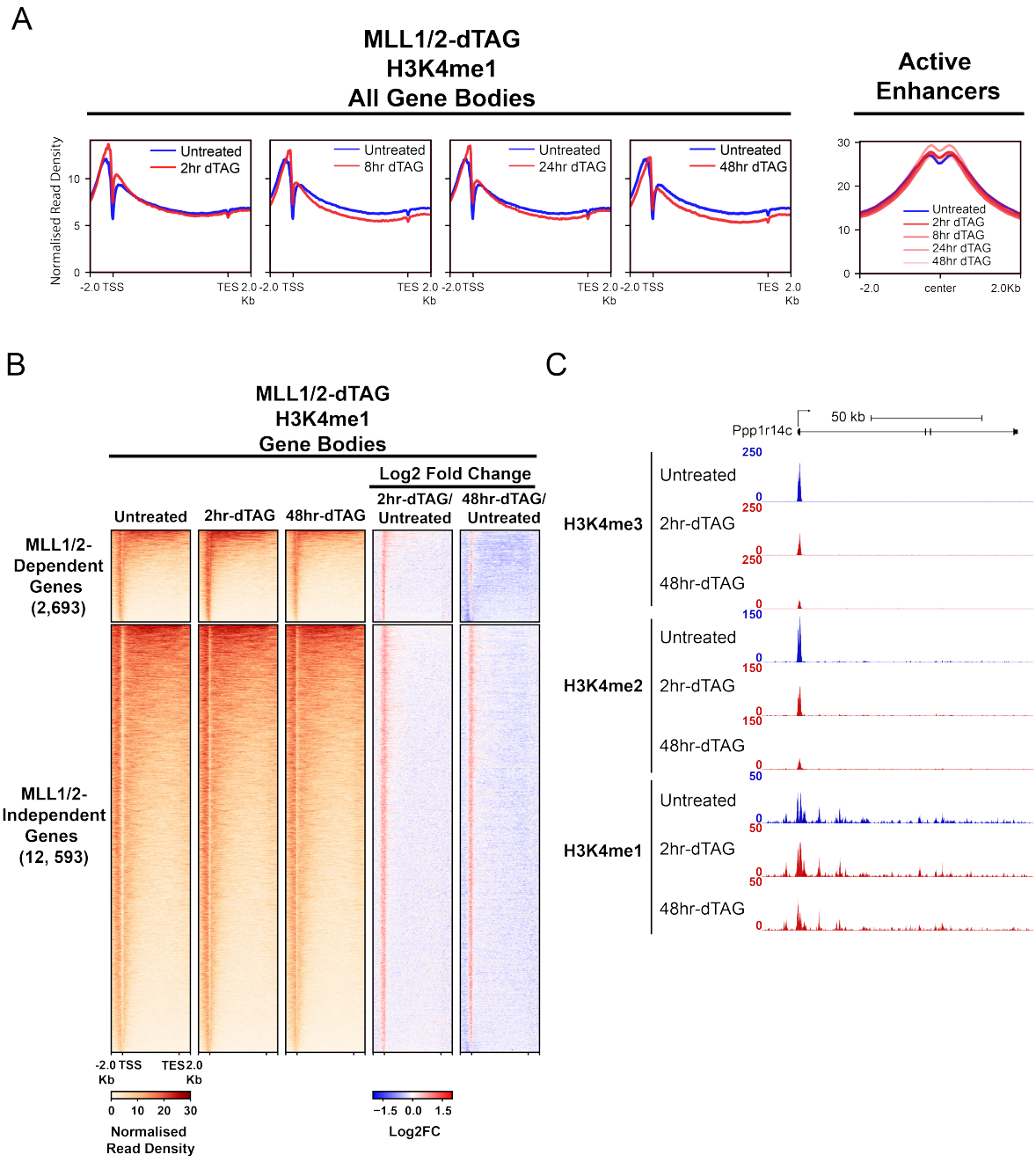

Figure 3.12: Removal of MLL1/2 has minor effects on H3K4me1.

(A) Metaplots showing H3K4me1 cChIP-seq signal over all gene bodies (n=20,633) and active enhancers (n=12,006).
(B) Heatmaps showing H3K4me1 cChIP-seq signal over bodies of MLL1/2-dependent and MLL1/2-independent genes after MLL1/2-depletion for 2 hours and 48 hours. Also shown are log2 fold change heatmaps of H3K4me1 cChIP-seq signal between dTAG-treated and untreated conditions. All heatmaps are sorted by descending H3K4me1 signal in the untreated sample.
(C) Genome coverage tracks of H3K4me1/2/3 over a representative MLL1/2-dependent gene after MLL1/2-depletion (*Ppp1r14c*).

3.3 SET1A/B and MLL1/2 synergise to deposit H3K4me3 at CpG islands

3.3.1 Most CpG-island promoters require both SET1A/B and MLL1/2 for H3K4me3 deposition

Given that neither SET1A/B- nor MLL1/2-depletion caused profound reductions in H3K4me3 I reasoned that they may collaborate to deposit H3K4me3 at the majority of promoters. This hypothesis was especially compelling since both SET1A/B- and MLL1/2-depletion resulted in an immediate reduction followed by a sustained recovery in H3K4me3, suggesting potential compensation between these methyltransferases. Therefore, to understand if SET1A/B and MLL1/2 synergise to shape H3K4me3 deposition at CpG island promoters, I first analysed bulk levels of H3K4me3 using a previously generated cell line in which SET1A/B and MLL1/2 can be simultaneously depleted after dTAG treatment (Fig. 3.13A) (Amy Hughes and Anna Lastuvkova). In contrast to depleting SET1A/B or MLL1/2 on their own, removing SET1A/B and MLL1/2 simultaneously resulted in a large reduction in bulk H3K4me3 levels to approximately 50% after 2 hours of dTAG treatment as assessed by immunoblotting (Fig. 3.13B). Surprisingly, removing the four H3K4-trimethyltransferases did not remove all H3K4me3. Additionally, H3K4me3 levels did not decline further, but instead recovered from 8 hours of dTAG treatment onwards. This suggests that another methyltransferase capable of tri-methylating H3K4 might compensate for loss of SET1A/B/MLL1/2.

The surprising finding that removing the four CGI-associated H3K4 methyltransferases did not result in the removal of all H3K4me3 led me to investigate where this remaining H3K4me3 was distributed across the genome. I performed cChIP-seq for H3K4me3 in SET1A/B/MLL1/2-dTAG cells and found that approximately 50% of H3K4me3 was retained across promoters (Fig. 3.13C). Analysis of H3K4me3 across all promoters after SET1A/B/MLL1/2-depletion broadly recapitulated the observations from western blots of bulk H3K4me3: H3K4me3 levels were reduced to approximately 50% after 2 hours of dTAG treatment, followed by a modest recovery after 8 hours and 24 hours of dTAG treatment (Fig. 3.13C).

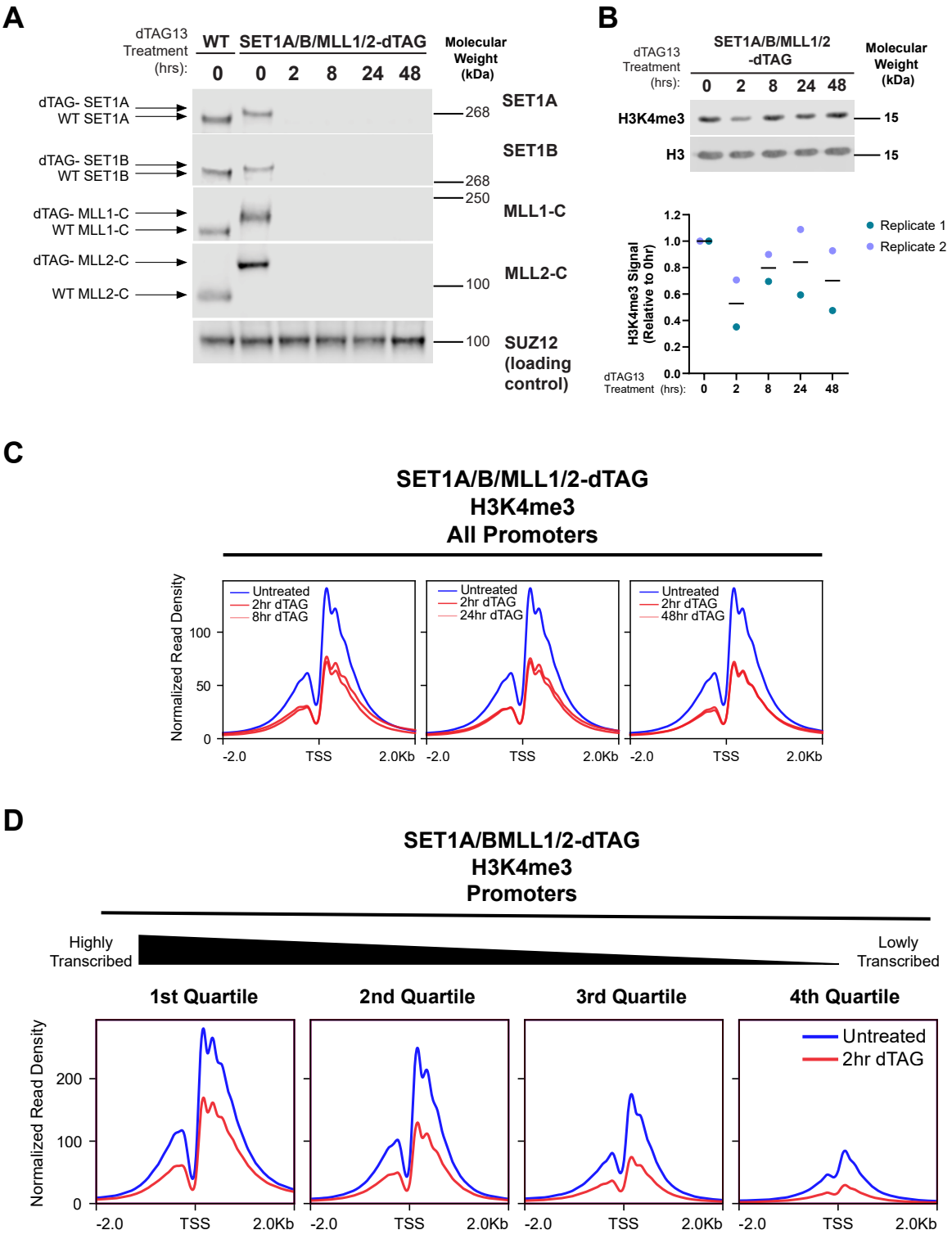

Figure 3.13: Removal of SET1A/B/MLL1/2 leads to larger, but incomplete reductions in H3K4me3.

(A) Western blots showing robust degradation of SET1A, SET1B, MLL1, and MLL2 after dTAG treatment. SUZ12 is used as a loading control.
(B) Western blot analysis of bulk H3K4me3 levels after SET1A/B/MLL1/2 depletion. Histone H3 is used as a loading control. Quantification of H3K4me3 signal is shown below - H3K4me3 signal is normalised to total histone H3 signal and shown as relative to untreated. Individual values from two biological replicates are plotted with a line showing the mean of the two values.
(C) Metaplots showing H3K4me3 cChIP-seq signal at all promoters (n=20,633) after SET1A/B/MLL1/2-depletion.
(D) Metaplots showing H3K4me3 cChIP-seq signal at H3K4me3-associated promoters (n=14,724) divided into quartiles based on transcription after 2 hours of SET1A/B/MLL1/2-depletion.

However, I noted that this pattern was not uniform across the genome and was highly dependent on the transcription levels of the gene. Indeed, dividing promoters into quartiles according to transcription revealed that promoters could broadly be divided into two classes – MLL1/2-dependent promoters, and SET1A/B/MLL1/2-dependent promoters (Fig. 3.13D). MLL1/2-dependent promoters comprised the most lowly-transcribed genes of the fourth quartile and exhibited reductions in H3K4me3 after SET1A/B/MLL1/2-depletion in a manner that is identical to that after MLL1/2-depletion. SET1A/B/MLL1/2-dependent promoters comprised genes from the first three quartiles of most highly transcribed genes where individual removal of SET1A/B or MLL1/2 did not lead to major reductions in H3K4me3. At these promoters, H3K4me3 levels were reduced to approximately 50% after 2 hours SET1A/B/MLL1/2-depletion, although the more highly transcribed genes of the first quartile retained higher levels of H3K4me3 compared to the less highly transcribed genes of the second and third quartiles (Fig. 3.13D).

A comparative analysis of H3K4me3 levels after SET1A/B-, MLL1/2-, or SET1A/B/MLL1/2-depletion demonstrates that SET1A/B and MLL1/2 synergise to deposit H3K4me3 (Fig. 3.14). While removal of SET1A/B or MLL1/2 on their own results in moderate reductions in H3K4me3 across all promoters on average, simultaneous removal of SET1A/B/MLL1/2 results in much more substantial reductions at all promoters (Fig. 3.14A, B, C). Consequently, removal of SET1A/B/MLL1/2 results in a larger number of promoters that exhibit significant reductions in H3K4me3. Indeed, I find that upon SET1A/B/MLL1/2-depletion, 7,456 promoters, approximately half of all promoters containing an H3K4me3 cChIP-seq peak, exhibit H3K4me3 reductions of over 2-fold (Fig. 3.14 D). This is greater than the 706 promoters where H3K4me3 is reduced by more than 2-fold upon SET1A/B-depletion and the 2,693 promoters where H3K4me3 is reduced by more than 2-fold upon MLL1/2-depletion, combined (Fig. 3.14D). These findings demonstrate that the effect of removing SET1A/B/MLL1/2 on H3K4me3 are not merely additive of the effects seen when removing SET1A/B and MLL1/2 on their own, but represents synergy between these complexes.

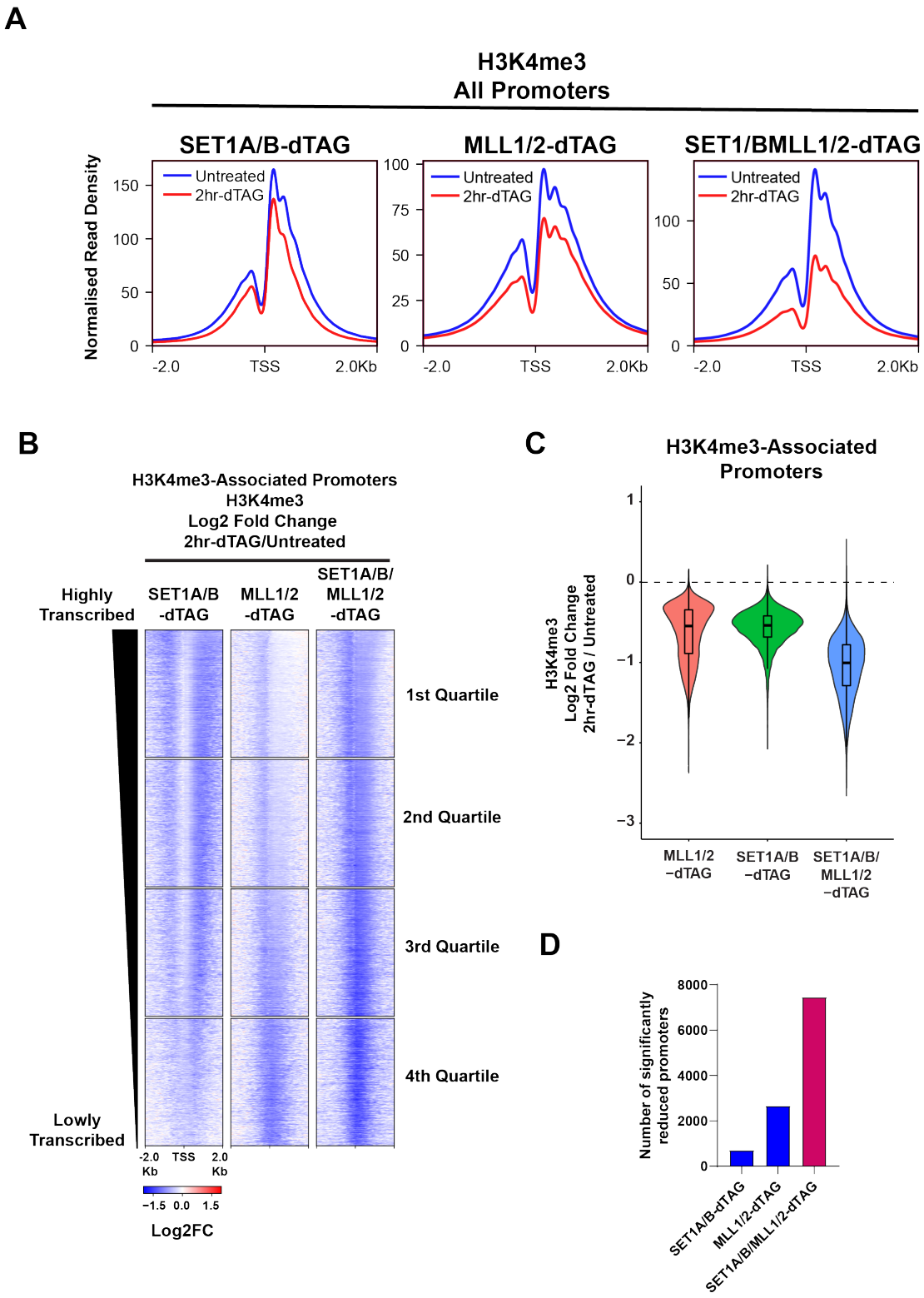

Figure 3.14: SET1A/B and MLL1/2 synergise to deposit H3K4me3.

(A) Metaplots showing H3K4me3 cChIP-seq signal at all promoters (n=20,633) after 2 hours of SET1A/B-, MLL1/2, or SET1A/B/MLL1/2-depletion.
(B) Heatmaps showing log2 fold change of H3K4me3 cChIP-seq signal between 2-hour-dTAG and untreated conditions. Heatmaps are centred at H3K4me3-associated promoters divided into quartiles based on transcription and sorted in descending order by transcription.
(C) Violin plots showing log2 fold change of H3K4me3 cChIP-seq signal at H3K4me3-associated promoters between 2-hour-dTAG and untreated conditions.
(D) Bar graph showing the number of promoters where H3K4me3 is reduced by more than 2-fold after 2 hours of dTAG treatment in the indicated cell lines.

Interestingly, extending the analysis of H3K4me3 distribution in the gene body uncovers a second pattern of recovery in which the low levels of H3K4me3 found within the gene body immediately downstream of the promoter gradually increase from 8 hours of dTAG treatment onwards (Fig. 3.15A). This spreading of H3K4me3 into the gene body is also correlated with transcription, with the increase being more pronounced in highly transcribed genes (Fig. 3.15A, B). Indeed, the lowly-transcribed genes of the 4th quartile did not exhibit spreading of H3K4me3 into the gene body (Fig. 3.15A-C). This suggests that at highly-transcribed genes, an alternative mechanism deposits H3K4me3 in the absence of SET1A/B/MLL1/2, possibly in a transcription-coupled manner.

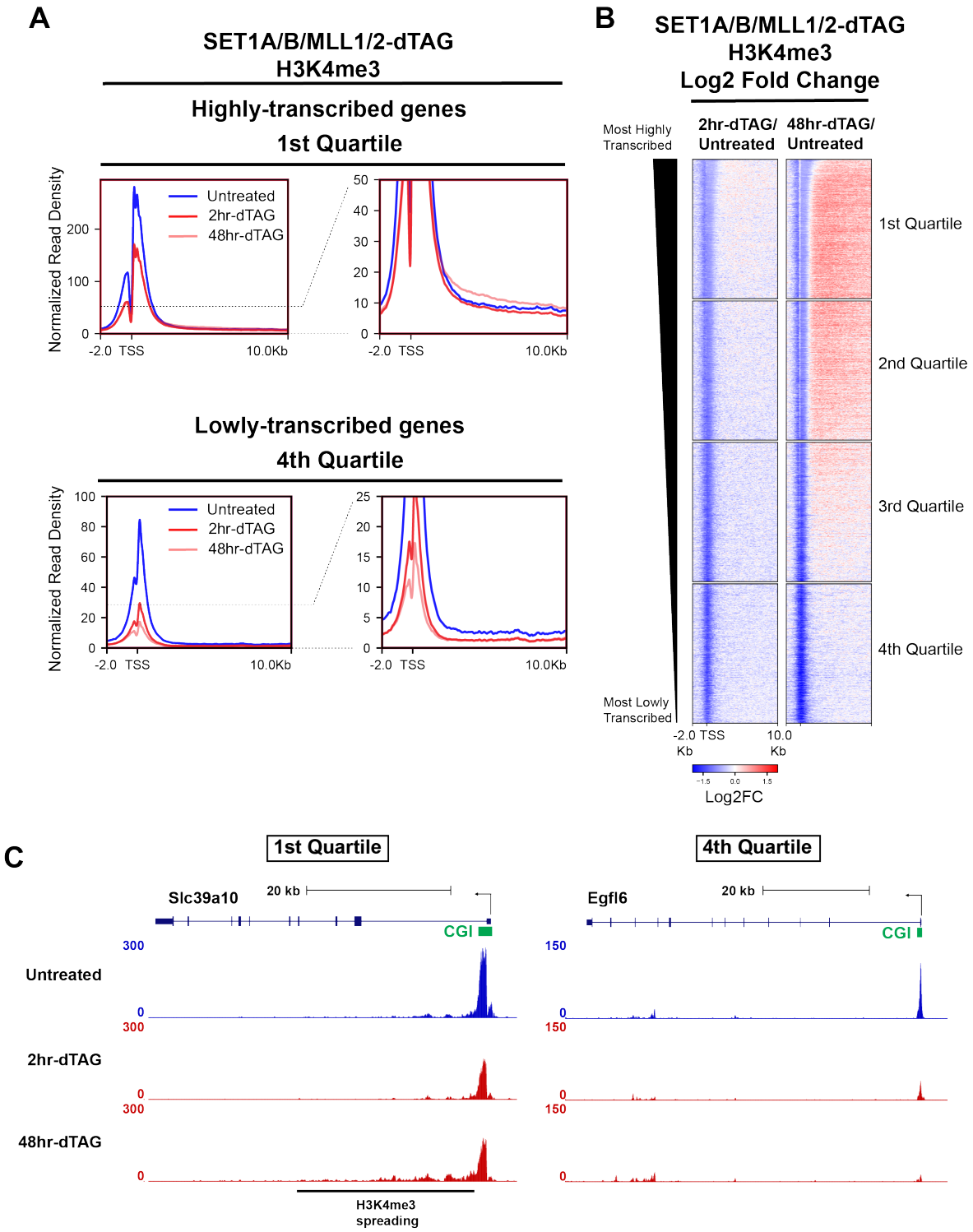

Figure 3.15: H3K4me3 spreads into gene bodies of highly-transcribed genes after prolonged SET1A/B/MLL1/2-depletion.

(A) Metaplots showing H3K4me3 cChIP-seq signal at highly-transcribed 1st Quartile genes (top) and lowly-transcribed 4th Quartile genes (bottom) after 2-hour and 48-hour dTAG treatment .
(B) Heatmaps showing log2 fold change of H3K4me3 cChIP-seq signal between dTAG-treated and untreated conditions. Genes are divided into quartiles based on transcription and sorted in descending order by transcription.
(C) Genome coverage tracks of H3K4me3 after 2 hours and 48 hours of SET1A/B/MLL1/2-depletion at a representative gene from the 1st Quartile (*Slc39a10*) and the 4th Quartile (*Egfl6*). Spreading of H3K4me3 into the gene body of *Slc39a10* is highlighted. The underlying CpG island is also shown.

3.3.2 SET1A/B and MLL1/2 do not synergise to deposit H3K4me2

I next asked whether SET1A/B and MLL1/2 also co-operate to deposit H3K4me2. In contrast to the effects seen on H3K4me3, removing SET1A/B in addition to MLL1/2 did not result in reductions in H3K4me2 that were significantly larger than those seen after removing MLL1/2 on their own. Interestingly, the effects of removing SET1A/B/MLL1/2 on H3K4me2 appeared to be a combination of the effects seen when removing SET1A/B or MLL1/2 on their own. Broadly speaking, levels of H3K4me2 were reduced upon SET1A/B/MLL1/2-depletion at enhancers and at the promoters of lowly-transcribed genes (Fig. 3.16A, B). This is similar to the effects seen on H3K4me2 after MLL1/2-depletion. At the promoters of highly-transcribed genes, H3K4me2 increased around the transcription start site at the +1 and +2 nucleosomes after 2 hours of dTAG treatment, similar to the effect seen after SET1A/B-depletion (Fig. 3.16B, C). These findings indicate that SET1A/B and MLL1/2 do not synergise to deposit H3K4me2 and that the effects of removing SET1A/B and MLL1/2 on H3K4me2 are distinct.

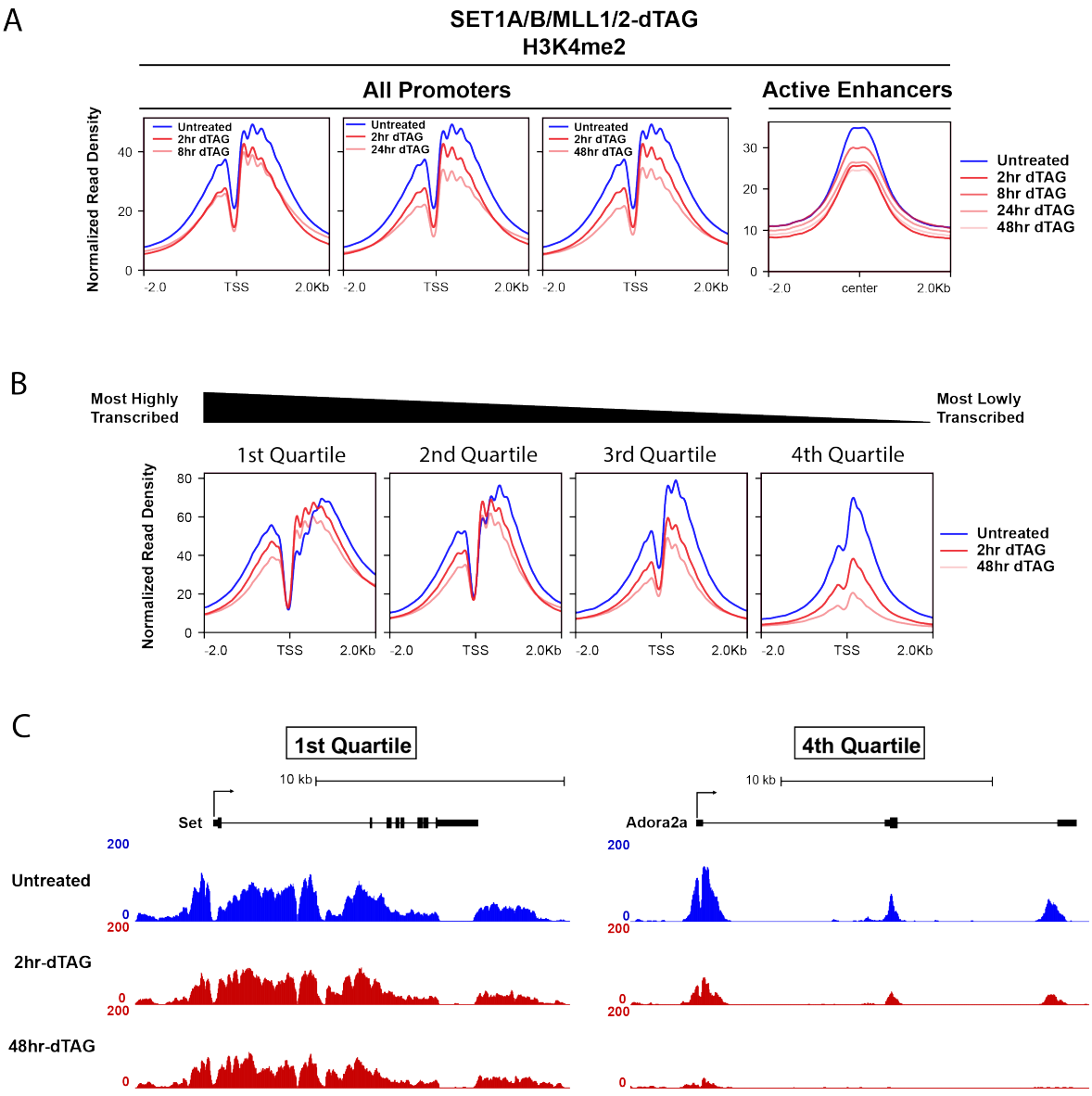

Figure 3.16: Removing SET1A/B/MLL1/2 leads to large reductions in H3K4me2

(A) Metaplots showing H3K4me2 cChIP-seq signal across all promoters (n=20,633) and active enhancers (n= 12,006) after 2hr-depletion of SET1A/B/MLL1/2.
(B) Heatmaps showing log2 fold change of H3K4me2 cChIP-seq signal between 2hr-dTAG-treated and untreated cells across all H3K4me2-associated promoters (n=15,379) divided into quartiles based on transcription. Heatmaps are ranked in descending order by transcription levels in untreated cells.
(C) Genome coverage tracks of H3K4me2 at a representative 1st Quartile gene (Set) and a representative 4th Quartile gene (Adora2a) after SET1A/B/MLL1/2-depletion.

A comparative analysis of H3K4me2 reductions at promoters after depletion of SET1A/B, MLL1/2, and SET1A/B/MLL1/2 demonstrate that simultaneous depletion of SET1A/B/MLL1/2 does not lead to further reductions comapred to removal of MLL1/2 on their own (Fig. 3.17A). Indeed, analysis of H3K4me2 levels across promoters divided into quartiles based on transcription indicate that promoters where H3K4me2 increase upon SET1A/B-depletion largely exhibit the same increase with the additional removal of MLL1/2. Similarly, promoters where H3K4me2 decrease upon MLL1/2-depletion are also exhibit reductions of a similar magnitude with the additional removal of SET1A/B (Fig. 3.17A). This effect was also seen clearly at enhancers, where reductions in H3K4me2 after MLL1/2 remain largely the same upon the additional removal

of SET1A/B (Fig. 3.17B). Taken together, these findings indicate that SET1A/B and MLL1/2 have largely distinct, non-overlapping roles in shaping H3K4me2 deposition.

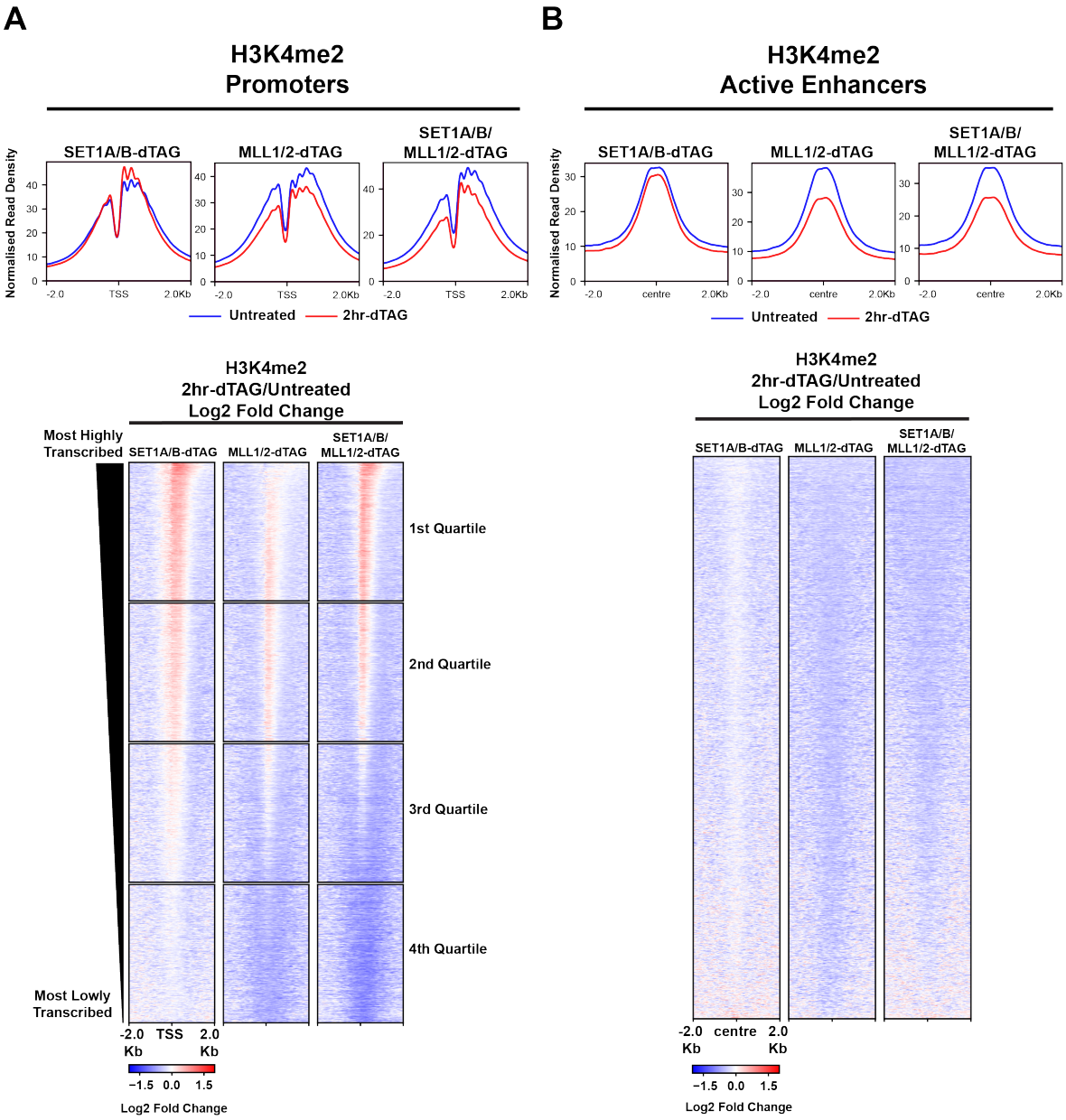

Figure 3.17: SET1A/B and MLL1/2 do not co-operate to deposit H3K4me2.

(A) Top - Metaplots showing H3K4me2 cChIP-seq signal across all promoters (n=20,633) after 2hr-depletion of SET1A/B/MLL1/2. Bottom - Heatmaps showing log2 fold change of H3K4me2 cChIP-seq signal between 2hr-dTAG-treated and untreated cells across all H3K4me2-associated promoters divided into quartiles based on transcription. Heatmaps are ranked in descending order by transcription levels in untreated cells.
(B) As in (A) but for all active enhancers (n=12,006).

3.3.3 H3K4me1 broadly increases upon depletion of SET1A/B/MLL1/2

Surprisingly, removing the four H3K4 tri-methyltransferases simultaneously did not remove all H3K4me3. This suggests that another H3K4 methyltransferase may compensate after their removal to maintain levels of H3K4me3 at the promoters of highly-transcribed genes. Additionally, prolonged removal of SET1A/B/MLL1/2 led to H3K4me3 spreading out of the CGI-promoter and into the gene body. This suggests that this compensating H3K4 methyltransferase does not bind specifically to CGIs. Additionally, highly-transcribed genes retain more H3K4me3 and experience more pronounced recovery of H3K4me3 upon prolonged SET1A/B/MLL1/2-depletion, suggesting that this methyltransferase has a preference for methylating H3K4 at highly-transcribed genes.

The other major H3K4 methyltransferases in mammals are MLL3 and MLL4, also members of the SET1/MLL family. MLL3/4 primarily bind enhancers and are thought to be major H3K4 mono- methyltransferases. Although H3K4me1 has traditionally been associated with enhancers, H3K4me1 is a very pervasive modification and is also deposited along the gene bodies of actively transcribed genes. Therefore, it was conceivable that MLL3/4 could be the compensating methyltransferase. To examine this hypothesis, I analysed changes in H3K4me1 after SET1A/B/MLL1/2-depletion by cChIP-seq. Indeed, upon SET1A/B/MLL1/2-depletion, H3K4me1 levels increased broadly throughout the genome around transcription start sites, across gene bodies, and at enhancers (Fig. 3.17A-C). H3K4me1 levels increased rapidly around transcription start sites, and was evident after 2 hours of dTAG treatment. This increase was maintained for the remainder of the dTAG treatment time-course, suggesting that a methyltransferase actively maintains this enrichment of H3K4me1 around promoters in the absence of SET1A/B/MLL1/2. The increase in H3K4me1 at enhancers was also rapid and sustained; as is evident after 2 hours of dTAG treatment and maintained after 48 hours of dTAG treatment. Interestingly, the increase in H3K4me1 across gene bodies was slower, with H3K4me1 levels progressively accumulating across gene bodies throughout the dTAG treatment time-course. This pattern of increase in H3K4me1 after SET1A/B/MLL1/2-depletion was similar to that observed after SET1A/B-depletion, albeit in much larger magnitude. This strongly suggests that the removal of SET1A/B/MLL1/2 leads to an elevated H3K4-monomethyltransferase activity within the cell.

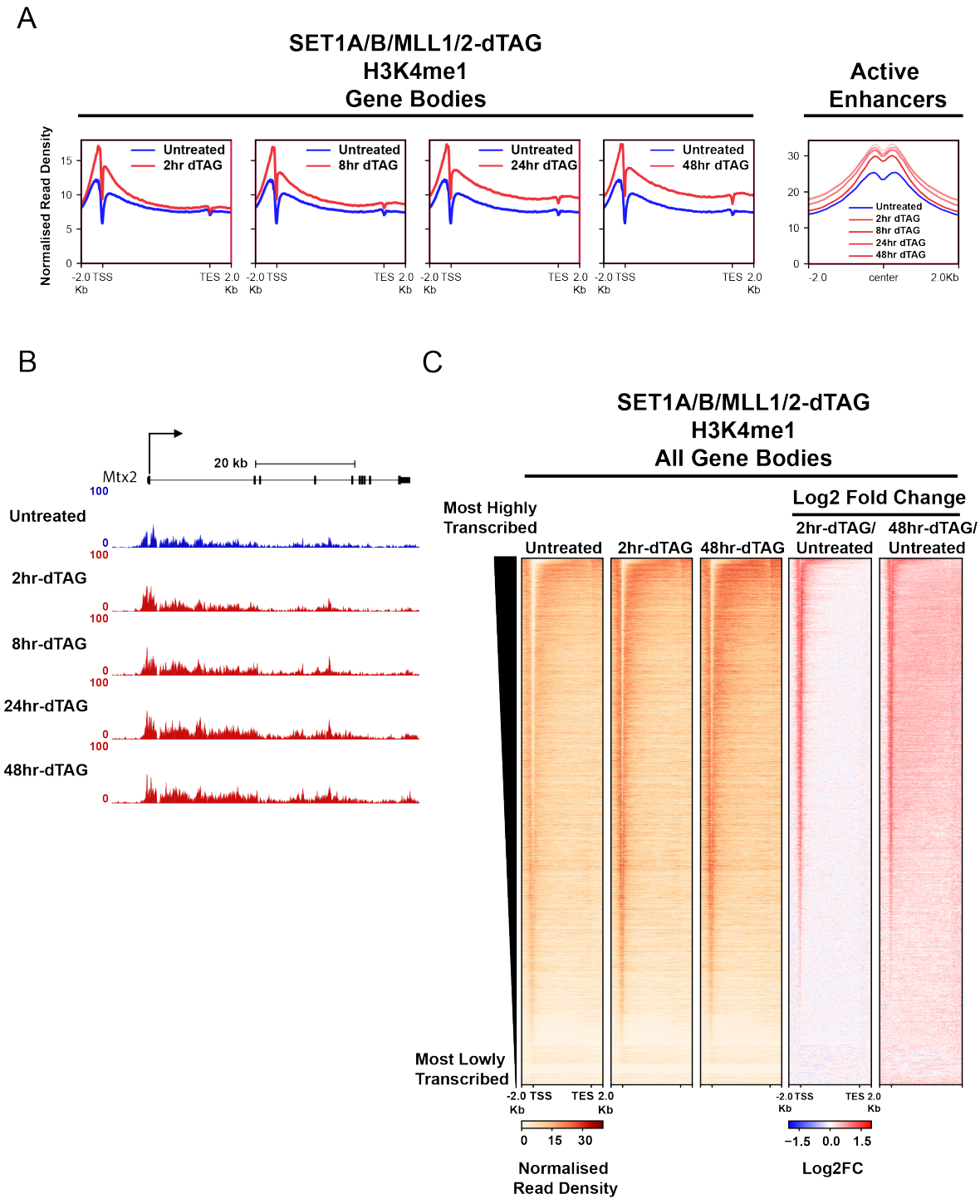

Figure 3.18: H3K4me1 increases broadly after SET1A/B/MLL1/2-depletion.

(A) Metaplots showing H3K4me1 cChIP-seq signal over all gene bodies (n=20,633) and all active enhancers (n=12,006).
(B) Genome coverage tracks of H3K4me1 at a representative gene (*Mtx2*) after SET1A/B/MLL1/2-depletion.
(C) Heatmaps illustrating H3K4me1 cChIP-seq signal over all gene bodies in untreated, 2hr-dTAG-treated, and 48hr-dTAG-treated cells. Also shown are heatmaps of log2 fold changes in ChIP-seq signal between dTAG-treated and untreated conditions. All heatmaps are sorted in decreasing order by transcription.

3.4 Summary and Discussion

In this chapter, I sought to understand how the CGI-associated SET1A/B and MLL1/2 complex shape H3K4 methylation. I have first demonstrated that, in accordance with previous work, removal of SET1A/B does not result in major reductions in H3K4me3 genome-wide. I find that very few genes are entirely dependent on SET1A/B for H3K4me3 deposition, and H3K4me3 levels at most of those genes recover after prolonged SET1A/B removal. Additionally, my characterization of H3K4me1/2 levels after SET1A/B depletion also demonstrates that SET1A/B do not contribute to the deposition of H3K4me1/2 genome wide. This suggests that the primary role of SET1A/B is the conversion of existing H3K4me2 to H3K4me3. Additionally, I find that in the absence of SET1A/B, H3K4me1/2 levels increase in a manner which suggests that other methyltransferases contribute to H3K4 methyltransfease activity in the absence of SET1A/B. Taken together, these findings show that SET1A/B have a broad role in depositing H3K4me3 at the promoters of highly-transcribed genes but have minor contributions toward H3K4me1/2.

On the other hand, I find that MLL1/2 plays a central role in H3K4me2/3. In addition to the well-characterized role of MLL1/2 in depositing H3K4me3 at lowly-transcribed genes, I find that MLL1/2 also deposits H3K4me3 at highly-transcribed genes, although other methyltransferases compensate for loss of MLL1/2 at these genes over time. Additionally, I find that MLL1/2 contribue centrally to H3K4me2 deposition across the genome, in particular at lowly-transcribed genes but also at enhancers. I find that promoters of lowly-transcribed genes which depend entirely on MLL1/2 for H3K4me3 levels also depend on MLL1/2 for H3K4me2, but do not depend on MLL1/2 for H3K4me1. This demonstrates that at MLL1/2 target genes, the primary role of MLL1/2 is to convert existing H3K4me1 to H3K4me2/3.

Most importantly, I find that SET1A/B and MLL1/2 synergise to shape H3K4me3 deposition at promoters of moderately- and highly-transcribed genes. I find that the individual contributions of SET1A/B/ and MLL1/2 to these genes are moderate, but their simultaneous removal results in a far greater reduction in H3K4me3. This indicates that both SET1A/B and MLL1/2 function at the promoters of highly transcribed genes to deposit H3K4me3, and can compensate for the removal of each other at these promoters. This finding resolves an enigma in the field where depletion of SET1A/B or MLL1/2 alone do not greatly impact levels of H3K4me3 at these promoters despite high occupancy of SET1A/B and MLL1/2. I find that H3K4me3 levels at virtually all CpG-island associated promoters are significantly reduced after SET1A/B and MLL1/2, indicating that these CpG-island associated methyltransferases collectively shape H3K4me3 deposition at CpG islands. However, I find that this cooperation between SET1A/B and MLL1/2 is limited to H3K4me3 deposition. Removal of SET1A/B and MLL1/2 simultaneously does not result in further reductions in H3K4me2 compared to removing just MLL1/2. Therefore, a major role of SET1A/B and MLL1/2 is to cooperatively convert H3K4me2 to H3K4me3 at highly transcribed genes.

An outstanding and intriguing question would be whether SET1A/B and MLL1/2 also cooperate to regulate transcription at these genes. Acute depletion of SET1A/B have been shown to result in the premature termination of moderately-transcribed and lowly-transcribed genes, but not highly-transcribed genes where occupancy

of SET1A is at its highest (Hughes et al., 2023). On the other hand, constitutive knockout of MLL1/2 typically results in modest reductions in transcription and gene expression in mESCs, and these reductions are typically limited to genes which are already lowly-transcribed. Whether SET1A/B and MLL1/2 regulate transcription at highly-transcribed genes through a shared mechanism remains unknown and would be the focus of future studies.

Surprisingly, I find that depleting all CXXC-domain containing H3K4 methyltransferases does not remove all H3K4me3 at CpG islands, although I find that in their absence, H3K4me3 becomes less correlated with CpG islands and more correlated with active transcription. Crucially, in the absence of SET1A/B and MLL1/2, H3K4me3 enrichment begins to extend beyond the CpG island boundary and into the gene body in a manner that is correlated with active transcription. Indeed, in the absence of SET1A/B/MLL1/2, the distribution of H3K4me3 becomes reminiscent of that of H3K4me1/2, which are both enriched over gene bodies. Additionally, H3K4me1 levels become greatly increased after SET1A/B/MLL1/2-depletion. This strongly suggests that an additional methyltransferase that primarily deposits H3K4me1/2 becomes more prevalent in the absence of SET1A/B/MLL1/2 and is able to deposit H3K4me3 in a transcription-coupled manner. The most obvious candidates for this additional methyltransferase are MLL3 and MLL4, which are both H3K4 mono-methyltransferases (J. E. Lee et al., 2013). Given that highly transcribed genes typically have broad H3K4me1 domains over their gene bodies, it is conceivable that MLL3/4 could adopt the role of an H3K4-trimethyltransferase at these genes in the absence of SET1A/B/MLL1/2. Quantitative proteomic studies suggest that well over half of all SET1/MLL complexes contain SET1A/B/MLL1/2 as the methyltransferase subunit (van Nuland et al., 2013). It is possible that upon the depletion of these proteins, a large proportion of shared core subunits are liberated and may lead to elevated MLL3/4 function. In the next chapter of this thesis, I will address the role of MLL3/4 in H3K4 methylation and investigate whether they can account for the H3K4me3 remaining after SET1A/B/MLL1/2 depletion.

Chapter 4

SET1/MLL-complexes collectively implement a large portion of H3K4 methylation

Introduction

In the previous chapter, I found that SET1A/B and MLL1/2 each contribute modestly to H3K4 methylation on their own, but synergise and collaborate to deposit H3K4me3. However, I found that this synergy does not encompass all H3K4me3 throughout the genome; approximately 50% of H3K4me3 is retained after removing SET1A/B/MLL1/2. This is surprising given that SET1A/B/MLL1/2 have traditionally been thought to account for all H3K4me3 in mammals, with their association with CpG islands also accounting for the close correlation between H3K4me3 and CpG islands. I find that in the absence of SET1A/B/MLL1/2, H3K4me3 begins to spread out of the CpG island into the gene body, suggesting that an H3K4 methyltransferase that is not associated with CpG islands may be responsible for depositing H3K4me3 in the absence of SET1A/B/MLL1/2. Additionally, I find that SET1A/B and MLL1/2 do not synergise to deposit H3K4me2, and that H3K4me1 increases broadly throughout the genome upon depletion of SET1A/B/MLL1/2. Taken together, these findings strongly suggest that monomethyltransferases such as MLL3/4 may have a greater role in depositing H3K4me3 than previously understood, and could compensate for SET1A/B/MLL1/2 removal.

