Investigating the role of SET1A and SET1B in the regulation of gene expression

Amy Laura Hughes



WOLFSON COLLEGE, UNIVERSITY OF OXFORD

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

HILARY TERM, 2021

Abstract

The development of complex multicellular organisms relies on the precise spatiotemporal control of gene expression. The post-translational modification of histone proteins within chromatin has emerged as an important mechanism by which the regulation of gene expression can be achieved. One such histone modification, H3K4me3, is intrinsically linked to transcriptional activity; levels of H3K4me3 at gene promoters are positively correlated with gene expression and numerous H3K4me3 'reader' proteins have been identified that could elicit a positive transcriptional outcome. Despite this, H3K4me3 appears to be dispensable for gene transcription in simple eukaryotes. Studying H3K4me3 in mammalian cells is complicated by the diversification in the H3K4 methylation machinery during evolution. Mammalian cells produce six complexes proposed to methylate H3K4: MLL1-4 and SET1A/B. SET1A/B broadly associate with gene promoters and have been proposed to place the majority of H3K4me3 in mammalian genomes. However, the effects of SET1A/B removal on gene expression have typically been subtle and lacked directionality. Therefore, the contribution of SET1A/B to the regulation of gene expression, either through the deposition of H3K4me3 or otherwise, has remained largely unclear.

To address the role of SET1A/B in mammalian gene expression I have developed cell lines in which SET1A and SET1B can be rapidly depleted from cells either individually or in combination. By using these rapid depletion systems coupled with quantitative gene expression analysis I find that SET1A/B function to support active gene expression in embryonic stem cells. In contrast to the prevailing views in the field, I reveal that SET1A/B make a rather subtle contribution to H3K4me3. Furthermore, I find that SET1A/B are transcriptional activators that are able to activate gene transcription independently of their catalytic activity. I then begin to dissect the catalysis-independent mechanisms by which SET1A/B regulate gene expression and find that their interaction with WDR82 is essential for transcriptional activation. Furthermore, I discover that SET1A/B influence the occupancy and phosphorylation of RNA polymerase II. Together, these observations reveal an unexpected role for SET1A/B in supporting active transcription independently of their catalytic activity.

Acknowledgements

I would like to thank Rob Klose for his encouragement, guidance and enthusiasm throughout my project. I have learnt so much from being a member of your lab and I am extremely grateful to have had you as a mentor.

Thank you to Neil, Emilia and Anne for the constant experimental advice and guidance and to Nadya, Paula and Angelika for your patience in helping me with analysis. Thank you to Miles, Jess, Mafalda, Anne, Emma, and the rest of the Klose and Brockdorff labs for creating a wonderful environment to work in during my DPhil. Thank you to Tatyana and Emma for tirelessly ensuring that the lab environment is running smoothly and to Anna for helping to generate a couple of the cell lines used in this thesis. Thank you to Paula and Jess for reading this thesis.

A special thank you to Jess for being my SET1 partner in crime. A review that we wrote together in 2020 forms the basis for my introduction.

Thank you to all of my friends for the adventures and laughter over the last five years.

Most of all thank you to my family. Mum, Dad, Nan and Grandad; all I ever want is to make you proud.

Declaration of Authorship

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

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

Amy Laura Hughes

April 2021

Associated Publications

Hughes, A. L., Kelley, J. R. and Klose, R. J. (2020) 'Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1863: p. 194567. doi: 10.1016/j.bbagrm.2020.194567.

Some sections of this review are repurposed for use within the introduction of this thesis, with permission from Jessica Kelley and Rob Klose.