MLL3/4 are thought to be the primary H3K4-monomethyltransferases that function at enhancers. ChIP-seq experiments have shown that they bind to distal regulatory regions where H3K4me1 is enriched, and their knockout results in loss of H3K4me1 at these sites (D. Hu, Gao, et al., 2013, J. E. Lee et al., 2013). Additionally, MLL3/4 have been shown to recruit the H3K27 acetyltransferases CBP/P300 and facilitate the loading of Mediator and RNA polymerase II onto enhancers (Lai et al., 2017, C. Wang et al., 2016). However, studies using catalytically-inactive mutants of MLL3/4 have also shown that many of these activites are not dependent on their methyltransferase activities (Dorighi et al., 2017). This suggests that deposition of H3K4me1 and enhancer activation by MLL3/4 may largely be decoupled and represent different cellular activities. Therefore, the immense focus on their functions at enhancers may be overlooking broader roles they may have in methylating H3K4 throughout the genome. Indeed, the SET domain of MLL4 have been shown to be able

to generate all three H3K4 methylation states on nucleosomes in vitro (Kwon et al., 2020). Conversely, although MLL3/4 have a major role in depositing H3K4me1, many studies have shown that they do not account of all H3K4me1. Indeed, analysis of bulk H3K4me1 levels by both immunoblotting and mass spectrometry from mES cells in which MLL3/4 are knocked out or cells in which their catalytic activity is abolished have consistently shown a 50% reduction (Dorighi et al., 2017, Boileau et al., 2023, J. E. Lee et al., 2013). This suggests that the other H3K4 methyltransferases deposit the remaining H3K4me1. Indeed, all SET1/MLL methyltransferases can generate H3K4me1 *in vitro*, although my experiments indicate that simultaneous depletion of SET1A/B/MLL1/2 does not lead to reductions in H3K4me1. These observations indicate that there is currently a lack of understanding of how MLL3/4 may cooperate with other SET/MLL complexes to shape H3K4 methylation throughout the genome.

How SET1/MLL complexes function as a system to shape H3K4 methylation throughout the genome is unclear. My previous findings have shown that the CGI-associated SET1/MLL complexes function together to shape H3K4me3 deposition at CGIs. Whether all SET1/MLL complexes have a shared role in H3K4 methylation is currently not well understood. Studies have thus far mostly focused on studying SET1/MLL complexes individually. However, my investigations using rapid perturbation of SET1A/B and MLL1/2 complexes have shown that they have major roles in shaping H3K4me3 is at shared sites where they synergise. Whether MLL3/4 have more prominent functions in H3K4 methylation at sites shared with SET1A/B and MLL1/2 is unknown. Similarly, it is unknown whether SET1A/B/ and MLL1/2 have greater roles in H3K4 mono- and di-methylation at sites shared with MLL3/4 that are masked by redundancy between these complexes.

In this chapter, I investigate the role of MLL3/4 in H3K4 methylation using degron technology to rapidly deplete MLL3/4 in mESCs. I then investigate whether SET1/MLL complexes as a whole have a greater shared role in H3K4 methylation by using degon technology to rapidly deplete all SET1/MLL proteins in mESCs.

4.1 MLL3/4 primarily deposit H3K4me1 broadly across the genome

4.1.1 A degron system to rapidly deplete MLL3/4

To understand how MLL3/4 contribute to H3K4 methylation, I first sought to develop a degron system to enable rapid degradation of MLL3/4 in mESCs. As with all SET1/MLL methyltransferases, MLL3 and MLL4 are large proteins with the catalytic SET domain required for H3K4 methylation situated at the C-terminus. To avoid perturbing the SET domains of MLL3 and MLL4, I decided to tag the N-termini of MLL3 and MLL4 with the FKBP12^{F36V} domain (Fig. 4.1A). Additionally, I decided to insert three T7 tags in tandem (3xT7) to enable ChIP experiments using antibodies against the T7 epitope should commercially available antibodies against endogenous MLL3 and MLL4 prove to be of insufficient quality for ChIP (Fig. 4.1A). Furthermore, I decided to insert a double-StrepII tag (2xStrepII) should I wish to affinity-purify MLL3 or MLL4 in the future (Fig. 4.1A).

I first generated a cell line in which MLL3 could be rapidly depleted using this degron strategy. I found that, similarly to SET1A/B and MLL1/2, MLL3 could be depleted within 2 hours of dTAG treatment (Fig. 4.1B). However, attempts to generate a cell line in which MLL4 could be rapidly depleted proved unsuccessful. After 2 hours of dTAG treatment, a substantial amount of MLL4 was still detectable by immunoblotting (Fig. 4.1C). Furthermore, increasing the concentration of the dTAG-13 compound ten-fold from 100nM to 1μM and the duration of the dTAG treatment to 24 hours also did not result in complete degradation of MLL4 (Fig. 4.1D). Additionally, MLL4 was also detectable by immunoblotting using an antibody against the T7 tag, indicating that the tagged MLL4 protein was not amenable to targeted degradation (Fig. 4.1E). This was surprising given that this degron strategy was efficient for MLL3, which is highly homologous to MLL4. I hypothesized that there may be unique structural features of MLL4 at the N-terminus which may result in occlusion of the FKBP12^{F36V} tag and render it inaccessible by the dTAG-13 compound. Given that FKBP12^{F36V} was only separated from MLL4 by a 9 amino acid glycine-serine linker, I decided to redesign the tagging strategy for MLL4 and insert eGFP between FKBP12^{F36V} and MLL4 (Fig. 4.1F). I reasoned that inserting a large folded domain between MLL4 and FKBP12^{F36V} would ensure that they would be well-separated. Indeed, applying this new tagging strategy resulted in rapid degradation of MLL4 within 2 hours of dTAG treatment (Fig. 4.1G).

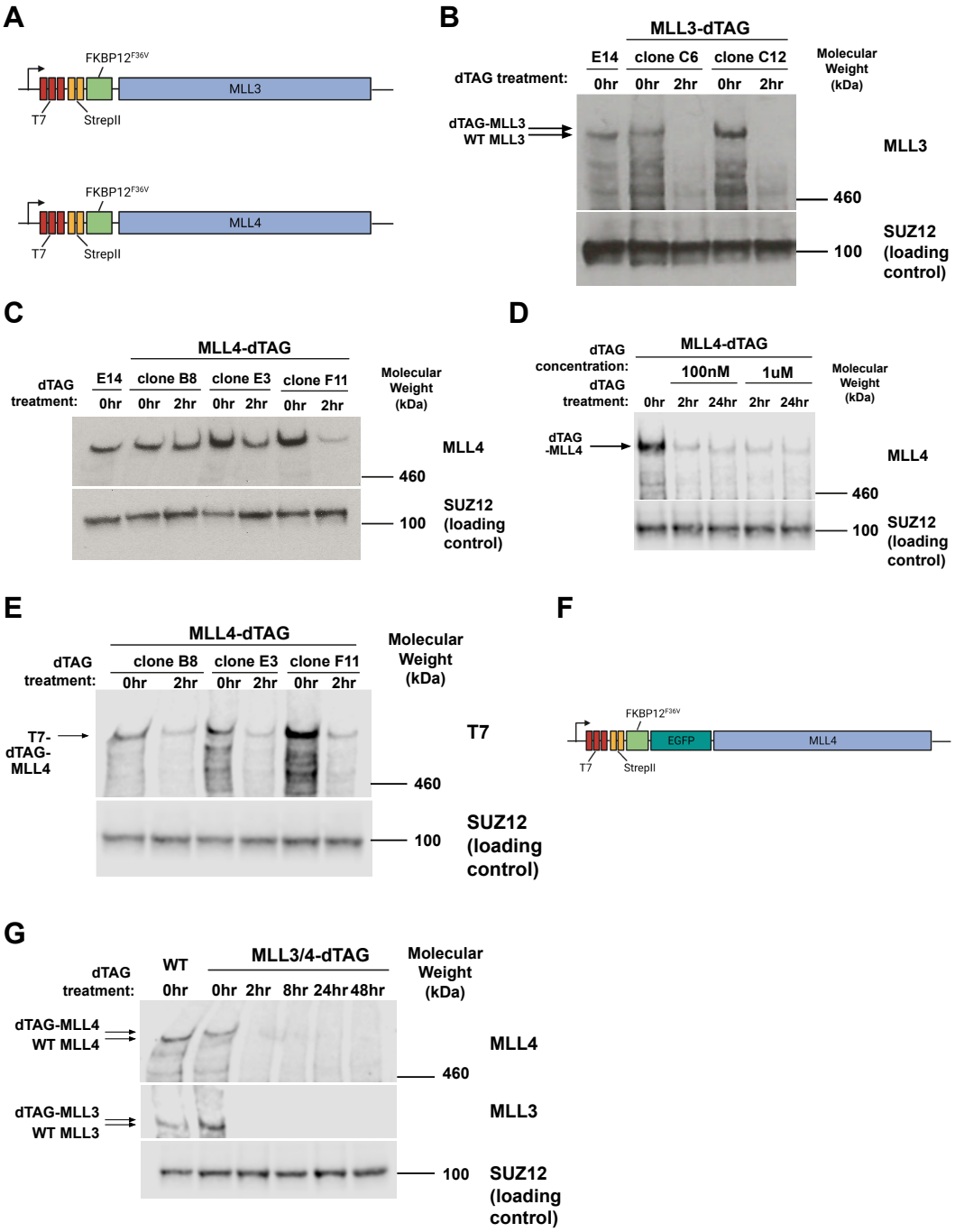

Figure 4.1: Generating an MLL3/4-dTAG cell line.

(A) Schematic illustrating tagging strategy for generating MLL3-dTAG and MLL4-dTAG proteins.
(B) Western blot demonstrating degradation of MLL3 after 2 hours of dTAG treatment in 2 MLL3-dTAG clones. SUZ12 is used as a loading control.
(C) Western blot demonstrating incomplete degradation of MLL4 after 2 hours of dTAG treatment in 3 MLL4-dTAG clones. SUZ12 is used as a loading control.
(D) Western blot demonstrating incomplete degradation of MLL4 after 2 hours and 24 hours of dTAG treatment in one MLL4-dTAG clone with an additional 10-fold increase in dTAG concentration. SUZ12 is used as a loading control.
(E) Western blot demonstrating incomplete degradation of MLL4 after 2 hours of dTAG treatment in 3 MLL4-dTAG clones using an antibody against the T7 epitope. SUZ12 is used as a loading control.
(F) Schematic illustrating MLL4-EGFP-dTAG tagging strategy.
(G) Western blot demonstrating robust degradation of MLL3 and MLL4 after 2 hours of dTAG treatment in the MLL3/4-dTAG cell line. SUZ12 is used as a loading control.

4.1.2 MLL3/4 deposits H3K4me1 broadly across the genome

I first sought to understand the contributions of MLL3/4 to global levels of H3K4me1 by immunoblotting. Bulk levels of H3K4me1 declined gradually and steadily after MLL3/4-depletion to approximately 60% after 48 hours of dTAG treatment (Fig. 4.2A). This confirms that MLL3/4 are indeed required for a large proportion of H3K4me1 in mESCs. Additionally, this demonstrates that H3K4me1 turnover is considerably slower than that of H3K4me2/3, which experience major reductions after 2 hours of SET1A/B/MLL1/2-depletion.

I next sought to comprehensively characterize the contributions of MLL3/4 to H3K4me1 deposition throughout the genome. cChIP-seq analysis showed that H3K4me1 levels were substantially reduced at both gene bodies and enhancers after MLL3/4-depletion (Fig. 4.2B, C). Reductions in H3K4me1 at gene bodies and enhancers followed a similar pattern to reductions in bulk H3K4me1 - a gradual and steady decline from 2 hours of dTAG treatment onwards (Fig. 4.2B, C, D). Interestingly, cChIP-seq analysis indicated that H3K4me1 levels largely remained steady after 24 hours of MLL3/4-depletion, with no further reductions seen after 48 hours of dTAG treatment (Fig. 4.2B, C). This is surprising given that these cells would have gone through multiple rounds of cell division in the absence of MLL3/4, and therefore, if MLL3/4 were responsible for all H3K4me1 you would expect it to be diluted by each round of replication. This strongly suggests that the H3K4me1 remaining after 24 hours of MLL3/4 depletion is being deposited by another methyltransferase, and that MLL3/4 only account for a portion of H3K4me1. Interestingly, I found that reductions in H3K4me1 at gene bodies appeared uniform across the genome, with highly-transcribed and lowly-transcribed genes equally affected (Fig. 4.3E). Taken together, these findings indicate that MLL3/4 have a much wider role in H3K4-monomethylation beyond their functions at enhancers and deposit H3K4me1 broadly across the genome.

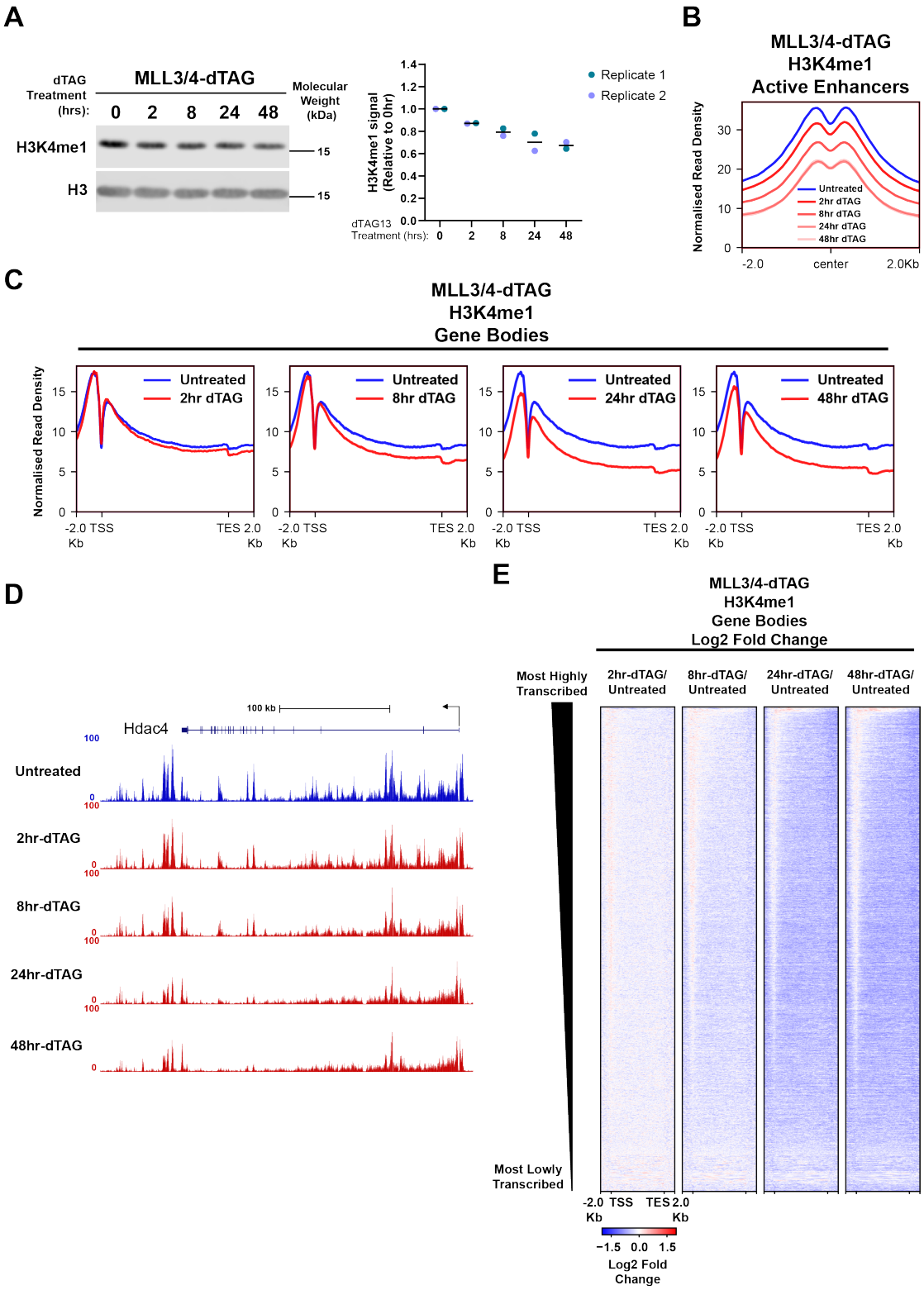

Figure 4.2: MLL3/4 deposit H3K4me1 broadly across the genome.

(A) Analysis of bulk H3K4me1 levels after MLL3/4-depletion by immunoblotting. Histone H3 is used as a loading control. Quantification of bulk H3K4me1 levels is shown on the right. H3K4me1 signal is normalised to total H3 signal and shown as relative to the untreated condition. Individual values from 2 biological replicates are plotted with a line showing the mean of the two values.
(B) Metaplots showing H3K4me1 cChIP-seq signal at active enhancers (n=12,006) after MLL3/4-depletion.
(C) Metaplots showing H3K4me1 cChIP-seq signal at all gene bodies (n=20,633) after MLL3/4-depletion.
(D) Genome coverage tracks showing H3K4me1 after MLL3/4-depletion at a representative gene (*Hdac4*).
(E) Heatmaps showing the log2 fold change in H3K4me1 cChIP-seq signal at all gene bodies between dTAG treated and untreated MLL3/4-dTAG cells. Heatmaps are sorted in descending order by transcription.

4.1.3 MLL3/4 occupancy does not reflect all catalytic activity

Previous studies using knockout or conditional deletion approaches have extensively characterized the role of MLL3/4 in depositing H3K4me1 at enhancers. This is substantiated by ChIP-seq analyses which show that MLL3/4 primarily bind at distal enhancers and are largely absent from active promoters (Jozwik et al., 2016, C. Wang et al., 2016, Dorighi et al., 2017). However, these studies have also shown that MLL3/4-knockout cells have much lower levels of bulk H3K4me1 compared to wild type cells. This is surprising given that H3K4me1 at enhancers only represent a minority of its total distribution throughout the genome. I therefore decided to revisit our understanding of MLL3/4 occupancy throughout the genome by performing double-crosslinked cChIP-seq for MLL3/4 using an antibody against the T7 tag introduced into endogenous MLL3 and MLL4.

In agreement with previous findings, I found that the average MLL3/4 ChIP-seq read density was higher at enhancers compared to promoters (Fig. 4.3A). Additionally, I found that the average MLL3/4 ChIP-seq read density at intergenic enhancers was slightly higher compared to intragenic enhancers, indicating that MLL3/4 preferentially binds distal regulatory regions (Fig. 4.3A). To investigate genomic features that were enriched at MLL3/4 binding sites, I characterized 6,553 T7-MLL3/4 peaks (Fig. 4.3B). Of these 6,553 T7-MLL3/4 peaks, 1,364 (20.81%) were located within 2kb of an annotated TSS, while 1,977 (30.17%) were located within intragenic regions. In contrast, 3,212 (49.02%) peaks were found at intergenic regions. These 6,554 T7-MLL3/4 peaks were highly enriched in accessible chromatin, H3K27ac, H3K4me1, and H3K4me2 (Fig. 4.3C). In contrast, T7-MLL3/4 peaks contained relatively low levels of H3K4me3 and non-methylated CpG dinucleotides (Fig. 4.3C). Interestingly, I noted that although T7-MLL3/4 peaks tended to be flanked by H3K4me1-enriched nucleosomes, they tended not to overlap each other (Fig. 4.3C, D). Instead, T7-MLL3/4 peaks tended to overlap with ATAC-seq peaks. This suggests that MLL3/4 peaks from cross-linked ChIP-seq experiments may represent sites at which MLL3/4 are stably bound to nucleosome-depleted regions rather than sites at which they are methylating H3K4, where their association with chromatin may be transient and therefore difficult to capture through cross-linking. This is similar to previous work by the Klose lab which demonstrated that although RING1B deposits all H2AK119ub1 throughout the genome, RING1B binding is only captured by ChIP-seq at sites where they are stably bound and only represent a small fraction of its association with chromatin (Fursova et al., 2019). Indeed, I noted that there were sites containing MLL3/4-dependent H3K4me1 peaks that did not contain MLL3/4 peaks (Fig. 4.3D). These findings strongly suggest that crosslinked ChIP-seq analysis of MLL3/4 binding does not capture all MLL3/4 catalytic activity throughout the genome.

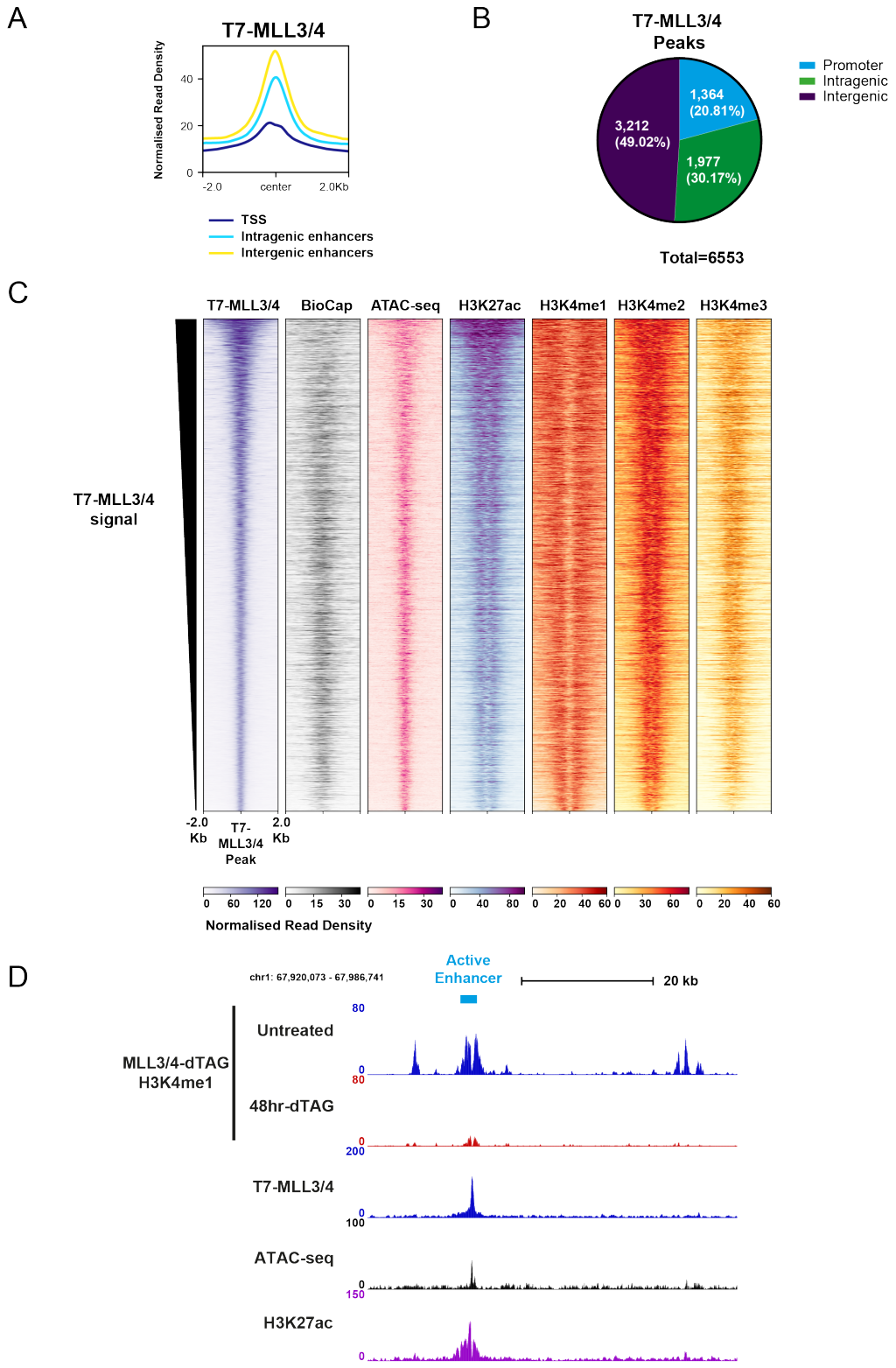

Figure 4.3: MLL3/4 ChIP-seq peaks are primarily located at distal regulatory elements.

(A) Metaplot showing distribution of T7-MLL3/4 ChIP-seq signal at promoters (n=20,634), intragenic enhancers (n=7,850), and intergenic enhancers (n=4,156).
(B) Pie-chart showing the distribution of T7-MLL34 peaks (n=6,553) at promoters, intragenic regions, and intergenic regions.
(C) Heatmaps showing T7-MLL3/4, BioCap-seq, ATAC-seq, H3K27ac, H3K4me1, H3K4me2, and H3K4me3 ChIP-seq signal at the 6,553 T7-MLL3/4 peaks. Heatmaps are sorted in descending order by T7-MLL3/4 ChIP-seq read density.
(D) Genome coverage tracks of H3K4me1 ChIP-seq from untreated and 48hr-dTAG-treated MLL3/4-dTAG cells, T7-MLL3/4 ChIP-seq, ATAC-seq, and H3K27ac ChIP-seq at a representative intergenic region on chromosome 1 containing an active enhancer. ATAC-seq and H3K27ac ChIP-seq data are from Fursova et al., 2021.

4.1.4 MLL3/4 have major contributions to H3K4me2 at enhancers but not at promoters

Having shown that MLL3/4 have a much broader role in depositing H3K4me1 throughout the genome beyond their well-characterized functions at enhancers, I next sought to understand whether they may also contribute to the deposition of H3K4me2/3 throughout the genome. I first examined the role of MLL3/4 in depositing H3K4me2. Analysis of bulk H3K4me2 levels by immunoblotting indicated that H3K4me2 levels were moderately reduced upon MLL3/4-depletion (Fig. 4.4A). To understand if this reduction was widespread across the genome and whether it could account for H3K4me2 at promoters of highly-transcribed genes, I performed cChIP-seq for H3K4me2 after MLL3/4-depletion. Analysis of H3K4me2 levels at promoters indicated that MLL3/4 had minor contributions toward H3K4me2 across promoters, which was evident after 2 hours of MLL3/4-depletion (Fig. 4.4B). Surprisingly, H3K4me2 at promoters began to recover after 8 hours of MLL3/4-depletion, suggesting that another methyltransferase, likely MLL1/2, rapidly compensates for MLL3/4-depletion (Fig. 4.4B) Unsurprisingly, I found that MLL3/4 had the largest effects on H3K4me2 at enhancers, where H3K4me2 reductions were similar to that seen after MLL1/2-depletion (Fig. 4.4B). Interestingly, I found that H3K4me2 levels at enhancers did not recover after prolonged MLL3/4-depletion, suggesting that MLL1/2 and MLL3/4 have non-redundant roles in H3K4me2 deposition at enhancers (Fig. 4.4B). On the other hand, H3K4me2 levels across gene bodies were not reduced significantly after MLL3/4-depletion (Fig. 4.4B). Taken together, these observations indicate that MLL3/4 primarily contribute to H3K4me2 at enhancers, with only minor contributions at promoters and gene bodies.

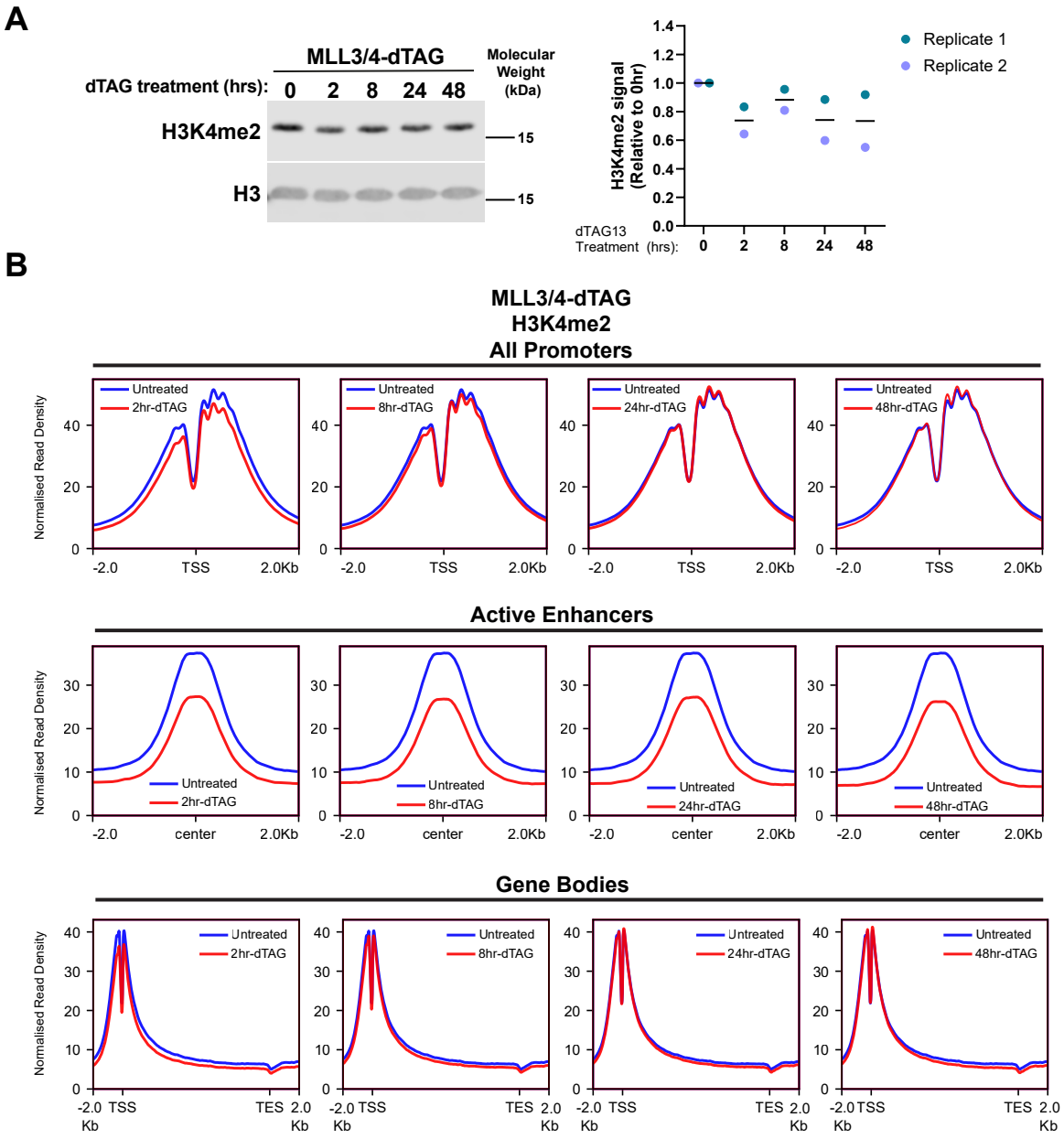

Figure 4.4: MLL3/4 primarily contributes to H3K4me2 at enhancers.

(A) Analysis of bulk H3K4me2 levels after MLL3/4-depletion by immunoblotting. Histone H3 is used as a loading control. Quantification of bulk H3K4me2 levels is shown on the right. H3K4me2 signal is normalised to total H3 levels and shown as relative to the untreated condition. Individual values from 2 biological replicates are plotted with a line showing the mean of the two values.
(B) Metaplots showing H3K4me2 cChIP-seq signal at all promoters (n=20,633), active enhancers (n=12,006), and gene bodies (n=20,633) after MLL3/4-depletion.

4.1.5 MLL3/4 do not have major contributions to H3K4me3

Having found that MLL3/4 can contribute toward H3K4me2, I next wanted to understand if MLL3/4 can also contribute toward H3K4me3. Analysis of bulk H3K4me3 levels by immunoblotting did not reveal any significant changes after MLL3/4-depletion (Fig. 4.5A). Interestingly, cChIP-seq revealed that MLL3/4-depletion resulted in a minor reduction in H3K4me3 across all promoters which did not recover after prolonged MLL3/4-depletion (Fig. 4.5B). Surprisingly, I noted that there were a small number of exceptions, including 116 genes where H3K4me3 was reduced by more than 1.5-fold after 2 hours of MLL3/4-depletion (Fig. 4.5B, C). Interestingly, H3K4me3 at some of these genes are also affected when SET1A/B were depleted, indicating that there are at least some promoters where MLL3/4 deposits H3K4me3, possibly having redundancy with SET1A/B (Fig. 4.5D). It is also possible that SET1A/B requires MLL3/4-deposited H3K4me1 at these promoters for conversion to H3K4me3. Interestingly, analysis of H3K4me3 cChIP-seq signal at active enhancers demonstrated that MLL3/4 have a greater effect on H3K4me3 at these regions, in line with their contributions to H3K4me1/2 at these sites (Fig. 4.5D). However, it is worth noting that H3K4me3 levels at enhancers are typically very low, and the contribution of MLL3/4 towards H3K4me3 only represent a small fraction of all H3K4me3 throughout the genome. Taken together, these observations indicate that although MLL3/4 do contribute H3K4me3, their contributions are minimal with the exception of a small number of genes.

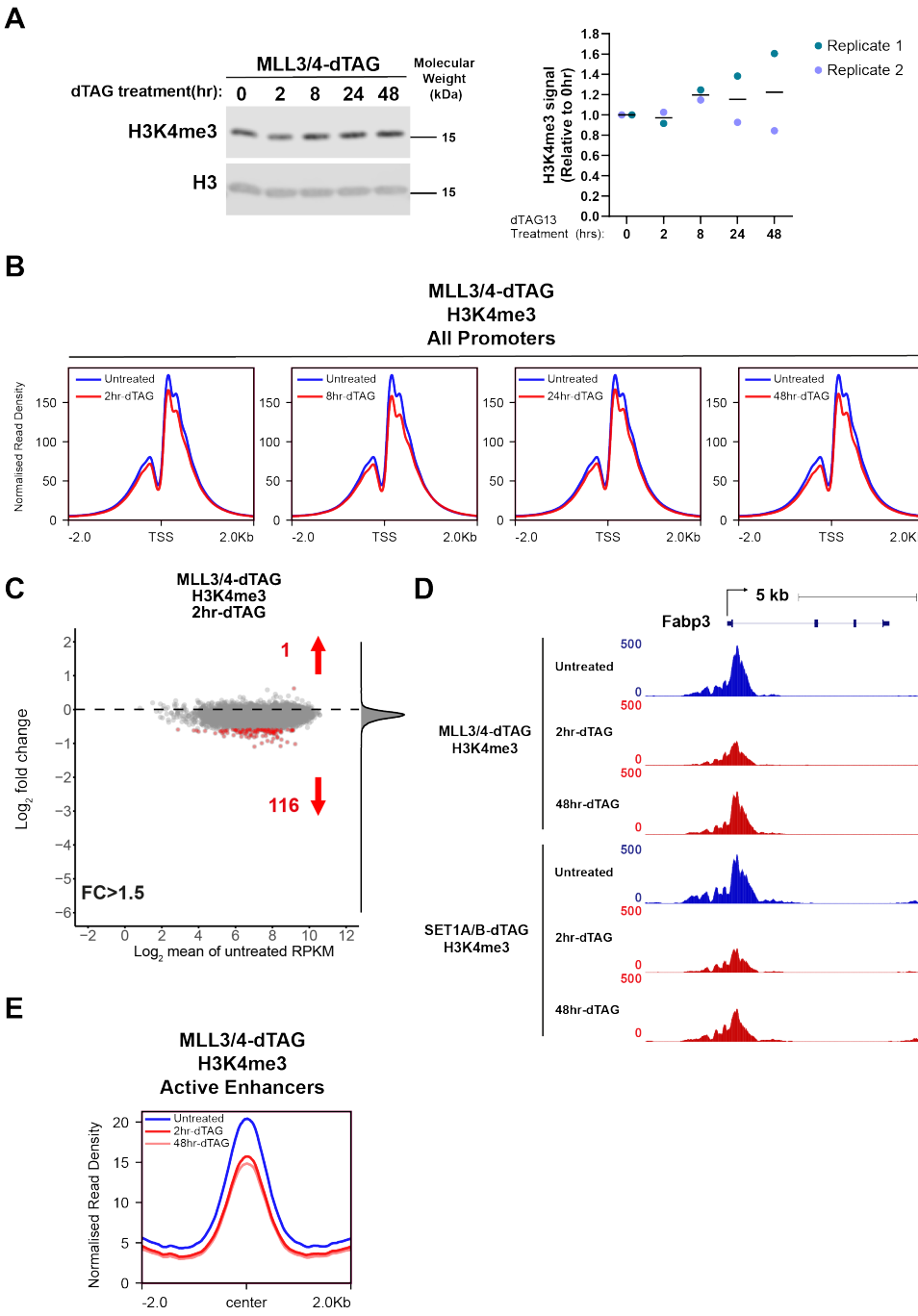

Figure 4.5: MLL3/4 does not have major contributions to H3K4me3.

(A) Analysis of bulk H3K4me3 levels after MLL3/4-depletion by immunoblotting. Histone H3 is used as a loading control. Quantification of bulk H3K4me3 levels is shown on the right. H3K4me3 signal is normalised to total H3 levels and shown as relative to the untreated condition. Individual values from 2 biological replicates are plotted with a line showing the mean of the two values.
(B) Metaplots showing H3K4me3 cChIP-seq signal at all promoters (n=20,633) after MLL3/4-depletion.
(C) MA plot showing the Log_2 Fold Change in H3K4me3 at all H3K4me3-associated promoters after MLL3/4 depletion for 2 hours. The number of promoters with H3K4me3 reduced or increased by more than 1.5-fold are shown in red.
(D) Genome coverage track of H3K4me3 at a representative gene (*Fabp3)* after 2 hours and 48 hours of MLL3/4 depletion and after 2 hours and 48 hours of SET1A/B depletion.
(E) Metaplots showing H3K4me3 cChIP-seq signal at all active enhancers (n=12,006) after 2 hours and 48 hours of MLL3/4-depletion.

4.2 SET1/MLL complexes do not account for all H3K4 methylation

4.2.1 Simultaneous depletion of SET1A/B/MLL1/2/3/4 removes 50% of global H3K4 methylation

I next sought to build on my previous finding that SET1A/B and MLL1/2 synergise to deposit H3K4me3 by investigating whether MLL3/4 also synergise with SET1A/B and MLL1/2 to deposit H3K4 methylation. To do this, I used our previously generated SET1A/B/MLL1/2-dTAG cell line and targeted MLL3 and MLL4 with the tagging strategy described above to generate a cell line in which SET1A/B, MLL1/2, and MLL3/4 could all be depleted within 2 hours of dTAG treatment (Fig. 4.6A). Using this SET1A/B/MLL1/2/3/4-dTAG cell line, I sought to investigate whether removing MLL3/4 in addition to SET1A/B/MLL1/2 would result in loss of all H3K4me3. Surprsingly, analysis of bulk H3K4me3 by immunoblotting revealed that removing all six SET1/MLL methyltransferases did not lead to a loss of all H3K4me3 (Fig. 4.6B). H3K4me3 was reduced to approximately 50% after 2 hours of dTAG treatment (Fig. 4.6B). This reduction in H3K4me3 levels is comparable to that observed after 2 hours of SET1A/B/MLL1/2-depletion, indicating that MLL3/4 do not synergise with SET1A/B/MLL1/2 to deposit H3K4me3. Additionally, I noted that prolonged removal of all SET1/MLL methyltransferases resulted in a dramatic recovery of bulk H3K4me3 levels from 8 hours of dTAG treatment onwards, indicating that the recovery of H3K4me3 levels after prolonged SET1A/B/MLL1/2-depletion could not be attributed to MLL3/4 (Fig. 4.6B). I next analysed bulk H3K4me2 levels after SET1A/B/MLL1/2/3/4-depletion (Fig. 4.6B). Interestingly, I found that H3K4me2 levels decreased after 2 hours of dTAG treatment to 50%, but recovered substantially from 8 hours of dTAG treatment onwards, suggesting that a common mechanism was responsible for the recovery of H3K4me2/3 in the absence of SET1/MLL methyltransferases (Fig. 4.6B). Finally, I sought to understand whether SET1A/B/MLL1/2 could synergise with MLL3/4 to deposit H3K4me1. However, I found that removal of all six SET1/MLL methyltransferases resulted in an approximately 50% reduction in H3K4me1 by 8 hours of dTAG treatment, which was broadly similar to the reductions seen after depletion of MLL3/4 on their own (Fig. 4.6B). Taken together, these findings demonstrate that SET1/MLL complexes account for only approximately 50% of all H3K4 methylation, and that a SET1/MLL-independent methyltransferase deposits the remainder of all H3K4 methylation and is able to partially compensate for SET1/MLL loss in H3K4me2/3 deposition.

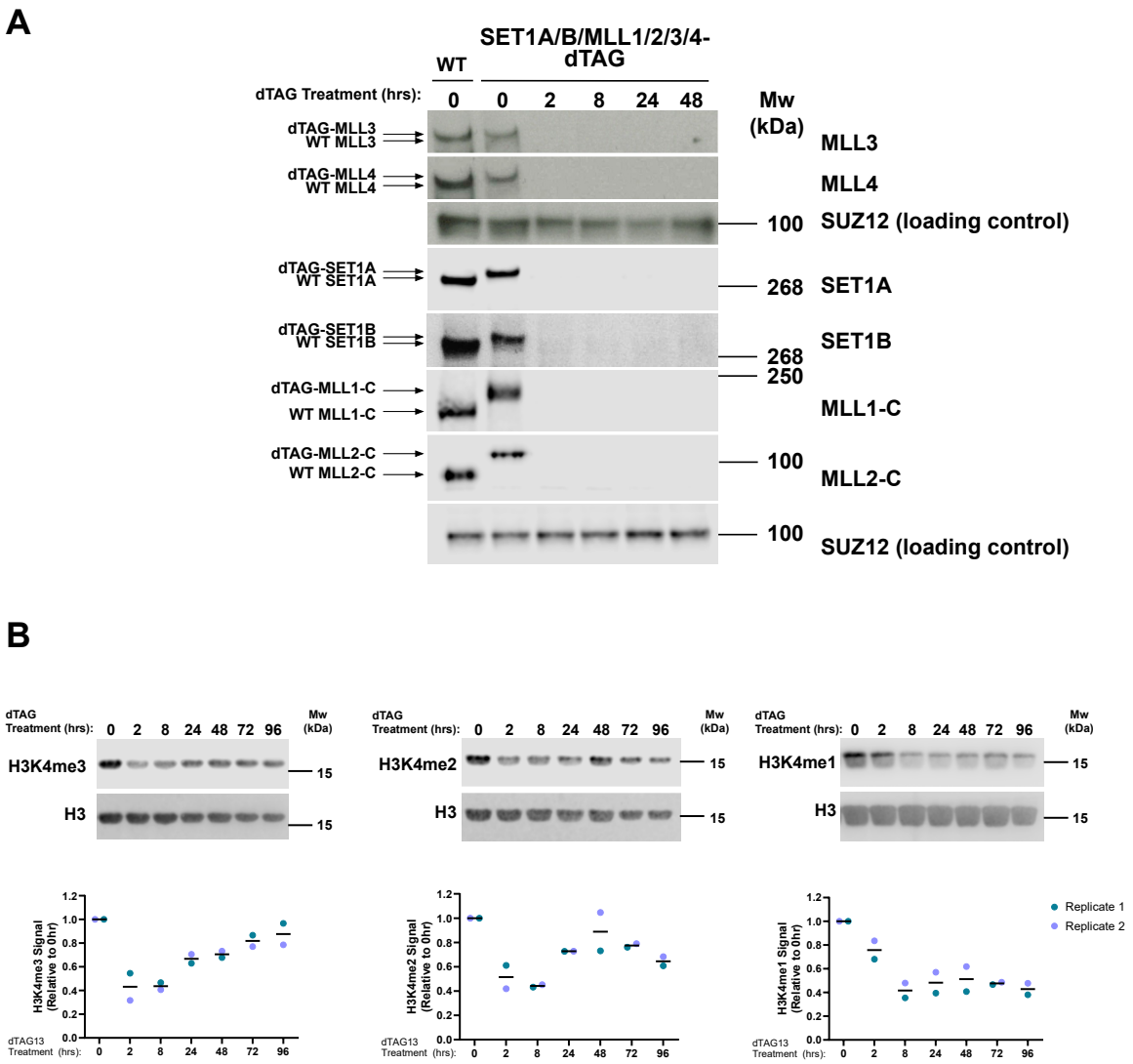

Figure 4.6: SET1A/B/MLL1/2/3/4-depletion only removes 50% of H3K4 methylation.

(A) Western blots showing robust degrdation of SET1/MLL proteins in the SET1A/B/MLL1/2/3/4-dTAG cell line after dTAG treatment. SUZ12 is used as a loading control.
(B) Western blot analysis of bulk H3K4me1/2/3 levels after SET1A/B/MLL1/2/3/4-depletion. Histone H3 is used as a loading control. Quantifications are shown below with H3K4me1/2/3 signal normalised to total H3 signal. Individual values are plotted from 2 biological replicates with a line showing the mean of the two values.

4.2.2 MLL3/4 does not additionally contribute to H3K4me3 at promoters

Bulk western blot analysis suggests that H3K4me3 reductions were similar after SET1A/B/MLL1/2-depletion and SET1A/B/MLL1/2/3/4-depletion. I wanted to examine if there may be more subtle or specific differences in H3K4me3 when MLL3/4 was depleted in addition to SET1A/B/ML1/2. I performed cChIP-seq for H3K4me3 in SET1A/B/MLL1/2/3/4-dTAG cells and found that across all promoters, H3K4me3 were reduced by approximately 50% after 2 hours of dTAG treatment, similar to that seen in the SET1A/B/MLL1/2-dTAG cell line (Fig. 4.7A). I next analysed H3K4me3 across promoters divided into quartiles based on transcription and found that similarly to SET1A/B/MLL1/2-depletion, removing all six methyltransferases resulted in stronger reductions in lowly-transcribed genes compared to highly-transcribed genes (Fig. 4.7B). These findings indicate that the primary effects of removing SET1A/B/MLL1/2/3/4 on H3K4me3 are not significantly different than removing only SET1A/B/MLL1/2, demonstrating that MLL3/4 do not contribute greatly to H3K4me3 deposition.

I next analysed H3K4me3 levels after more prolonged depletion of SET1A/B/MLL1/2/3/4. I found that similarly to removing SET1A/B/MLL1/2, removing SET1A/B/MLL1/2/3/4 led to a recovery in H3K4me3 at the promoter after 8 hours of dTAG treatment, followed by a gradual decrease from 24 hours onwards (Fig. 4.7A). However, I noted that this recovery was more pronounced when MLL3/4 was removed in addition the SET1A/B/MLL1/2, and H3K4me3 levels remained elevated after 48 hours of dTAG treatment compared to after 2 hours of dTAG treatment (Fig. 4.7A).

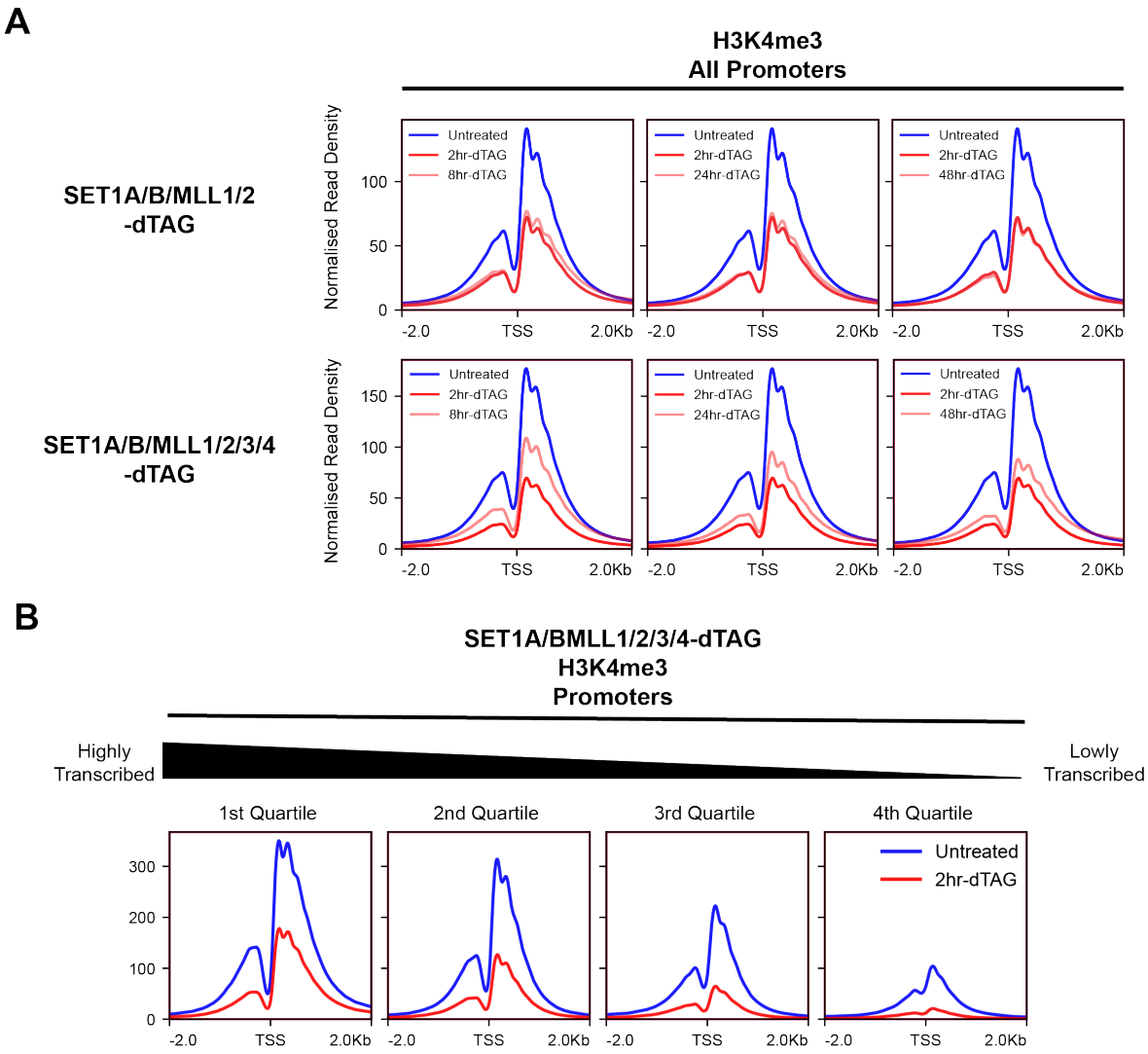

Figure 4.7: Removal of MLL3/4 in addition to SET1A/B/MLL1/2 does not lead to greater reductions in H3K4me3.
(A) Metaplots showing H3K4me3 cChIP-seq at all promoters (n=20,633) after SET1A/B/MLL1/2-depletion (top) and after SET1A/B/MLL1/2/3/4-depletion (bottom).
(B) Metaplots showing H3K4me3 cChIP-seq at all H3K4me3-associated promoters (n=14,724) divided into quartiles based on transcription after SET1/MLL-depletion.

I next sought to characterize the spreading of H3K4me3 into the bodies of highly-transcribed genes as seen after prolonged SET1A/B/MLL1/2-depletion. I found that this effect was not only preserved with the additional removal of MLL3/4, but became more pronounced (Fig. 4.8A, B). Taken together, these findings indicate that the major effect of removing MLL3/4 in addition to removing SET1A/B/MLL1/2 is a more pronounced and widespread recovery of H3K4me3 after prolonged depletion, rather than any greater reductions.

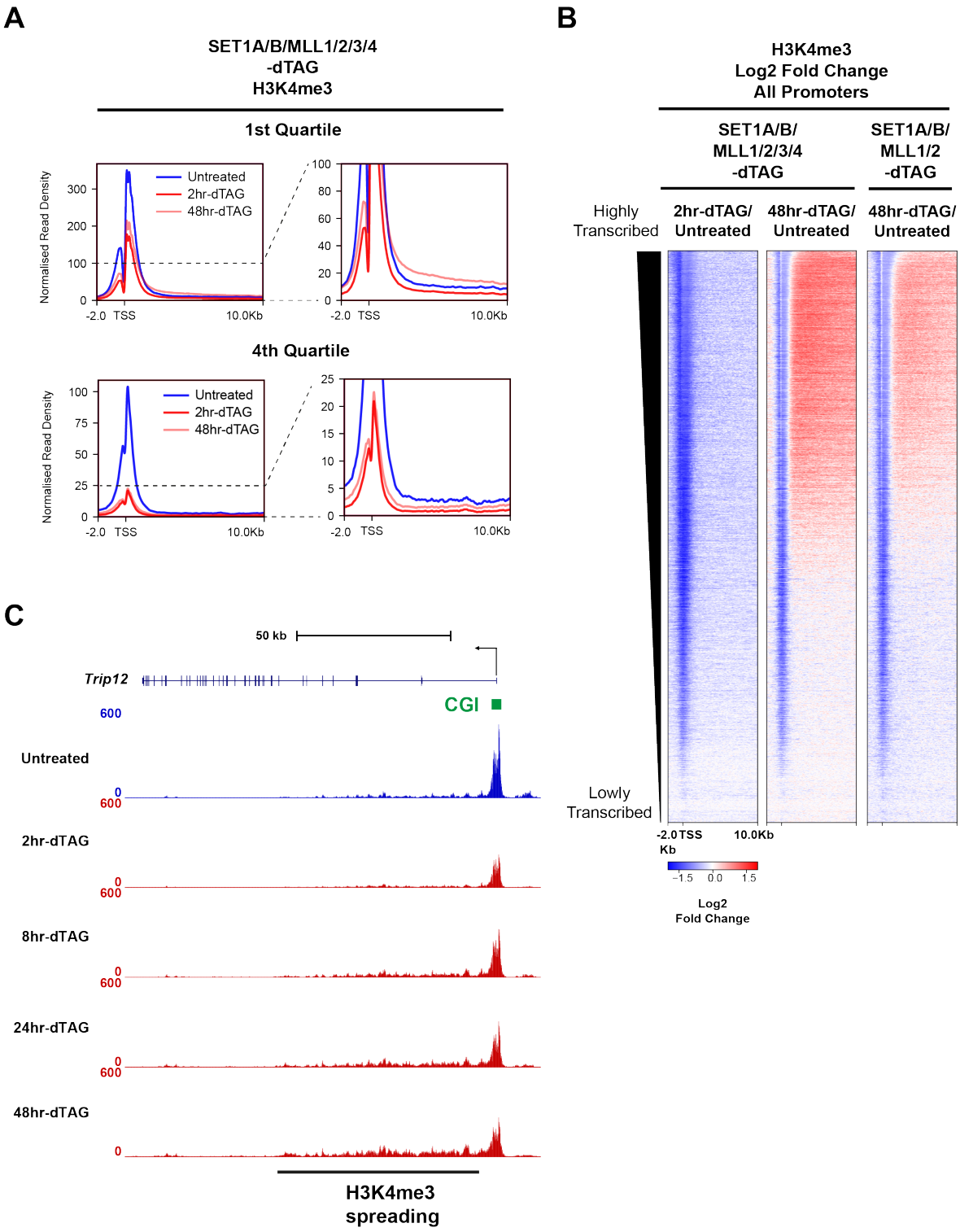

Figure 4.8: H3K4me3 recovers broadly in gene bodies upon extended SET1/MLL-depletion

(A) Metaplots showing H3K4me3 cChIP-seq signal at highly-transcribed 1st Quartile promoters (top) and lowly-transcribed 4th Quartile promoters (bottom) after 2 hours and 48 hours of SET1/MLL-depletion.
(B) Heatmaps showing log2 fold change in H3K4me3 levels between 2hr-dTAG-treated or 48hr-dTAG-treated and untreated SET1A/B/MLL1/2/3/4-dTAG cells at all genes (n=20,633) from the transcription start site extending 10kb into the gene body. Also shown is an equivalent heatmap showing log2 fold change in H3K4me3 levels between 48hr-dTAG-treated and untreated SET1A/B/MLL1/2-dTAG cells. Heatmaps are sorted in descending order by transcription.
(C) Genome coverage tracks for H3K4me3 at a representative gene (*Trip12*) after SET1/MLL-depletion illustrating spreading of H3K4me3 into the gene body after prolonged SET1/MLL-depletion. The CpG island at the promoter as defined by BioCap-seq data in mESCs is shown in green.

4.2.3 SET1A/B, MLL1/2, and MLL3/4 synergise to deposit H3K4me2

In Chapter 3, I found that MLL1/2 are major H3K4-dimethyltransferases reponsible for H3K4me2 at the promoters of lowly transcribed genes and a large proportion of H3K4me2 at enhancers. I found that MLL3/4 do not deposit H3K4me2 at promoters, but are also responsible for a large proportion of H3K4me2 at enhancers. I first sought to understand if MLL3/4 would synergise with SET1A/B and MLL1/2 to deposit H3K4me2 at promoters. Analysis of H3K4me2 levels after 2 hours of SET1A/B/MLL1/2/3/4 depletion indicated that the additional removal MLL3/4 led to slightly greater reductions in H3K4me2 at the promoters, although H3K4me2 at the +1 nucleosome remained high, likely due to demethylation of H3K4me3 (Fig. 4.9A).

I next sought to understand whether the additional removal of MLL3/4 would lead to greater reductions at enhancers. Interestingly, H3K4me2 levels at enhancers were significantly reduced after 2 hours of SET1A/B/MLL1/2/3/4-depletion, with reductions being considerably greater than those observed upon depletion of MLL1/2 or MLL3/4 on their own (Fig. 4.9B). Interestingly, at certain enhancers, H3K4me2 reductions are far greater when all six SET1/MLL methyltransferases are removed than the additive effects of removing SET1A/B, MLL1/2, or MLL3/4 their own, indicating they synergise to deposit H3K4me2 (Fig. 4.9C, D). Taken together, these findings indicate that while MLL3/4 do not contribute to H3K4me2 at promoters, they collectively deposit the vast majority of H3K4me2 at enhancers.

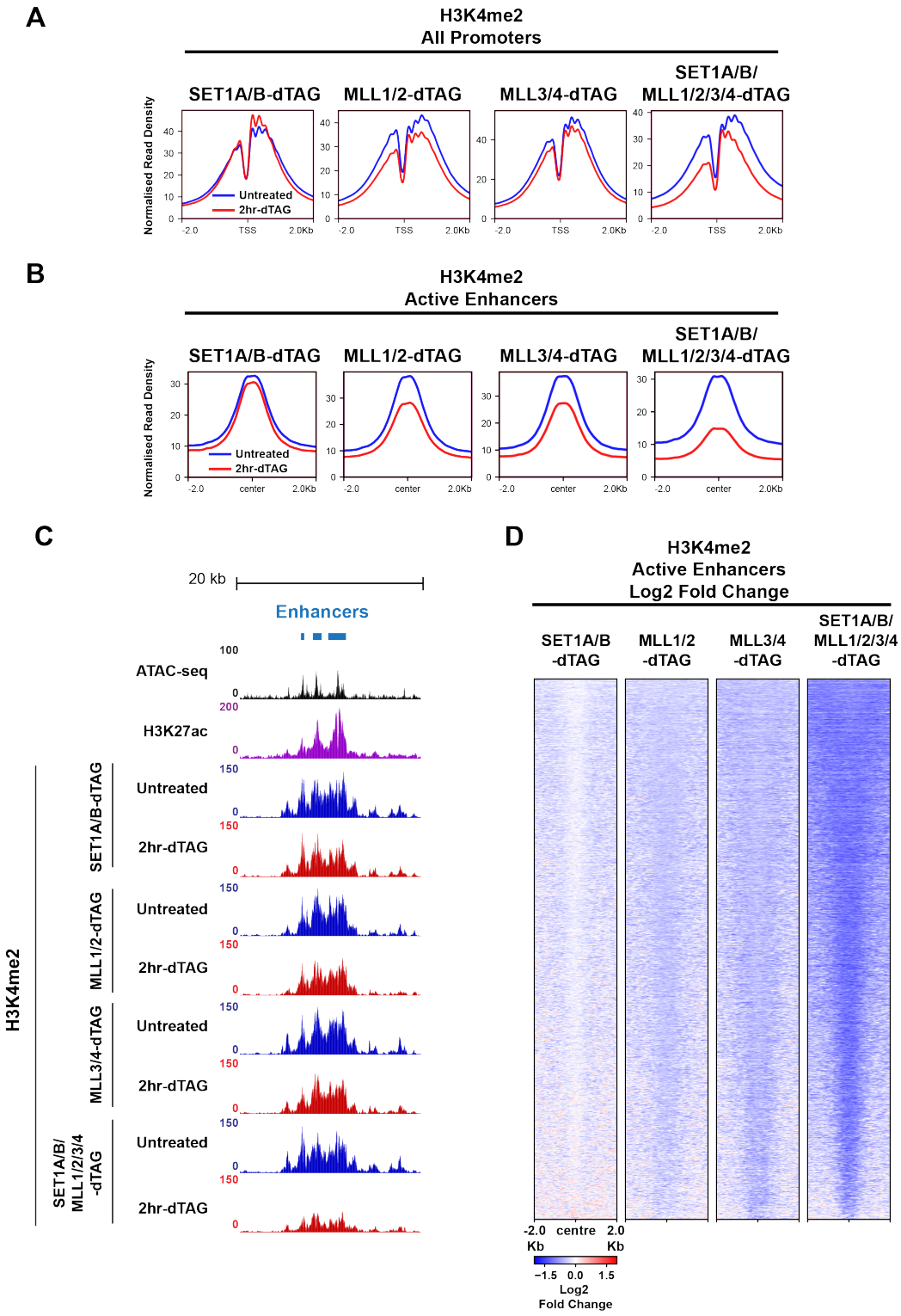

Figure 4.9: SET1A/B/MLL1/2/3/4 synergise to deposit H3K4me2 at enhancers but not promoters.

(A) Metaplots showing H3K4me2 cChIP-seq at all promoters (n=20,633) after 2 hours of dTAG treatment in the indicated cell lines.
(B) Metaplots showing H3K4me2 cChIP-seq at active enhancers (n=12,006) after 2 hours of dTAG treatment in the indicated cell lines.
(C) Genome coverage tracks showing H3K4me2 after 2 hours of dTAG treatment in the indicated cell lines at a representative cluster of enhancers. Genome coverage tracks of ATAC-seq and H3K27ac ChIP-seq data used to annotate enhancer regions are also shown (Fursova et al., 2021).
(D) Heatmaps showing log2 fold change in H3K4me2 cChIP-seq signal at active enhancers between 2 hour dTAG-treated and untreated cells of the indicated cell lines. Heatmaps are sorted by H3K4me2 cChIP signal in untreated SET1A/B-dTAG cells.

4.2.4 Removing MLL3/4 in addition to SET1A/B/MLL1/2 leads to recovery of H3K4me2 in bodies of highly-transcribed genes

Analysis of bulk H3K4me2 levels by immunoblotting in the SET1A/B/MLL1/2/3/4-dTAG cell line indicated that H3K4me2 recovered substantially from 8 hours of dTAG treatment onwards. I wanted to understand where this recovery occurred throughout the genome. Interestingly, analysis of H3K4me2 levels across all promoters demonstrated that H3K4me2 levels at the +1 and +2 nucleosomes did not change significantly throughout the dTAG treatment timecourse (Fig. 4.10A). Interestingly, from 8 hours of dTAG treatment onwards, H3K4me2 beyond the +1 and +2 nucleosomes recovered significantly (Fig. 4.10A). Based on these observations, I extended the analysis of H3K4me2 into the gene body. I found that after 2 hours of SET1A/B/MLL1/2/3/4-depletion, H3K4me2 levels broadly decreased throughout the genome, but started to recover from 8 hours of dTAG treatment onwards (Fig. 4.10A). Interestingly, this recovery was highly correlated with transcription levels of the gene, with highly-transcribed genes exhibiting a more pronounced recovery (Fig. 4.10B). Indeed, for some highly-transcribed genes, H3K4me2 levels in the gene body recovered beyond untreated levels after 8 hours of dTAG treatment (Fig. 4.10C). Additionally, none of the genes in the 4th quartile and only approximately half of the genes in the 3rd quartile exhibited any recovery in H3K4me2 levels in the gene body (Fig. 4.10B, C). Additionally, this recovery behaviour was strictly confined to regions downstream of the transcription start site, strongly suggesting that this increase is coupled to active transcription in the gene (Fig. 410.B). Interestingly, I also noted that this increase in H3K4me2 in the gene body was not merely an increase in enrichment of existing H3K4me2 peaks, but was accompanied by a change in the distribution of H3K4me2. In the gene bodies of most genes, H3K4me2 distribution is characterized by a combination of sharp, punctate peaks and a low level of H3K4me2 that forms a domain over the gene body (Fig. 4.10C). These domains are more prominent at highly-transcribed genes, where they comprise a larger proportion of the total H3K4me2 in the gene body, suggesting that these broad H3K4me2 domains are related to transcription. I noted that upon prolonged depletion of SET1/MLL proteins, the increase in H3K4me2 was most pronounced in these broad H3K4me2 domains (Fig. 4.10C). Taken together, these observations strongly suggest that upon prolonged depletion of SET1/MLL proteins, a novel methyltransferase deposits H3K4me2 in a transcription-dependent manner.

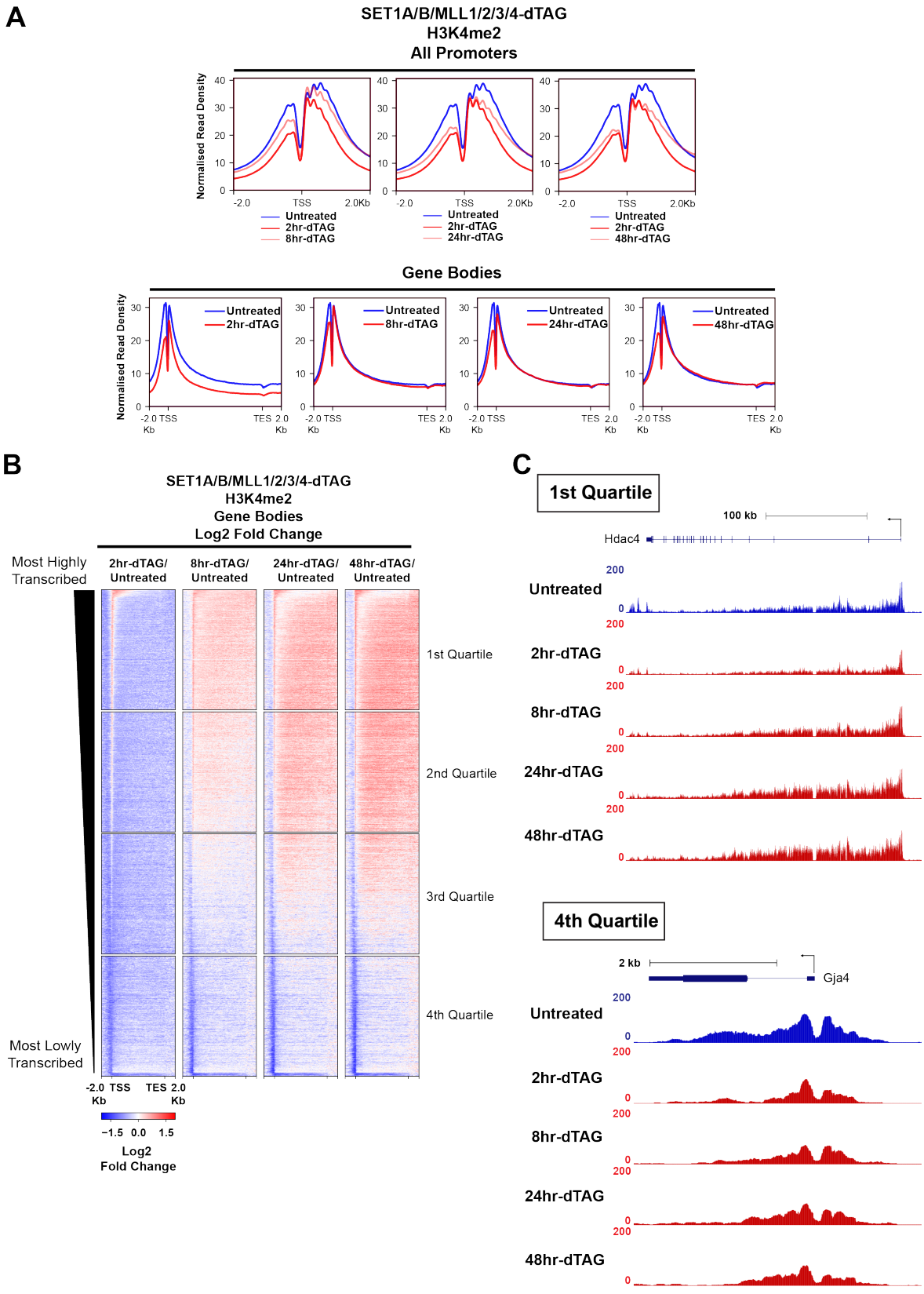

Figure 4.10: H3K4me2 recovers in gene bodies after prolonged SET1/MLL-depletion.

(A) Metaplots showing H3K4me2 cChIP-seq at all promoters (top) and gene bodies (bottom) after SET1A/B/MLL1/2/3/4-depletion (n=20,633).
(B) Heatmaps showing log2 fold change in H3K4me2 cChIP-seq signal at H3K4me2-associated gene bodies (n=15,379) divided into quartiles based on transcription between dTAG-treated and untreated SET1A/B/MLL1/2/3/4-dTAG cells.
(C) Representative genome coverage tracks showing H3K4me2 at a 1st quartile gene (*Hdac4*) and a 4th quartile gene (*Gja4*) after SET1/MLL-depletion.

4.2.5 Removing MLL3/4 in addition to SET1A/B/MLL1/2 leads to recovery of H3K4me1 in bodies of highly-transcribed genes

Given that removal of MLL3/4 did not result in removal of all H3K4me1, I next wanted to understand if there was redundancy among SET1A/B, MLL1/2, and MLL3/4 in the maintenance of H3K4me1. Analysis of H3K4me1 levels by cChIP-seq revealed that H3K4me1 levels around the transcription start site increased upon 2 hours of SET1A/B/MLL1/2/3/4-depletion, likely reflecting the rapid demethylation of H3K4me3 (Fig. 4.11A). In contrast, H3K4me1 levels in the gene body began a gradual decrease from 2 hours of dTAG treatment onwards, similar to the effect seen after MLL3/4-depletion (Fig. 4.11A). Surprisingly, analysis of H3K4me1 levels across all gene bodies indicated that H3K4me1 levels begin to recover after 24 hours of dTAG treatment and continued to recover after 48 hours (Fig. 4.11A). Interestingly, this recovery was not uniform across the genome, but was largely confined to highly-transcribed genes, with H3K4me1 levels in the gene bodies of lowly-transcribed genes continuing to decline after prolonged SET1/MLL-depletion (Fig. 4.11B, C).

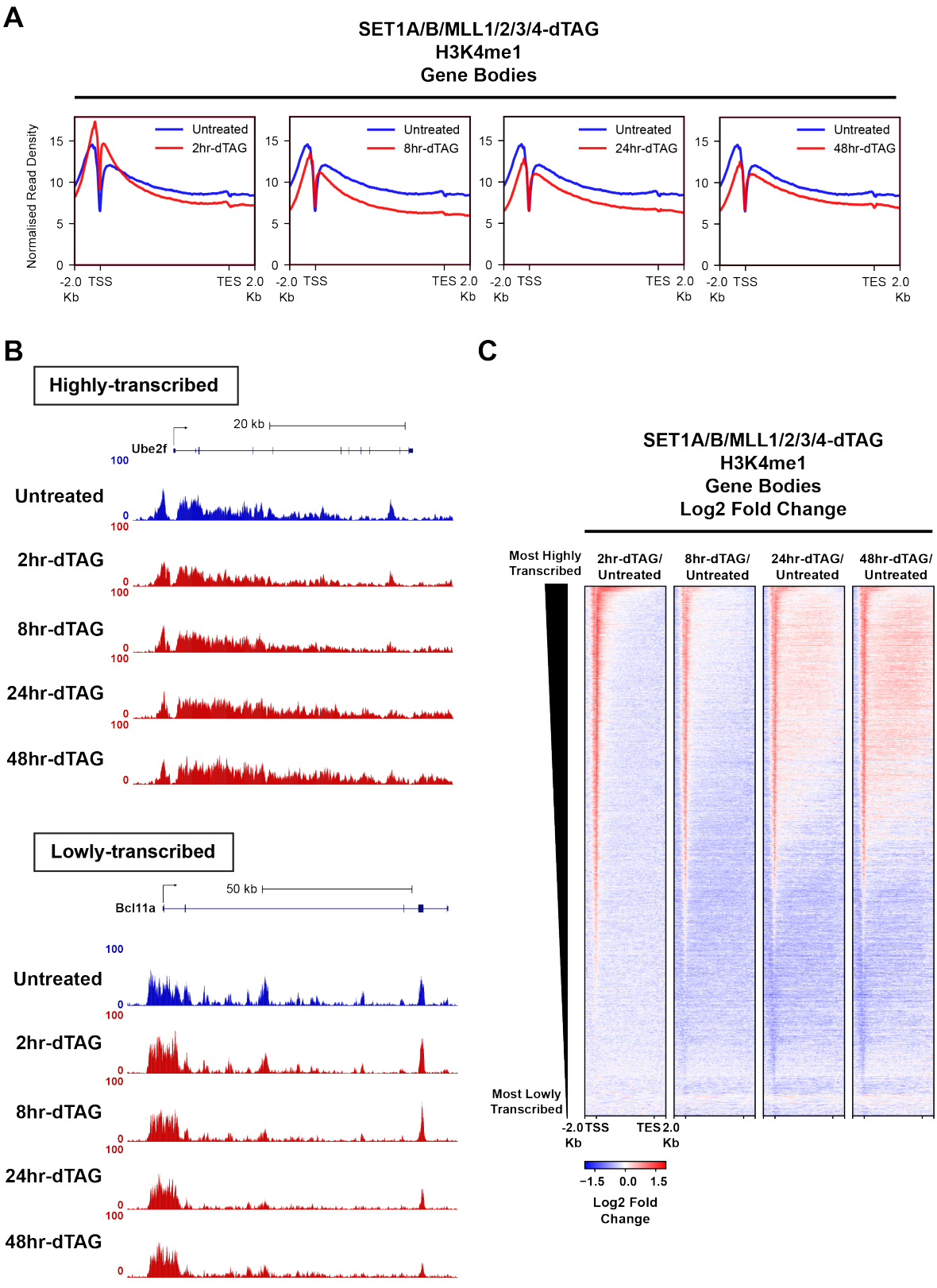

Figure 4.11: H3K4me1 recovers at highly-transcribed genes after SET1/MLL-depletion.

(A) Metaplots showing H3K4me1 cChIP-seq signal across all gene bodies (n=20,633) after SET1A/B/MLL1/2/3/4-depletion.
(B) Genome coverage tracks of H3K4me1 at a representative gene that is highly transcribed (*Ube2f*) and a representative gene that is lowly-transcribed (*Bcl11a*) demonstrating recovery of H3K4me1 after prolonged SET1/MLL-depletion.
(C) Heatmaps showing log2 fold change in H3K4me1 cChIP-seq signal at all gene bodies between dTAG-treated and untreated SET1A/B/MLL1/2/3/4-dTAG cells. Heatmaps are sorted in descending order by transcription.

A closer examination of the recovery of H3K4me1 at highly-transcribed genes after SET1/MLL-depletion revealed that this recovery bore many similarities to the recovery in H3K4me2 after SET1/MLL-depletion. Firstly, both recovery in H3K4me1 and H3K4me2 are largely confined to highly-transcribed genes. Secondly, both recovery of H3K4me1 and H3K4me2 are restricted to the gene body; H3K4me1/2 upstream of the transcription start site exhibited no recovery. Thirdly, I noted that the distribution of H3K4me1 in the gene body also appeared altered after prolonged SET1/MLL-depletion, with sharp H3K4me1 peaks either decreasing or remaining unchanged, but with broad H3K4me1 domains the cover the gene body at low levels increasing. Furthermore, comparative analysis of H3K4me1/2/3 recovery after 48 hours of SET1/MLL-depletion in the gene body demonstrates that the genes which exhibit recovery in H3K4me1/2 and spreading of H3K4me3 in the gene body show a high degree of overlap (Fig. 4.12A, B). These observations strongly suggest that in the absence of SET1/MLL proteins, there is a novel methyltransferase which deposits H3K4me1/2/3 broadly acrosss the gene bodies of highly-transcribed genes.

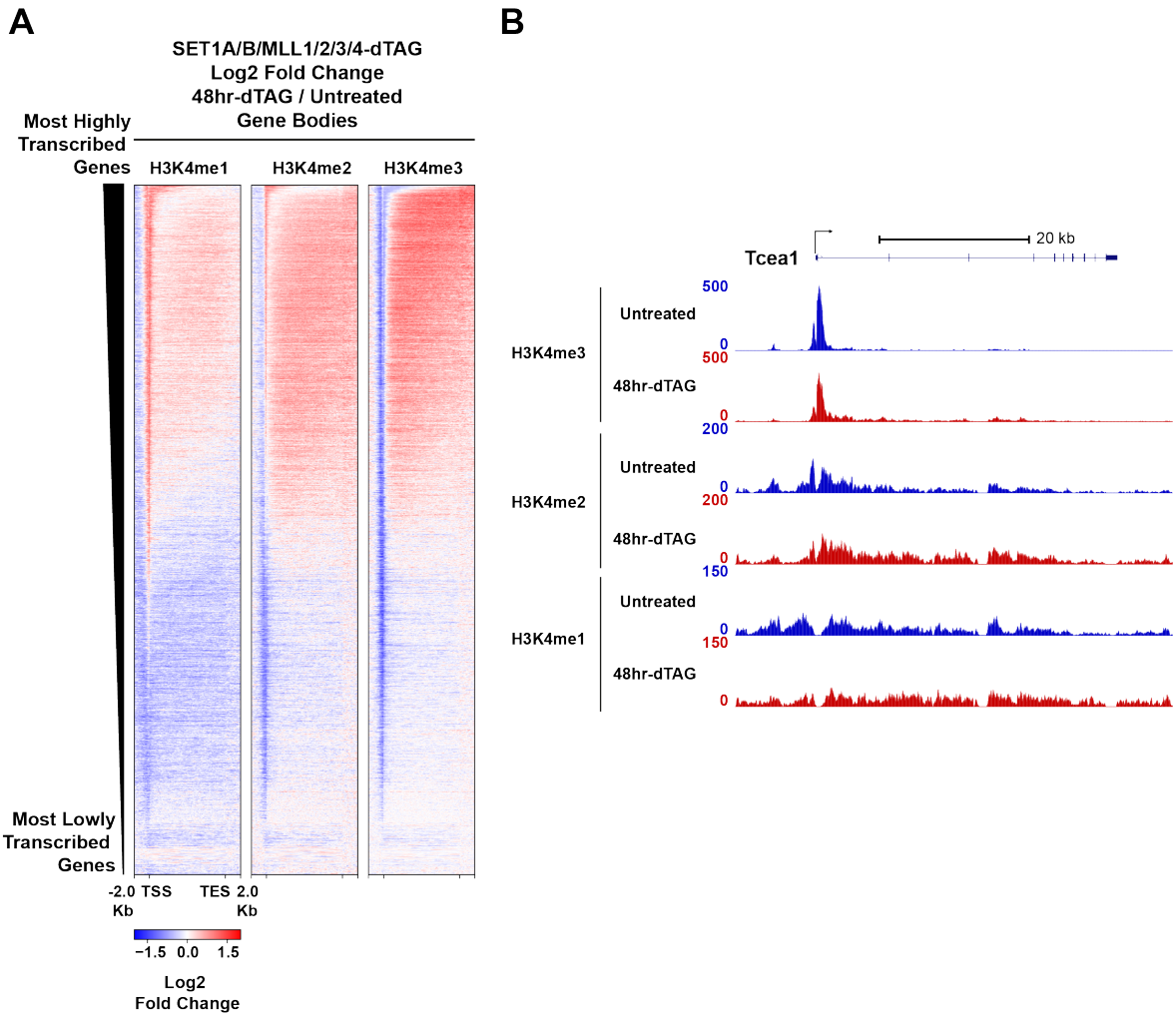

Figure 4.12: H3K4me1/2/3 recovery after SET1/MLL-depletion is highly correlated.

(A) Heatmaps showing log2 fold change in H3K4me1/2/3 levels across all gene bodies between 48hr-dTAG treated and untreated SET1A/B/MLL1/2/3/4-dTAG cells.
(B) Genome coverage tracks of a representative gene (*Tcea1*) demonstrating recovery of H3K4me1/2 and spreading of H3K4me3 into gene bodies after 48 hours of SET1/MLL-depletion.

4.3 A SET1/MLL-independent H3K4 methyltransfease deposits H3K4me2/3 in a transcription-coupled manner

4.3.1 Inhibition of transcription initiation demonstrates a SET1/MLL-independent H3K4 methyltransferase

Given that highly-transcribed genes retain more H3K4 methylation after SET1/MLL-depletion and experience a more pronounced recovery in H3K4 methylation across the gene body, I reasoned that a SET1/MLL-independent methyltransferase was depositing H3K4 methylation in a transcription-coupled manner. To test this hypothesis, I decided to investigate the effect of inhibiting transcription on H3K4 methylation. To do this, I employed triptolide (TRP), a small molecule inhibitor of the general transcription factor TFIIH that has been shown to effectively inhibit transcription initiation by RNA polymerase II (Jonkers et al., 2014, Titov et al., 2011). Using TRP to inhibit transcription and dTAG to deplete SET1/MLL proteins in SET1A/B/MLL1/2/3/4-dTAG cells, I wanted to understand how much H3K4 methylation is dependent on transcription and whether the H3K4 methylation remaining after SET1/MLL-depletion was dependent on transcription (Fig. 4.13A). I first sought to validate the use of triptolide to inhibit transcription initiation by RT-qPCR. To detect nascent pre-mRNA, I used primers amplifying either across an exon-intron boundary or within intronic regions. After 2 hours of triptolide treatment, transcription at *Oct4*, *Hspg2*, and *Gapdh* was largely undetectable, indicating that triptolide effectively inhibits transcription (Fig. 4.13B).

Having validated the use of triptolide to inhibit transcription, I next analysed bulk levels of H3K4me2/3 after 2 hours of triptolide treatment. Surprisingly, H3K4me2/3 was reduced to approximately 50% after 2 hours of triptolide treatment, similar to the reductions observed after depletion of SET1A/B/MLL1/2/3/4 (Fig. 4.13C). H3K4me2/3 further declined to approximately 25% when triptolide treatment was used in conjunction with dTAG to deplete SET1A/B/MLL1/2/3/4 (Fig. 4.13C). This indicates that in the absence of SET1/MLL proteins, there is an additional H3K4 methyltransferase activity that is dependent on transcription. I next wanted to understand if the recovery of H3K4me2/3 upon prolonged depletion of SET1A/B/MLL1/2/3/4 was dependent on transcription. H3K4me2/3 both recovered substantially after 24 hours of SET1A/B/MLL1/2/3/4-depletion. However, when triptolide was added for the final 2 hours of this 24 hour dTAG treatment, H3K4me2/3 levels remained at 25% of untreated levels (Fig. 4.13C). This indicates that transcription inhibition is sufficient to abrogate recovery of H3K4me2/3 levels after SET1A/B/MLL1/2/3/4-depletion.

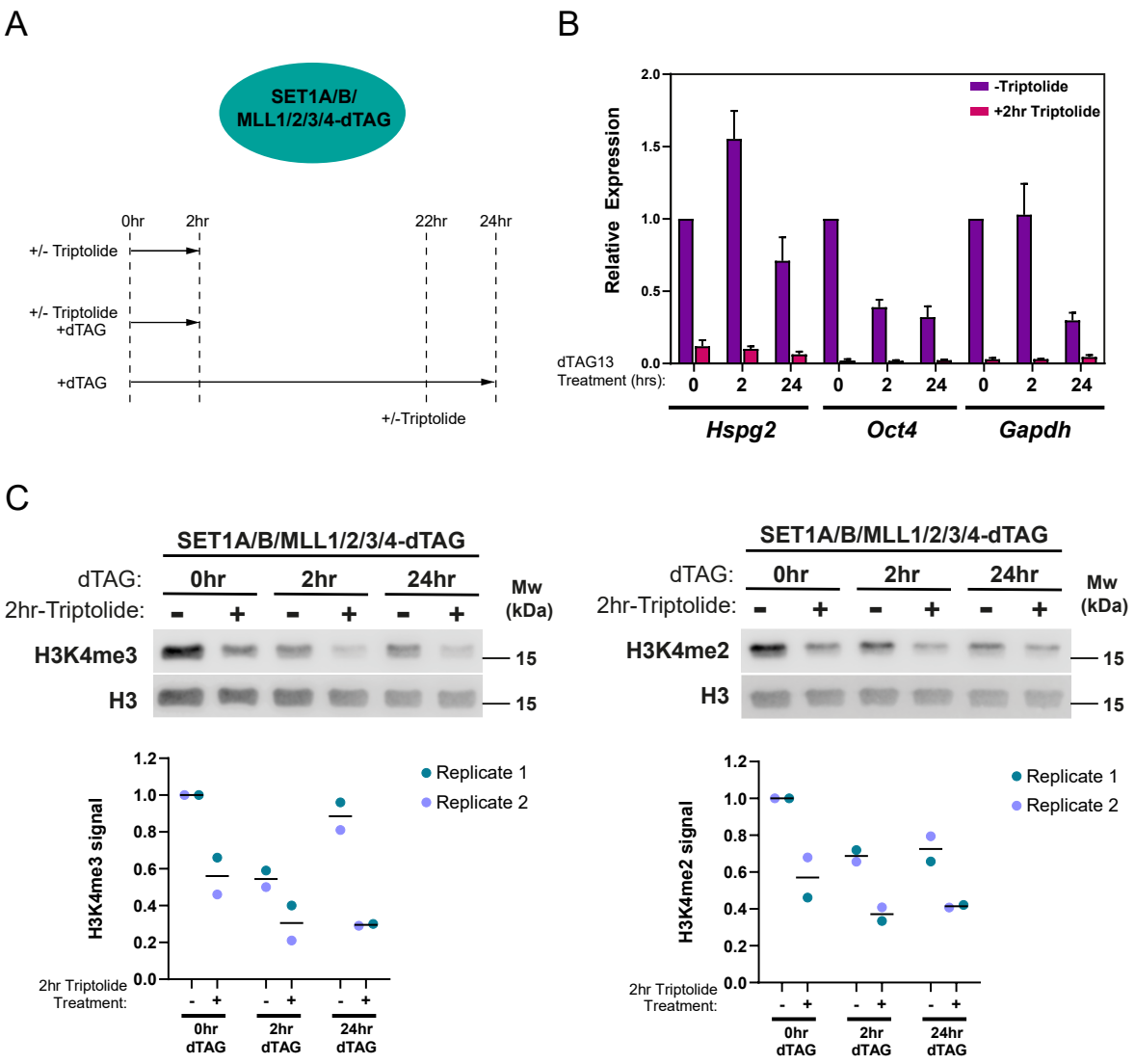

Figure 4.13: A SET1/MLL-independent, transcription-dependent H3K4 methyltransferase activity.

(A) Schematic illustrating the design of the transcription-inhibition experiment.
(B) RT-qPCR data demonstrating robust inhibition of transcription after triptolide treatment. To assess levels of nascent mRNA, primers were designed to amplify either across an exon-intron boundary (Hspg2) or amplify a region entirely within an intron (Oct4, Gapdh). Gene expression was normalized to U6 snRNA and shown as relative to the untreated condition. Mean values from 3 biological replicates are plotted. Error bars show standard error of the mean.
(C) Western blots illustrating reductions in H3K4me3 (left) and H3K4me2 (right) after triptolide and dTAG treatment. Histone H3 was used as a loading control. Quantifications are shown with H3K4me2/3 signal normalized to total H3 signal and shown as relative to the untreated condition. Individual values are plotted from 2 biological replicates with a line showing the mean of the two values.

4.3.2 A subset of genes are highly-dependent on transcription for H3K4me3

I next wanted to understand where this transcription-dependent methyltransferase activity was found in the genome by performing cChIP-seq for H3K4me3 after triptolide and dTAG treatment in SET1A/B/MLL1/2/3/4-dTAG cells. First, I analysed H3K4me3 levels genome-wide after a 2 hour tripdolide treatment. Surprisingly, I found a modest decrease in H3K4me3 across all promoters after triptolide treatment, in contrast to the global 50% decrease in bulk levels as seen by immunoblotting (Fig. 4.14A). However, I noted that at highly-transcribed genes, this decrease in H3K4me3 was not limited to the 2kb region around the TSS, but extended further into the gene body. H3K4me3 tends to form broad domains at the promoters of highly-transcribed genes, extending well into the gene body and in some instances encompassing the entirety of the gene. This suggests that H3K4me3 is often deposited as a consequence of active, ongoing transcription at highly transcribed genes. Indeed, I noted that H3K4me3 at a subset of 581 promoters were very sensitive to transcription inhibition, with H3K4me3 being reduced by more than 2-fold after 2 hours of triptolide treatment (Fig. 4.14B). These genes were primarily highly-transcribed genes, with over half of them belong to the first 2 quartiles of highly-transcribed genes and having substantially higher levels of RNA polymerase II binding (Fig. 4.14C, D). At these transcription-dependent promoters, H3K4me3 reductions at the +1 and +2 nucleosomes after 2 hours of triptolide treatment are comparable to reductions seen after 2 hours of dTAG treatment, and reductions further downstream of the +1 and +2 nucleosomes are considerably more severe (Fig. 4.14E). Due to the fact that transcription inhibition does not distinguish between H3K4me3 deposited by SET1/MLL proteins recruited to these promoters by transcription and H3K4me3 deposited by a novel transcription-dependent methyltransferase, I next analysed levels of H3K4me3 at these promoters when SET1/MLL proteins are depleted simultaneously with transcription inhibition. Upon simultaneous depletion of SET1/MLL proteins and transcription inhibition, H3K4me3 close to the transcription start site at these genes declined considerably more than with either SET1/MLL-depletion or transcription inhibition on their own, while H3K4me3 further into the gene body did not decline more than when only triptolide was used (Fig. 4.14E, F). Taken together, these findings indicate that at a subset of highly-transcribed genes, there is a SET1/MLL-independent H3K4 methyltransferase that deposits H3K4me3 at the promoters of highly-transcribed genes along with SET1/MLL proteins, but continues to deposit H3K4me3 further into the gene body independently of SET1/MLL proteins.

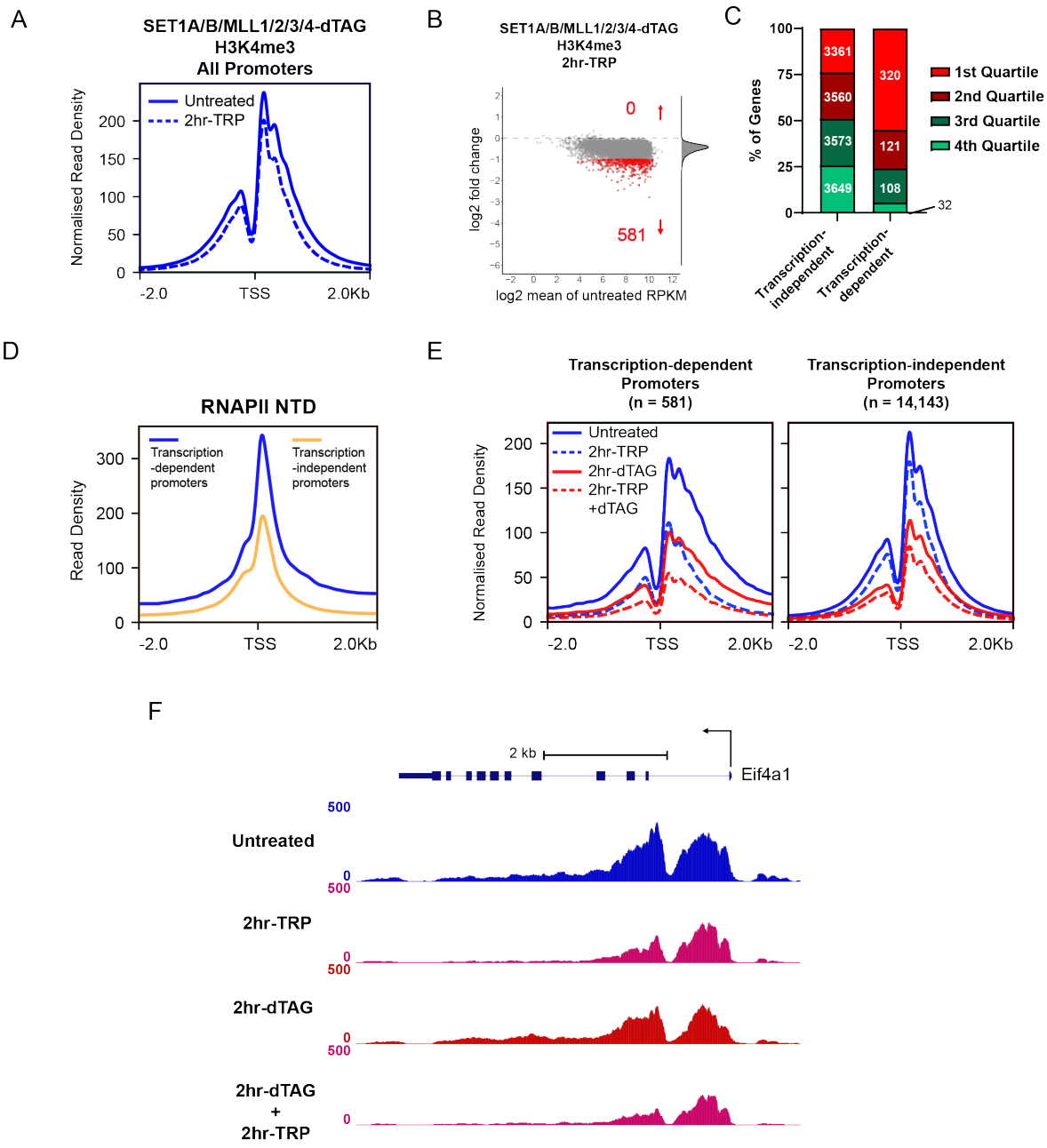

Figure 4.14: A subset of genes are highly-dependent on transcription for H3K4me3

(A) Metaplot showing H3K4me3 cChIP-seq signal at all H3K4me3-associated promoters (n=14,724) following 2 hours of triptolide treatment.
(B) MA plot showing log2 fold change between H3K4me3 cChIP-seq signal at H3K4me3-associated promoters (n=14,724) between 2hr-triptolide-treated cells and untreated cells. Log2 fold change values are plotted against H3K4me3 levels (RPKM) in untreated cells. Promoters that exhibited 2-fold increase or decrease in H3K4me3 levels are highlighted in red and their numbers are shown.
(C) Bar graph showing number of transcription-dependent (n=581) or transcription-independent (n=14,143) genes that fall into each quartile of genes based on transcription.
(D) Metaplot showing cChIP-seq signal of total RNA polymerase II in untreated cells at transcription-dependent or transcription-independent promoters.
(E) Metaplots showing H3K4me3 cChIP-seq signal at transcription-dependent and transcription-independent promoters. Shown are H3K4me3 in untreated cells, cells treated for 2 hours with triptolide, cells treated for 2 hours with dTAG, and cells treated for 2 hours with dTAG and triptolide.
(F) Genome coverage track showing H3K4me3 at a representative transcription-dependent gene (*Eif4a1*) after 2 hours of triptolide treatment, 2 hours of dTAG treatment, and 2 hours of triptolide and dTAG treatment in SET1A/B/MLL1/2/3/4-dTAG cells.

4.3.3 A SET1/MLL-independent methyltransferase deposits H3K4me3 broadly across the genome

Having identified a transcription-dependent, SET1/MLL-independent H3K4 methyltransferase activity that was particularly pronounced at a subset of highly-transcribed genes, I next sought to characterize this activity more broadly across the genome. Interestingly, I found that although reductions in H3K4me3 after transcription inhibition were more pronounced at highly-transcribed genes, lowly-transcribed genes also exhibited reductions in H3K4me3 upon transcription inhibition, albeit of a smaller magnitude (Fig. 4.15). This indicates active transcription contributes to H3K4me3 deposition at all genes, although this contribution is far greater at some highly-transcribed genes.

I next examined how much of this transcription-dependent H3K4me3 was due to a novel methyltransferase activity by analysing H3K4me3 levels after simultaneous SET1/MLL-depletion and transcription inhibition. Broadly across the genome, simultaneous SET1/MLL-depletion and transcription inhibition led to greater reductions than either transcription inhibition or SET1/MLL-depletion on their own, indicating that this SET1/MLL-independent, transcription-dependent methyltransferase activity is widespread and not only localised to a subset of genes (Fig. 4.15). Additionally, I found that this transcription-dependent methyltransferase activity is more prominent at highly-transcribed genes, where H3K4me3 levels are more sensitive to triptolide treatment and less sensitive to SET1/MLL-depletion (Fig. 4.15). Taken together, these observations indicate that at highly-transcribed genes, there is a SET1/MLL-independent methyltransferase that deposits H3K4me3 in a transcription-dependent manner.

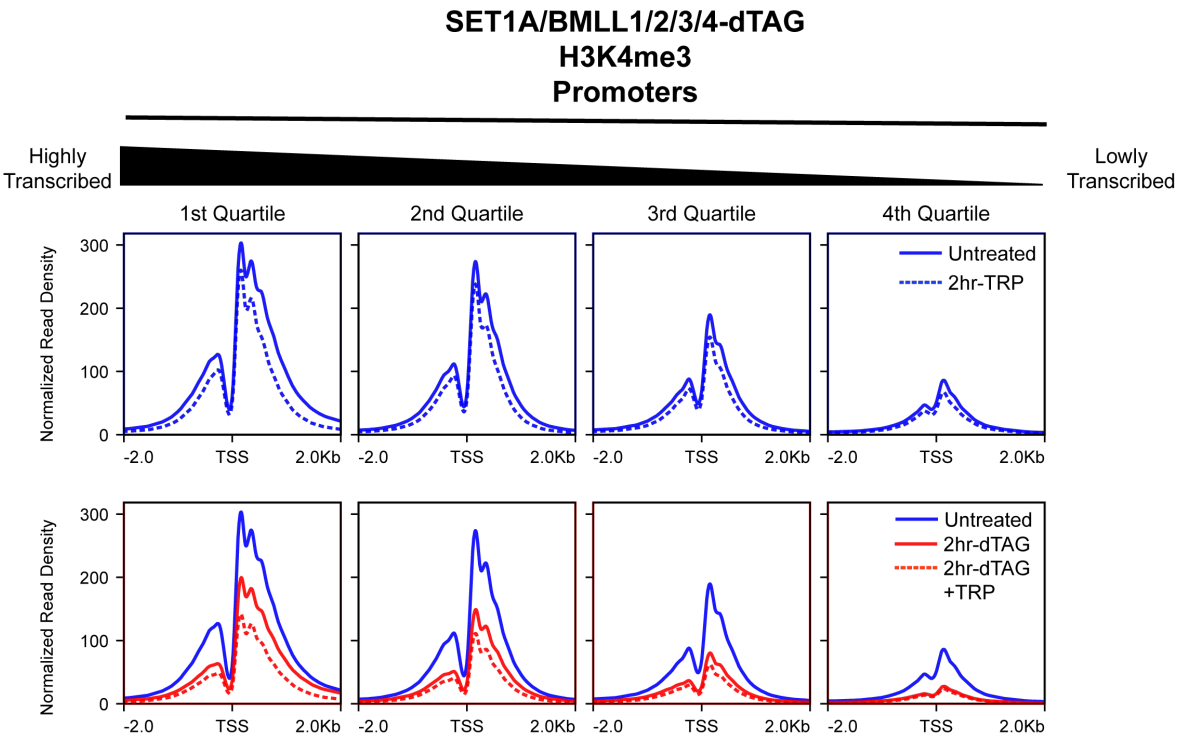

Figure 4.15: H3K4me3 at highly transcribed genes are more sensitive to triptolide treatment.

Metaplots showing H3K4me3 cChIP-seq at all H3K4me3-associated promoters (n=14,724) divided into quartiles based on transcription after 2 hours of triptolide treatment (top) or 2 hours of dTAG treatment, with or without triptolide (bottom).

4.3.4 A SET1/MLL-independent methyltransferase compensates for SET1/MLL removal

I next sought to further characterize the transcription-dependent compensation of H3K4me3 upon prolonged SET1/MLL-depletion. Western blot analysis indicated that bulk H3K4me3 levels recover significantly after 24 hours of dTAG treatment in SET1A/B/MLL1/2/3/4-dTAG cells, but that this recovery is abolished upon the addition of triptolide for the last 2 hours of this 24-hour dTAG treatment. cChIP-seq analysis confirms that recovery of H3K4me3 levels at the promoter was also largely abolished (Fig. 4.16A). Broadly speaking, H3K4me3 levels at promoters after inhibiting transcription for the last 2 hours of the 24 hour dTAG treatment were very similar to those observed after a 2 hour dTAG and triptolide treatment. cChIP-seq of H3K4me3 after prolonged depletion of SET1A/B/MLL1/2/3/4 had previously shown that H3K4me3 spreads from the promoter into the gene body. Indeed, I found that inhibiting transcription for the last 2 hours of the 24 hour dTAG treatment resulted in virtually complete removal of gene-body-associated H3K4me3 (Fig. 4.16B, C). These findings indicate that the transcription-dependent H3K4 methyltransferase is also largely responsible for compensating SET1/MLL loss.

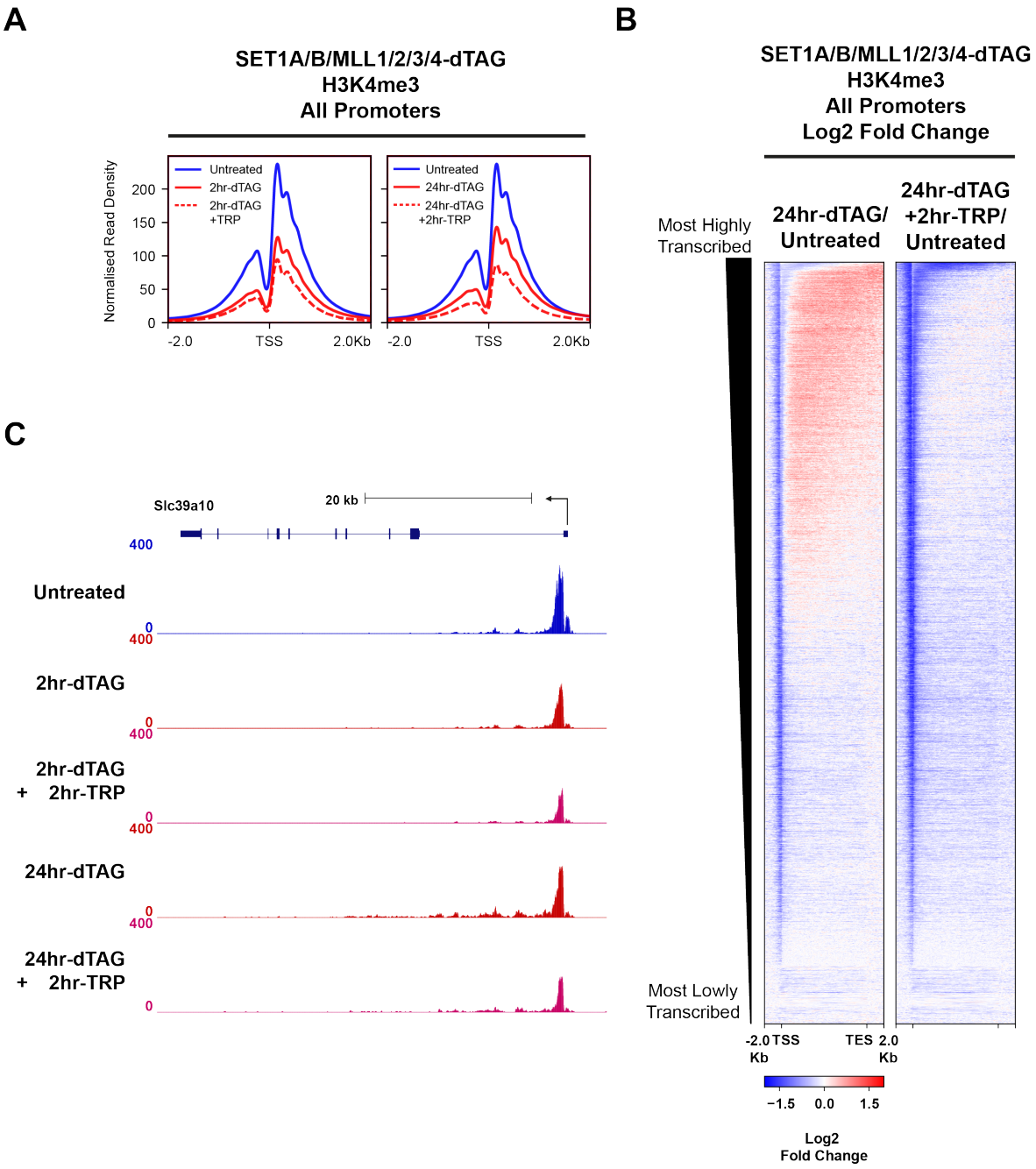

Figure 4.16: H3K4me3 recovery after prolonged SET1/MLL-depletion is partially dependent on transcription.

(A) Metaplots showing H3K4me3 cChIP-seq at all H3K4me3-associated promoters (n=14,724) after 2 hours (left) and 24 hours (right) of SET1/MLL-depletion with or without triptolide.
(B) Heatmaps showing log2 fold change in H3K4me3 cChIP-seq signal between 24hr-dTAG treated and untreated cells, with or without triptolide at all promoters (n=20,633). Heatmaps are sorted in descending order according to transcription.
(C) Genome coverage tracks showing H3K4me3 after 2 hours and 24 hours of dTAG treatment, with or without triptolide, in the SET1A/B/MLL1/2/3/4-dTAG cell line at a representative gene (*Slc39a10*), which shows spreading of H3K4me3 into the gene body.

4.4 Summary and Discussion

In this chapter, I first analysed the contributions of MLL3/4 to H3K4 methylation by generating a cell line in which MLL3 and MLL4 are epitope-tagged and can be rapidly depleted by dTAG. Using this cell line, I show that crosslinking ChIP-seq primarily captures MLL3/4 binding at enhancers, but that this does not represent the entirety of their activity on chromatin. Indeed, I show that MLL3/4 deposit H3K4me1 pervasively across transcribed regions of the genome in addition to enhancers. I next show that MLL3/4 have minimal contributions to H3K4me2/3 except at enhancers. Based on these observations, I hypothesise that there are two binding modalities of MLL3/4: a stably-bound fraction which deposits all three methylation states of H3K4 at enhancers and a dynamic fraction which interacts transiently with chromatin to deposit H3K4me1 only. Future studies should focus on studying these two populations of MLL3/4 using single-particle tracking analysis in live cells. Additionally, future studies should focus on how MLL3/4 are recruited to chromatin. While it has been demonstrated that transcription factors such as GATA6 and NEUROD1 can recruit MLL3/4 to enhancers, how MLL3/4 activity is guided to other regions of the genome are yet to be elucidated (Xie et al., 2023).

I next sought to understand whether MLL3/4 synergise with SET1A/B/MLL1/2 to deposit H3K4 methylation by generating a cell line in which all six methyltransferases can be rapidly depleted. Surprisingly, I found removing all six SET1/MLL methyltransferases did not result in removal of all H3K4 methylation. Indeed, I found that while SET1A/B/MLL1/2/3/4 synergise to deposit H3K4me2, especially at enhancers, removing MLL3/4 in addition to SET1A/B/MLL1/2 did not lead to any further reductions in H3K4me1 or H3K4me3. Indeed, the most prominent effect of removing all six SET1/MLL methyltransferases was a recovery of H3K4me1/2/3 levels after prolonged dTAG treatment that is highly-correlated with transcription. Additionally, the spreading of H3K4me3 into the gene bodies of highly-transcribed genes seen after prolonged SET1A/B/MLL1/2-depletion was also seen after prolonged SET1A/B/MLL1/2/3/4-depletion. In fact, this spreading of H3K4me3 into gene bodies was more pronounced when MLL3/4 was removed in addition to SET1A/B/MLL1/2. These findings indicate that MLL3/4 are not responsible for depositing H3K4me3 in the absence of SET1A/B/MLL1/2, and that the increase in H3K4me1 after SET1A/B/MLL1/2-depletion is only partially dependent on MLL3/4. Similar to my findings here, previous studies conditional and constitutive knockout models have shown that MLL3/4 are only responsible for approximately half of all H3K4 monomethylation as assessed by western blotting or by mass spectrometry (Dorighi et al., 2017, C. Wang et al., 2016, D. Hu, Gao, et al., 2013). Here, I show that the other members of the SET1/MLL family do not account for the remainder of the H3K4me1, and that MLL3/4 do not account for the remainder of H3K4me3 seen after SET1A/B/MLL1/2-depletion. Taken together, these findings demonstrate that there is a novel, SET1/MLL-independent H3K4 methyltransferase whose activity is highly-correlated with active transcription. Building on these observations, I demonstrate that a large proportion of H3K4me2/3 that remain after SET1/MLL-depletion are dependent on transcription. I also demonstrate that the recovery of H3K4me2/3 after prolonged SET1/MLL-depletion is dependent on transcription, suggesting that there is a transcription-dependent H3K4 methyltransferase that acts independently of SET1/MLL proteins.

There are various mechanistic explanations for how this novel methyltransferase might compensate for SET1/MLL loss. Assumming that this novel methyltransferase is directly recruited to chromatin by the transcriptional machinery, one hypothesis would be that its ability to interact with the transcriptional machinery may be enhanced upon prolonged SET1/MLL removal and therefore result in the recovery of H3K4 methylation that is observed. Another hypothesis is that transcription is drastically reduced upon SET1/MLL-depletion but recovers after prolonged SET1/MLL-depletion, with this recovery being more pronounced at highly-transcribed genes. Previous work from the Klose lab has shown that removal of SET1A/B for 2 hours leads to larger reductions in gene expression compared to removal of SET1A/B for 24 hours, suggesting that other mechanisms exist to compensate for SET1A/B removal (Hughes et al., 2023). Whether this trend holds true for the combinatorial removal of all SET1/MLL proteins remains to be seen, but if so could explain recovery in H3K4 methylation after prolonged SET1/MLL-depletion.

The relationship between transcription and H3K4me3 is controversial; although there is a correlation between transcription and H3K4me3, it has been difficult to experimentally demonstrate any causative relationship between the two due to the difficulty in abolishing H3K4me3 and the difficulty in abolishing transcription itself. In this chapter, I demonstrate that a large proportion of H3K4me3 throughout the genome is deposited as a consequence of active transcription. However, I note that the use of triptolide to broadly inhibit transcription does not give mechanistic insight as to how the process of transcription may influence H3K4 deposition. The rapid onset of H3K4me3 reductions after triptolide treatment imply that there is indeed a direct link between the transcriptional machinery and H3K4 methyltransferases, as has been demonstrated to be the case in yeast (Soares et al., 2017). However, it is also possible other transcription-dependent process influence H3K4 methylation. For example, H2BK120ub1 deposition is tightly linked to transcription and has been demonstrated to stimulate H3K4 methylation (Kim et al., 2009). Depletion of PAF1 has been shown to result in a rapid reduction in both H2BK120ub1 and H3K4me3 (Z. Wang et al., 2023). Whether transcription-coupled H3K4 methylation is directly dependent on interactions with the transcription machinery or is dependent on other transcription-related processes remains to be determined.

Chapter 5

WRAD-containing complexes deposit all H3K4 methylation

Introduction

The finding that depleting all six SET1/MLL methyltransferases does not lead to a loss of all H3K4 methylation is surprising given that they are proposed to be the primary H3K4 methyltransferases in mammals (Shilatifard, 2012). This was proposed to be the case in part because their homologs in yeast and drosophila have been shown to deposit all H3K4 methylation, but also in part because depleting subunits of the WRAD subcomplex in mammalian cells have been shown to result in major reductions in bulk levels of H3K4 methylation. The WRAD subcomplex consists of four regulatory subunits: WDR5, ASH2L, RBBP5, and DPY30, and is required for the methyltransferase activity of SET1/MLL proteins (Ernst and Vakoc, 2012). *In vitro* biochemistry experiments have shown that SET1/MLL proteins in isolation are weak H3K4 mono-methyltransferases; they are only able to exhibit robust methyltransferase activity when in complex with WRAD (Shinsky et al., 2014). Early studies using RNAi-mediated knockdown or conditional knockout of WRAD components consistently reported major reductions in H3K4 methylation, although they rarely reported complete loss of H3K4 methylation, likely due to incomplete depletion of WRAD subunits (Jiang et al., 2011, Benayoun et al., 2014, Bochyńska et al., 2022, Vanderkruk et al., 2023). However, recent studies using degron approaches have shown that depleting any of the four subunits of WRAD results in massive reductions in H3K4 methylation. Indeed, degron-mediated depletion of RBBP5 and DPY30 in mESCs have been reported to result in rapid loss of H3K4me3 followed by loss of H3K4me1/2 upon extended depletion (H. Wang et al., 2023, S. Hu et al., 2023). Similarly, depletion of ASH2L in mouse embryonic fibroblasts have resulted in near-complete loss of all H3K4 methylation, and depletion of WDR5 in human cancer cells have resulted in loss of all H3K4me3 (Barsoum et al., 2023, Siladi et al., 2022). Therefore, these findings strongly suggest that SET1/MLL complexes are the primary contributors to H3K4 methylation in most mammalian cell types.

One explanation for the incomplete loss of H3K4 methylation I observed after depletion of SET1/MLL proteins is that dTAG-mediated depletion is incomplete. However, there is strong evidence to suggest that this is not the case. Previous work from the Klose lab has demonstrated by ChIP-seq that dTAG-SET1A is largely removed from chromatin after dTAG treatment, and that its depletion leads to widespread transcriptional defects (Hughes et al., 2023). Although the removal of SET1B from chromatin has not been confirmed by ChIP,

depleting SET1B in addition to SET1A leads to greater reductions in gene expression, strongly suggesting that SET1B depletion is complete. Additionally, previous studies using conditional and constitutive knockout approaches have indicated that SET1B does not have major contributions to H3K4 methylation in mESCs (Bledau et al., 2014, Sze et al., 2023). Whether MLL1-4 are removed from chromatin after dTAG-mediated depletion is not yet confirmed, although reductions in H3K4 methylation are broadly in agreement with previous studies employing conditional or constitutive knockout approaches. Furthermore, the rapid turnover of H3K4me2/3 observed suggests that H3K4 methylation is very dynamic and would likely require constant maintenance by H3K4 methyltransferases, making it unlikely that small amounts of SET1/MLL proteins would be able to maintain the levels of H3K4 methylation retained after SET1/MLL-depletion.

I hypothesized that WRAD is required for a novel, SET1/MLL-independent methyltransferase activity. In this chapter, I first seek to understand whether depleting WRAD components leads to greater reductions in H3K4 methylation compared to depleting SET1/MLL proteins, and proceed to investigate the effect of removing SET1/MLL proteins on the WRAD subcomplex. I then confirm that SET1/MLL proteins are removed from chromatin after dTAG treatment using ChIP-seq, and then proceed to confirm that they no longer interact with the WRAD subcomplex using using mass spectrometry. Surprisingly, I find that SET1B continues to interact with WRAD after dTAG treatment, although mass spectrometry only recovers peptides mapping to the C-terminal region of SET1B. I therefore speculate that an as-yet-uncharacterized alternative isoform of SET1B is the methyltransferase reponsible for maining H3K4 methylation levels after dTAG treatment.

5.1 WDR5 depletion results in greater reduction of H3K4me2/3 than depletion of SET1/MLL methyltransferases

5.1.1 WDR5-depletion leads to profound reductions in H3K4me2/3 but not H3K4me1

To complement my findings that depletion of SET1/MLL methyltransferases cannot account for all H3K4 methylation, I wanted to develop an orthogonal approach to ablate SET1/MLL methyltransferase activity. I reasoned this could be achieved by depleting a subunit of the WRAD-subcomplex which should render all SET1/MLL complexes catalytically-incompetent. WDR5 is a key component of the WRAD subcomplex and has been proposed to act as a scaffold for the SET1/MLL complexes by bridging the SET1/MLL proteins with the remainder of the WRAD-subcomplex (Guarnaccia and Tansey, 2018). Therefore, I generated a cell line in which WDR5 was endogenously tagged at the C-terminus with FKBP12^{F36V} (Fig. 5.1A). Using this WDR5-dTAG cell line, I was able to deplete WDR5 within 2 hours of dTAG treatment, comparable to the depletion of SET1/MLL proteins in the SET1A/B/MLL1/2/3/4-dTAG cell line (Fig. 5.1A).

I first analysed bulk levels of H3K4me1/2/3 after WDR5-depletion in a 72-hour dTAG treatment time-course. After 2 hours of WDR5-depletion, bulk levels of H3K4me2/3 were reduced to 60%, which is comparable to the reductions seen after SET1/MLL-depletion (Fig. 5.1B). However, as the time-course treatment progressed, bulk H3K4me2/3 levels continued to decline, reaching a minimum after 24 hours of dTAG treatment to approximately 25% for H3K4me3 and after 48 hours to approximately 15% for H3K4me2. This stands in stark contrast to the effect seen after SET1/MLL-depletion, where H3K4me2/3 recover substantially from 8 hours of dTAG treatment onwards. This indicates that the SET1/MLL-independent methyltransferase compensating for SET1/MLL removal is also dependent on WDR5 for methyltransferase activity. Surprisingly, bulk H3K4me1 levels did not decline significantly after WDR5-depletion, for which I propose several explanations (Fig. 5B). I previously showed that turnover of H3K4me1 is slow compared to that of H3K4me2/3, which suggests that the primary mechanism for H3K4me1 removal is by dilution during chromatin replication rather than by demethylases or variant histone turnover. WDR5 is implicated in mitosis and I noted that prolonged depletion of WDR5 severely affected cell viability (Guarnaccia and Tansey, 2018, Oh et al., 2020). It is possible that prolonged WDR5 depletion led to impaired cell division and therefore a lack of turnover of H3K4me1. Another possibility is that WDR5 depletion does not affect the ability of SET1/MLL methyltransferases to deposit H3K4me1 due to the intrinsic H3K4 mono-methyltransferase activity of the SET1/MLL proteins. Additionally, MLL3 has been reported to exhibit increased H3K4 mono-methyltransferase activity *in vitro* when WDR5 is not included in the complex (Shinsky and Cosgrove, 2015). It is also possible that the large reductions in H3K4me2/3 resulted in an accumulation of H3K4me1 due to demethylation of the higher methylation states. Lastly, it is also possible that there is a SET1/MLL-independent methyltransferase that does not require WDR5 for H3K4me1 deposition.

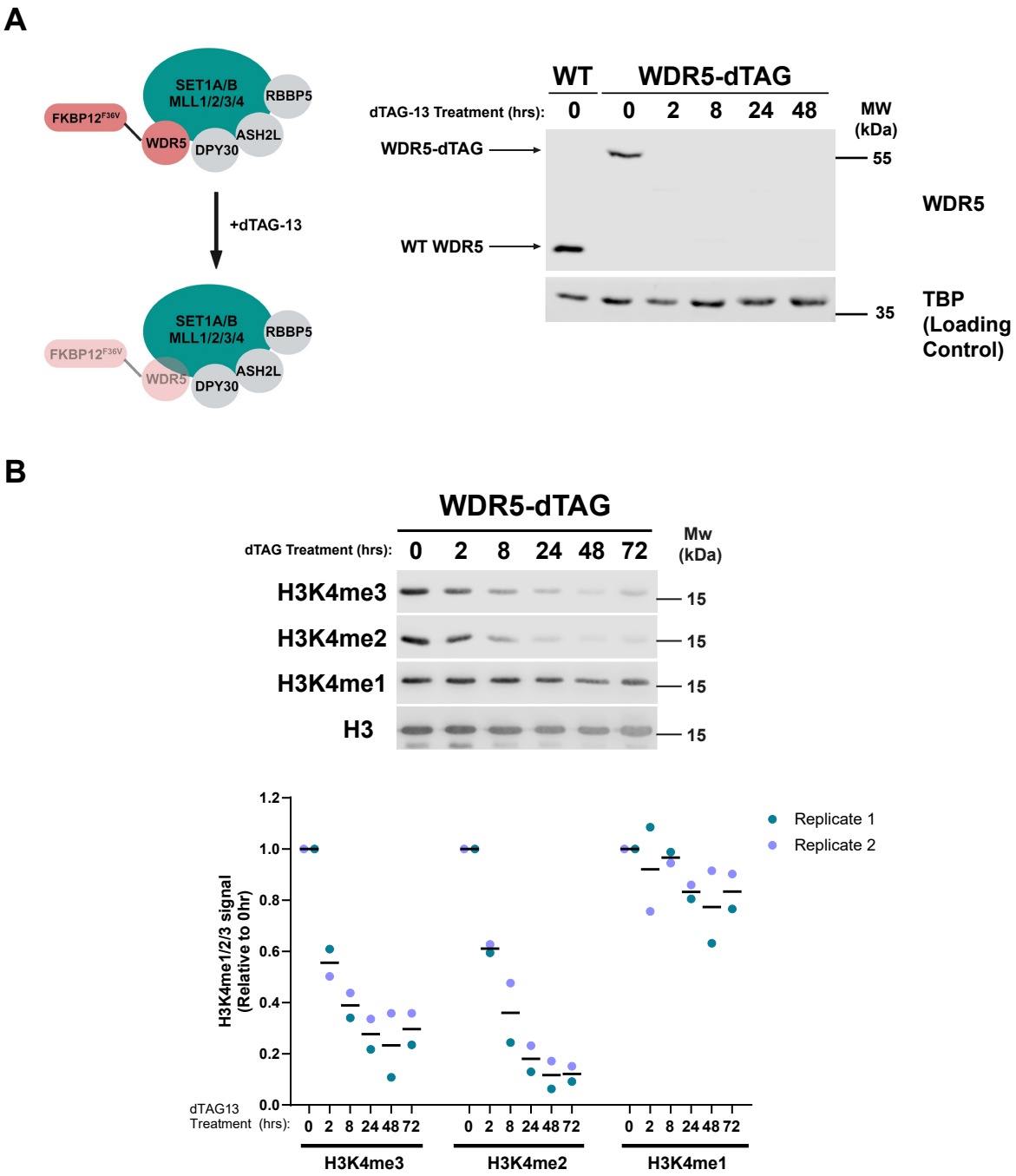

Figure 5.1: WDR5 depletion results in greater reductions in H3K4me2/3 but not H3K4me1

(A) Immunoblotting of WDR5 from nuclear extract of WDR5-dTAG cells following dTAG13 treatment. TBP was used as a loading control. A cartoon illustrating depletion of WDR5 is shown on the left.
(B) Immunoblotting of H3K4me1/2/3 from histone extracts of WDR5-dTAG cells following dTAG13 treatment. Histone H3 was used as a loading control. Quantitative analysis of bulk H3K4me1/2/3 levels are shown below. H3K4me1/2/3 signal was normalised to total H3 signal and shown as relative to the untreated sample. Individual values from 2 biological replicates are plotted along with a line showing the mean of the two values.

5.1.2 WDR5 is required for the SET1/MLL-independent H3K4 methyltransferase activity

Immunoblot analysis indicated that removing WDR5 for 2 hours did not result in greater reductions in bulk levels of H3K4me2/3 compared to removing SET1/MLL proteins for 2 hours. However, the recovery of H3K4me2/3 after 8 hours of SET1/MLL-depletion was not observed 8 hours after WDR5-depletion. I next wanted to understand how H3K4me3 was affected across the genome after acute WDR5-depletion and after prolonged WDR5-depletion. To do this, I performed cChIP-seq for H3K4me3 after 2 hours and 24 hours of WDR5-depletion to capture early and late changes in H3K4me3, respectively. H3K4me3 across all promoters was reduced to approximately 50% after 2 hours of WDR5-depletion, and continued to decline to approximately 30% after 24 hours of WDR5-depletion (Fig. 5.2A). Interestingly, I noted that reductions in H3K4me3 were less severe after 2 hours of WDR5-depletion compared to 2 hours of SET1/MLL-depletion (Fig. 5.2A). Furthermore, H3K4me3 reductions were more pronounced after 2 hours of SET1/MLL-depletion compared to 2 hours of WDR5-depletion across all promoters irrespective of the levels of transcription of the associated gene, suggesting that 2 hours of WDR5-depletion does not completely ablate SET1/MLL methyltransferase activity (Fig. 5.2A). However, In contrast to the effect seen after SET1/MLL-depletion, H3K4me3 continued to decline after 24 hours of WDR5-depletion across promoters of highly- and lowly-transcribed genes, indicating that the recovery in H3K4me3 seen after 24 hours of SET1/MLL-depletion is dependent on WDR5 (Fig. 5.2A). Additionally, the spreading of H3K4me3 into gene bodies after 24 hours of SET1/MLL-depletion was not observed after 24 hours of WDR5 depletion, and all gene-body-associated H3K4me3 was also severely reduced (Fig. 5.2B). Therefore, these findings indicate that WDR5 is required for a SET1/MLL-independent H3K4 methyltransferase activity that causes the recovery of H3K4me3 seen after prolonged SET1/MLL-depletion. However, a substantial amount of H3K4me3 was still retained at the promoters of highly-transcribed genes after 24 hours of WDR5-depletion (Fig. 5.2A), indicating that depletion of WDR5 alone does not cause a complete loss of H3K4me3.

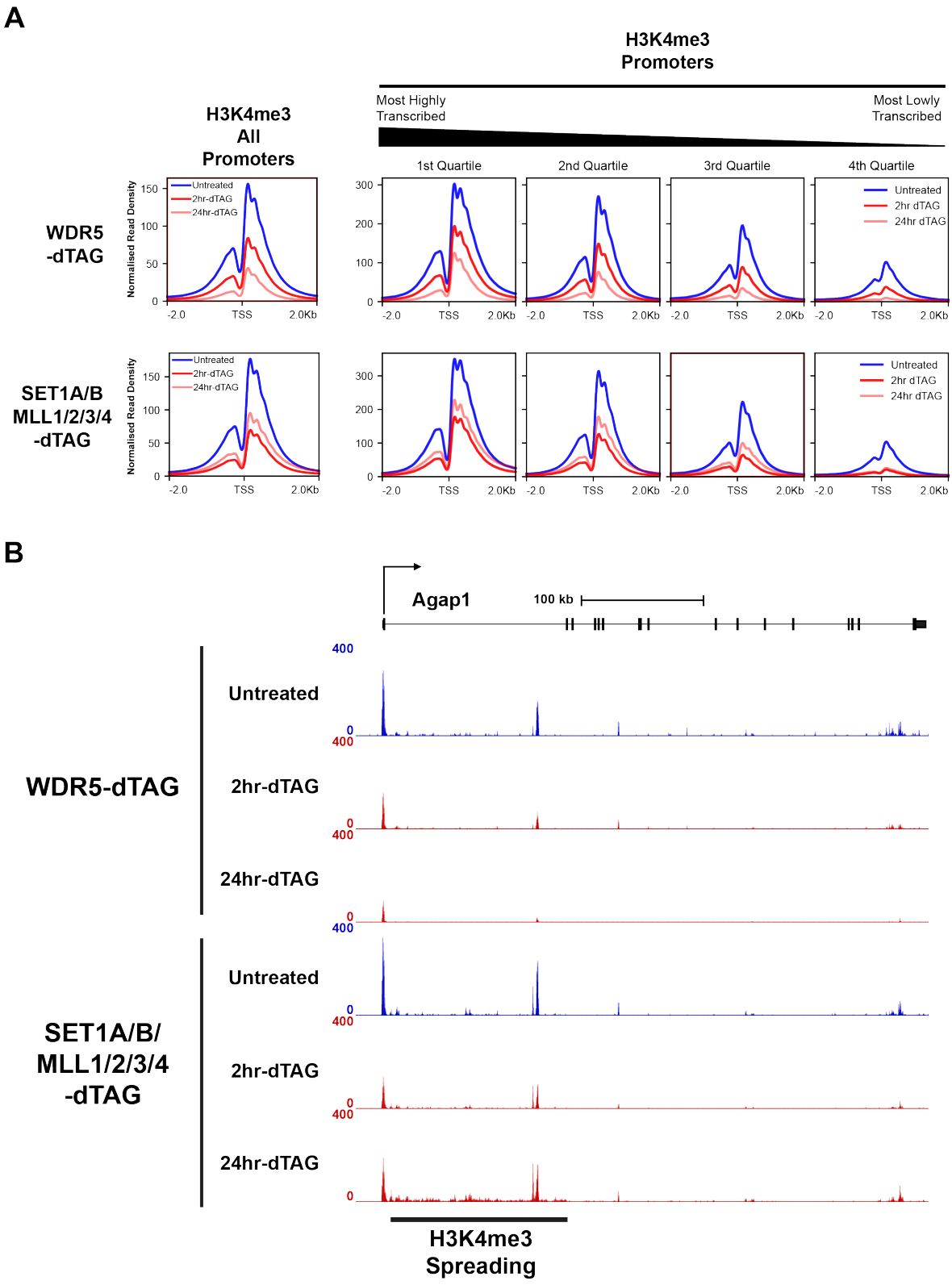

Figure 5.2: WDR5 depletion leads to genome-wide reductions in H3K4me3 that are greater than SET1/MLL-depletion

(A) Metaplots showing H3K4me3 cChIP-seq signal at all promoters (n=20,633) and all H3K4me3-associated promoters (14,724) divided into quartiles based on transcription after WDR5-depletion (top) or SET1/MLL-depletion (bottom).
(B) Genome coverage tracks of H3K4me3 at a representative gene (*Agap1*) after 2 hours and 24 hours of WDR5-depletion or SET1/MLL-depletion.

I next sought to understand how H3K4me2 was influenced genome-wide after WDR5-depletion for 2 hours and 24 hours. Interestingly, I noted that reductions in H3K4me2 after 2 hours of WDR5-depletion were more moderate compared to 2 hours of SET1/MLL-depletion (Fig. 5.3A, B). This was particularly evident at the promoters of lowly-transcribed genes, where H3K4me2 levels are reduced by more than 50% after 2 hours of SET1/MLL-depletion (Fig. 5.3A). In contrast, after 2 hours of WDR5-depletion, H3K4me2 levels are only moderately reduced at these promoters (Fig. 5.3A). Additionally, H3K4me2 upstream of the transcription start site of highly- and moderately-transcribed genes were also more moderately reduced after 2 hours of WDR5-depletion compared to 2 hours of SET1/MLL-depletion (Fig. 5.3A). The more moderate reduction in H3K4me2 after WDR5-depletion was also seen at enhancers, where H3K4me2 was reduced by approximately 30% after 2 hours of dTAG treatment, comparable to the reductions seen after 2 hours of MLL1/2-depletion or 2 hours of MLL3/4-depletion (Fig. 5.3B). In contrast, H3K4me2 levels were reduced by over 50% after 2 hours of SET1/MLL-depletion at enhancers (Fig. 5.3B). Given that reductions in H3K4me3 levels are broadly similar after 2 hours of SET1/MLL-depletion and 2 hours of WDR5-depletion, the more moderate reductions in H3K4me2 after 2 hours of WDR5-depletion cannot be explained by an increase in demethylation of H3K4me3. Instead, a more likely explanation for these more moderate reductions is that 2 hours of WDR5-depletion does not completely remove H3K4-dimethyltransferase activity from SET1/MLL complexes.

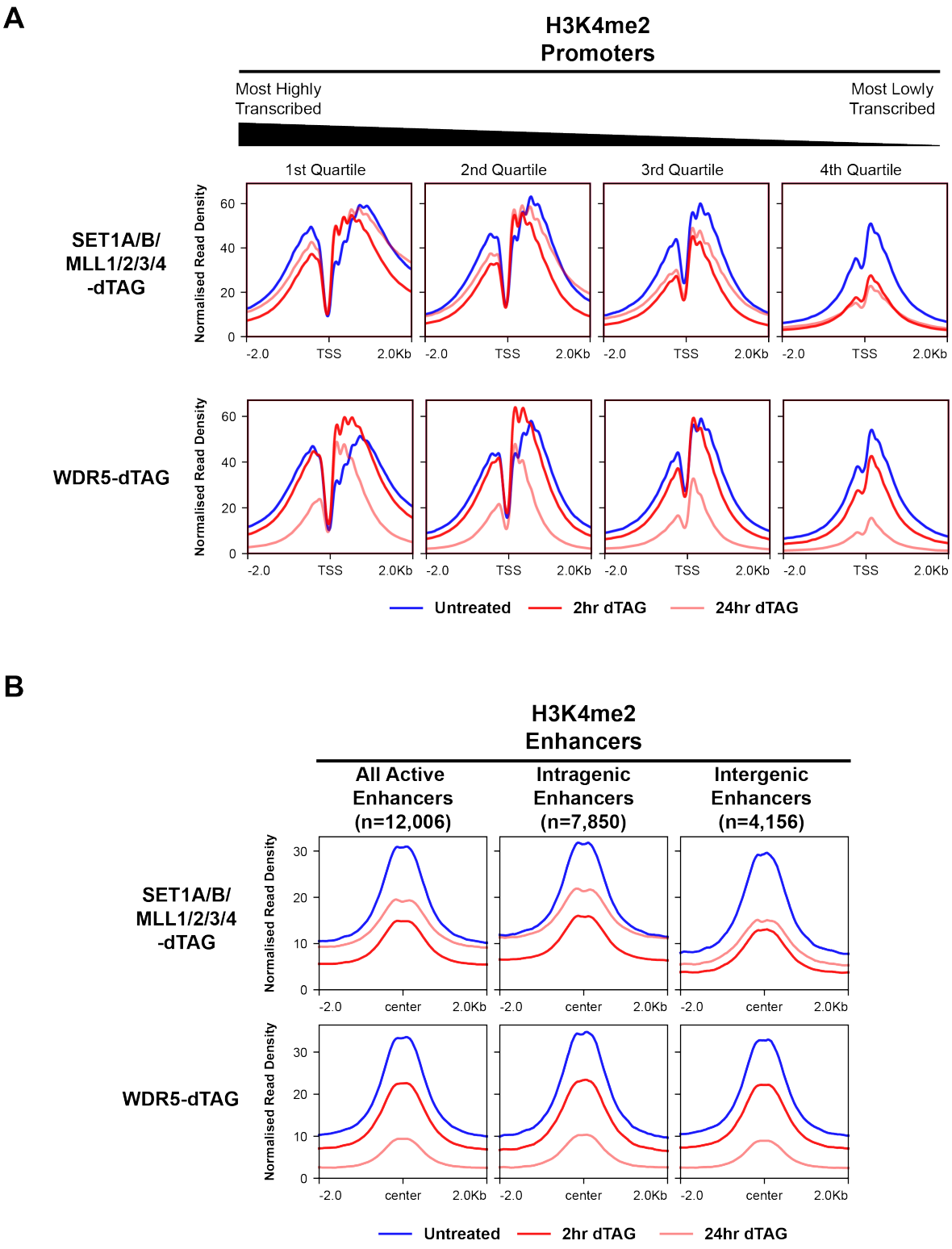

Figure 5.3: WDR5 depletion leads to genome-wide reductions in H3K4me2 that are greater than SET1/MLL-depletion

(A) Metaplots showing H3K4me2 cChIP-seq signal at all H3K4me2-associated promoters (n=15,379) divided into quartiles based on transcription after SET1.MLL-depletion (top) or WDR5-depletion (bottom).
(B) Metaplots showing H3K4me2 cChIP-seq signal at all active enhancers, active intragenic enhancers, and active intergenic enhancers after SET1.MLL-depletion (top) or WDR5-depletion (bottom).

As expected, H3K4me2 was severely reduced after 24 hours of WDR5 depletion. Interestingly, this reductions was very prominent at enhancers and at the promoters of lowly-transcribed genes (Fig. 5.3A, B). Additionally, H3K4me2 in gene bodies was also severely reduced (Fig. 5.4A). Given that H3K4me2 recovers broadly across gene bodies after prolonged SET1/MLL-depletion, and in some cases exceeds leves in untreated cells, this indicates that WDR5 is required for this SET1/MLL-independent H3K4 di-methyltransferase activity. Interestingly, after 24 hours of WDR5-depletion, the only sites where H3K4me2 was still retained were promoters, with highly-transcribed genes retaining more H3K4me2 than lowly-transcribed genes (Fig. 5.3A, Fig. 5.4B). Indeed, the distribution of H3K4me2 was severely altered at these sites after 24 hours of WDR5 depletion and resembled that of H3K4me3, with an enrichment at the +1 and +2 nucleosomes where H3K4me2 levels are ordinarily low (Fig. 5.4A, B). This strongly suggests that the H3K4me2 remaining after WDR5-depletion is mainly a product of demethylation of H3K4me3 and that removal of WDR5 largely removes H3K4 di-methyltransferase activity from mESCs. Taken together, these findings indicate that WDR5 is necessary for the recovery of H3K4me2/3 seen after prolonged SET1/MLL-depletion. However, the finding that highly-transcribed genes still retain a considerable amount of H3K4me2/3 at their promoters suggests that WDR5-depletion may not completely ablate SET1/MLL methyltransferase activity.

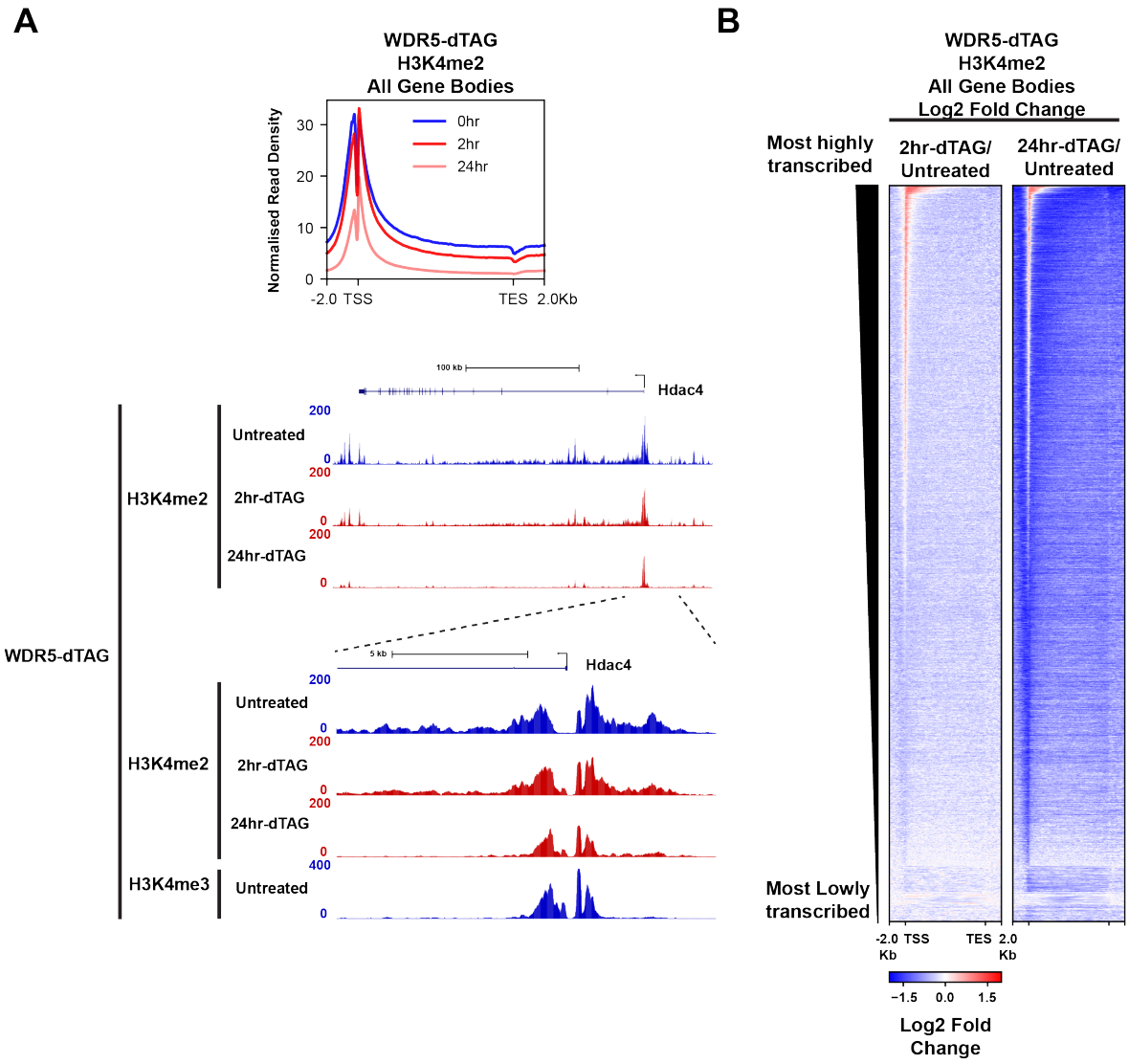

Figure 5.4: H3K4me2 is largely removed from gene bodies following WDR5-depletion

(A) Metaplot showing H3K4me2 cChIP-seq signal at all gene bodies (n=20,633) after WDR5-depletion. **(B)** Genome coverage tracks showing H3K4me2 at an example gene (*Hdac4*) after 2 hours and 24 hours of WDR5-depletion. Also shown is a zoomed-in view of H3K4me2 at the promoters of *Hdac4* after WDR5-depletion and H3K4me3 from untreated cells.

Depleting WDR5 for 24 hours led to the removal of virtually all H3K4me2/3 in gene bodies and abrogated the recovery of H3K4me2/3 seen after prolonged SET1/MLL depletion, suggesting that the SET1/MLL-independent H3K4me2/3 activity is completely dependent on WDR5. Paradoxically, depletion of WDR5 did not lead to complete loss of H3K4me2/3 at the CGI promoters of highly-transcribed genes, suggesting that depleting WDR5 did not completely ablate SET1/MLL methyltransferase activity. Furthermore, depleting WDR5 for 2 hours led to more modest reductions in H3K4me2/3 compared to depleting SET1/MLL proteins for 2 hours. Therefore, I hypothesized that depleting WDR5 did not completely ablate SET1/MLL methyltransferase activity, but prolonged depletion of WDR5 resulted in major defects in transcription, on which the SET1/MLL-independent methyltransferase activity is dependent (see Chapter 4). To test this hypothesis, I generated a cell line in which SET1A/B, MLL1/2, MLL3/4, and WDR5 can be simultaneously depleted within 2 hours of dTAG treatment (Fig. 5.5A, B). Using this cell line, I analysed levels of bulk H3K4me1/2/3 after 2 hours of dTAG treatment and after 24 hours of dTAG treatment to see if removing WDR5 and SET1/MLL proteins together would lead to greater reductions in H3K4 methylation than removing SET1/MLL proteins on their own. Interestingly, I found that removing SET1/MLL proteins and WDR5 simultaneously led to greater reductions in H3K4me2/3 compared to removing WDR5 or SET1/MLL proteins on their own (Fig. 5.5C). After 2 hours of SET1/MLL/WDR5-depletion, H3K4me2/3 levels were reduced to approximately 25% of untreated levels, which is broadly comparable to levels seen after 2 hours of SET1/MLL-depletion combined with a 2 hour triptolide treatment. After 24 hours of SET1/MLL/WDR5-depletion, H3K4me2/3 levels further declined to less than 10% of untreated levels. In contrast, H3K4me1 levels did not change appreciably after 2 hours of SET1/MLL/WDR5-depletion. However, after 24 hours of dTAG treatment, bulk H3K4me1 levels were reduced to approximately 25% of untreated levels, greater than the 50% reduction seen after 24 hours of SET1/MLL-depletion. This also suggests that the lack of reduction in H3K4me1 after prolonged WDR5-dTAG treatment was not due to defects in cell division or replication but may be due to inherent H3K4 mono-methyltransferase activity of SET1/MLL proteins. Taken together, these findings indicate that WDR5 contributes to an H3K4 methyltransferase activity that is independent of SET1/MLL proteins.

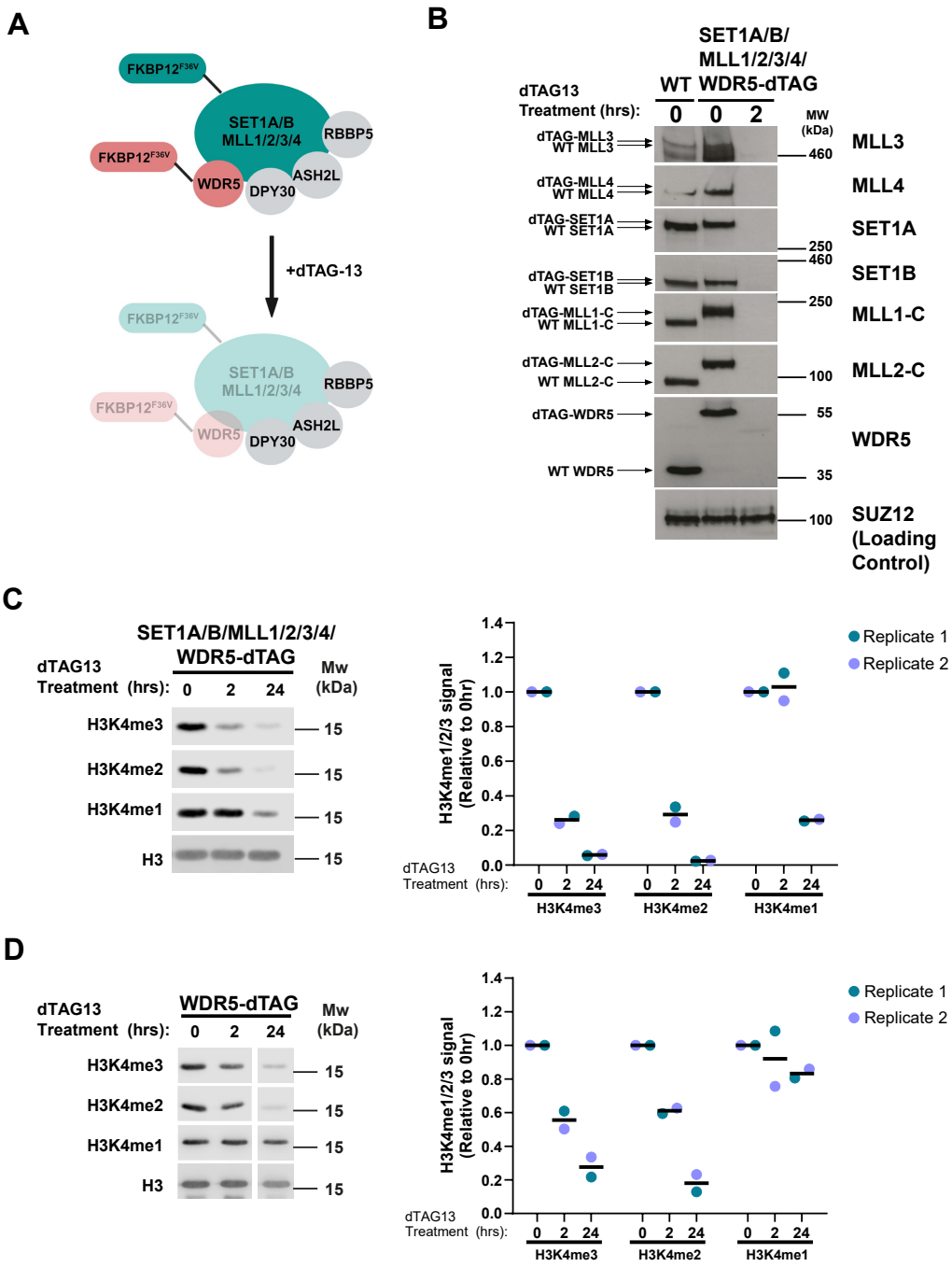

Figure 5.5: WDR5 regulates an H3K4 methyltransferase independent of SET1/MLL proteins.

(A) Cartoon illustrating depletion of SET1/MLL proteins and WDR5 after dTAG treatment.
(B) Immunoblotting of SET1A, SET1B, MLL1-C, MLL2-C, MLL3, MLL4, WDR5 after 2 hours of dTAG treatment in the SET1A/B/MLL1/2/3/4/WDR5-dTAG cell line. SUZ12 is used as a loading control.
(C) Immunoblotting of H3K4me1/2/3 from histone extracts of SET1A/B/MLL1/2/3/4/WDR5-dTAG cells following dTAG-13 treatment. Histone H3 was used as a loading control. Quantitative analysis of bulk H3K4me1/2/3 levels are shown on the right. H3K4me1/2/3 signal was normalised to total H3 signal and shown as relative to the untreated sample. Individual values from 2 biological replicates are plotted, along with a line showing the mean of the two values.
(D) Immunoblotting of H3K4me1/2/3 from histone extracts of WDR5-dTAG cells following dTAG13 treatment. Histone H3 was used as a loading control. Quantitative analysis of bulk H3K4me1/2/3 levels are shown on the right. H3K4me1/2/3 signal was normalised to total H3 signal and shown as relative to the untreated sample. Individual values from 2 biological replicates are plotted, along with a line showing the mean of the two values.

5.2 The SET1/MLL-independent H3K4 methyltransferase requires the WRAD subcomplex

5.2.1 A WRAD-containing complex remains intact after SET1/MLL-depletion

Given that WDR5 plays a centrol role in regulating SET1/MLL-complexes, the finding that removing WDR5 in addition to SET1/MLL proteins leads to greater reductions in H3K4me1/2/3 compared to removing SET1/MLL proteins on their own suggests that there is an H3K4 methyltransferase activity remaining in the cell after SET1/MLL-depletion that relies on WDR5. To understand the nature of this methyltransferase activity, I first investigated if the subunits of the WRAD subcomplex were still present after SET1/MLL-depletion by immunoblotting (Fig. 5.6). I found that WDR5 levels were largely unchanged after SET1/MLL-depletion. This is unsurprising given that WDR5 has been reported to be a component in many other protein complexes, and its inclusion in SET1/MLL complexes likely only represents a small fraction of all WDR5 in the nucleus. I did not analyse levels of DPY30 by western blot due to technical difficulties in blotting for DPY30. Additionally, DPY30 only interacts with ASH2L and its fate is most likely tied to that of ASH2L. Furthermore, DPY30 is a component of the NURF complex and is therefore not unique to SET1/MLL complexes, unlike RBBP5 and ASH2L, which have not been reported to be part of any other complexes. Interestingly, both RBBP5 and ASH2L levels were reduced substantially after 2 hours of SET1/MLL-depletion, but did not decline further upon extended SET1/MLL-depletion. An explanation for this reduction in RBBP5 and ASH2L levels could be that WRAD components complexed with SET1/MLL proteins are also degraded along with SET1/MLL proteins after dTAG treatment, and the remaining ASH2L and RBBP5 represent their monomeric forms. A second explanation is that the liberation of ASH2L and RBBP5 from SET1/MLL complexes after dTAG treatment leads to a faster turnover rate. Nevertheless, these results indicate that there is substantial amounts of WRAD components remaining in the nucleus that could form a complex in the absence of SET1/MLL proteins.

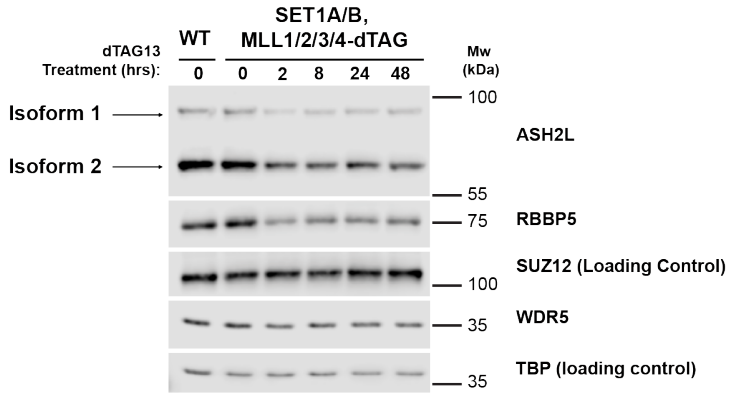

Figure 5.6: WRAD subunits remainin the nucleus after SET1/MLL-depletion

Western blots of RBBP5, ASH2L, and WDR5 in nuclear extract from SET1A/B/MLL1/2/3/4-dTAG cells after prolonged dTAG treatment. SUZ12 was used as a loading control for RBBP5 and ASH2L western blots and TBP was used as a loading control for the WDR5 western blot.

I next investigated whether a WRAD-containing complex existed in the nucleus after SET1/MLL-depletion. To do this, I fractionated nuclear proteins using size exclusion chromatography from SET1A/B/MLL1/2/3/4-dTAG cells and analysed the distribution of SET1/MLL complex components across the different fractions by immunoblotting (Fig. 5.7). I first analysed the distribution of SET1/MLL complex components in untreated cells to gain an understanding of how SET1/MLL complexes were distributed across the different fractions.

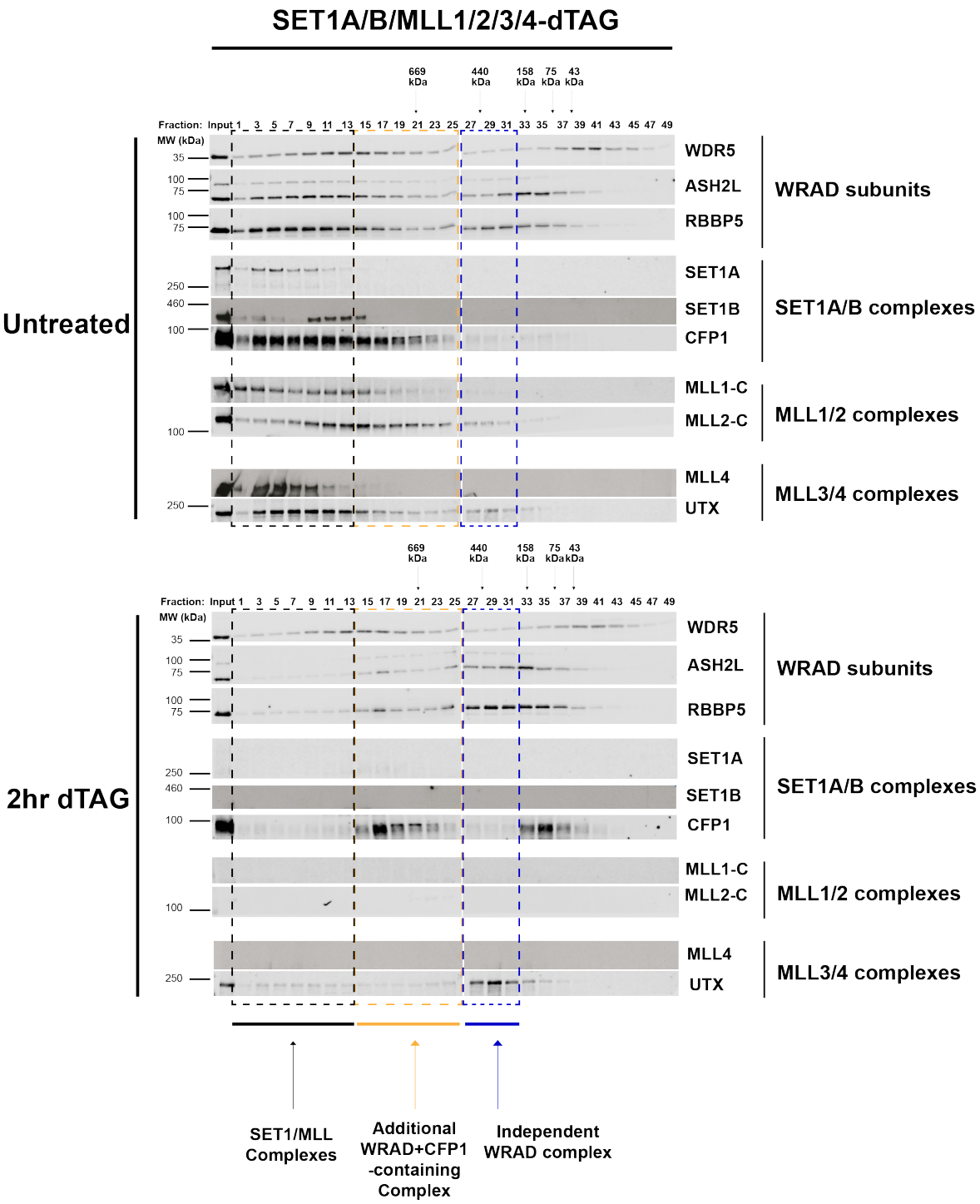

Figure 5.7: WRAD-containing complexes exist independently of SET1/MLL proteins

Western blots of SET1/MLL complex subunits from nuclear extract of SET1A/B/MLL1/2/3/4-dTAG cells fractionated by size exclusion chromatography.

First, I noted that SET1/MLL proteins and their specific interactor primiarily eluted in high molecular weight fractions that were too large to represent their monomieric forms, indicating that they are primarily in complexes. As had been shown previously, SET1A exclusively eluted in very high molecular weight fractions above the 669kDa standard, indicating that the vast majority of SET1A exists within a complex, likely with WRAD and SET1A/B-specific interactors such as CFP1 (Hughes et al., 2023). A similar distribution was seen for SET1B, although SET1B abundance peaked in fractions that were of slightly smaller molecular weight than SET1A, indicating that there are some differences in the composition of SET1A and SET1B complexes. CFP1 also primarily eluted in high molecular weight fractions and was undetectable in lower molecular weight fractions that would correspond to monomeric CFP1 (82kDa), indicating that CFP1 was primarily found in large multi-subunit complexes. Indeed, CFP1 eluted in the same high molecular weight fractions as SET1A and SET1B. Surprisingly, I found that CFP1 also eluted in fractions of intermediate molecular weight (fractions 15-25) in which SET1A and SET1B were absent, but were too large to represent monomeric CFP1, implying that CFP1 exists in complexes that did not contain SET1A/B. Similar to SET1A and SET1B, MLL4 also elutde exclusively in the highest molecular weight fractions, with virtually no MLL4 eluting in fractions corresponding to its monomeric form (600kDa), indicating that MLL4 is mostly in a complex with WRAD and other interactors. Accordingly, UTX, an MLL3/4-specific interactor, eluted in the same fractions as MLL4, indicating that UTX is also primarily in MLL3/4-complexes, although some UTX also eluted in lower molecular weight fractions corresponding to monomeric UTX (180kDa). Interestingly, although MLL1-C and MLL2-C mostly eluted in high molecular weight fractions above 669kDa, they had much broader distributions across the different fractions compared to SET1A and SET1B, including fractions between 440kDa and 669kDa. This suggests that although the majority of MLL1 and MLL2 are in complexes, some of these complexes are likely composed of only MLL1 or MLL2 in complex with WRAD without their specific interactors MENIN or PSIP.

Secondly, I noted that WRAD components primarily eluted in a peak of high molecular weight fractions and a second peak of low molecular weight fractions, suggesting that they exist both in complexes and in their monomeric forms. As expected, WDR5 eluted across a wide range of fractions, encompassing all 25 fractions analysed. Within this broad distribution, WDR5 was particularly enriched in high molecular weight fractions above 669kDa, likely representing WDR5 in large mulit-subunit complexes, and low molecular weight fractions below 43kDa, representing monomeric WDR5 (37kDa). Both RBBP5 and ASH2L also eluted in high molecular weight fractions in which SET1/MLL also eluted, likely representing RBBP5 and ASH2L in complex with SET1/MLL proteins. Interestingly, RBBP5 and ASH2L also eluted in lower molecular weight fractions between 440kDa and 75kDa that were too large to represent monomeric RBBP5 (70kDa) and ASH2L (80kDa for isoform 1, 65kDa for isoform 2). Interestingly, I noted that WDR5, ASH2L, and RBBP5 all eluted in fractions between the 158kDa and 440kDa standards. Mass spectrometry studies have suggested that the WRAD subcomplex primarily exists with a 1:1:1:2 stoichiometry between WDR5, ASH2L, RBBP5, and DPY30. My co-immunoprecipitations experiments demonstrate that the smaller isoform 2 (65kDa) of ASH2L is the primary ASH2L species in WRAD (see later, Fig. 5.8). Therefore, the sum of RBBP5 (70kDa), ASH2L (65kDa), WDR5 (37kDa), and two copies of DPY30 (19kDa x2) 210kDa, making it likely that these fractions

between 158kDa and 440kDa contain an independent WRAD complex. Although it is important to note that the migration of proteins through the gel filtration column depends on both the shape of the protein as well as size, it is possible that some of these fractions contain WDR5, ASH2L, RBBP5, and DPY30 in complex on their own, without SET1/MLL proteins.

I next analysed the distribution of SET1/MLL complexes after 2 hours of SET1/MLL-depletion. As expected, SET1A, SET1B, MLL1-C, MLL2-C, and MLL4 were not detected in any of the fractions. Accordingly, UTX no longer eluted in high molecular weight fractions, and instead eluted in low moleuclar weight fractions corresponding to monomeric UTX, indicating that MLL3/4 complexes are fully disrupted. Similarly, CFP1 no longer eluted in high molecular weight fractions, and also eluted in low molecular weight fractions corresponding to monomeric CFP1. Surprisingly, CFP1 continued to elute in intermediate molecular weight fractions. Interestingly, SET1A and SET1B did not elute in these intermediate molecular weight fractions in untreated cells. This suggests that SET1A/B complexes are fully disrupted after dTAG treatment, but CFP1 is part of another complex which is not disrupted. Surprisingly, I noted that RBBP5 and ASH2L also eluted in these fractions after SET1/MLL-depletion, while no longer eluting in higher molecular weight fractions. In untreated cells, WDR5, RBBP5, ASH2L, CFP1, MLL1-C, and MLL2-C all elute in these fractions. After dTAG treatment, MLL1-C and MLL2-C are no longer detectable in nuclear extract, but WRD5, RBBP5, ASH2L, CFP1 all continue to elute in these intermediate molecular weight fractions. Taken together, these observations suggest that a complex consisting of WRAD and CFP1 exists independently of SET1A/B/MLL1/2/3/4 that is not perturbed upon depletion of SET1/MLL proteins. Interestingly, the sum of their molecular weights is 292kDa, which is far too small to account for their elution in fractions of over 669kDa. This indicates that there are possibly two WRAD-type complexes left in mESCs after SET1/MLL depletion - one containing WRAD on its own, one containing WRAD + CFP1 and another subunit that is much larger.

To directly test whether a WRAD + CFP1 - containing complex existed in mESCs, I performed co-immunprecipitation experiments in wild-type ESCs and SET1A/B/MLL1/2/3/4-dTAG cells using antibodies against WDR5 (Fig. 5.8A, B). In wild-type cells and untreated cells, SET1/MLL proteins, along with WRAD components and complex-specific interactors CFP1 and UTX could be precipitated with WDR5 (Fig. 5.8B). As expected, after 2 hours of dTAG treatment, SET1/MLL proteins no longer precipitated with WDR5. Interestingly, RBBP5 and CFP1 could be precipitated with WDR5, indicating that they continue to interact with each other in the absence of SET1/MLL proteins. Surprisingly, I noted that a small amount of UTX precipitated with WDR5 after SET1/MLL-depletion. Size exclusion chromatography demonstrated that UTX does not co-elute with CFP1 in the same fractions after SET1/MLL depletion. Rather, it elutes in fractions of lower molecular weight in which WDR5, ASH2L, and RBBP5 also elute. Therefore, it is possible that there are there are three WRAD-containing complexes in the absence of SET1/MLL proteins: an independent WRAD complex, a WRAD-CFP1 complex, and a WRAD-UTX complex. Surprisingly, I found that ASH2L no longer precipitated with WDR5 in the absence of SET1/MLL proteins. Given that WDR5 is a highly abundant protein with many interactors, only a small fraction of WDR5 is likely to interact with ASH2L. Additionally, ASH2L levels are reduced substantially after SET1/MLL depletion, and it is possible that the co-immunprecipitation

experiment performed did not have the required sensitivity to detect the remaining ASH2L-WDR5 interaction.

To confirm these findings, I next performed co-immunoprecipitation experiments in SET1A/B/MLL1/2/3/4-dTAG cells for RBBP5 (Fig. 5.8C). Indeed, I found that after 2 hours of dTAG treatment, ASH2L, WDR5, and CFP1 all co-precipitated with RBBP5, indicating that a WDR5-ASH2L-RBBP5-CFP1 complex existed in the absence of SET1/MLL proteins. Additionally, the RBBP5-UTX interaction was not retained after SET1/MLL-depletion, indicating that the WDR5-UTX interaction likely exists independently of WRAD. Taken together, these results indicate that in the absence of SET1/MLL proteins, at least two WRAD-containing complex exists in the cell, an independent WRAD complex, and a WRAD+CFP1 complex that contains additional subunits. Given that WDR5, ASH2L, RBBP5, DPY30, and CFP1 themselves do not have SET domains and have no reported methyltransferase activity on their own, these findings suggest that there are additional components of this WRAD-containing complex, possibly including a methyltransferase that is responsible for the SET1/MLL-independent H3K4 methylation.

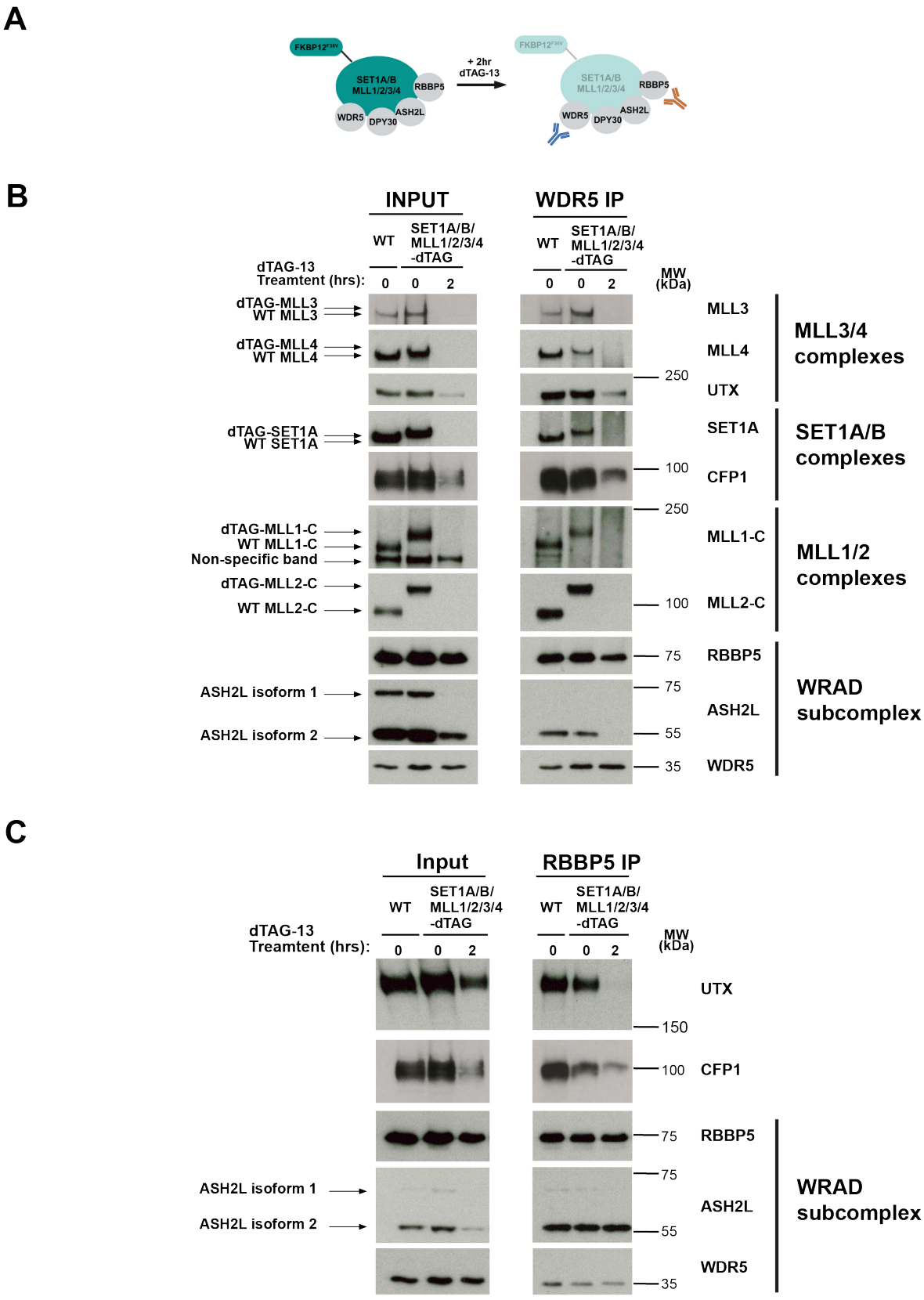

Figure 5.8: WRAD components continue to interact after SET1/MLL-depletion

(A) Cartoon illustrating schematic of co-immunoprecipitation experiments.
(B-C) Western blots analysis of SET1/MLL complex subunits after immunoprecipitation of WDR5 (B) and RBBP5 (C) from nuclear extracts from WT and SET1A/B/MLL1/2/3/4-dTAG cells.

5.2.2 The WRAD-containing complex is SET1/MLL-independent

In this chapter, I have demonstrated that the WRAD-subcomplex remains intact after removal of SET1/MLL proteins, and that this WRAD-subcomplex accounts for the H3K4 methylation remaining in the cell after SET1/MLL removal. I have suggested that this WRAD-subcomplex is part of a SET1/MLL-independent H3K4 methyltransferase complex. However, another possibility is that the depletion of SET1/MLL proteins in the SET1A/B/MLL1/2/3/4-dTAG is incomplete, with a sub-population of SET1/MLL proteins persisting in the nucleus and continuing to methylate H3K4. Throughout this thesis, I have used immunoblotting of SET1/MLL proteins from nuclear extracts to demonstrate their depletion after dTAG treatment. However, it is worth noting that the fraction of chromatin-interacting proteins that are bound to chromatin typically represents a small minority its entire population in the nucleus. Therefore, it is possible that SET1/MLL proteins that are not depleted after dTAG treatment and are bound to chromatin may not be detected by immunoblotting. I therefore sought to investigate whether there were SET1/MLL proteins still bound to chromatin after dTAG treatment by performing double-crosslinked cChIP-seq in the SET1A/B/MLL1/2/3/4-dTAG cell line. The Klose lab had previously performed double-crosslinked cChIP-seq for SET1A using an antibody against the T7 epitope in a cell line in which SET1A was tagged with both T7 and FKBP12^{F36V} and demonstrated that virtually all SET1A is removed from chromatin after 2 hours of dTAG treatment (Hughes et al., 2023). Given the limited roles that SET1B has in supporting gene expression and depositing H3K4me3, and the lack of high-quality antibodies against SET1B available for ChIP, I decided not to perform cChIP-seq for SET1B, instead focusing on MLL1/2/3/4.

I first decided to perform cChIP-seq for MLL3 and MLL4. In the SET1A/B/MLL1/2/3/4-dTAG cell line, both MLL3 and MLL4 are endogeneously tagged with a 3xT7 tag in addition to FKBP12^{F36V}. Therefore, I performed cChIP-seq for MLL3/4 simultaneously using an antibody against the T7 epitpope in SET1A/B/MLL1/2/3/4-dTAG cells after 2 hours of dTAG treatment. As described in Chapter 4, 6,553 peaks were called for MLL3/4, and at all 6,553 peaks, T7-MLL3/4 signal was largely undetectable after 2 hours of dTAG treatment (Fig. 5.9A, B, D). At enhancers, where T7-MLL3/4 signal is strongest, T7-MLL3/4 signal was also largely undetectable after 2 hours of dTAG treatment (Fig. 5.9A, C, D). Therefore, cChIP-seq indicates that MLL3/4 binding on chromatin is largely undetectable after dTAG treatment.

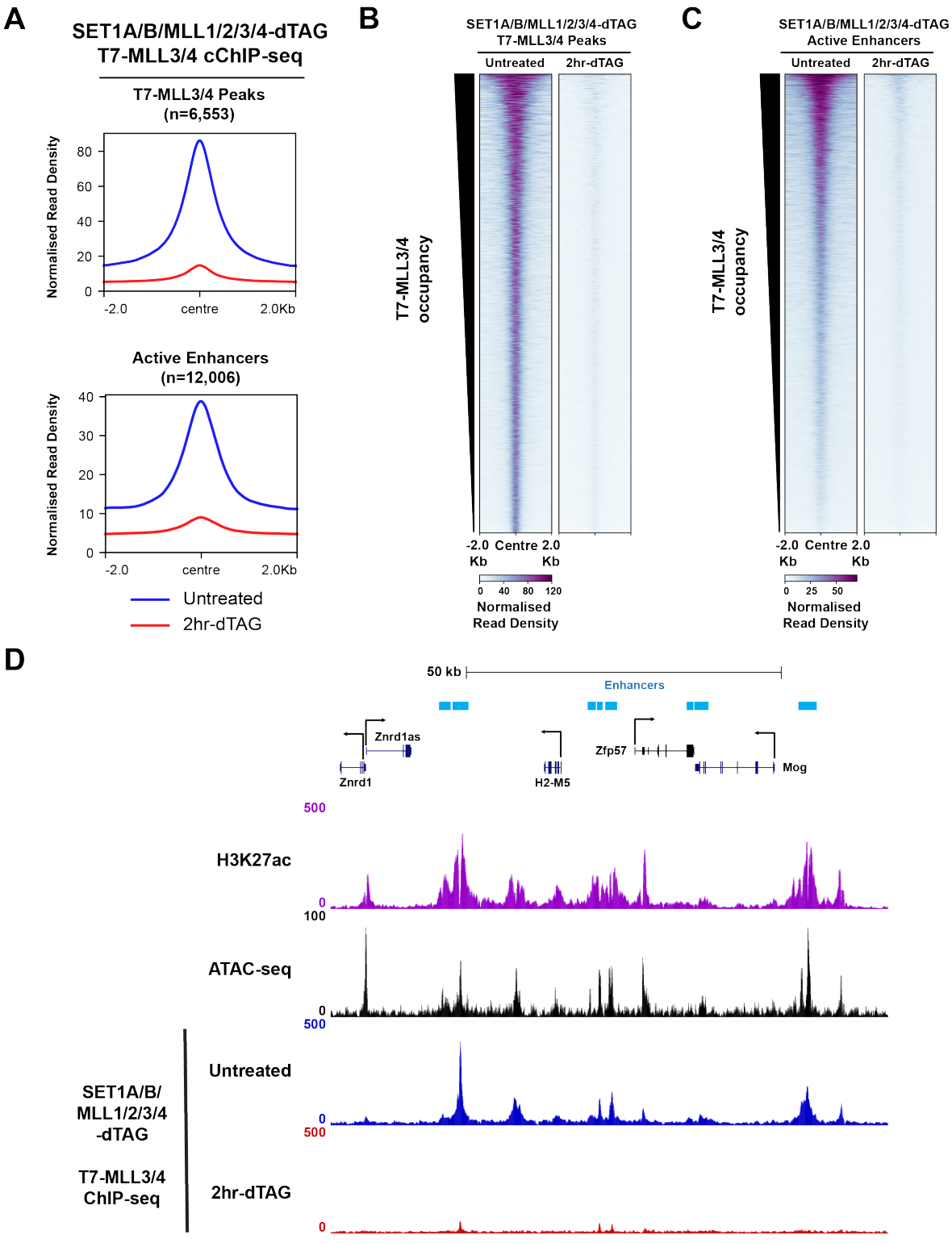

Figure 5.9: MLL3/4 are removed from chromatin following dTAG treatment in SET1A/B/MLL1/2/3/4-dTAG cells

(A) Metaplots showing T7-MLL3/4 cChIP-seq signal at T7-MLL3/4 peaks (n=6,553) (top) and active enhancers (n=12,006) (bottom).
(B-C) Heatmaps showing T7-MLL3/4 cChIP-seq signal at (B) T7-MLL3/4 peaks (n=6,553), and (C) active enhancers (n=12,006). Heatmaps are sorted in descending order by T7-MLL3/4 cChIP-seq signal.
(D) Genome coverage track of H3K27ac, ATAC-seq, and T7-MLL3/4 at a representative cluster of genes and active enhancers. H3K27ac and ATAC-seq tracks are taken from Fursova et al., 2021. T7-MLL3/4 cChIP-seq tracks are from untreated and 2hr-dTAG-treated SET1A/B/MLL1/2/3/4-dTAG cells.

I next decided to perform cChIP-seq for MLL1 and MLL2 in the SET1A/B/MLL1/2/3/4-dTAG cell line after 2 hours of dTAG treatment. Since MLL1 and MLL2 were not epitope-tagged with the exception of the $FKBP12^{F36V}$ tag, for which no high-quality commercially available antibodies are available, I decided to use antibodies against endogeneous MLL1 and MLL2. Unfortunately, commercially available antibodies against MLL1-C did not yield ChIP signal with adequate signal-noise ratio. Therefore, I decided to generate a cell line in which MLL1-C was tagged with 3xT7 and $FKBP12^{F36V}$ using the same internal tagging strategy described previously, but cChIP-seq using the T7 antibody also did not produced ChIP-seq profiles of sufficient quality. Fortunately, I was able to produce high-quality ChIP-seq data using our commercially-available antibody against MLL2-C (Fig. 5.10). As expected, MLL2 binding was highly correlated with CpG islands and was found primarily at promoters (Fig. 5.10A). Accordingly, across all promoters and all CGIs, MLL2-C signal was virtually undetectable after 2 hours of dTAG treatment (Fig. 5.10A, B, C, D). Therefore, cChIP-seq demonstrates that MLL2-C is largely liberated from chromatin after dTAG treatment.

With the exception of MLL1 and SET1B, I have confirmed that all SET1/MLL proteins can be robustly removed from chromatin after 2 hours of dTAG treatment. While I cannot confirm that MLL1 is fully removed from chromatin after dTAG treatment, I note that constitutive knockout of MLL1 in mESCs has yielded very mild reductions in H3K4me3, and is therefore unlikely to account for the H3K4 methylation seen after SET1/MLL-depletion (Denissov et al., 2014). It is most likely a novel methyltransferase activity, that is dependent on WRAD, that accounts for the SET1/MLL-independent H3K4 methylation.

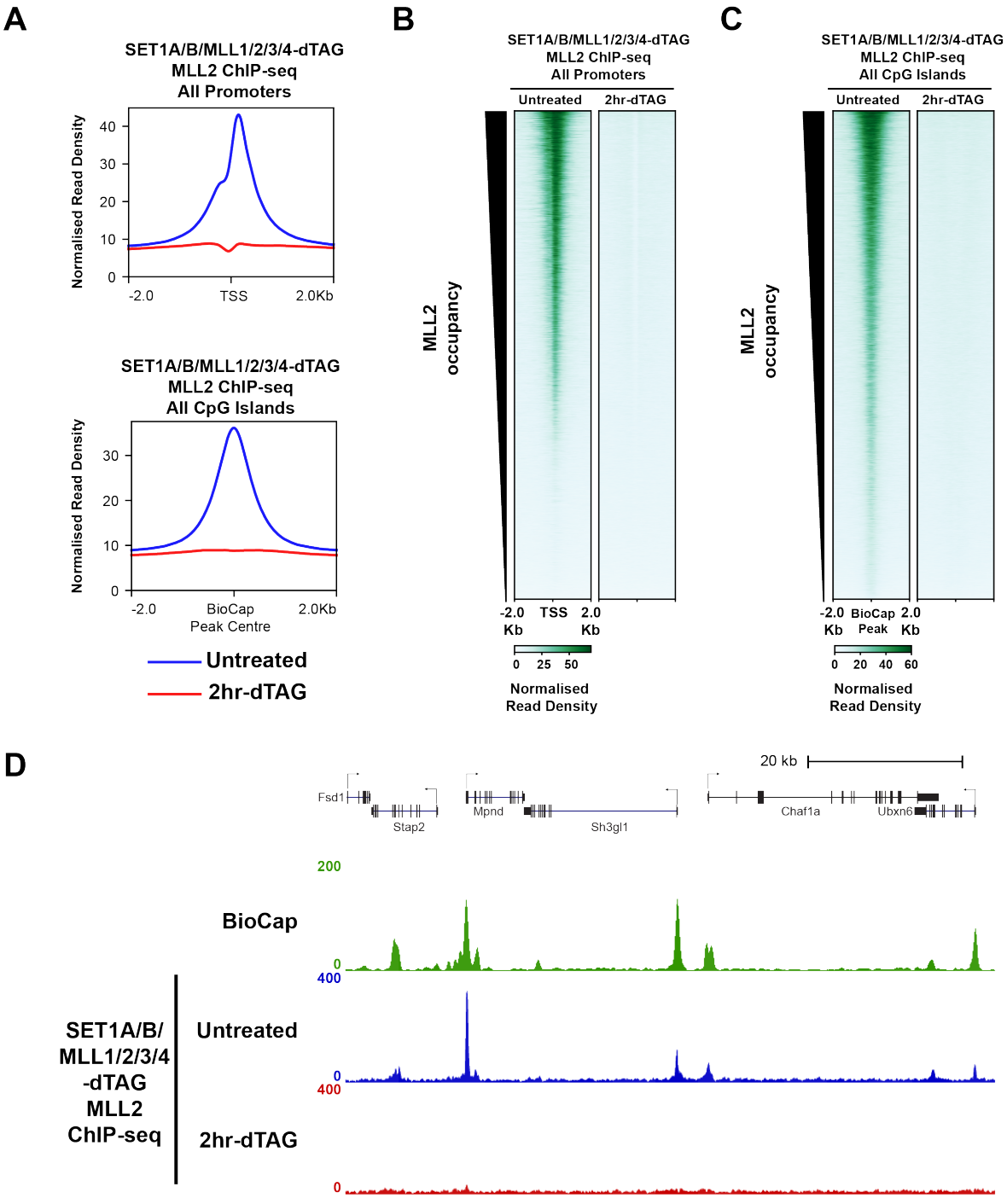

Figure 5.10: MLL2 is removed from chromatin after dTAG treatment of SET1A/B/MLL1/2/3/4-dTAG cells

(A) Metaplots of MLL2 cChIP-seq signal at all promoters (n=20,633) and CpG islands (n=27,047) after 2 hours of dTAG treatment in SET1A/B/MLL1/2/3/4-dTAG cells.
(B-C) Heatmaps of MLL2 cChIP-seq signal at (B) all promoters and (C) all CpG islands after 2 hours of dTAG treatment in SET1A/B/MLL1/2/3/4-dTAG cells. Heatmaps are sorted in descending order by MLL2 cChIP-seq signal.
(D) Genome coverage track of MLL2 cChIP-seq in SET1A/B/MLL1/2/3/4-dTAG cells and BioCap in wild-type cells at a representative gene cluster. BioCap data is taken from Long et al., 2013.

5.2.3 The SET1/MLL-independent, WRAD-containing complex may interact with chromatin in a highly-dynamic manner

Having demonstrated that a WRAD-containing complex existed in the cell after SET1A/B/MLL1/2/3/4 were removed, I sought to understand if this complex remained bound on chromatin by performing double-crosslinked cChIP-seq for RBBP5. I chose RBBP5 as a proxy for this WRAD-containing complex because purifications of RBBP5 from mammalian cells have indicated that it is not part of any other complex other than WRAD (van Nuland et al., 2013, Alsulami et al., 2019). In contrast, WDR5 and DPY30 are both components of other chromatin-associated complexes, and ASH2L has been reported to associated with chromatin independently of WRAD through MYC (van Nuland et al., 2013, Ullius et al., 2014). Therefore, I reasoned that RBBP5 ChIP-seq peaks would be representative of WRAD binding on chromatin.

To investigate if WRAD was still bound to chromatin after SET1/MLL proteins were depleted, I performed cChIP-seq for RBBP5 in SET1A/B/MLL1/2/3/4-dTAG cells after 2 hours and 24 hours of dTAG treatment.As expected, RBBP5 binding had a strong preference for CpG island promoters (Fig. 5.11A, C) At promoters, RBBP5 binding was highly correlated with H3K4me3, with promoters having strong enrichment of H3K4me3 generally having stronger RBBP5 ChIP-seq signal (Fig. 5.11A). Accordingly, RBBP5 binding was also highly correlated with RNA polymerase II, SET1A, and MLL2 binding. Furthermore, at enhancers, RBBP5 binding correlated strongly with MLL3/4 binding (Fig. 5.11B, C). Taken together, these observations confirm that RBBP5 ChIP-seq signal largely represents SET1/MLL occupancy on chromatin.

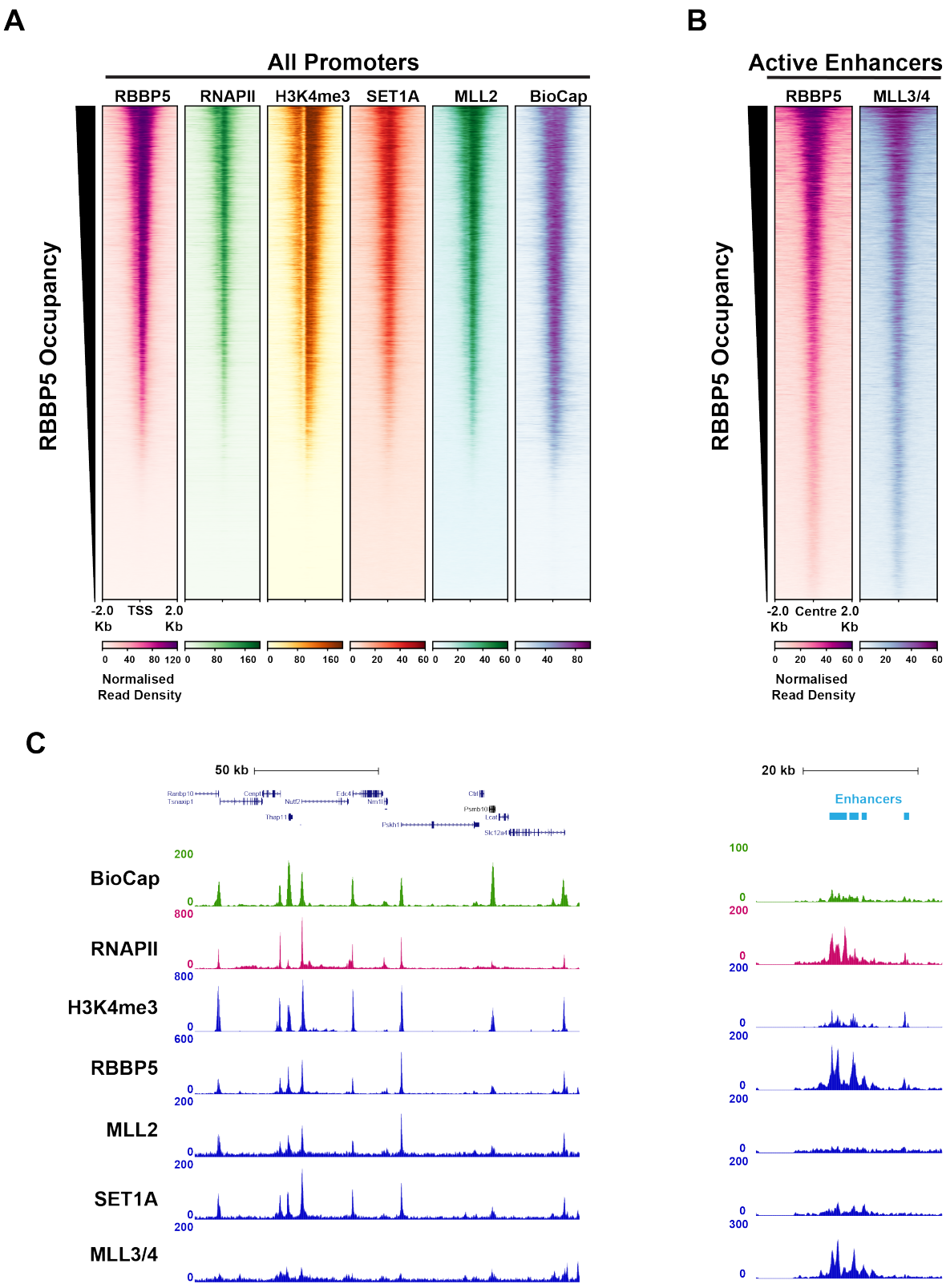

Figure 5.11: RBBP5 occupancy correlates with SET1/MLL occupancy

(A) Heatmaps illustrating RBBP5, RNAPII, H3K4me3, SET1A, MLL2 ChIP-seq, and BioCap signal at all promoters, sorted in descending order by RBBP5 cChIP-seq signal (n=20,633).
(B) Heatmaps illustrating RBBP5 and MLL3/4 ChIP-seq signal at active enhancers (n=12,006), sorted in descending order by RBBP5 cChIP-seq signal.
(C) Representative genome coverage tracks of BioCap data, RNAPII, H3K4me3, RBBP5, MLL2, SET1A, and MLL3/4 ChIP-seq data at a gene cluster (left) and an enhancer cluster (right).

Interestingly, RBBP5 was almost completely removed from chromatin after 2 hours of SET1/MLL-depletion, and remained depleted from chromatin after 24 hours (Fig. 5.12A). This was surprising because immunoblotting of nuclear extracts demonstrate that although levels of RBBP5 are reduced after 2 hours of SET1/MLL-depletion, RBBP5 is still clearly detectable, and levels do not appear to decline further after prolonged SET1/MLL-depletion (see earlier, Fig. 5.6). In contrast, RBBP5 was virtually undetectable across all RBBP5 peaks, all promoters, CGIs, and enhancers (Fig. 5.12A, B). This indicates that the RBBP5 that remains in the nucleus after SET1/MLL-depletion is not stably bound to chromatin, at least as measured by ChIP.

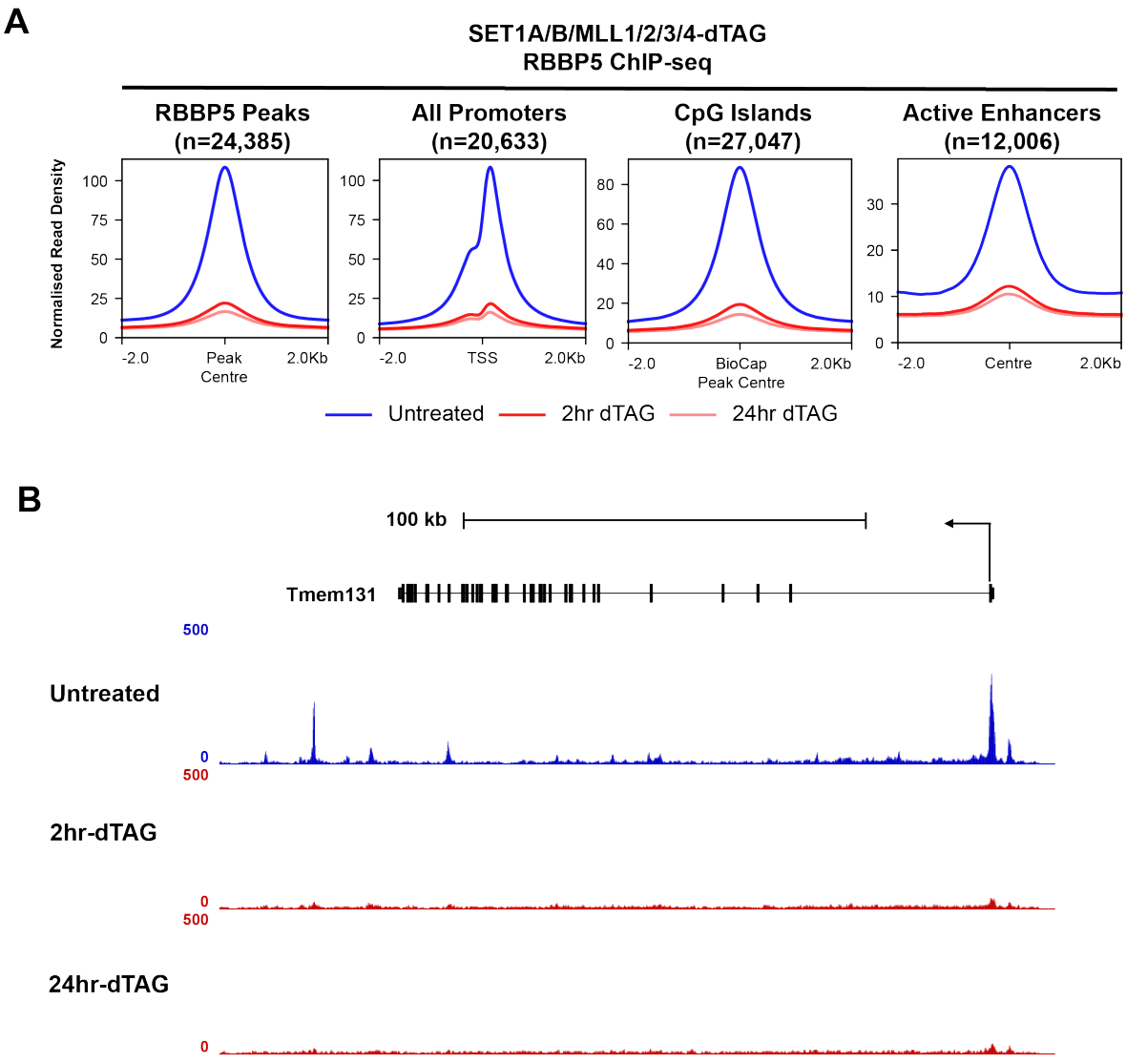

Figure 5.12: RBBP5 is removed from chromatin after SET1/MLL-depletion

(A) Metaplots showing normalised RBBP5 cChIP-seq signal at RBBP5 peaks, promoters, CpG islands, and active enhancers after 2 hours and 24 hours of SET1/MLL-depletion.
(B) Genome coverage tracks showing RBBP5 at an example gene (*Tmem131*) after 2 hours and 24 hours of SET1/MLL-depletion

Interestingly, I noted that RBBP5 ChIP-seq peaks at enhancers tended to be centered around ATAC-seq peaks flanked by H3K4me1-enriched nucleosomes (Fig. 5.13A, B). This pattern was also observed for T7-MLL3/4 ChIP-seq peaks, as described in Chapter 4. This indicates that RBBP5 ChIP-seq peaks at enhancers largely captured the stable binding of MLL3/4 complexes at nucleosome depleted regions rather than the interaction of MLL3/4 on nucleosomes. Additionally, I found that although RBBP5 peaks correlated very well with H3K4me3 at promoters, broad domains of H3K4me1/2 that extend into the gene body did not have strong RBBP5 enrichment (Fig. 5.13C, D). These findings suggest that crosslinking ChIP-seq primarily captures binding sites where SET1/MLL complexes are stably bound to chromatin, and may not capture SET1/MLL complexes which are engaged in transient interactions with chromatin. Taken together, these findings suggest that the WRAD-containing complex that remains after SET1/MLL-depletion may be highly dynamic and only interacts with chromatin transiently to perform H3K4 methylation.

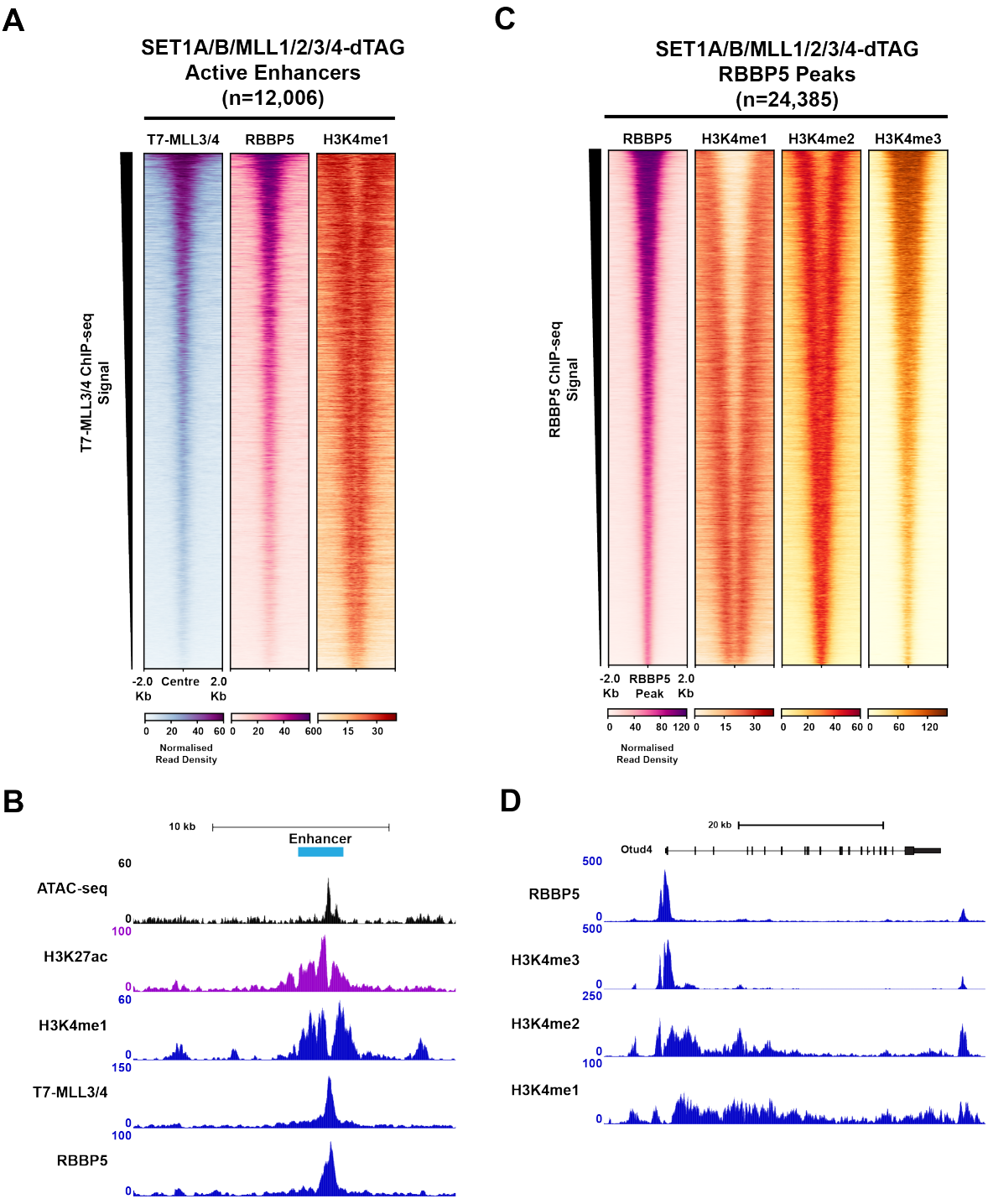

Figure 5.13: RBBP5 ChIP-seq only captures stable WRAD binding

(A) Heatmaps of RBBP5, T7-MLL3/4, and H3K4me1 at all active enhancers (n=12,006), sorted by T7-MLL3/4 cChIP-seq signal.
(B) Genome coverage track at a representative enhancer element showing ATAC-seq, H3K27ac, H3K4me1, T7-MLL3/4, and RBBP5 cChIP-seq. ATAC-seq and H3K27ac cChIP-seq are from Fursova et al., 2021).
(C) Heatmaps of RBBP5, H3K4me1/2/3 cChIP-seq at RBBP5 cChIP-seq peaks (n=24,385), sorted by RBBP5 cChIP-seq signal.
(D) Genome coverage track at a representative gene (*Otud4*) containing an RBBP5 peak showing RBBP5, H3K4me1/2/3.

5.2.4 The SET1/MLL-independent, WRAD-containing complex accounts for the vast majority of remaining H3K4 methylation

The finding that RBBP5 no longer stably bound to chromatin after SET1/MLL-depletion was surprising because high-level occupancy of the H3K4 methyltransferase was thought to be required for H3K4 trimethylation. For example, depletion of CFP1 largely removes SET1A from chromatin without depleting the overall SET1A protein levels, and leads to loss of H3K4me3 at SET1A-dependent genes (Brown et al., 2017). Therefore, I next sought to confirm that the SET1/MLL-independent H3K4 methyltransferase activity was dependent on WRAD by targeting another subunit of WRAD. At the beginning of this chapter, I suggested that there may be a WRAD-complex that regulates H3K4 methylation in the absence of SET1/MLL proteins because removing WDR5 in addition to removing SET1/MLL proteins led to greater reductions in H3K4 methylation than removing SET1/MLL proteins on their own. However, WDR5 is a highly abundant protein that is part of multiple chromatin-modifying complexes, including SET1/MLL complexes, the ATAC-complex, the NSL complex, a mESC-specific NuRD complex, and the PCGF6-containing variant PRC1 complex (Guarnaccia and Tansey, 2018). WDR5 has also been reported to interact with MYC and components of the TFIID general transcription factor (Guarnaccia and Tansey, 2018, Oh et al., 2020, Thomas et al., 2015). It is therefore a possibility that it was the disruption of one of these complexes, rather than WRAD, that led to the greater reductions in H3K4me1/2/3 after SET1/MLL/WDR5-depletion.

In contrast to WDR5, RBBP5 has not be reported to be a component of other chromatin-associated complexes (van Nuland et al., 2013). Therefore, I next sought to confirm that the SET1/MLL-independent WRAD complex is indeed responsible for the remaining H3K4me1/2/3 by depleting RBBP5, instead of WDR5, in addition to SET1/MLL proteins (Fig. 5.14A). To do this, I endogeneously tagged RBBP5 at the C-terminus with $FKBP12^{F36V}$ in the SET1A/B/MLL1/2/3/4-dTAG cell line to generate a SET1A/B/MLL1/2/3/4/RBBP5-dTAG cell line in which SET1A/B/MLL1/2/3/4/RBBP5 can be depleted within 2 hours of dTAG treatment (Fig. 5.14B). Using this cell line, I analysed bulk levels of H3K4me1/2/3 after 2 hours and 24 hours of dTAG treament (Fig. 5.14C). Indeed, depleting RBBP5 in addition to SET1/MLL poteins produced the same effect on H3K4 methylation as removing WDR5 in addition to SET1/MLL proteins. After 2 hours of dTAG treatment, H3K4me2/3 were reduced to 25% of untreated levels, and after 24 hours, were reduced to less than 10%. Interestingly, removing RBBP5 instead of WDR5 in addition to SET1/MLL proteins had a greater effect on H3K4me1 levels. As expected, H3K4me1 levels did not change appreciably after 2 hours of dTAG treatment, but after 24 hours of dTAG treatment, levels were reduced to less than 10%, suggesting that RBBP5 has a greater role in the H3K4-monomethyltransferase activity of this WRAD complex. Taken together, these results indicate that a SET1/MLL-independent WRAD complex accounts for the vast majority of H3K4 methylation remaining after SET1/MLL-depletion, and that WRAD-containing complexes account for virtually all H3K4 methylation in mESCs.

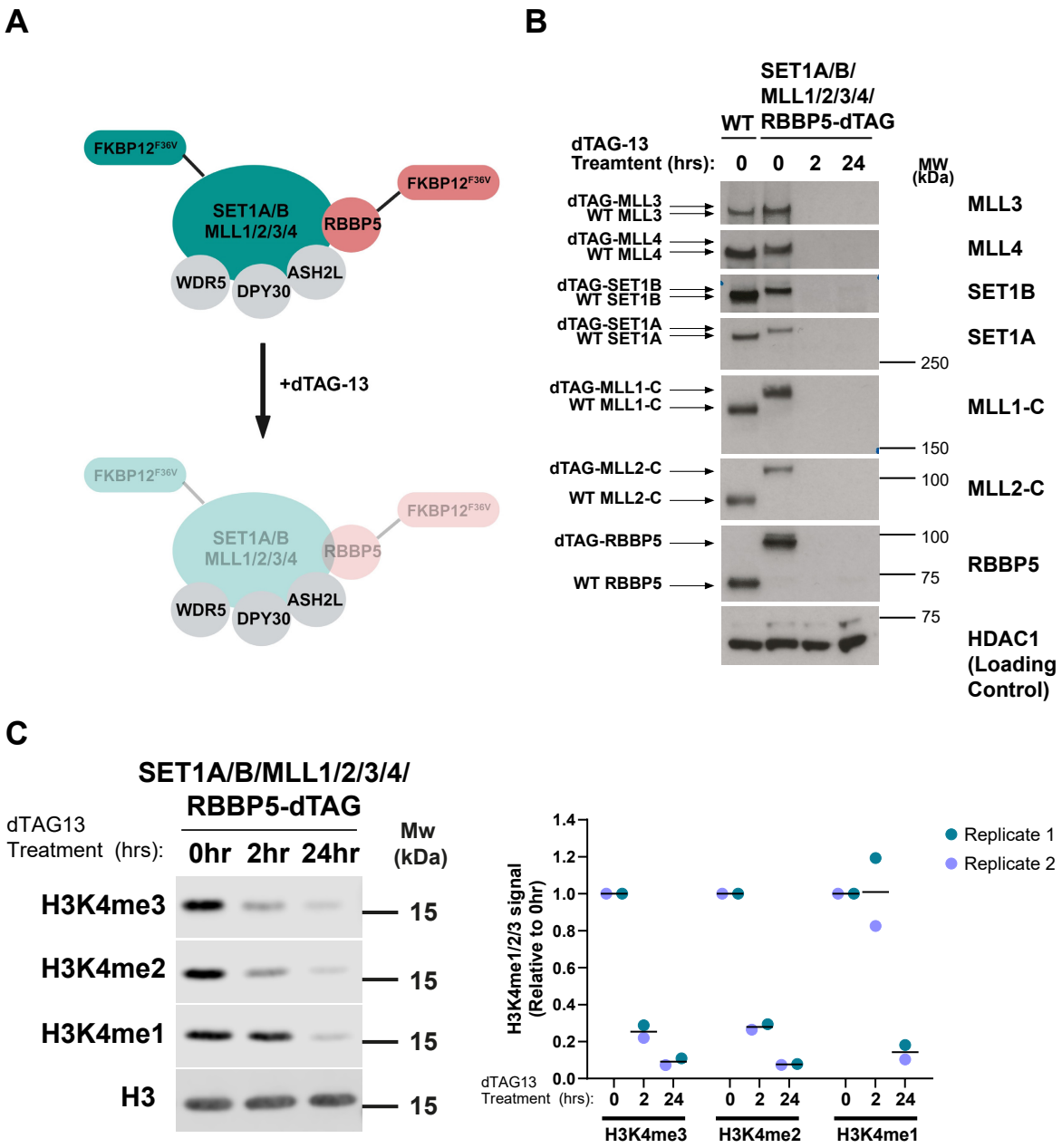

Figure 5.14: RBBP5 depletion in addition to SET1/MLL depletion results in near-complete loss of H3K4 methylation
(A) Cartoon illustrating schematic of SET1A/B/MLL1/2/3/4/RBBP5-depletion by dTAG.
(B) Western blots demonstrating robust depletion of SET1/MLL proteins and RBBP5 after dTAG treatment. HDAC1 is used as a loading control.
(C) Immunoblotting of H3K4me1/2/3 from histone extracts of SET1A/B/MLL1/2/3/4/RBBP5-dTAG cells following dTAG-13 treatment. Histone H3 was used as a loading control. Quantitative analysis of bulk H3K4me1/2/3 levels are shown on the right. H3K4me1/2/3 signal was normalised to total H3 signal and shown as relative to the untreated sample. Individual values from 2 biological replicates are plotted, along with a line showing the mean of the two values.

Having found that removing RBBP5 instead of WDR5 in addition to SET1/MLL proteins appeared to have a greater effect on H3K4me1, I next sought to understand whether RBBP5 played a greater role than WDR5 in regulating the methyltransferase activity of WRAD-containing complexes. To do this, I generated a cell line in which RBBP5 could be depleted on its own, and analysed bulk levels of H3K4 methlation after 2 hours and 24 hours of dTAG treatment (Fig. 5.15A, B, C). Surprisingly, I found that removing RBBP5 on its own

had virtually the same effect on H3K4 methylation as removing RBBP5 and SET1/MLL proteins simultaneously (Fig. 5.15C). H3K4me2/3 was reduced to 25% after 2 hours of dTAG treatment and subsequently to less than 10% after 24 hours, and H3K4me1 was reduced to less than 10% after 24 hours of dTAG treatment with little change after 2 hours. This indicates that removing RBBP5 completely removes both the methyltransferase activity of SET1/MLL proteins along with the methyltransferase activity of the SET1/MLL-independent methyltransferase. Taken together, these findings indicate that RBBP5 is the central determinant of H3K4 methylation in mESCs.

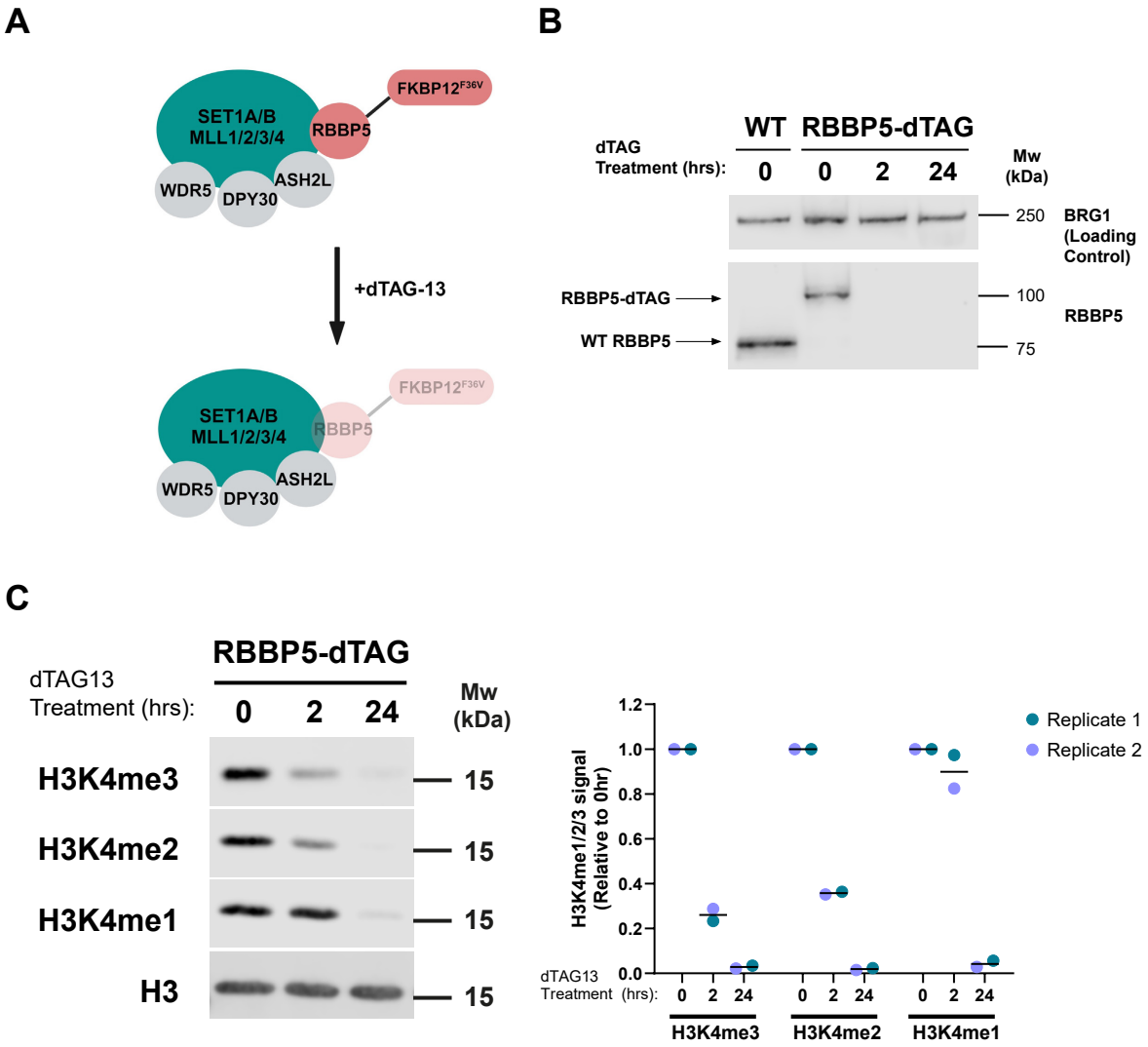

Figure 5.15: RBBP5 is the central determinant of H3K4 methylation in mESCs

(A) Cartoon illustrating depletion of RBBP5 by dTAG.
(B) Western blotting illustrating robust depletion of RBBP5 by dTAG. BRG1 is used as a loading control.
(C) Immunoblotting of H3K4me1/2/3 from histone extracts of RBBP5-dTAG cells following dTAG-13 treatment. Histone H3 was used as a loading control. Quantitative analysis of bulk H3K4me1/2/3 levels are shown on the right. H3K4me1/2/3 signal was normalised to total H3 signal and shown as relative to the untreated sample. Individual values from 2 biological replicates are plotted, along with a line showing the mean of the two values.

5.3 Identification of a putative alternative SET1B isoform

5.3.1 Analysing the WDR5 interactome

I next sought to identify the methyltransferase responsible for the SET1/MLL-independent H3K4 methylation. Given that this methyltransferase interacts with WDR5 and RBBP5 and requires these two proteins for methyltransferase activity, I sought to identify this protein by immunoprecipitating WDR5 from SET1A/B/MLL1/2/3/4-dTAG cells after dTAG treatment and identifying interacting proteins by mass spectrometry (Fig. 5.16A). I chose WDR5 because it is an abundant nuclear protein and because WDR5 protein levels do not decrease after SET1/MLL depletion, in contrast to RBBP5 (Fig. 5.6). Additionally, WDR5 is part of a multitude of other protein complexes in the nucleus which can provide a background for quantitative comparisons of the WDR5 interactome before and after SET1/MLL depletion. In order to efficiently purify endogenous WDR5 from nuclear extracts and reduce non-specific immunoprecipitation, I decided to tag endogenous WDR5 with a 3xFLAG tag at the C-terminus in the SET1A/B/MLL1/2/3/4-dTAG cell line. The resulting SET1A/B/MLL1/2/3/4-dTAG-WDR5-FLAG cell line would allow for the depletion of SET1/MLL proteins and subsequent purification of WDR5 using an anti-FLAG affinity gel. Immunoblotting confirmed that SET1/MLL complex components could be recovered following immunoprecipitation of WDR5 using anti-FLAG affinity gel from this cell line (Fig. 5.16B). A control FLAG immunoprecipitation of WDR5 from the parental SET1A/B/MLL1/2/3/4-dTAG cell line, in which WDR5 is not tagged, indicated that the FLAG immunoprecipitation was highly-specific (Fig. 5.16B).

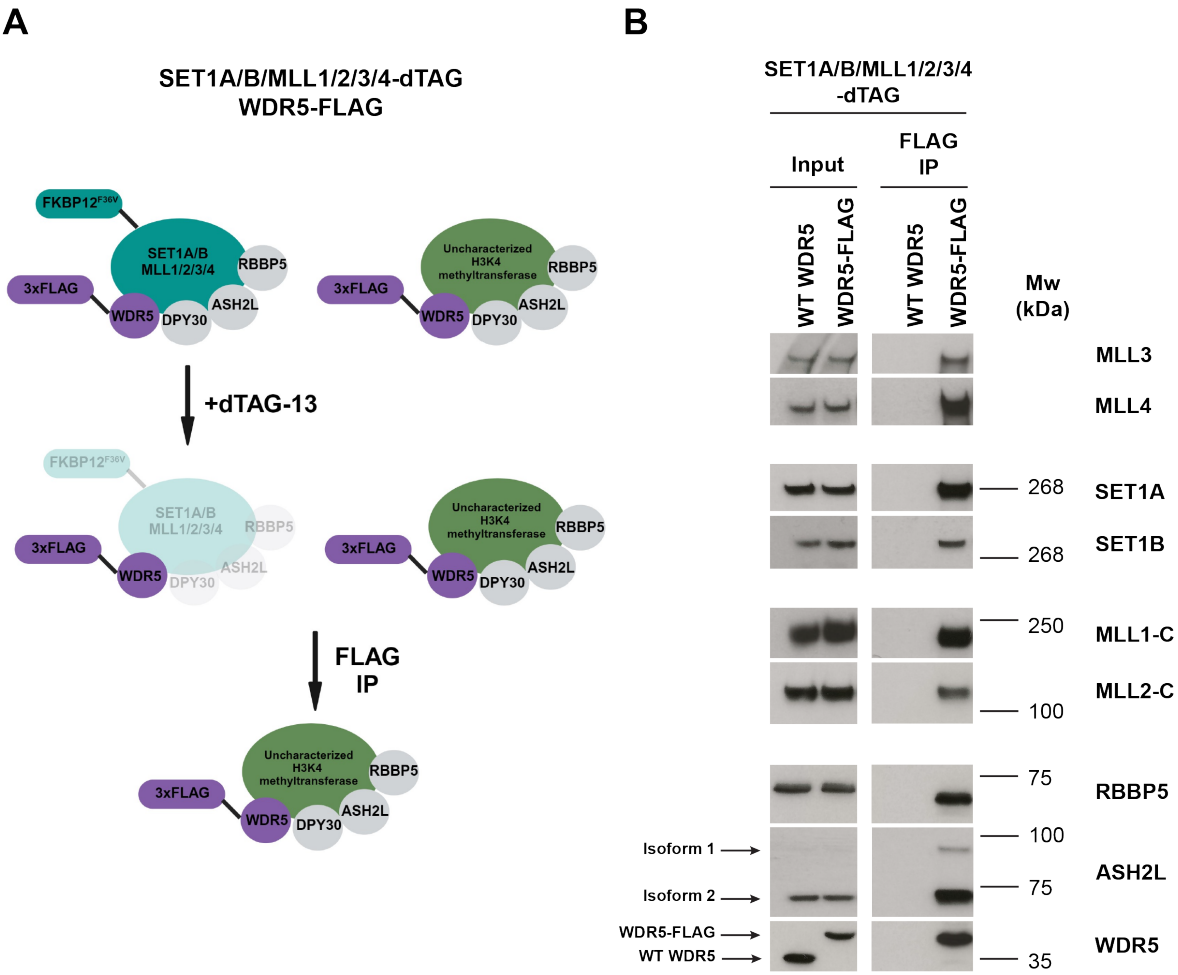

Figure 5.16: Generating a SET1A/B/MLL1/2/3/4-dTAG-WDR5-FLAG cell line

(A) Cartoon illustrating WDR5-FLAG IP using a SET1A/B/MLL1/2/3/4-dTAG-WDR5-FLAG cell line
(B) Western blots illustrating successful immunoprecipitation of SET1/MLL complex components after FLAG IP.

Having validated the purification of WDR5 using the 3xFLAG tag, I next purified WDR5 from nuclear extracts of untreated cells and cells after 2 hours and 24 hours of SET1/MLL depletion, and subjected the immunoprecipitated proteins to mass spectrometry (Liquid chromatography tandem mass spectrometry and the subsequent data analysis was performed by Inge de Krijger). I decided to analyze the WDR5 interactome after 2 hours and 24 hours of SET1/MLL depletion in order to filter out false positives; given that the SET1/MLL-independent H3K4 methyltransferase is active after prolonged SET1/MLL depletion, only proteins that interacted with WDR5 after both 2 hours and 24 hours of SET1/MLL depletion would be considered. Additionally, I performed FLAG immunoprecipitations from the parental SET1A/B/MLL1/2/3/4-dTAG cell line, in which WDR5 is not tagged, as a negative control; only proteins with a significant enrichment over the negative control would be considered a valid interactor.

An analysis of the WDR5 interactome in untreated cells demonstrated that the majority of known WDR5 interactors were identified by mass spectrometry (Fig. 5.17). In total, 144 proteins were identified as WDR5 interactors. All six SET1/MLL proteins and all four WRAD subunits were identified, in addition to complex-specific interactors, such as WDR82, BOD1L and CFP1 (CXXC1) for SET1A/B complexes, MENIN and

LEDGF for MLL1/2 complexes, and UTX (KDM6A), NCOA6, PTIP, and PA1 (PAGR1A) for MLL3/4 complexes. Additionally, other known SET1/MLL-complex interactors such as HCFC1/2 and OGT were also identified. Interestingly, BOD1, which is a SET1B-specific interactor, was not identified. BOD1 is a small protein of only 173 amino acids and has been reported to be difficult to identify by mass spectrometry due to its small size and relatively low abundance, therefore, its omission could possibly be attributed to technical limitations (Ciotta et al., 2023). Nonetheless, almost all previously reported SET1/MLL complex subunits and SET1/MLL-interactors were identified, which substantiates the notion that WDR5 is an integral component of all SET1/MLL complexes. Furthermore, components of other WDR5-containing complexes were also identified. These include subunits of the non-specific lethal (NSL) complex, PCGF6-PRC1 complex, NuRD complex, WHHERE complex, ATAC complex, and SAGA complex, which have all been reported in other studies (Appendices Table A.1)(Oh et al., 2020, Fischer et al., 2021, Vilhais-Neto et al., 2017, Bode et al., 2016, Guarnaccia et al., 2021). Taken together, these observations demonstrate that this WDR5 purification comprehensively captures the WDR5 interactome.

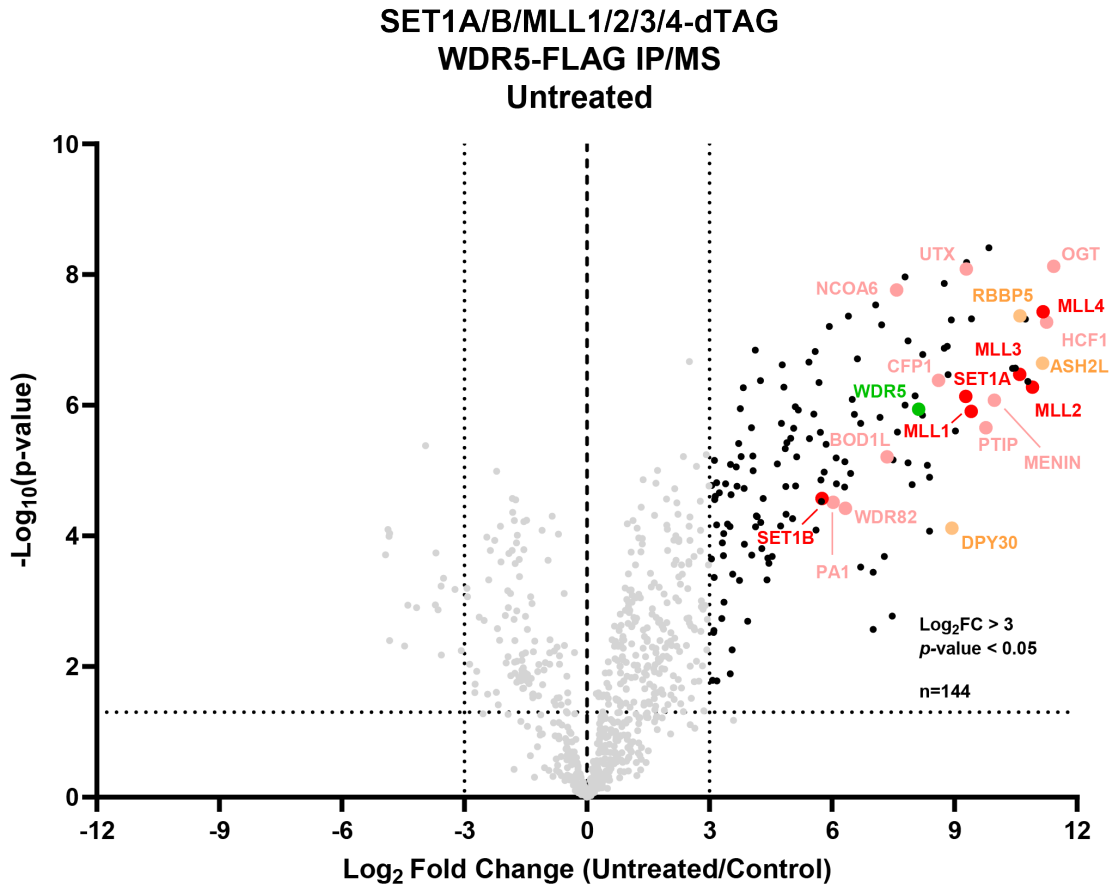

Figure 5.17: WDR5-FLAG immunoprecipitation followed by mass spectrometry identifies known WDR5 interactors.
Volcano plot showing all proteins identified as interacting with WDR5 in untreated SET1A/B/MLL1/2/3/4-dTAG-WDR5-FLAG cells. A significance threshold of p<0.05 and a cut-off of LFC > 3 was applied to identify proteins that were enriched in the FLAG-IP sample relative to the mock IP control. SET1/MLL proteins are highlighted in red, WRAD components are highlighted in yellow, WDR5 is highlighted in green, and complex-specific interactors are highlighted in pink.

5.3.2 SET1B continues to interact with WDR5 after SET1/MLL-depletion

I next compared the WDR5-interactome after 2 hours and 24 hours of SET1/MLL depletion (Fig. 5.18). As expected, the interactions between WDR5 and the remaining subunits of WRAD, as well as CFP1, were not significantly reduced after SET1/MLL depletion, confirming my previous findings by co-IP (Fig. 5.8, Fig. 5.18). The interactions between WDR5 and five of the six methyltransferases, SET1A and MLL1-4, were significantly reduced after dTAG treatment (Fig. 5.18). Similarly, the interactions between WDR5 and many complex-specific interactors were also significantly reduced. For example, none of the four MLL3/4-specific interactors, UTX, NCOA6, PTIP, and PA1 continued to interact with WDR5 after dTAG treatment. Similarly, Menin, an MLL1/2-specific interactor, no longer interacted with WDR5 after dTAG treatment. The only SET1A-specific interactor, BOD1L, also no longer interacted with WDR5 after dTAG treatment. Interestingly, I noted that the interaction between WDR5 and WDR82 was significantly reduced after 2 hours of dTAG treatment, although this reduction dropped below the $\log_2$ fold change > 3 cut-off after 24 hours of dTAG treatment (Fig. 5.18).

Surprisingly, the interaction between WDR5 and SET1B was not significantly reduced after 2 hours and 24 hours of dTAG treatment (Fig. 5.18). In fact, the interaction between SET1B and WDR5 appeared to be slightly increased in the dTAG-treated conditions, although this was well below the significance cut off of $\text{Log}_2\text{FC} < 3$.

5.3.3 A putative alternative C-terminal SET1B isoform

The interaction between SET1B and WDR5 after dTAG treatment is surprising because SET1B is undetectable from nuclear extracts by western blotting after dTAG treatment (Fig. 3.1, 3.13, 4.6). Additionally, previous work in the Klose lab had demonstrated that dTAG-mediated depletion of SET1B leads to greater reductions in gene expression, both in the number of genes and the magnitude of the reductions, compared to depleting SET1A or SET1B on their own (Hughes et al., 2023). This SET1B-mediated effect on gene expression only requires a short N-terminal fragment which encompasses the WDR82 interaction region; tethering this N-terminal fragment to a luciferase reporter gene is sufficient to induce its expression *in vivo*. In contrast, SET1B interacts with WDR5 through the WIN motif located in the n-SET domain, which is located near the C-terminus of the protein. Furthermore, CFP1 also interacts with SET1B via the n-SET domain. Interestingly, I noted that the WDR5-WDR82 interaction was reduced after dTAG treatment, although this reductions missed the fold change cut-off in the 24 hours dTAG-treated condition (Fig. 5.18). Therefore, I hypothesized that there may be a previously uncharacterized isoform of SET1B which included the C-terminal n-SET and SET domains but did not include the N-terminal region of SET1B. Given that the $\text{FKBP12}^{\text{F36V}}$ tag is fused to the N-terminus of SET1B, this hypothetical C-terminal isoform would not contain the $\text{FKBP12}^{\text{F36V}}$ tag and therefore not be susceptible to dTAG-mediated degradation. Additionally, the antibody used for immunoblotting SET1B was raised using a peptide surrounding Val208 of human SET1B, located near the N-terminus of SET1B. Therefore, immunoblotting of SET1B using this antibody would only detect the SET1B isoform which contained the N-terminal region of the protein.

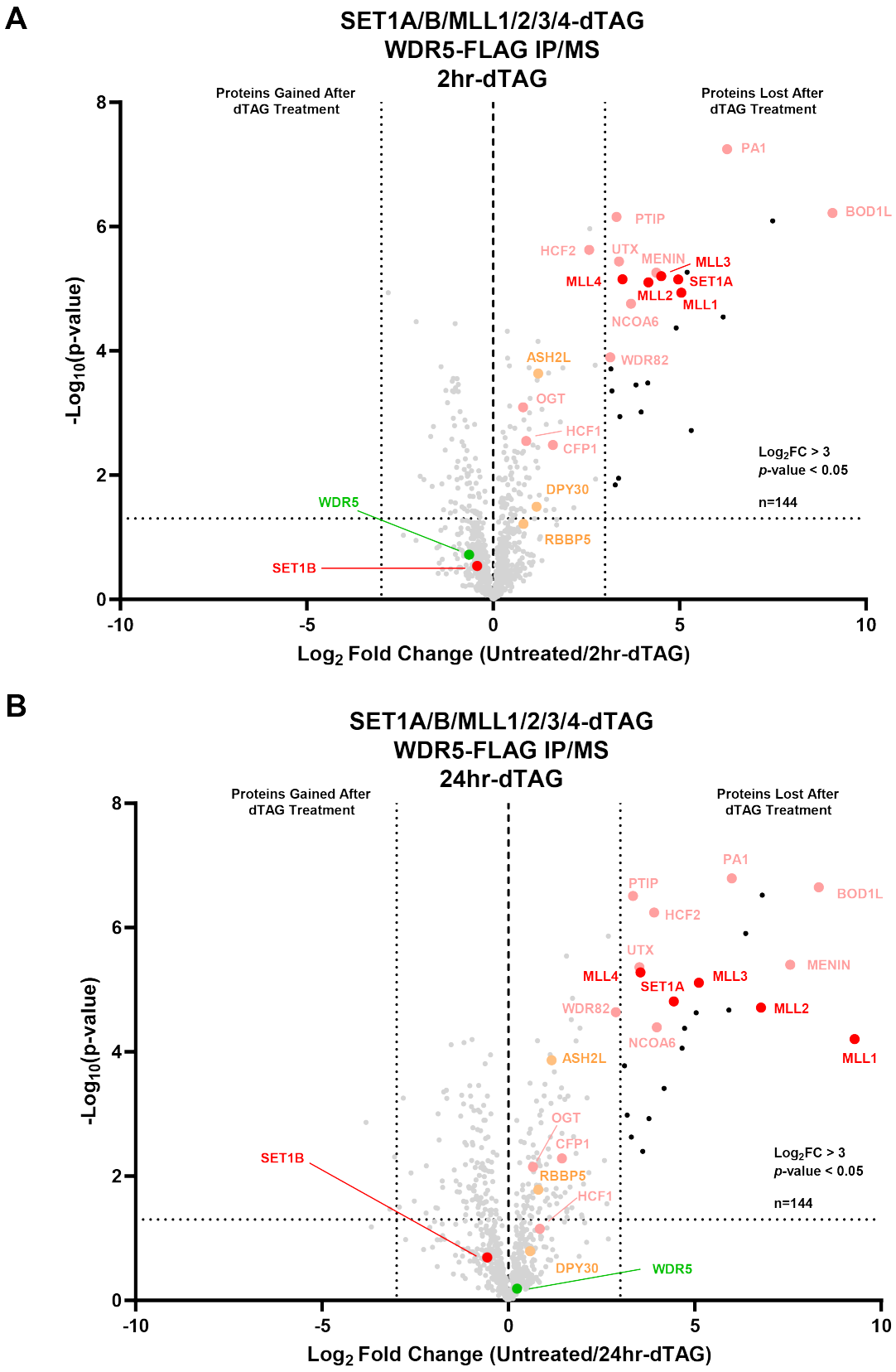

Figure 5.18: SET1B continues to interact with WDR5 after dTAG treatment.
Volcano plot showing all WDR5-interacting proteins that are differentially enriched in untreated cells compared to cells treated with dTAG-13 for **(A)** 2 hours and **(B)** 24 hours. A significance threshold of $p < 0.05$ and a cut-off of LFC > 3 was applied to identify proteins that were enriched in the untreated sample relative to the dTAG-treated sample. SET1/MLL proteins are highlighted in red, WRAD components are highlighted in yellow, WDR5 is highlighted in green, and complex-specific interactors are highlighted in pink.

In order to understand whether the SET1B-WDR5 interaction detected after dTAG treatment by mass spectrometry was between WDR5 and a truncated C-terminal isoform of SET1B, we mapped the identified SET1B peptides to the SET1B amino acid sequence (Fig. 5.19). Indeed, we found that all SET1B peptides identified in the WDR5-IP/MS experiment were localized to a C-terminal region of SET1B between amino acids 1382-1970, encompassing the n-SET and SET domains (Fig. 5.19). Surprisingly, however, no peptides mapping beyond this C-terminal region were identified even in the untreated condition, with the same peptides identified in both the untreated and dTAG-treated contitions. This is surprising because full-length SET1B was clearly identified by immunoblotting of WDR5-IP immunoprecipitates using the SET1B antibody which was raised against an N-terminal peptide (Fig. 5.16B). Additionally, there are enough trypsin cleavage sites in SET1B to yield peptides covering the N-terminal region of SET1B; indeed, analysis of an independent mouse whole proteome dataset indicates that SET1B peptides from the N-terminus and C-terminus are recovered (Inge de Krijger, Michiel Vermeulen, unpublished data). Furthermore, SET1A, which shares high sequence similarity with SET1B, yielded peptides which encompassed the entirety of the protein. A possible explanation for this discrepancy is that the abundance of the full length SET1B is too low to be detected by mass spectrometry. Interestingly, the Klose lab had previously analysed the interactome of WDR82 by mass spectrometry and failed to detect SET1B, despite identifying most known WDR82 interactors including SET1A (Jessica Kelley and Rob Klose, unpublished data). Given that WDR82 only interacts with the N-terminus of SET1B, only full-length SET1B would be recovered in WDR82 purifications. Therefore, the failure to identify SET1B in WDR82 purifications may indicate that full-length SET1B is not present at high levels in the nucleus. This also corroborates with the observation that the WDR5-WDR82 interaction is largely lost after SET1/MLL-depletion.

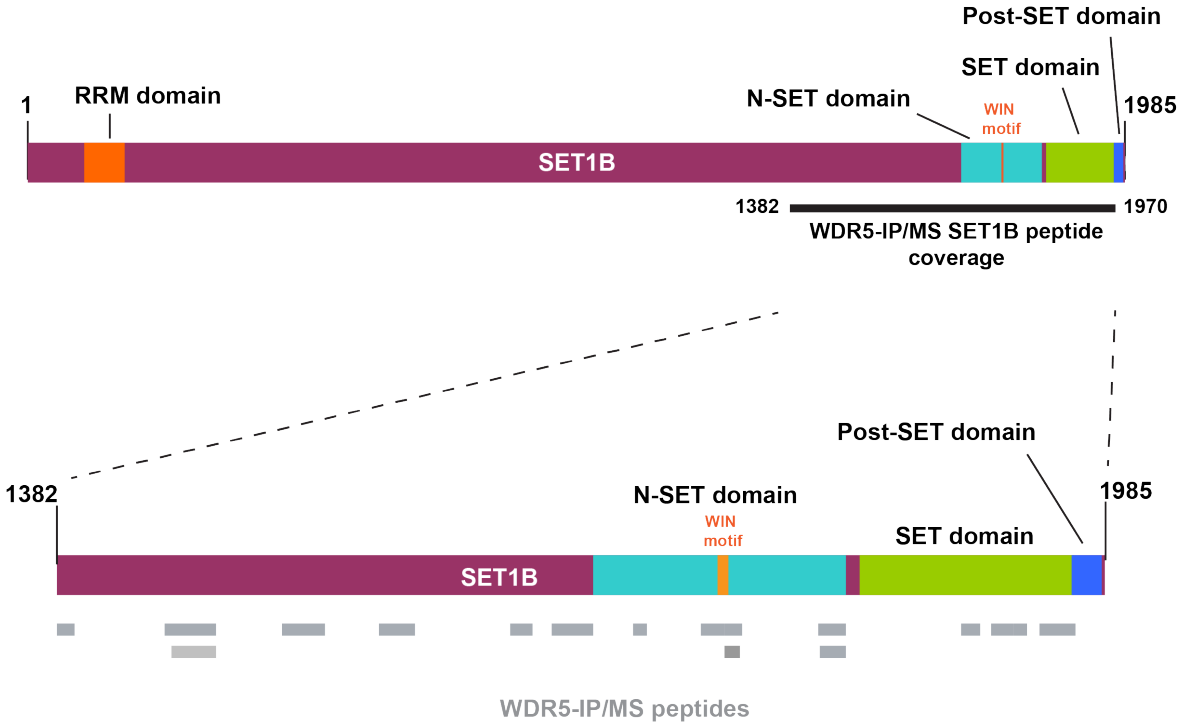

Figure 5.19: WDR5-IP/MS peptdies map to a C-terminal region of SET1B

Schematic of the SET1B protein with characterized domains annotated. Also shown is a zoomed-in view of the C-terminal region to which peptides from the WDR5-FLAG IP/MS experiments mapped. Peptides are shown as grey boxes.

I next sought to understand how such an alternative SET1B isoform would be produced from the *Setd1b* gene. I reasoned that the most likely explanation for an alternative SET1B isoform would be the existence of an uncharacterized alternative transcription start site within the *Setd1b* gene that would result in the production of a second transcript encompassing exons that code for the C-terminal alternative isoform. To investigate this possibility, I analysed the distribution of histone modifications and occupancy of the transcriptional machinery at the *Setd1b* gene (Fig. 5.20A, B). I first investigated ongoing transcription at the *Setd1b* gene using a TT-seq dataset from untreated SET1A/B-dTAG cells (Hughes et al., 2023). Interestingly, I noted that the TT-seq read density at the 3' half of the gene appeared to be higher than the 5' half. This 3' half region began just upstream of Exon 11 and continued to the 3' end of the gene (Fig. 5.20B). Furthermore, I noted that there was significant antisense transcription initiating just upstream of this region. Eukaryotic protein-coding genes are typically bidirectional, with a short antisense transcript produced by a separate RNAPII PIC that is then rapidly degraded (Haberle and Stark, 2018). Therefore, I hypothesised that this was an alternative transcription start site from which an alternative C-terminal SET1B isoform could be produced. I next examined a publicly-available CAGE-seq data set from OS25 mESCs (Forrest et al., 2014). CAGE (Cap analysis gene expression) is a method that specifically detects the capped 5' ends of mRNAs and is used to map transcription start sites. I noted the presence of a small CAGE-seq peak at this putative alternative transcription start site, further suggesting that this is a site of transcription initation. I next examined RNAPII ChIP-seq data from untreated SET1A/B-dTAG cells, but did not observe an RNAPII peak at this alternative transcription start site, as would be expected.

However, I noted that the enrichment of RNAPII ChIP-seq signal is high in the gene body of the full length *Setd1b* gene; therefore a small peak of RNAPII at the alternative TSS may have been obscured by this broad enrichment (Fig. 5.20B). However, I noted that there was an ATAC-seq peak and a MED14 ChIP-seq peak at this putative alternative TSS, indicating that there is a nucleosome-depleted region and binding of the Mediator coactivator complex at this site. Additionally, I noted that there was elevated H3K27ac enrichment flanking the ATAC-seq peak and the MED14 ChIP-seq peak. H3K27ac enrichment, a nucleosome depleted region, and Mediator occupancy are seen at both promoters enhancers. Therefore, I sought to differentiate between the two possibilities by examining H3K4me3 enrichment, which is high at promoters but very low at enhancers. Indeed, the nucleosomes surrounding the alternative TSS were enriched in H3K4me3. Interestingly, using my previously-presented dataset of H3K4me3 in untreated cells and cells treated with triptolide for 2 hours (see Chapter 4), I noted that the H3K4me3 peaks surrounding the alternative TSS were highly-dependent on transcription and were virtually completely lost after triptolide treatment. Therefore, despite the lack of a clear RNAPII ChIP-seq peak, this alternative TSS has many of the genomic hallmarks of a promoter.

I next mapped the annotated n-SET and SET domains to the *Setd1b* gene (Fig. 5.20C). THe n-SET domain is encoded by exons 12-14, which would place it downstream of the alternative transcription start site, which is upstream of exon 11. Additionally, SET1B peptides that were identified in the WDR5-FLAG IP/MS experiment mapped to a region between exon 12 and the exon 17 of *Setd1b*. Therefore, transcription initiation from this alternative TSS could produce a transcript that coded for the n-SET domain, which is required for WDR5 and CFP1 interactions, the SET domain that is required for methyltransferase activity, and would account for all the SET1B peptides that were identified in the WDR5-FLAG IP/MS experiment.

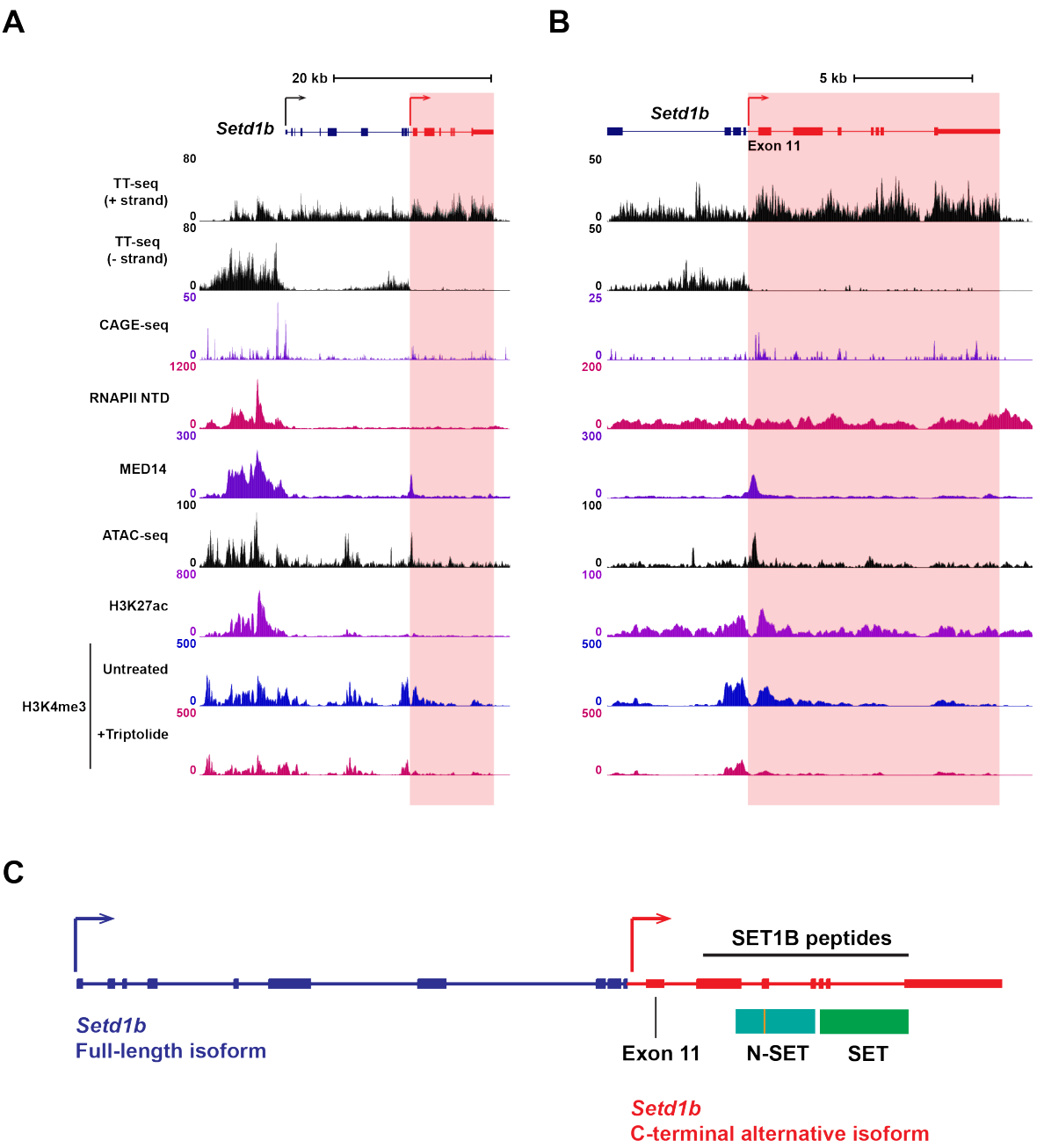

Figure 5.20: A putative C-terminal SET1B alternative isoform

(A) Genome coverage tracks of TT-seq, CAGE-seq, RNAPII (N-terminal domain) ChIP-seq, MED14 ChIP-seq, ATAC-seq, H3K27ac ChIP-seq, and H3K4me3 ChIP-seq at the *Setd1b* locus. TT-seq and RNAPII ChIP-seq are from Hughes et al., 2023, MED14 ChIP-seq is from Dimitrova et al., 2022, ATAC-seq and H3K27ac ChIP-seq are from Fursova et al., 2021. CAGE-seq is from Forrest et al., 2014. H3K4me3 ChIP-seq is from SET1A/B/MLL1/2/3/4-dTAG cells that are untreated or treated with triptolide for 2 hours (see Chapter 4). The region coding for the putative alternative isoform is highlighted in red.
(B) Genome coverage tracks as shown in (A) but with a zoomed-in view focusing on the region coding for the putative alternative isoform.
(C) *Setd1b* gene structure with the putative alternative isoform in red. Regions coding for the n-SET and SET domains are shown. The region encompassing mapped SET1B peptides from the WDR5 FLAG-IP/MS data is also shown.

Discussion

In this chapter, I demonstrate there is a previously uncharacterized WRAD-containing complex in mESCs, and that this complex is responsible for the H3K4 methylation that remains after depleting the six characterized SET1/MLL complexes. Using mass spectrometry, I identify SET1B as a protein that interacts with WDR5 after SET1/MLL-depletion. Interestingly, only peptides mapping to a 588-amio acid C-terminal region were identified as interacting with WDR5. I therefore hypothesize that these SET1B peptides originate from a previously uncharacterized SET1B isoform encompassing a C-terminal region of SET1B that is approximately 588 amino acids in length. Crucially, this region contains the n-SET, SET, and post-SET domains, all of which are required for maximum methyltransferase activity of the SET1B complex on nucleosomes *in vitro*. Interestingly, a previous study has shown that a SET1B fragment comprising the n-SET, SET, and post-SET domains is sufficient to form a complex with WRAD and CFP1 *in vitro*, and is capable of robust H3K4 mono-, di-, and tri-methyltransferase activity on nucleosomes (Kwon et al., 2020). Furthermore, this SET1B complex has been shown to exhibit considerably greater methyltransferase activity compared to complexes containing an equivalent fragment of SET1A, MLL1, MLL2, MLL3, or MLL4 (Kwon et al., 2020). Therefore, I hypothesize that there is a previously uncharacterized SET1B complex containing this alternative SET1B isoform, WRAD, and CFP1, which is responsible for all H3K4 methylation remaining after the six conventional SET1/MLL proteins are depleted.

I hypothesize that this alternative SET1B isoform is produced from an alternative transcription start site immediately upstream of exon 11 of the *Setd1b* gene. This hypothesis is driven by the presence of genomic features upstream of exon 11 of the *Setd1b* gene which are characteristic of a promoter, including enrichment of H3K4me3 and H3K27ac, the presence of a nucleosome depleted region, and occupancy of the Mediator complex. Additionally, TT-seq data indicates that nascent transcription is elevated downstream of this putative alternative TSS, and that there is signficant antisense transcription originating from this site, suggesting alternative transcription initiation occurs from this site. Future work would be focused on defining this alternative mRNA transcript, using techniques such as rapid amplification of cDNA ends (RACE) to map the 5' end of this transcript. The use of an internal, alternative transcription start site to produce a truncated isoform of SET1B is reminiscent of KDM2A and KDM2B, which both have a long isoform and a short isoform, with the short isoform being produced from an internal alternative start site (Turberfield et al., 2019, Tanaka et al., 2010). Interestingly, a recent study reported that *Kdm5b*, which codes for an H3K4 demethylase, also contains an alternative transcription start site 222bp downstream of the conventional TSS (Di Nisio et al., 2023). In both of these instances, alternative isoforms lack the demethylase domains, suggesting that these alternative isoforms serve to create a functional separation of their catalytic-dependent and catalytic-independent functions. Similarly, this alternative SET1B isoform lacks the N-terminal domain required for interaction with WDR82, which the Klose lab has previously demonstrated as essential for activating transcription (Hughes et al., 2023). Therefore, the use of alternative isoforms for functional separation of chromatin modifiers may be more widespread than previously understood.

The existence of this uncharacterized SET1B complex merits a reappraisal of our understanding of the function of SET1B. The SET1B-dTAG strategy inserts the $FKBP12^{F36V}$ domain to the N-terminus of SET1B, which would deplete the full-length isoform of SET1B, but not the alternative C-terminal isoform. Similarly, previous studies have employed constitutive or conditional knockout strategies which remove early exons of the *Setd1b* gene. For example, Bledau et al., 2014 employ Cre recombination to remove exon 5 and Sze et al., 2020 use CRISPR-Cas9 to remove exons 1-4. These knockout strategies would not remove the alternative isoform. Additionally, this alternative isoform, comprising at least 588 amino acids, is likely to be considerably smaller than the full-length SET1B isoform, which is 1,985 amino acids long. Therefore, its presence in western blots as a small molecular weight band is likely to escape attention. Additionally, the use of antibodies against the N-terminal portion of SET1B would not detect this alternative isoform. Throughout this study, I have used an antibody against the N-terminal portion of SET1B for immunoblotting, which would not detect this alternative C-terminal SET1B isoform. Future work should focus on investigating the role of the alternative SET1B isoform in depositing H3K4 methylation in mESCs, although the use of an alternative transcription start site is likely to complicate efforts to employ the dTAG system on this alternative isoform.

Interestingly, size exclusion chromatography indicates that there are likely two WRAD-containing complexes in mESCs, differentiated by the inclusion of CFP1 (Fig. 5.7). An outstanding question is whether the alternative SET1B isoform is part of both complexes. If the alternative SET1B isoform is part of both complexes, it may provide a molecular basis through which it shapes H3K4 methylation. In addition to its role in targeting SET1 complexes to CGIs of highly-transcribed genes, CFP1 has been suggested to have a role in stimulating the methyltransferase activity of SET1A/B, specifically to promote H3K4me2/3 deposition (Kwon et al., 2020). This is reminiscent of the role Spp1, the yeast homolog of CFP1, plays in promoting H3K4 trimethyltransferase activity of the yeast Set1 complex: deletion of Spp1 results in major reductions in H3K4me3 but not H3K4me1/2 (Dehé et al., 2006, Soares et al., 2017). Therefore, it is possible that the CFP1-containing complex serves to recruit the alternative SET1B isoform to CGI promoters to deposit H3K4me3 and the CFP1-absent complex engages with chromatin in a more promiscuous manner to deposit H3K4me1/2. Future studies should focus on understanding whether the alternative SET1B isoform is part of both complexes by performing Co-IP experiments using antibodies against the C-terminal region of SET1B to understand if it interacts with CFP1. Future studies should also aim to understand whether CFP1 is required for H3K4 methylation by the CFP1-WRAD complex.

Interestingly, the CFP1-containing complex elutes in high molecular weight fractions of 669kDa and above while the CFP1-absent complex elutes in lower molecular weight fractions around the 440kDa standard. Given that the molecular weight of WRAD is approximately 210kDa, the inclusion of the alternative SET1B isoform, which is approximately 600 amino acids in length, is unlikely to account for the apparent molecular weight of either complex. Therefore, there are likely to be additional subunits to both complexes which may serve to recruit them to chromatin. It is also worth noting that the alternative SET1B isoform would not interact with WDR82, which interacts with the N-termini of SET1A/B (J. H. Lee and Skalnik, 2008, Ebmeier et al., 2017). This is interesting because I demonstrated in Chapter 4 that a large proportion of the H3K4me3 re-

maining after depleting the six full-length SET1/MLL proteins is dependent on transcription. WDR82 interacts with RNAPII that is phosphorylated on serine 5 of the C-terminal domain, and has been proposed to recruit SET1 complexes to the transcriptional machinery. Therefore, it is possible that additional subunits mediate an interaction between the alternative SET1B complex with the transcriptional machinery. Alternatively, it is possible that the transcription-dependent H3K4 methyltransferase activity is not directly dependent on an interaction with the transcriptional machinery. It is also possible that transcription-coupled processes, such as the deposition of H2BK120ub1, link H3K4 methylation by the alternative SET1B isoform with transcription. Accordingly, depletion of PAF1 has been shown to rapidly reduce global H2BK120ub1 and H3K4me3 levels (Z. Wang et al., 2023). Future work should aim to understand the relationship between the transcription and the H3K4 methylation deposited by the alternative SET1B isoform.

Conclusions and Future Directions

SET1/MLL complexes methylate H3K4 and have therefore been proposed to positively regulate transcription through H3K4 methylation. Much effort has been dedicated to understanding where the different SET1/MLL complexes bind throughout the genome and how they contribute to H3K4 methylation and to transcription. Studies using constitutive or conditional knockout approaches to study individual SET1/MLL proteins have led to the view that SET1 complexes are the major H3K4 tri-methyltransfeases, MLL1/2 complexes are gene-specific H3K4 tri-methyltransferases, and that MLL3/4 complexes are the primary H3K4 di- and mono-methyltransferases (Bledau et al., 2014, Denissov et al., 2014, D. Hu, Gao, et al., 2013). However, previous work from the Klose lab using a rapid depletion approach has demonstrated that depleting SET1 complexes on their own only results in modest reductions in H3K4me3 and that SET1 complexes can support transcription through an H3K4 methylation-independent mechanism (Hughes et al., 2023). Here, I show that SET1A/B do have major contributions to H3K4me3, but MLL1/2 is able to compensate for their loss (Chapter 3). I also show that MLL1/2 have a greater contribution to H3K4me3 across the genome, but SET1A/B are able to similarly compensate for their loss in depositing H3K4me3 at highly-transcribed genes. Strikingly, I find that reductions in H3K4me3 upon SET1A/B and MLL1/2 are not merely additive, and that SET1A/B and MLL1/2 synergise to deposit a large proportion of H3K4me3 across the genome. These findings resolve a long-standing contradiction in the field, wherein the occupancy of SET1A/B and MLL1/2 on chromatin strongly correlate with H3K4me3, and yet depletion of either pair on their own only result in modest reductions in H3K4me3 at genes where their occupancy and H3K4me3 levels are high (Denissov et al., 2014, Brown et al., 2017, Hughes et al., 2023).

My findings here strongly suggest that highly-transcribed genes represent sites of shared activity by SET1A/B and MLL1/2. An important outstanding question is whether they also have shared roles in influencing transcription at these genes. Previous work from the Klose lab has demonstrated that almost a quarter of all transcribed genes require SET1 complexes for transcription (Hughes et al., 2023). Paradoxically, these genes are lowly- or moderately- transcribed genes, where the occupancy of SET1 complexes is relatively low. Indeed, highly-transcribed genes where SET1 occupancy is high exhibit very little transcriptional defects upon loss of SET1 complexes. In light of my findings that SET1A/B and MLL1/2 synergise to deposit H3K4me3 at these genes, it is possible that they similarly synergise to regulate the transcription of these highly-transcribed genes. Indeed, I observe that simultaneous depletion of SET1A/B and MLL1/2 results in more severe effects on cell viability compared to depletion of SET1A/B on their own. This is interesting given that SET1A is the only SET1/MLL protein required for ES cell viability, and suggests that simultaneous depletion of SET1A/B/MLL1/2 results in a much more severe effect on the transcriptome which in turn affects cell viability. At lowly- and moderately-

transcribed genes, SET1A/B utilize a catalysis-independent mechanism to support transcription by opposing premature transcription termination by ZC3H4 in a catalysis-independent manner (Hughes et al., 2023). This mechanism relies on an interaction with WDR82, an interaction not shared with MLL1/2. Similarly, unique interaction partners of MLL1/2 such as MENIN have been suggested to play a role in transcriptional regulation (Sparbier et al., 2023). Whether SET1A/B and MLL1/2 cooperatively regulate transcription at highly-transcribed genes through a shared mechanism such as H3K4me3-deposition or through distinct, catalysis-independent mechanisms will be the focus of future studies.

In spite of the synergistic effects of SET1A/B and MLL1/2 on H3K4me3 deposition, I found that they are not the only H3K4 tri-methyltransferases. Indeed, using a combinatorial degron approach to simultaneously deplete all SET1/MLL proteins, I have found that the six SET1/MLL methyltransferases do not deposit all H3K4 methylation in mESCs, and that an uncharacterized methyltransfease may be responsible for as much as 50% of all H3K4 methylation (Chapters 4 and 5). Surprisingly, I find that subunits of the WRAD subcomplex are required for the remaining H3K4 methylation. Indeed, depleting RBBP5 results in a near complex loss of H3K4 methylation, which is in accordance with recently-published findings (S. Hu et al., 2023, H. Wang et al., 2023). Using a combination of biochemical approaches, I find that a WRAD-containing complex remains after SET1/MLL-depletion, and that CFP1 and a putative alternative SET1B isoform may be part of this complex. Interestingly, in contrast to the characterized SET1/MLL proteins, this truncated SET1B isoform consists of only the protein domains required for methyltransferase activity. It therefore seems unlikely that this isoform has catalysis-independent functions and implies that H3K4 methytransferase activity may be the primary role of this isoform. Interestingly, the existence of multiple, functionally distinct isoforms has also been reported for other chromatin-modifying proteins. This is most prominently illustrated by KDM2A and KDM2B, which are both CGI-binding proteins that contain a long isoform with H3K36 demethylase activity and a short isoform without catalytic activity (Turberfield et al., 2019). Therefore, the use of multiple protein isoforms to separate the different functions of a single chromatin-modifying protein may be more widespread than previously recognized.

A central feature of chromatin-modifying proteins is that their chromatin occupancy is reflective of their activity. Indeed, the chromatin occupancy of H3K4 methyltransferases is highly-correlated with H3K4 methylation: the promoters of highly-transcribed genes have high levels of H3K4me3 and correspondingly high levels of SET1A, MLL2, and RBBP5 binding (Fig. 5.11A). Additionally, enhancers with high levels of H3K4me1 also have correspondingly high levels of MLL3/4 binding (Fig. 5.13A). A prevailing view is that longer residency time on chromatin is required for the methyltransferase to catalyse higher methylation states. This hypothesis is in part based on observations that the higher H3K4 methylation states in yeast are dependent on the amount of time Set1 spends near the nucleosome. This was demonstrated by the fact that tethering Set1 to RNAPII leads to enrichment of H3K4me2/3 throughout the gene body, and mutants in which the RNAPII elongation rate is reduced exhibit less spreading of H3K4me2/3 into the gene body (Soares et al., 2017). My findings here mostly uphold this view. I find that MLL3/4 do not require stable-binding for H3K4me1 deposition, which is apparent when comparing ChIP-seq analysis of MLL3/4 occupancy and the effect that MLL3/4-

depletion has on H3K4me1 across the genome. MLL3/4 occupancy is primarily enriched at distal regulatory sites, but depletion of MLL3/4 results in broad reductions in H3K4me1 across the genome, both at distal regulatory sites but also at gene bodies and other broad H3K4me1 domains where MLL3/4 occupancy as assessed by ChIP-seq is low. However, removal of MLL3/4 does not result in reductions in H3K4me2/3 throughout the genome except at enhancers where they have high-level occupancy. This suggests that MLL3/4 is able to monomethylate H3K4 when tranisently interacting with chromatin but requires stable binding at enhancers to produce H3K4me2/3.

A striking feature of the SET1/MLL-independent H3K4 methyltransferase complex is that it does not appear to require high-level chromatin occupancy to deposit H3K4me1, H3K4me2, or H3K4me3. This is evident when analysing the chromatin occupancy of RBBP5 after SET1/MLL-depletion - RBBP5 is virtually completely removed from chromatin as assessed by crosslinking ChIP-seq. However, approximately 50% of H3K4 methylation remains after SET1/MLL-depletion, and this H3K4 methylation is dependent on RBBP5. This suggests that approximately 50% of all H3K4me3 is deposited by a methyltransferase complex which does not stably bind to chromatin. This is surprising given that the likely candidate for this methyltransferase complex appears to consist of a truncated SET1B isoform in complex with CFP1 and WRAD. CFP1 is the primary determinant of SET1A occupancy throughout the genome, and CFP1-dependent interactions with chromatin are required for SET1A-mediated H3K4 methylation (Brown et al., 2017). One possible explanation is that the function of CFP1 in the SET1B complex is different from that in the SET1A complex. To date, a detailed analysis of SET1B occupancy has not been carried out, and our understanding of how SET1B engages with the genome mostly relies on the assumption that SET1A and SET1B bind to chromatin in an identical fashion. Indeed, there is no direct evidence so far to indicate that SET1B has high-level occupancy at CGI promoters through CFP1-mediated interactions with chromatin. It is also possible that in the alternative SET1B complex, other subunits have a more dominant role in mediating SET1B interactions with chromatin, or that CFP1 plays a fundamentally different role in the alternative SET1B complex that is not primarily related to chromatin binding. Interestingly, early studies of the SET1A and SET1B complexes in HEK293 cells demonstrated that they display distinct nuclear localisation patterns as determined by immunofluorescence microscopy, suggesting that these two complexes may indeed have very different mechanisms of interacting with chromatin despite near-identical subunit composition (J. H. Lee et al., 2007, J. H. Lee and Skalnik, 2002). Indeed, my findings indicate that this putative alternative SET1B complex may require transcription for H3K4 di- and trimethylation. It is possible that the putative alternative SET1B complex primarily relies on interactions with transcription elongation complexes for engagement with chromatin, which are not typically well-captured by crosslinking ChIP-seq.

How H3K4 methylation regulates transcription is still a subject of debate. Much difficulty in understanding the function of H3K4 methylation stems from the difficulty in removing all H3K4 methylation in mammals due to the presence of six H3K4 methyltransferases which redundantly methylate H3K4. Additionally, it is difficult to separate the catalysis-independent functions of the H3K4 methyltransferases in regulating transcription from their catalysis-dependent functions. I envisage that the SET1/MLL depletion approach presented in

this thesis can be combined with rescue experiments using catalytically-dead mutants, which will allow for the rapid ablation of SET1/MLL catalytic activity but the retention of catalysis-independent functions. Interestingly, SET1/MLL complexes themselves bind H3K4me3, and it is therefore possible that an inherent function of SET1/MLL complexes is to deposit H3K4me3 to stabilize their own interactions with chromatin to form a positive feedback loop. My finding that a significant proportion of H3K4me2/3 are dependent on active transcription lends support to this view, and it can be envisaged that SET1/MLL complexes may methylate H3K4 co-transcriptionally in order to maintain an active chromatin state at actively transcribed genes. It seems likely that at highly-transcribed genes, multiple mechanisms are at work to maintain such an active chromatin state, with a combination of CGI-targeting by multiple SET1/MLL complexes and a transcription-coupled mechanism to continually maintain high levels of H3K4me3 at the promoters highly active genes. Additionally, highly-transcribed genes also have high levels of H3K4me1/2 in the gene body, and it is possible that they may have a role in supporting transcription, perhaps by facilitating efficient transcription elongation. Indeed, I find that a large proportion of H3K4me2 also requires ongoing transcription, and it is possible that H3K4me1 in the gene bodies of highly-transcribed genes is also transcription-dependent. It is possible that SET1/MLL complexes also co-transcriptionally mono- and di-methylate the gene body of highly transcribed genes to contribute to an active chromatin state across the entire gene that can facilitate future rounds of transcription. How this chromatin state supports transcription remains to be determined, and will require a comprehensive understanding of how the transcriptome and the transcriptional machinery are affected upon loss of H3K4 methylation. In this study, I have extensively characterized H3K4 methylation by SET1/MLL complexes, and future work will build on my findings to understand how SET1/MLL complexes and H3K4 methylation regulate transcription.

Appendix A

WDR5-interacting proteins

Table A.1: List of WDR5-interacting proteins in untreated SET1A/B/MLL1/2/3/4-dTAG-WDR5-FLAG cells. List of proteins determined to be sigificantly enriched in WDR5-FLAG-IP compared to the negative control IP. Proteins were determined to be significantly enriched if the t-test difference between WDR5-FLAG-IP and negative control IP was larger than 3 and if the t-test p value in the WDR5-FLAG-IP was less than 0.05. 144 proteins were identified as WDR5-interactors.

WDR5-interacting proteins		
Gene Name	-Log_{10} t-test p value	t-test Difference
Ogt	8.12688	11.4322
Hcfc1	7.27252	11.2544
Kmt2d	7.4307	11.1705
Ash2l	6.64558	11.1568
Kmt2b	6.27711	10.9124
Mga	6.36331	10.7941
Yeats2	7.32001	10.7387
Rbbp5	7.36822	10.6012
Kmt2c	6.4732	10.5968
Zfp532	6.56874	10.49
Kansl1	6.5634	10.4224
Men1	6.07613	9.97352
Phf20	8.40996	9.84336
Paxip1	5.65421	9.76988
Rnf2	7.32314	9.41986
Kmt2a	5.9061	9.40951
L3mbtl2	8.18541	9.29951
Kdm6a	8.08421	9.29229
Setd1a	6.13631	9.27967
Hdac2	5.60536	9.02631
Dpy30	4.11746	8.93697
Mcrs1	7.30461	8.92344

Continuation of Table A.1		
Gene Name	-Log_{10} t-test p value	t-test Difference
Kat8	6.47	8.84453
Mbip	6.90264	8.82395
Kansl2	6.87097	8.75239
Hcfc2	7.86343	8.74708
Cxxc1	6.38094	8.60948
Kansl3	4.89982	8.39044
Csrp2bp	4.07392	8.38811
Hdac1	5.08114	8.33369
E2f6	6.77818	8.22045
Tada2a	5.84931	8.21903
Wdr5	5.93929	8.12579
Trrap	6.14483	8.03386
Ccdc101	4.78344	7.96287
Csnk2a2	5.11983	7.86244
Dr1	6.98481	7.86085
Mbd3	6.00222	7.79284
Ep400	7.96266	7.78714
Tada3	5.59013	7.6051
Ncoa6	7.7641	7.58238
Mta1	5.16608	7.49456
Zzz3	2.7739	7.47957
Bod1l	5.21068	7.35046
Pcgf6	3.68722	7.28028
Kat2b	7.22919	7.21782
Tfdp1	5.81557	7.17708
Gatad2b	7.53366	7.06788
Cbx3	3.44279	7.01262
Rbbp4	2.57074	7.0072
Kat2a	5.72485	6.70615
Gatad2a	3.5229	6.70302
Cse1l	6.70857	6.62585
Zfp462	5.86201	6.54712
Mta2	6.08937	6.4981

Continuation of Table A.1		
Gene Name	-Log_{10} t-test p value	t-test Difference
Chd4	4.95636	6.45421
mKIAA0304	7.36296	6.40574
Wdr82	4.42172	6.33
Fkbp1a	5.13538	6.31942
Mta3	4.74643	6.31185
Uty	4.79805	6.11197
Usp7	5.19177	6.10614
PA1	4.51301	6.02839
Bend3	7.20621	5.93471
Sall1	5.40211	5.85576
Cdk2ap1	4.97639	5.81239
Setd1b	4.57249	5.75701
Max	4.52897	5.74332
Zfp423	4.85663	5.73265
Rhox10	5.58459	5.71374
Znf131	6.34662	5.68443
Rybp	4.0905	5.6124
Zfp296	6.82145	5.5931
Zfp219	5.86527	5.55586
Csnk2b	5.49001	5.45176
Dmap1	6.65998	5.43849
Wiz	5.92713	5.17437
Ehmt1	5.20846	5.1394
Zfx	4.76663	5.11215
Tet1	5.97569	5.10475
Rbbp7	5.64857	5.06137
Brd3	4.26309	5.03418
Csnk2a1	5.49415	4.98863
Kdm1b	5.42853	4.89666
Foxk2	4.33302	4.87733
Kat5	4.75712	4.86851
Yeats4	5.33376	4.85904
Zfp609	6.27566	4.82222

Continuation of Table A.1		
Gene Name	-Log_{10} t-test p value	t-test Difference
Brd8	6.61995	4.78226
Hdgfrp2	5.72156	4.76636
Jade1	4.15278	4.74307
Asxl2	5.10164	4.66453
Kpna3	3.68414	4.53451
Rere	3.5806	4.45389
Ehmt2	3.65863	4.43046
Sall4	3.32823	4.41329
Ctr9	4.57455	4.31655
Meaf6	3.80674	4.28055
Drosha	4.20559	4.2543
Brd2	6.37648	4.24716
Rfc4	4.29185	4.16925
Kat7	4.30824	4.1492
Bap1	4.13985	4.12781
Ring1	6.84219	4.12031
Psip1	4.99888	4.06948
Cbx1	5.22389	4.06093
Rfc5	3.70557	4.03376
Mbtd1	5.6549	4.03038
Sin3a	2.69487	3.935
Kpna4	3.87347	3.85673
Ep400	4.72744	3.85327
Polr2g	6.2703	3.83137
Epc2	5.21588	3.77473
Vps72	5.94738	3.75794
Elmsan1	3.31902	3.73504
Hspa8	5.4142	3.71629
Ruvbl2	4.75926	3.6796
Ing3	5.05663	3.65897
Ruvbl1	3.41456	3.56668
Mdn1	2.25601	3.55708
Arid5b	4.62965	3.52684

Continuation of Table A.1		
Gene Name	-Log_{10} t-test p value	t-test Difference
Phf20l1	5.09246	3.51698
Orc2	1.88972	3.50712
Taf2	4.14312	3.50565
Las1l	4.18286	3.44011
Npat	4.79868	3.394
Brd4	2.9843	3.35593
Csnk1a1	4.03513	3.34755
Sall3	3.69865	3.34151
Znf512b	3.89251	3.31384
Pelp1	2.73706	3.30642
Tex10	4.65863	3.24624
H2afv	1.78244	3.18033
Foxk1	4.81413	3.17559
Nr0b1	4.16723	3.17547
Esrrb	4.60753	3.14546
Sap30	4.55477	3.12964
Rfc1	5.15452	3.12397
Zbtb17	3.36495	3.11338
Kpna1	2.55311	3.10709
Ssbp1	2.52368	3.10254
Ddb1	1.78801	3.07512
Cdyl	4.77367	3.05684
Kpna2	3.64691	3.04686
End of Table		

Bibliography

Alsulami, M., Munawar, N., DIllon, E., Oliviero, G., Wynne, K., Alsolami, M., Moss, C., Gaora, P., O'Meara, F., Cotter, D., & Cagney, G. (2019). SETD1A methyltransferase is physically and functionally linked to the DNA damage repair protein RAD18. *Molecular and Cellular Proteomics*, *18*(7), 1428–1436. https://doi.org/10.1074/mcp.RA119.001518

Andersson, R., & Sandelin, A. (2020, February). Determinants of enhancer and promoter activities of regulatory elements. https://doi.org/10.1038/s41576-019-0173-8

Ashokkumar, D., Zhang, Q., Much, C., Bledau, A. S., Naumann, R., Alexopoulou, D., Dahl, A., Goveas, N., Fu, J., Anastassiadis, K., Stewart, A. F., & Kranz, A. (2020). MLL4 is required after implantation, whereas MLL3 becomes essential during late gestation. *Development (Cambridge)*, *147*(12). https://doi.org/10.1242/dev.186999

Austenaa, L. M., Barozzi, I., Simonatto, M., Masella, S., Della Chiara, G., Ghisletti, S., Curina, A., de Wit, E., Bouwman, B. A., 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*(3), 460–474. https://doi.org/10.1016/j.molcel.2015.09.018

Bach, C., Mueller, D., Buhl, S., Garcia-Cuellar, M. P., & Slany, R. K. (2009). Alterations of the CxxC domain preclude oncogenic activation of mixed-lineage leukemia 2. *Oncogene*, *28*(6), 815–823. https://doi.org/10.1038/onc.2008.443

Bae, H. J., Dubarry, M., Jeon, J., Soares, L. M., 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. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-16082-2

Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell*, *129*(4), 823–837. https://doi.org/10.1016/j.cell.2007.05.009

Barsoum, M., Sayadi-Boroujeni, R., Stenzel, A. T., Bussmann, P., Lüscher-Firzlaff, J., & Lüscher, B. (2023). Sequential deregulation of histone marks, chromatin accessibility and gene expression in response to PROTAC-induced degradation of ASH2L. *bioRxiv*, 2023.09.08.556832. https://doi.org/10.1101/2023.09.08.556832

Benayoun, B. A., Pollina, E. A., Ucar, D., Mahmoudi, S., Karra, K., Wong, E. D., Devarajan, K., Daugherty, A. C., Kundaje, A. B., Mancini, E., Hitz, B. C., Gupta, R., Rando, T. A., Baker, J. C., Snyder, M. P., Cherry, J. M., & Brunet, A. (2014). H3K4me3 Breadth Is Linked to Cell Identity and Transcriptional Consistency. *Cell*, *158*(3), 673–688. https://doi.org/https://doi.org/10.1016/j.cell.2014.06.027

Bernstein, B. E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D. K., Huebert, D. J., McMahon, S., Karlsson, E. K., Kulbokas, E. J., Gingeras, T. R., Schreiber, S. L., & Lander, E. S. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, *120*(2), 169–181. https://doi.org/10.1016/j.cell.2005.01.001

Bian, C., Xu, C., Ruan, J., Lee, K. K., Burke, T. L., Tempel, W., Barsyte, D., Li, J., Wu, M., Zhou, B. O., Fleharty, B. E., Paulson, A., Allali-Hassani, A., Zhou, J.-Q., Mer, G., Grant, P. A., Workman, J. L., Zang, J., & Min, J. (2011). Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation. *The EMBO Journal*, *30*(14), 2829–2842. https://doi.org/https://doi.org/10.1038/emboj.2011.193

Bird, A. (2002, January). DNA methylation patterns and epigenetic memory. https://doi.org/10.1101/gad.947102

Birke, M., Schreiner, S., García-Cuéllar, M.-P., Mahr, K., Titgemeyer, F., & Slany, R. K. (2002). *The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation* (tech. rep. No. 4).

Blackledge, N. P., Farcas, A. M., Kondo, T., King, H. W., McGouran, J. F., Hanssen, L. L., Ito, S., Cooper, S., Kondo, K., Koseki, Y., Ishikura, T., Long, H. K., Sheahan, T. W., Brockdorff, N., Kessler, B. M., Koseki, H., & Klose, R. J. (2014). Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. *Cell*, *157*(6), 1445–1459. https://doi.org/10.1016/j.cell.2014.05.004

Blackledge, N. P., Fursova, N. A., Kelley, J. R., Huseyin, M. K., Feldmann, A., & Klose, R. J. (2020). PRC1 Catalytic Activity Is Central to Polycomb System Function. *Molecular Cell*, *77*(4), 857–874. https://doi.org/10.1016/j.molcel.2019.12.001

Blackledge, N. P., & Klose, R. (2011). CpG island chromatin. *Epigenetics*, *6*(2), 147–152. https://doi.org/10.4161/epi.6.2.13640

Blackledge, N. P., & Klose, R. J. (2021, December). The molecular principles of gene regulation by Polycomb repressive complexes. https://doi.org/10.1038/s41580-021-00398-y

Blackledge, N. P., Long, H. K., Zhou, J. C., Kriaucionis, S., Patient, R., & Klose, R. J. (2012). Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic Acids Research*, *40*(4), e32–e32. https://doi.org/10.1093/nar/gkr1207

Blackledge, N. P., Thomson, J. P., & Skene, P. J. (2013). CpG island chromatin is shaped by recruitment of ZF-CxxC proteins. *Cold Spring Harbor Perspectives in Biology*, *5*(11). https://doi.org/10.1101/cshperspect.a018648

Bledau, A. S., Schmidt, K., Neumann, K., Hill, U., Ciotta, G., Gupta, A., Torres, D. C., Fu, J., Kranz, A., Stewart, A. F., & Anastassiadis, K. (2014). The H3K4 methyltransferase Setd1a is first required at the epiblast stage, whereas Setd1b becomes essential after gastrulation. *Development (Cambridge)*, *141*(5), 1022–1035. https://doi.org/10.1242/dev.098152

Bochyńska, A., Stenzel, A. T., Sayadi Boroujeni, R., Kuo, C.-C., Barsoum, M., Liang, W., Bussmann, P., Costa, I. G., Lüscher-Firzlaff, J., & Lüscher, B. (2022). Induction of senescence upon loss of the Ash2l core subunit of H3K4 methyltransferase complexes. *Nucleic Acids Research*, *50*(14), 7889–7905. https://doi.org/10.1093/nar/gkac591

Bode, D., Yu, L., Tate, P., Pardo, M., & Choudhary, J. (2016). Characterization of two distinct nucleosome remodeling and deacetylase (NuRD) complex assemblies in embryonic stem cells. *Molecular and Cellular Proteomics*, *15*(3), 878–891. https://doi.org/10.1074/mcp.M115.053207

Boileau, R. M., Chen, K. X., & Blelloch, R. (2023). Loss of MLL3/4 decouples enhancer H3K4 monomethylation, H3K27 acetylation, and gene activation during embryonic stem cell differentiation. *Genome Biology*, *24*(1). https://doi.org/10.1186/s13059-023-02883-3

Brici, D., Zhang, Q., Reinhardt, S., Dahl, A., Hartmann, H., Schmidt, K., Goveas, N., Huang, J., Gahurova, L., Kelsey, G., Anastassiadis, K., Stewart, A. F., & Kranz, A. (2017). Setd1b, encoding a histone 3 lysine 4 methyltransferase, is a maternal effect gene required for the oogenic gene expression program. *Development (Cambridge)*, *144*(14), 2606–2617. https://doi.org/10.1242/dev.143347

Brickner, J. H. (2023). Inheritance of epigenetic transcriptional memory through read–write replication of a histone modification. *Annals of the New York Academy of Sciences*, *1526*(1), 50–58. https://doi.org/https://doi.org/10.1111/nyas.15033

Brown, D. A., Di Cerbo, V., Feldmann, A., Ahn, J., Ito, S., Blackledge, N. P., Nakayama, M., McClellan, M., Dimitrova, E., Turberfield, A. H., Long, H. K., King, H. W., Kriaucionis, S., Schermelleh, L., Kutateladze, T. G., Koseki, H., & Klose, R. J. (2017). The SET1 Complex Selects Actively Transcribed Target Genes via Multivalent Interaction with CpG Island Chromatin. *Cell Reports*, *20*(10), 2313–2327. https://doi.org/https://doi.org/10.1016/j.celrep.2017.08.030

Cao, F., Chen, Y., Cierpicki, T., Liu, Y., Basrur, V., Lei, M., & Dou, Y. (2010). An Ash2L/RbBP5 Heterodimer Stimulates the MLL1 Methyltransferase Activity through Coordinated Substrate Interactions with the MLL1 SET Domain. *PLOS ONE*, *5*(11), e14102–. https://doi.org/10.1371/journal.pone.0014102

Cao, K., Collings, C. K., Marshall, S. A., Morgan, M. A., Rendleman, E. J., Wang, L., Sze, C. C., Sun, T., Bartom, E. T., & Shilatifard, A. (2017). SET1A/COMPASS and shadow enhancers in the regulation of homeotic gene expression (20170509th ed.). *Genes Dev*, *31*(8), 787–801. https://doi.org/10.1101/gad.294744.116

Cao, K., Collings, C. K., Morgan, M. A., Marshall, S. A., Rendleman, E. J., Ozark, P. A., Smith, E. R., & Shilatifard, A. (2018). *C A N C E R An Mll4/COMPASS-Lsd1 epigenetic axis governs enhancer function and pluripotency transition in embryonic stem cells* (tech. rep.). https://www.science.org

Cenik, B. K., & Shilatifard, A. (2021). COMPASS and SWI/SNF complexes in development and disease. *Nature Reviews Genetics*, *22*(1), 38–58. https://doi.org/10.1038/s41576-020-0278-0

Champagne, K. S., Saksouk, N., Peña, P. V., Johnson, K., Ullah, M., Yang, X.-J., Côté, J., & Kutateladze, T. G. (2008). The crystal structure of the ING5 PHD finger in complex with an H3K4me3 histone

peptide. *Proteins: Structure, Function, and Bioinformatics*, *72*(4), 1371–1376. https://doi.org/https://doi.org/10.1002/prot.22140

Chang, P. Y., Hom, R. A., Musselman, C. A., Zhu, L., Kuo, A., Gozani, O., Kutateladze, T. G., & Cleary, M. L. (2010). Binding of the MLL PHD3 Finger to Histone H3K4me3 Is Required for MLL-Dependent Gene Transcription. *Journal of Molecular Biology*, *400*(2), 137–144. https://doi.org/10.1016/j.jmb.2010.05.005

Cho, Y. W., Hong, S. H., Jin, Q., Wang, L., Lee, J. E., Gavrilova, O., & Ge, K. (2009). Histone Methylation Regulator PTIP Is Required for PPARγ and C/EBPα Expression and Adipogenesis. *Cell Metabolism*, *10*(1), 27–39. https://doi.org/10.1016/j.cmet.2009.05.010

Cho, Y. W., Hong, T., Hong, S. H., Guo, H., Yu, H., Kim, D., Guszczynski, T., Dressler, G. R., Copeland, T. D., Kalkum, M., & Ge, K. (2007). PTIP associates with MLL3- and MLL4-containing histone H3 lysine 4 methyltransferase complex. *Journal of Biological Chemistry*, *282*(28), 20395–20406. https://doi.org/10.1074/jbc.M701574200

Ciotta, G., Singh, S., Gupta, A., Torres, D. C., Fu, J., Choudhury, R., Chu, W. K., Choudhary, C., Gahurova, L., Al-Fatlawi, A., Schroeder, M., Aasland, R., Poetsch, A. R., Anastassiadis, K., & Stewart, A. F. (2023). The BOD1L subunit of the mammalian SETD1A complex sustains the expression of DNA damage repair genes despite restraining H3K4 trimethylation. *bioRxiv*, 2023.04.06.535882. https://doi.org/10.1101/2023.04.06.535882

Clouaire, T., Webb, S., & Bird, A. (2014). Cfp1 is required for gene expression-dependent H3K4 trimethylation and H3K9 acetylation in embryonic stem cells. *Genome biology*, *15*(9), 451. https://doi.org/10.1186/s13059-014-0451-x

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J. H., Skalnik, D., & Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes and Development*, *26*(15), 1714–1728. https://doi.org/10.1101/gad.194209.112

Collins, R. E., Tachibana, M., Tamaru, H., Smith, K. M., Jia, D., Zhang, X., Selker, E. U., Shinkai, Y., & Cheng, X. (2005). In vitro and in vivo analyses of a Phe/Tyr switch controlling product specificity of histone lysine methyltransferases. *Journal of Biological Chemistry*, *280*(7), 5563–5570. https://doi.org/10.1074/jbc.M410483200

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

Cortazar, M. A., Sheridan, R. M., Erickson, B., Fong, N., Glover-Cutter, K., Brannan, K., & Bentley, D. L. (2019). Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism. *Mol Cell*, *76*(6), 896–908. https://doi.org/10.1016/j.molcel.2019.09.031

Crump, N. T., & Milne, T. A. (2019). Why are so many MLL lysine methyltransferases required for normal mammalian development? (20190516th ed.). *Cellular and molecular life sciences : CMLS*, *76*(15), 2885–2898. https://doi.org/10.1007/s00018-019-03143-z

Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes and Development*, *25*(10), 1010–1022. https://doi.org/10.1101/gad.2037511

Dehé, P. M., Dichtl, B., Schaft, D., Roguev, A., Pamblanco, M., Lebrun, R., Rodríguez-Gil, A., Mkandawire, M., Landsberg, K., Shevchenko, A., Shevchenko, A., Rosaleny, L. E., Tordera, V., Chávez, S., Stewart, A. F., & Géli, V. (2006). Protein interactions within the Set1 complex and their roles in the regulation of histone 3 lysine 4 methylation. *Journal of Biological Chemistry*, *281*(46), 35404–35412. https://doi.org/10.1074/jbc.M603099200

Denissov, S., Hofemeister, H., Marks, H., Kranz, A., Ciotta, G., Singh, S., Anastassiadis, K., Stunnenberg, H. G., & Stewart, A. F. (2014). Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant (20140114th ed.). *Development*, *141*(3), 526–537. https://doi.org/10.1242/dev.102681

Dhar, S. S., Lee, S. H., Kan, P. Y., Voigt, P., Ma, L., Shi, X., Reinberg, D., & Lee, M. G. (2012). Trans-tail regulation of MLL4-catalyzed H3K4 methylation by H4R3 symmetric dimethylation is mediated by a tandem PHD of MLL4 (2012/12/20). *Genes Dev*, *26*(24), 2749–2762. https://doi.org/10.1101/gad.203356.112

Dhar, S. S., Zhao, D., Lin, T., Gu, B., Pal, K., Wu, S. J., Alam, H., Lv, J., Yun, K., Gopalakrishnan, V., Flores, E. R., Northcott, P. A., Rajaram, V., Li, W., Shilatifard, A., Sillitoe, R. V., Chen, K., & Lee, M. G. (2018). MLL4 Is Required to Maintain Broad H3K4me3 Peaks and Super-Enhancers at Tumor Suppressor Genes (2018/06/05). *Mol Cell*, *70*(5), 825–841. https://doi.org/10.1016/j.molcel.2018.04.028

Di Nisio, E., Licursi, V., Mannironi, C., Buglioni, V., Paiardini, A., Robusti, G., Noberini, R., Bonaldi, T., & Negri, R. (2023). A truncated and catalytically inactive isoform of KDM5B histone demethylase accumulates in breast cancer cells and regulates H3K4 tri-methylation and gene expression. *Cancer Gene Therapy*, *30*(6), 822–832. https://doi.org/10.1038/s41417-022-00584-w

Dillon, S. C., Zhang, X., Trievel, R. C., & Cheng, X. (2005). The SET-domain protein superfamily: Protein lysine methyltransferases. https://doi.org/10.1186/gb-2005-6-8-227

Dimitrova, E., Feldmann, A., van der Weide, R. H., Flach, K. D., Lastuvkova, A., de Wit, E., & Klose, R. J. (2022). Distinct roles for CKM–Mediator in controlling Polycomb-dependent chromosomal interactions and priming genes for induction. *Nature Structural and Molecular Biology*, *29*(10), 1000–1010. https://doi.org/10.1038/s41594-022-00840-5

Dobrinić, P., Szczurek, A. T., & Klose, R. J. (2021). PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency. *Nature Structural and Molecular Biology*, *28*(10), 811–824. https://doi.org/10.1038/s41594-021-00661-y

Dorighi, K. M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B. S., Nady, N., Still, C. D., Garcia, B. A., Adelman, K., & Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Molecular Cell*, *66*(4), 568–576. https://doi.org/10.1016/j.molcel.2017.04.018

Douillet, D., Sze, C. C., Ryan, C., Piunti, A., Shah, A. P., Ugarenko, M., Marshall, S. A., Rendleman, E. J., Zha, D., Helmin, K. A., Zhao, Z., Cao, K., Morgan, M. A., Singer, B. D., Bartom, E. T., Smith, E. R., & Shilatifard, A. (2020). Uncoupling histone H3K4 trimethylation from developmental gene expression via an equilibrium of COMPASS, Polycomb and DNA methylation. *Nature Genetics*, *52*(6), 615–625. https://doi.org/10.1038/s41588-020-0618-1

D'urso, A., Takahashi, Y.-H., Xiong, B., Marone, J., Coukos, R., Randise-Hinchliff, C., Wang, J.-P., Shilatifard, A., & Brickner, J. H. (2016). Set1/COMPASS and Mediator are repurposed to promote epigenetic transcriptional memory. https://doi.org/10.7554/eLife.16691.001

Ebmeier, C. C., Erickson, B., Allen, B. L., Allen, M. A., Kim, H., Fong, N., Jacobsen, J. R., Liang, K., Shilatifard, A., Dowell, R. D., Old, W. M., Bentley, D. L., & Taatjes, D. J. (2017). Human TFIIH Kinase CDK7 Regulates Transcription-Associated Chromatin Modifications. *Cell Reports*, *20*(5), 1173–1186. https://doi.org/https://doi.org/10.1016/j.celrep.2017.07.021

Elmendorf, B. J., Shilatifard, A., Yan, Q., Conaway, J. W., & Conaway, R. C. (2001). Transcription Factors TFIIF, ELL, and Elongin Negatively Regulate SII-induced Nascent Transcript Cleavage by Non-arrested RNA Polymerase II Elongation Intermediates. *Journal of Biological Chemistry*, *276*(25), 23109–23114. https://doi.org/10.1074/jbc.M101445200

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

Ernst, P., & Vakoc, C. R. (2012). WRAD: enabler of the SET1-family of H3K4 methyltransferases. *Briefings in Functional Genomics*, *11*(3), 217–226. https://doi.org/10.1093/bfgp/els017

Estell, C., Davidson, L., Steketee, P. C., Monier, A., & West, S. (2021). Zc3h4 restricts non-coding transcription in human cells. *eLife*, *10*. https://doi.org/10.7554/ELIFE.67305

Fang, L., Zhang, J., Zhang, H., Yang, X., Jin, X., Zhang, L., Skalnik, D. G., Jin, Y., Zhang, Y., Huang, X., Li, J., & Wong, J. (2016). H3K4 Methyltransferase Set1a Is A Key Oct4 Coactivator Essential for Generation of Oct4 Positive Inner Cell Mass. *Stem Cells*, *34*(3), 565–580. https://doi.org/10.1002/stem.2250

Farcas, A. M., Blackledge, N. P., Sudbery, I., Long, H. K., McGouran, J. F., Rose, N. R., Lee, S., Sims, D., Cerase, A., Sheahan, T. W., Koseki, H., Brockdorff, N., Ponting, C. P., Kessler, B. M., & Klose, R. J. (2012). KDM2B links the polycomb repressive complex 1 (PRC1) to recognition of CpG islands. *eLife*, *2012*(1). https://doi.org/10.7554/eLife.00205

Farrelly, L. A., Thompson, R. E., Zhao, S., Lepack, A. E., Lyu, Y., Bhanu, N. V., Zhang, B., Loh, Y. H. E., Ramakrishnan, A., Vadodaria, K. C., Heard, K. J., Erikson, G., Nakadai, T., Bastle, R. M., Lukasak, B. J., Zebroski, H., Alenina, N., Bader, M., Berton, O., … Maze, I. (2019). Histone serotonylation is

a permissive modification that enhances TFIID binding to H3K4me3. *Nature*, *567*(7749), 535–539. https://doi.org/10.1038/s41586-019-1024-7

Feng, W., Yonezawa, M., Ye, J., Jenuwein, T., & Grummt, I. (2010). PHF8 activates transcription of rRNA genes through H3K4me3 binding and H3K9me1/2 demethylation. *Nature Structural and Molecular Biology*, *17*(4), 445–450. https://doi.org/10.1038/nsmb.1778

Fischer, V., Plassard, D., Ye, T., Reina-San-Martin, B., Stierle, M., Tora, L., & Devys, D. (2021). The related coactivator complexes SAGA and ATAC control embryonic stem cell self-renewal through acetyltransferase-independent mechanisms. *Cell Reports*, *36*(8). https://doi.org/10.1016/j.celrep.2021.109598

Fisher, M. J., & Luse, D. S. (2023). Promoter-proximal nucleosomes attenuate RNA polymerase II transcription through TFIID. *Journal of Biological Chemistry*, *299*(7). https://doi.org/10.1016/j.jbc.2023.104928

Forrest, A. R. R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M. J. L., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., … CLST. (2014). A promoter-level mammalian expression atlas. *Nature*, *507*(7493), 462–470. https://doi.org/10.1038/nature13182

Fursova, N. A., Blackledge, N. P., Nakayama, M., Ito, S., Koseki, Y., Farcas, A. M., King, H. W., Koseki, H., & Klose, R. J. (2019). Synergy between Variant PRC1 Complexes Defines Polycomb-Mediated Gene Repression (2019/04/29). *Mol Cell*, *74*(5), 1020–1036. https://doi.org/10.1016/j.molcel.2019.03.024

Fursova, N. A., Turberfield, A. H., Blackledge, N. P., Findlater, E. L., Lastuvkova, A., Huseyin, M. K., Dobrinic, P., & Klose, R. J. (2021). BAP1 constrains pervasive H2AK119ub1 to control the transcriptional potential of the genome. *Genes and Development*, *35*(9-10), 749–770. https://doi.org/10.1101/GAD.347005.120

Glaser, S., Schaft, J., Lubitz, S., Vintersten, K., van der Hoeven, F., Tufftеland, K. R., Aasland, R., Anastassiadis, K., Ang, S. L., & Stewart, A. F. (2006). Multiple epigenetic maintenance factors implicated by the loss of MII2 in mouse development. *Development*, *133*(8), 1423–1432. https://doi.org/10.1242/dev.02302

Gregersen, L. H., Mitter, R., Ugalde, A. P., Nojima, T., Proudfoot, N. J., Agami, R., Stewart, A., & Svejstrup, J. Q. (2019). SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell*, *177*(7), 1797–1813. https://doi.org/10.1016/j.cell.2019.04.038

Guarnaccia, A. D., Rose, K. L., Wang, J., Zhao, B., Popay, T. M., Wang, C. E., Guerrazzi, K., Hill, S., Woodley, C. M., Hansen, T. J., Lorey, S. L., Shaw, J. G., Payne, W. G., Weissmiller, A. M., Olejniczak, E. T., Fesik, S. W., Liu, Q., & Tansey, W. P. (2021). Impact of WIN site inhibitor on the WDR5 interactome. *Cell Reports*, *34*(3). https://doi.org/10.1016/j.celrep.2020.108636

Guarnaccia, A. D., & Tansey, W. P. (2018). Moonlighting with WDR5: A cellular multitasker. https://doi.org/10.3390/jcm7020021

Haberle, V., & Stark, A. (2018, October). Eukaryotic core promoters and the functional basis of transcription initiation. https://doi.org/10.1038/s41580-018-0028-8

Hanna, C. W., Taudt, A., Huang, J., Gahurova, L., Kranz, A., Andrews, S., Dean, W., Stewart, A. F., Colomé-Tatché, M., & Kelsey, G. (2018). MLL2 conveys transcription-independent H3K4 trimethylation in oocytes (20180101st ed.). *Nat Struct Mol Biol*, *25*(1), 73–82. https://doi.org/10.1038/s41594-017-0013-5

Hanna, C. W., Huang, J., Belton, C., Reinhardt, S., Dahl, A., Andrews, S., Francis Stewart, A., Kranz, A., & Kelsey, G. (2022). Loss of histone methyltransferase SETD1B in oogenesis results in the redistribution of genomic histone 3 lysine 4 trimethylation. *Nucleic Acids Research*, *50*(4), 1993–2004. https://doi.org/10.1093/nar/gkac051

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., & Ching, C. W. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*(7243), 108–112.

Herz, H.-M., Mohan, M., Garruss, A. S., Liang, K., Takahashi, Y.-h., Mickey, K., Voets, O., Verrijzer, C. P., & Shilatifard, A. (2012). Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. *Genes Dev*, *26*(23), 2604–2620.

Howe, F. S., Fischl, H., Murray, S. C., & Mellor, J. (2017). Is H3K4me3 instructive for transcription activation? *BioEssays*, *39*(1), e201600095. https://doi.org/https://doi.org/10.1002/bies.201600095

Hsieh, J. J., Cheng, E. H., & Korsmeyer, S. J. (2003). Taspase1: a threonine aspartase required for cleavage of MLL and proper HOX gene expression (2003/11/26). *Cell*, *115*(3), 293–303. https://doi.org/10.1016/s0092-8674(03)00816-x

Hu, D., Gao, X., Cao, K., Morgan, M. A., Mas, G., Smith, E. R., Volk, A. G., Bartom, E. T., Crispino, J. D., Di Croce, L., & Shilatifard, A. (2017). Not All H3K4 Methylations Are Created Equal: Mll2/COMPASS Dependency in Primordial Germ Cell Specification. *Mol Cell*, *65*(3), 460–475. https://doi.org/10.1016/j.molcel.2017.01.013

Hu, D., Gao, X., Morgan, M. A., Herz, H.-M., Smith, E. R., & Shilatifard, A. (2013). The MLL3/MLL4 Branches of the COMPASS Family Function as Major Histone H3K4 Monomethylases at Enhancers. *Molecular and Cellular Biology*, *33*(23), 4745–4754. https://doi.org/10.1128/mcb.01181-13

Hu, D., Garruss, A. S., Gao, X., Morgan, M. A., Cook, M., Smith, E. R., & Shilatifard, A. (2013). The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells. *Nature Structural & Molecular Biology*, *20*(9), 1093–1097. https://doi.org/10.1038/nsmb.2653

Hu, S., Song, A., Peng, L., Tang, N., Qiao, Z., Wang, Z., Lan, F., & Chen, F. X. (2023). H3K4me2/3 modulate the stability of RNA polymerase II pausing. *Cell Research*, *33*(5), 403–406. https://doi.org/10.1038/s41422-023-00794-3

Hughes, A. L., Kelley, J. R., & Klose, R. J. (2020). Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation (2020/05/04). *Biochim Biophys Acta Gene Regul Mech*, *1863*(8), 194567. https://doi.org/10.1016/j.bbagrm.2020.194567

Hughes, A. L., Szczurek, A. T., Kelley, J. R., Lastuvkova, A., Turberfield, A. H., Dimitrova, E., Blackledge, N. P., & Klose, R. J. (2023). A CpG island-encoded mechanism protects genes from premature transcription termination. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-36236-2

Hyun, K., Jeon, J., Park, K., & Kim, J. (2017, April). Writing, erasing and reading histone lysine methylations. https://doi.org/10.1038/emm.2017.11

Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R., James, K. D., Turner, D. J., Smith, C., Harrison, D. J., Andrews, R., & Bird, A. P. (2010). Orphan CpG Islands Identify numerous conserved promoters in the mammalian genome. *PLoS Genetics*, *6*(9). https://doi.org/10.1371/journal.pgen.1001134

Jain, K., Marunde, M. R., Burg, J. M., Gloor, S. L., Joseph, F. M., Poncha, K. F., Gillespie, Z. B., Rodriguez, K. L., Popova, I. K., Hall, N. W., Vaidya, A., Howard, S. A., Taylor, H. F., Mukhsinova, L., Onuoha, U. C., Patteson, E. F., Cooke, S. W., Taylor, B. C., Weinzapfel, E. N., … Strahl, B. D. (2023). *An acetylation-mediated chromatin switch governs H3K4 methylation read-write capability* (tech. rep.).

Jiang, H., Shukla, A., Wang, X., Chen, W.-y., Bernstein, B. E., & Roeder, R. G. (2011). Role for Dpy-30 in ES Cell-Fate Specification by Regulation of H3K4 Methylation within Bivalent Domains. *Cell*, *144*(4), 513–525. https://doi.org/10.1016/j.cell.2011.01.020

Jonkers, I., Kwak, H., & Lis, J. T. (2014). Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons (K. Struhl, Ed.). *eLife*, *3*, e02407. https://doi.org/10.7554/eLife.02407

Jozwik, K. M., Chernukhin, I., Serandour, A. A., Nagarajan, S., & Carroll, J. S. (2016). FOXA1 Directs H3K4 Monomethylation at Enhancers via Recruitment of the Methyltransferase MLL3. *Cell Reports*, *17*(10), 2715–2723. https://doi.org/https://doi.org/10.1016/j.celrep.2016.11.028

Kerry, J., Godfrey, L., Repapi, E., Tapia, M., Blackledge, N. P., Ma, H., Ballabio, E., O'Byrne, S., Ponthan, F., Heidenreich, O., Roy, A., Roberts, I., Konopleva, M., Klose, R. J., Geng, H., & Milne, T. A. (2017). MLL-AF4 Spreading Identifies Binding Sites that Are Distinct from Super-Enhancers and that Govern Sensitivity to DOT1L Inhibition in Leukemia. *Cell Reports*, *18*(2), 482–495. https://doi.org/10.1016/j.celrep.2016.12.054

Kim, J., Guermah, M., McGinty, R. K., Lee, J.-S., Tang, Z., Milne, T. A., Shilatifard, A., Muir, T. W., & Roeder, R. G. (2009). RAD6-Mediated Transcription-Coupled H2B Ubiquitylation Directly Stimulates H3K4 Methylation in Human Cells. *Cell*, *137*(3), 459–471. https://doi.org/10.1016/j.cell.2009.02.027

King, H. W., & Klose, R. J. (2017). The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. https://doi.org/10.7554/eLife.22631.001

Kouzarides, T. (2007, February). Chromatin Modifications and Their Function. https://doi.org/10.1016/j.cell.2007.02.005

Kranz, A., & Anastassiadis, K. (2020, August). The role of SETD1A and SETD1B in development and disease. https://doi.org/10.1016/j.bbagrm.2020.194578

Kwon, M., Park, K., Hyun, K., Lee, J.-H., Zhou, L., Cho, Y.-W., Ge, K., Skalnik, D. G., Muir, T. W., & Kim, J. (2020). H2B ubiquitylation enhances H3K4 methylation activities of human KMT2 family complexes. *Nucleic Acids Research*, *48*(10), 5442–5456. https://doi.org/10.1093/nar/gkaa317

Lai, B., Lee, J. E., Jang, Y., Lifeng, W., Peng, W., & Ge, K. (2017). MLL3/MLL4 are required for CBP/p300 binding on enhancers and super-enhancer formation in brown adipogenesis. *Nucleic Acids Research*, *45*(11), 6388–6403. https://doi.org/10.1093/nar/gkx234

Lalonde, M. E., Avvakumov, N., Glass, K. C., Joncas, F. H., Saksouk, N., Holliday, M., Paquet, E., Yan, K., Tong, Q., Klein, B. J., Tan, S., Yang, X. J., Kutateladze, T. G., & Côté, J. (2013). Exchange of associated factors directs a switch in HBO1 acetyltransferase histone tail specificity. *Genes and Development*, *27*(18), 2009–2024. https://doi.org/10.1101/gad.223396.113

Lauberth, S. M., Nakayama, T., Wu, X., Ferris, A. L., Tang, Z., Hughes, S. H., & Roeder, R. G. (2013). H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell*, *152*(5), 1021–1036. https://doi.org/10.1016/j.cell.2013.01.052

Lee, J. H., & Skalnik, D. G. (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 (2007/11/14). *Mol Cell Biol*, *28*(2), 609–618. https://doi.org/10.1128/mcb.01356-07

Lee, J. H., & Skalnik, D. G. (2002). CpG-binding protein is a nuclear matrix- and euchromatin-associated protein localized to nuclear speckles containing human trithorax: Identification of nuclear matrix targeting signals. *Journal of Biological Chemistry*, *277*(44), 42259–42267. https://doi.org/10.1074/jbc.M205054200

Lee, J. H., Tate, C. M., You, J. S., & Skalnik, D. G. (2007). Identification and characterization of the human Set1B histone H3-Lys 4 methyltransferase complex. *Journal of Biological Chemistry*, *282*(18), 13419–13428. https://doi.org/10.1074/jbc.M609809200

Lee, J.-H., You, J., Dobrota, E., & Skalnik, D. G. (2010). Identification and Characterization of a Novel Human PP1 Phosphatase Complex *. *Journal of Biological Chemistry*, *285*(32), 24466–24476. https://doi.org/10.1074/jbc.M110.109801

Lee, J. E., Wang, C., Xu, S., Cho, Y. W., Wang, L., Feng, X., Baldridge, A., Sartorelli, V., Zhuang, L., Peng, W., & Ge, K. (2013). H3K4 mono- And di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *eLife*, *2013*(2). https://doi.org/10.7554/eLife.01503

Lee, J.-E., Cho, Y.-W., Deng, C.-X., & Ge, K. (2020). MLL3/MLL4-Associated PAGR1 Regulates Adipogenesis by Controlling Induction of C/EBP β and C/EBP δ. *Molecular and Cellular Biology*, *40*(17). https://doi.org/10.1128/mcb.00209-20

Leroy, G., Dimaggio, P. A., Chan, E. Y., Zee, B. M., Blanco, M. A., Bryant, B., Flaniken, I. Z., Liu, S., Kang, Y., Trojer, P., & Garcia, B. A. (2013). A quantitative atlas of histone modification signatures from human cancer cells. *Epigenetics and Chromatin*, *6*(1). https://doi.org/10.1186/1756-8935-6-20

Li, Y., Han, J., Zhang, Y., Cao, F., Liu, Z., Li, S., Wu, J., Hu, C., Wang, Y., Shuai, J., Chen, J., Cao, L., Li, D., Shi, P., Tian, C., Zhang, J., Dou, Y., Li, G., Chen, Y., & Lei, M. (2016). Structural basis for activity

regulation of MLL family methyltransferases. *Nature*, *530*(7591), 447–452. https://doi.org/10.1038/nature16952

Light, W. H., Freaney, J., Sood, V., Thompson, A., D'Urso, A., Horvath, C. M., & Brickner, J. H. (2013). A Conserved Role for Human Nup98 in Altering Chromatin Structure and Promoting Epigenetic Transcriptional Memory. *PLoS Biology*, *11*(3). https://doi.org/10.1371/journal.pbio.1001524

Ling, Y., Smith, A. J., & Morgan, G. T. (2006). A sequence motif conserved in diverse nuclear proteins identifies a protein interaction domain utilised for nuclear targeting by human TFIIS. *Nucleic Acids Res*, *34*(8), 2219–2229. https://doi.org/10.1093/nar/gkl239

Local, A., Huang, H., Albuquerque, C. P., Singh, N., Lee, A. Y., Wang, W., Wang, C., Hsia, J. E., Shiau, A. K., Ge, K., Corbett, K. D., Wang, D., Zhou, H., & Ren, B. (2018). Identification of H3K4me1-associated proteins at mammalian enhancers. *Nature Genetics*, *50*(1), 73–82. https://doi.org/10.1038/s41588-017-0015-6

Long, H. K., Sims, D., Heger, A., Blackledge, N. P., Kutter, C., Wright, M. L., Grützner, F., Odom, D. T., Patient, R., Ponting, C. P., & Klose, R. J. (2013). Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife*, *2013*(2). https://doi.org/10.7554/eLife.00348

Mahajan, M. A., & Samuels, H. H. (2008). Nuclear receptor coactivator/coregulator NCoA6(NRC) is a pleiotropic coregulator involved in transcription, cell survival, growth and development. https://doi.org/10.1621/nrs.06002

Mas, G., Blanco, E., Ballaré, C., Sansó, M., Spill, Y. G., Hu, D., Aoi, Y., Le Dily, F., Shilatifard, A., Marti-Renom, M. A., & Di Croce, L. (2018). Promoter bivalency favors an open chromatin architecture in embryonic stem cells (20180917th ed.). *Nat Genet*, *50*(10), 1452–1462. https://doi.org/10.1038/s41588-018-0218-5

Mellor, J. (2006, July). It Takes a PHD to Read the Histone Code. https://doi.org/10.1016/j.cell.2006.06.028

Millán-Zambrano, G., Burton, A., Bannister, A. J., & Schneider, R. (2022, September). Histone post-translational modifications — cause and consequence of genome function. https://doi.org/10.1038/s41576-022-00468-7

Milne, T. A., Briggs, S. D., Brock, H. W., Martin, M. E., Gibbs, D., Allis, C., & Hess, J. L. (2002). MLL Targets SET Domain Methyltransferase Activity to Hox Gene Promoters. *Molecular Cell*, *10*(5), 1107–1117. https://doi.org/https://doi.org/10.1016/S1097-2765(02)00741-4

Milne, T. A., Hughes, C. M., Lloyd, R., Yang, Z., Rozenblatt-Rosen, O., Dou, Y., Schnepp, R. W., Krankel, C., Livolsi, V. A., Gibbs, D., Hua, X., Roeder, R. G., Meyerson, M., & Hess, J. L. (2004). *Menin and MLL cooperatively regulate expression of cyclin-dependent kinase inhibitors* (tech. rep.). www.pnas.orgcgidoi10.1073pnas.0408836102

Milne, T. A., Kim, J., Wang, G. G., Stadler, S. C., Basrur, V., Whitcomb, S. J., Wang, Z., Ruthenburg, A. J., Elenitoba-Johnson, K. S., Roeder, R. G., & Allis, C. D. (2010). Multiple Interactions Recruit MLL1

and MLL1 Fusion Proteins to the HOXA9 Locus in Leukemogenesis. *Molecular Cell*, *38*(6), 853–863. https://doi.org/10.1016/j.molcel.2010.05.011

Mishra, B. P., Zaffuto, K. M., Artinger, E. L., Org, T., Mikkola, H. K., Cheng, C., Djabali, M., & Ernst, P. (2014). The histone methyltransferase activity of MLL1 is dispensable for hematopoiesis and leukemogenesis. *Cell Reports*, *7*(4), 1239–1247. https://doi.org/10.1016/j.celrep.2014.04.015

Miyamoto, R., Okuda, H., Kanai, A., Takahashi, S., Kawamura, T., Matsui, H., Kitamura, T., Kitabayashi, I., Inaba, T., & Yokoyama, A. (2020). Activation of CpG-Rich Promoters Mediated by MLL Drives MOZ-Rearranged Leukemia. *Cell Reports*, *32*(13). https://doi.org/10.1016/j.celrep.2020.108200

Morrison, E. A., Bowerman, S., Sylvers, K. L., Wereszczynski, J., & Musselman, C. A. (2018). The conformation of the histone H3 tail inhibits association of the BPTF PHD finger with the nucleosome (G. J. Narlikar, Ed.). *eLife*, *7*, e31481. https://doi.org/10.7554/eLife.31481

Muntean, A. G., Tan, J., Sitwala, K., Huang, Y., Bronstein, J., Connelly, J. A., Basrur, V., Elenitoba-Johnson, K. S., & Hess, J. L. (2010). The PAF Complex Synergizes with MLL Fusion Proteins at HOX Loci to Promote Leukemogenesis. *Cancer Cell*, *17*(6), 609–621. https://doi.org/10.1016/j.ccr.2010.04.012

Nabet, B., Roberts, J. M., Buckley, D. L., Paulk, J., Dastjerdi, S., Yang, A., Leggett, A. L., Erb, M. A., Lawlor, M. A., Souza, A., Scott, T. G., Vittori, S., Perry, J. A., Qi, J., Winter, G. E., Wong, K.-K., Gray, N. S., & Bradner, J. E. (2018). The dTAG system for immediate and target-specific protein degradation. *Nature chemical biology*, *14*(5), 431–441. https://doi.org/10.1038/s41589-018-0021-8

Oh, E., Mark, K. G., Mocciaro, A., Watson, E. R., Prabu, J. R., Cha, D. D., Kampmann, M., Gamarra, N., Zhou, C. Y., & Rape, M. (2020). Gene expression and cell identity controlled by anaphase-promoting complex. *Nature*, *579*(7797), 136–140. https://doi.org/10.1038/s41586-020-2034-1

Okuda, H., Kawaguchi, M., Kanai, A., Matsui, H., Kawamura, T., Inaba, T., Kitabayashi, I., & Yokoyama, A. (2014). MLL fusion proteins link transcriptional coactivators to previously active CpG-rich promoters. *Nucleic Acids Research*, *42*(7), 4241–4256. https://doi.org/10.1093/nar/gkt1394

Pachano, T., Sánchez-Gaya, V., Ealo, T., Mariner-Faulí, M., Bleckwehl, T., Asenjo, H. G., Respuela, P., Cruz-Molina, S., Muñoz-San Martín, M., Haro, E., van IJcken, W. F., Landeira, D., & Rada-Iglesias, A. (2021). Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. *Nature Genetics*, *53*(7), 1036–1049. https://doi.org/10.1038/s41588-021-00888-x

Proudfoot, N. J. (2016). Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science*, *352*(6291), aad9926. https://doi.org/10.1126/science.aad9926

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A., & Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, *470*(7333), 279–283.

Rahman, S., Hoffmann, N. A., Worden, E. J., Smith, M. L., Namitz, K. E. W., Knutson, B. A., Cosgrove, M. S., & Wolberger, C. (2022). Multistate structures of the MLL1-WRAD complex bound to H2B-ubiquitinated nucleosome. https://doi.org/10.1073/pnas

Richter, W. F., Nayak, S., Iwasa, J., & Taatjes, D. J. (2022, November). The Mediator complex as a master regulator of transcription by RNA polymerase II. https://doi.org/10.1038/s41580-022-00498-3

Rickels, R., Herz, H. M., Sze, C. C., Cao, K., Morgan, M. A., Collings, C. K., Gause, M., Takahashi, Y. H., Wang, L., Rendleman, E. J., Marshall, S. A., Krueger, A., Bartom, E. T., Piunti, A., Smith, E. R., Abshiru, N. A., Kelleher, N. L., Dorsett, D., & Shilatifard, A. (2017). Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. *Nature Genetics*, *49*(11), 1647–1653. https://doi.org/10.1038/ng.3965

Rodríguez-Molina, J. B., O'Reilly, F. J., Fagarasan, H., Sheekey, E., Maslen, S., Skehel, J. M., Rappsilber, J., & Passmore, L. A. (2022). Mpe1 senses the binding of pre-mRNA and controls 3 end processing by CPF. *Molecular Cell*, *82*(13), 2490–2504. https://doi.org/https://doi.org/10.1016/j.molcel.2022.04.021

Roguev, A., Schaft, D., Shevchenko, A., Pijnappel, W., Wilm, M., Aasland, R., & Stewart, A. (2001). The Saccharomyces cerevisiae Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *The EMBO Journal*, *20*(24), 7137–7148. https://doi.org/https://doi.org/10.1093/emboj/20.24.7137

Rose, N. R., King, H. W., Blackledge, N. P., Fursova, N. A., Ember, K. J., Fischer, R., Kessler, B. M., & Klose, R. J. (2016). RYBP stimulates PRC1 to shape chromatin-based communication between Polycomb repressive complexes. https://doi.org/10.7554/eLife.18591.001

Ruthenburg, A. J., Allis, C. D., & Wysocka, J. (2007, January). Methylation of Lysine 4 on Histone H3: Intricacy of Writing and Reading a Single Epigenetic Mark. https://doi.org/10.1016/j.molcel.2006.12.014

Saxonov, S., Berg, P., & Brutlag, D. L. (2005). *A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters* (tech. rep.). www.pnas.orgcgidoi10.1073pnas.0510310103

Schübeler, D. (2015, January). Function and information content of DNA methylation. https://doi.org/10.1038/nature14192

Schwab, K. R., Patel, S. R., & Dressler, G. R. (2011). Role of PTIP in Class Switch Recombination and Long-Range Chromatin Interactions at the Immunoglobulin Heavy Chain Locus. *Molecular and Cellular Biology*, *31*(7), 1503–1511. https://doi.org/10.1128/MCB.00990-10

Sen, P., Dang, W., Donahue, G., Dai, J., Dorsey, J., Cao, X., Liu, W., Cao, K., Perry, R., Lee, J. Y., Wasko, B. M., Carr, D. T., He, C., Robison, B., Wagner, J., Gregory, B. D., Kaeberlein, M., Kennedy, B. K., Boeke, J. D., & Berger, S. L. (2015). H3K36 methylation promotes longevity by enhancing transcriptional fidelity. https://doi.org/10.1101/gad.263707

Shah, R. N., Grzybowski, A. T., Cornett, E. M., Johnstone, A. L., Dickson, B. M., Boone, B. A., Cheek, M. A., Cowles, M. W., Maryanski, D., Meiners, M. J., Tiedemann, R. L., Vaughan, R. M., Arora, N., Sun, Z. W., Rothbart, S. B., Keogh, M. C., & Ruthenburg, A. J. (2018). Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies (20180920th ed.). *Mol Cell*, *72*(1), 162–177. https://doi.org/10.1016/j.molcel.2018.08.015

Shilatifard, A. (2012). The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis (2012/06/06). *Annu Rev Biochem*, *81*, 65–95. https://doi.org/10.1146/annurev-biochem-051710-134100

Shin Voo, K., Carlone, D. L., Jacobsen, B. M., Flodin, A., & Skalnik, D. G. (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. *Molecular and Cellular Biology*, *20*(6), 2108–2121. https://doi.org/10.1128/MCB.20.6.2108-2121.2000

Shinsky, S. A., & Cosgrove, M. S. (2015). Unique role of the WD-40 repeat protein 5 (WDR5) subunit within the mixed lineage leukemia 3 (MLL3) histone methyltransferase complex. *Journal of Biological Chemistry*, *290*(43), 25819–25833. https://doi.org/10.1074/jbc.M115.684142

Shinsky, S. A., Hu, M., Vought, V. E., Ng, S. B., Bamshad, M. J., Shendure, J., & Cosgrove, M. S. (2014). A non-active-site SET domain surface crucial for the interaction of MLL1 and the RbBP5/Ash2L heterodimer within MLL family core complexes. *Journal of Molecular Biology*, *426*(12), 2283–2299. https://doi.org/10.1016/j.jmb.2014.03.011

Shinsky, S. A., Monteith, K. E., Viggiano, S., & Cosgrove, M. S. (2015). Biochemical reconstitution and phylogenetic comparison of human SET1 family core complexes involved in histone methylation. *Journal of Biological Chemistry*, *290*(10), 6361–6375. https://doi.org/10.1074/jbc.M114.627646

Siladi, A. J., Wang, J., Florian, A. C., Thomas, L. R., Creighton, J. H., Matlock, B. K., Flaherty, D. K., Lorey, S. L., Howard, G. C., Fesik, S. W., Weissmiller, A. M., Liu, Q., & Tansey, W. P. (2022). WIN site inhibition disrupts a subset of WDR5 function. *Scientific Reports*, *12*(1), 1848. https://doi.org/10.1038/s41598-022-05947-9

Sims, R. J., Chen, C. F., Santos-Rosa, H., Kouzarides, T., Patel, S. S., & Reinberg, D. (2005). Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *Journal of Biological Chemistry*, *280*(51), 41789–41792. https://doi.org/10.1074/jbc.C500395200

Soares, L. M., He, P. C., Chun, Y., Suh, H., Kim, T., & Buratowski, S. (2017). Determinants of Histone H3K4 Methylation Patterns. *Molecular Cell*, *68*(4), 773–785. https://doi.org/https://doi.org/10.1016/j.molcel.2017.10.013

Sparbier, C. E., Gillespie, A., Gomez, J., Kumari, N., Motazedian, A., Chan, K. L., Bell, C. C., Gilan, O., Chan, Y. C., Popp, S., Gough, D. J., Eckersley-Maslin, M. A., Dawson, S. J., Lehner, P. J., Sutherland, K. D., Ernst, P., McGeehan, G. M., Lam, E. Y., Burr, M. L., & Dawson, M. A. (2023). Targeting Menin disrupts the KMT2A/B and polycomb balance to paradoxically activate bivalent genes. *Nature Cell Biology*, *25*(2), 258–272. https://doi.org/10.1038/s41556-022-01056-x

Stroynowska-Czerwinska, A. M., Klimczak, M., Pastor, M., Kazrani, A. A., Misztal, K., & Bochtler, M. (2023). Clustered PHD domains in KMT2/MLL proteins are attracted by H3K4me3 and H3 acetylation-rich active promoters and enhancers. *Cellular and Molecular Life Sciences*, *80*(1). https://doi.org/10.1007/s00018-022-04651-1

Sump, B., Brickner, D. G., D'urso, A., Kim, S. H., & Brickner, J. H. (2022). Mitotically heritable, RNA polymerase II-independent H3K4 dimethylation stimulates INO1 transcriptional memory. *eLife*, *11*. https://doi.org/10.7554/eLife.77646

Szczurek, A. T., Dimitrova, E., Kelley, J. R., & Klose, R. J. (2023). Polycomb sustains promoters in a deep OFF-state by limiting PIC formation to counteract transcription. *bioRxiv*, 2023.06.13.544762. https://doi.org/10.1101/2023.06.13.544762

Sze, C. C., Cao, K., Collings, C. K., Marshall, S. A., Rendleman, E. J., Ozark, P. A., Chen, F. X., Morgan, M. A., Wang, L., & Shilatifard, A. (2017). Histone H3K4 methylation-dependent and -independent functions of set1A/COMPASS in embryonic stem cell self-renewal and differentiation. *Genes and Development*, *31*(17), 1732–1737. https://doi.org/10.1101/gad.303768.117

Sze, C. C., Ozark, P. A., Cao, K., Ugarenko, M., Das, S., Wang, L., Marshall, S. A., Rendleman, E. J., Ryan, C. A., Zha, D., Douillet, D., Chen, F. X., & Shilatifard, A. (2020). Coordinated regulation of cellular identity–associated H3K4me3 breadth by the COMPASS family. *Science Advances*, *6*(26), eaaz4764. https://doi.org/10.1126/sciadv.aaz4764

Sze, C. C., Ozark, P. A., Cao, K., Ugarenko, M., Das, S., Wang, L., Marshall, S. A., Rendleman, E. J., Ryan, C. A., Zha, D., Douillet, D., Chen, F. X., & Shilatifard, A. (2023). Coordinated regulation of cellular identity–associated H3K4me3 breadth by the COMPASS family. *Science Advances*, *6*(26), eaaz4764. https://doi.org/10.1126/sciadv.aaz4764

Takahashi, Y. H., Lee, J. S., Swanson, S. K., Saraf, A., Florens, L., Washburn, M. P., Trievel, R. C., & Shilatifard, A. (2009). Regulation of H3K4 Trimethylation via Cps40 (Spp1) of COMPASS Is Monoubiquitination Independent: Implication for a Phe/Tyr Switch by the Catalytic Domain of Set1. *Molecular and Cellular Biology*, *29*(13), 3478–3486. https://doi.org/10.1128/mcb.00013-09

Takahashi, Y.-h., Westfield, G. H., Oleskie, A. N., Trievel, R. C., Shilatifard, A., & Skiniotis, G. (2011). Structural analysis of the core COMPASS family of histone H3K4 methylases from yeast to human. *Proceedings of the National Academy of Sciences*, *108*(51), 20526–20531. https://doi.org/10.1073/pnas.1109360108

Takeda, S., Chen, D. Y., Westergard, T. D., Fisher, J. K., Rubens, J. A., Sasagawa, S., Kan, J. T., Korsmeyer, S. J., Cheng, E. H. Y., & Hsieh, J. J. D. (2006). Proteolysis of MLL family proteins is essential for taspase1-orchestrated cell cycle progression. *Genes Dev*, *20*(17), 2397–2409. https://doi.org/10.1101/gad.1449406

Tanaka, Y., Okamoto, K., Teye, K., Umata, T., Yamagiwa, N., Suto, Y., Zhang, Y., & Tsuneoka, M. (2010). JmjC enzyme KDM2A is a regulator of rRNA transcription in response to starvation. *The EMBO Journal*, *29*(9), 1510–1522. https://doi.org/https://doi.org/10.1038/emboj.2010.56

Tazi, J., & Bird, A. (1990). Alternative chromatin structure at CpG islands. *Cell*, *60*(6), 909–920. https://doi.org/https://doi.org/10.1016/0092-8674(90)90339-G

Terranova, R., Agherbi, H., Boned, A., Meresse, S., & Djabali, M. (2006). Histone and DNA methylation defects at Hox genes in mice expressing a SET domain-truncated form of Mll. *Proceedings of the National Academy of Sciences*, *103*(17), 6629–6634. https://doi.org/10.1073/pnas.0507425103

Tettey, T. T., Gao, X., Shao, W., Li, H., Story, B. A., Chitsazan, A. D., Glaser, R. L., Goode, Z. H., Seidel, C. W., Conaway, R. C., Zeitlinger, J., Blanchette, M., & Conaway, J. W. (2019). A Role for FACT in RNA Polymerase II Promoter-Proximal Pausing. *Cell Reports*, *27*(13), 3770–3779. https://doi.org/10.1016/j.celrep.2019.05.099

Thomas, L. R., Wang, Q., Grieb, B. C., Phan, J., Foshage, A. M., Sun, Q., Olejniczak, E. T., Clark, T., Dey, S., Lorey, S., Alicie, B., Howard, G. C., Cawthon, B., Ess, K. C., Eischen, C. M., Zhao, Z., Fesik, S. W., & Tansey, W. P. (2015). Interaction with WDR5 promotes target gene recognition and tumorigenesis by MYC. *Molecular Cell*, *58*(3), 440–452. https://doi.org/10.1016/j.molcel.2015.02.028

Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A. R., Deaton, A., Andrews, R., James, K. D., Turner, D. J., Illingworth, R., & Bird, A. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, *464*(7291), 1082–1086. https://doi.org/10.1038/nature08924

Titov, D. V., Gilman, B., He, Q.-L., Bhat, S., Low, W.-K., Dang, Y., Smeaton, M., Demain, A. L., Miller, P. S., Kugel, J. F., Goodrich, J. A., & Liu, J. O. (2011). XPB, a subunit of TFIIH, is a target of the natural product triptolide. *Nature Chemical Biology*, *7*(3), 182–188. https://doi.org/10.1038/nchembio.522

Turberfield, A. H., Kondo, T., Nakayama, M., Koseki, Y., King, H. W., Koseki, H., & Klose, R. J. (2019). KDM2 proteins constrain transcription from CpG island gene promoters independently of their histone demethylase activity. *Nucleic Acids Research*, *47*(17), 9005–9023. https://doi.org/10.1093/NAR/GKZ607

Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nature Protocols*, *11*(12), 2301–2319. https://doi.org/10.1038/nprot.2016.136

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

Ullius, A., Lüscher-Firzlaff, J., Costa, I. G., Walsemann, G., Forst, A. H., Gusmao, E. G., Kapelle, K., Kleine, H., Kremmer, E., Vervoorts, J., & Lüscher, B. (2014). The interaction of MYC with the trithorax protein ASH2L promotes gene transcription by regulating H3K27 modification. *Nucleic Acids Research*, *42*(11), 6901–6920. https://doi.org/10.1093/nar/gku312

Vanderkruk, B., Maeshima, N., Pasula, D. J., An, M., McDonald, C. L., Suresh, P., Luciani, D. S., Lynn, F. C., & Hoffman, B. G. (2023). Methylation of histone H3 lysine 4 is required for maintenance of beta cell function in adult mice. *Diabetologia*, *66*(6), 1097–1115. https://doi.org/10.1007/s00125-023-05896-6

van Nuland, R., Smits, A. H., Pallaki, P., Jansen, P. W., Vermeulen, M., & Timmers, H. T. (2013). Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes (20130318th ed.). *Mol Cell Biol*, *33*(10), 2067–2077. https://doi.org/10.1128/mcb.01742-12

Vermeulen, M., Eberl, H. C., Matarese, F., Marks, H., Denissov, S., Butter, F., Lee, K. K., Olsen, J. V., Hyman, A. A., Stunnenberg, H. G., & Mann, M. (2010). Quantitative Interaction Proteomics and Genome-wide Profiling of Epigenetic Histone Marks and Their Readers. *Cell*, *142*(6), 967–980. https://doi.org/10.1016/j.cell.2010.08.020

Vermeulen, M., Mulder, K. W., Denissov, S., Pijnappel, W. W., van Schaik, F. M., Varier, R. A., Baltissen, M. P., Stunnenberg, H. G., Mann, M., & Timmers, H. T. M. (2007). Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4. *Cell*, *131*(1), 58–69. https://doi.org/10.1016/j.cell.2007.08.016

Vilhais-Neto, G. C., Fournier, M., Plassat, J. L., Sardiu, M. E., Saraf, A., Garnier, J. M., Maruhashi, M., Florens, L., Washburn, M. P., & Pourquié, O. (2017). The WHHERE coactivator complex is required for retinoic acid-dependent regulation of embryonic symmetry. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-00593-6

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

Wachter, E., Quante, T., Merusi, C., Arczewska, A., Stewart, F., Webb, S., & Bird, A. (2014). Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. *eLife*, *3*, e03397. https://doi.org/10.7554/eLife.03397

Wang, C., Lee, J.-E., Lai, B., Macfarlan, T. S., Xu, S., Zhuang, L., Liu, C., Peng, W., & Ge, K. (2016). Enhancer priming by H3K4 methyltransferase MLL4 controls cell fate transition. *Proceedings of the National Academy of Sciences*, *113*(42), 11871–11876.

Wang, H., Fan, Z., Shliaha, P. V., Miele, M., Hendrickson, R. C., Jiang, X., & Helin, K. (2023). H3K4me3 regulates RNA polymerase II promoter-proximal pause-release. *Nature*, *615*(7951), 339–348. https://doi.org/10.1038/s41586-023-05780-8

Wang, S. P., Tang, Z., Chen, C. W., Shimada, M., Koche, R. P., Wang, L. H., Nakadai, T., Chramiec, A., Krivtsov, A. V., Armstrong, S. A., & Roeder, R. G. (2017). A UTX-MLL4-p300 Transcriptional Regulatory Network Coordinately Shapes Active Enhancer Landscapes for Eliciting Transcription. *Molecular Cell*, *67*(2), 308–321. https://doi.org/10.1016/j.molcel.2017.06.028

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, F. X. (2023). Coordinated regulation of RNA polymerase II pausing and elongation progression by PAF1. *Science Advances*, *8*(13), eabm5504. https://doi.org/10.1126/sciadv.abm5504

Wang, Z., Chivu, A. G., Choate, L. A., Rice, E. J., Miller, D. C., Chu, T., Chou, S. P., Kingsley, N. B., Petersen, J. L., Finno, C. J., Bellone, R. R., Antczak, D. F., Lis, J. T., & Danko, C. G. (2022). Prediction of histone post-translational modification patterns based on nascent transcription data. *Nature Genetics*, *54*(3), 295–305. https://doi.org/10.1038/s41588-022-01026-x

Wen, H., Li, J., Song, T., Lu, M., Kan, P. Y., Lee, M. G., Sha, B., & Shi, X. (2010). Recognition of histone H3K4 trimethylation by the plant homeodomain of PHF2 modulates histone demethylation. *Journal of Biological Chemistry*, *285*(13), 9322–9326. https://doi.org/10.1074/jbc.C109.097667

Worden, E. J., & Wolberger, C. (2019). Activation and regulation of H2B-Ubiquitin-dependent histone methyltransferases. *Current Opinion in Structural Biology*, *59*, 98–106. https://doi.org/https://doi.org/10.1016/j.sbi.2019.05.009

Worden, E. J., Zhang, X., & Wolberger, C. (2020). Structural basis for COMPASS recognition of an H2B-ubiquitinated nucleosome (J. Kuriyan, C. P. Hill, & K. Luger, Eds.). *eLife*, *9*, e53199. https://doi.org/10.7554/eLife.53199

Wysocka, J., Swigut, T., Xiao, H., Milne, T. A., Kwon, S. Y., Landry, J., Kauer, M., Tackett, A. J., Chait, B. T., Badenhorst, P., Wu, C., & Allis, C. D. (2006). A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, *442*(7098), 86–90. https://doi.org/10.1038/nature04815

Xie, G., Lee, J. E., Senft, A. D., Park, Y. K., Jang, Y., Chakraborty, S., Thompson, J. J., McKernan, K., Liu, C., Macfarlan, T. S., Rocha, P. P., Peng, W., & Ge, K. (2023). MLL3/MLL4 methyltransferase activities control early embryonic development and embryonic stem cell differentiation in a lineage-selective manner. *Nature Genetics*, *55*(4), 693–705. https://doi.org/10.1038/s41588-023-01356-4

Xue, H., Yao, T., Cao, M., Zhu, G., Li, Y., Yuan, G., Chen, Y., Lei, M., & Huang, J. (2019). Structural basis of nucleosome recognition and modification by MLL methyltransferases. *Nature*, *573*(7774), 445–449. https://doi.org/10.1038/s41586-019-1528-1

Yan, J., Chen, S. A. A., Local, A., Liu, T., Qiu, Y., Dorighi, K. M., Preissl, S., Rivera, C. M., Wang, C., Ye, Z., Ge, K., Hu, M., Wysocka, J., & Ren, B. (2018). Histone H3 lysine 4 monomethylation modulates long-range chromatin interactions at enhancers. *Cell Research*, *28*(2), 204–220. https://doi.org/10.1038/cr.2018.1

Yokoyama, A., Ficara, F., Murphy, M. J., Meisel, C., Naresh, A., Kitabayashi, I., & Cleary, M. L. (2011). Proteolytically cleaved MLL subunits are susceptible to distinct degradation pathways. *Journal of Cell Science*, *124*(13), 2208–2219. https://doi.org/10.1242/jcs.080523

Yu, B. D., Hanson, R. D., Hess, J. L., Horning, S. E., & Korsmeyer, S. J. (1998). *MLL, a mammalian trithorax-group gene, functions as a transcriptional maintenance factor in morphogenesis* (tech. rep.). www.pnas.org.

Zhang, C., Lin, H., Zhang, Y., Xing, Q., Zhang, J., Zhang, D., Liu, Y., Chen, Q., Zhou, T., Wang, J., Shan, Y., & Pan, G. (2023). BRPF1 bridges H3K4me3 and H3K23ac in human embryonic stem cells and is essential to pluripotency. *iScience*, *26*(2). https://doi.org/10.1016/j.isci.2023.105939