Contents

Abstract i
Acknowledgements ii
Declaration of Authorship iii
Associated Publications iv
Contents v
List of Figures viii
List of Tables x
List of common abbreviations xi
1. Introduction 1
1.1 The regulation of eukaryotic gene expression 1
1.1.1 Transcription initiation 1
1.1.2 Dynamic post-translational modification of RNAPII 3
1.1.3 DNA is packaged into chromatin 6
1.1.4 DNA methylation 7
1.1.5 The post-translational modification of histones 9
1.2 H3K4 methylation in mammalian genomes 11
1.2.1 Distribution of H3K4 methylation 11
1.2.2 H3K4me3 and CpG islands 12
1.2.3 H3K4me3 correlates with transcriptional activity 13
1.3 The influence of H3K4me3 on chromatin architecture 14
1.3.1 Does H3K4me3 prevent methylation of CpG islands? 14
1.3.2 Does H3K4me3 inhibit transcriptional repression? 15
1.3.3 H3K4me3 supports the formation of a permissive chromatin state 16
1.3.4 H3K4me3 and gene transcription 18
1.4 The H3K4 methylation system 20
1.4.1 Evolution of the H3K4 HMTs from yeast and *Drosophila* 20
1.4.2 Introducing the mammalian H3K4 HMTs 22
1.4.3 MLL1/2 23
1.4.3.1 The association of MLL1/2 with chromatin 23
1.4.3.2 The role of MLL1/2 in mammalian cells 26
1.4.4 SET1A/B 27
1.4.4.1 The association of SET1A/B with chromatin 27
1.4.4.2 The role of SET1A/B in mammalian cells 29
1.4.5 Can the H3K4 HMTs function independently of H3K4me3? 32

1.5 Aims of thesis 35
2. Materials and Methods 37
2.1 Cell lines and cell culture 37
2.1.1 Cell culture conditions 37
2.1.2 Cell counting 37
2.1.3 Tamoxifen, dTAG-13 and Doxycycline treatments 37
2.1.4 Embryonic stem cell lines 38
2.1.5 CRISPR-mediated genome editing 39
2.1.6 Transient transfections 42
2.1.7 Luciferase reporter assays 43
2.1.8 Cell growth analysis 43
2.2 Protein Methods 44
2.2.1 Acid extraction of histones 44
2.2.2 Preparation of nuclear extract 44
2.2.3 Immunoprecipitation 45
2.2.4 SDS-PAGE 46
2.2.5 Western blotting 46
2.2.6 Antibodies 47
2.3 Calibrated chromatin immunoprecipitation (cChIP) 48
2.3.1 Native cChIP-sequencing 48
2.3.2 Crosslinked cChIP-sequencing 50
2.4 RNA methods 51
2.4.1 RNA extraction, reverse transcription and quantitative PCR 51
2.4.2 Calibrated total RNA-sequencing (cRNA-seq) 52
2.4.3 Transient-Transcriptome (TT)-sequencing 53
2.5 Massively parallel sequencing and data processing 54
2.5.1 Massively parallel sequencing 54
2.5.2 Read alignment 54
2.5.3 Data normalisation 55
2.5.4 Peak calling and annotation 56
2.5.5 Differential gene expression analysis 57
2.5.6 Readcount quantification and enrichment analysis for cChIP-seq 57
2.5.7 Box plots, correlation and statistical analysis 58
2.5.8 Metaplots and Heatmaps 58
2.5.9 Published datasets used in this study 58
3. SET1A/B have a widespread role in gene expression 60

3.1 Introduction 60
3.2 Results 61
3.2.1 Generating a SET1A auxin-inducible degron line 61
3.2.2 A system to induce controlled and rapid degradation of SET1A 63
3.2.3 SET1A positively regulates gene expression genome-wide 66
3.2.4 Genes regulated by SET1A are actively transcribed and associated with CpG islands 70
3.2.5 SET1B regulates the expression of a small number of genes 73
3.2.6 The removal of SET1A and SET1B results in a widespread reduction in gene expression 77
3.3 Summary and Discussion 82
4. SET1A/B regulate gene expression independently of H3K4me3 86
4.1 Introduction 86
4.2 Results 87
4.2.1 Removal of SET1A results in a subtle but widespread reduction in H3K4me3 87
4.2.2 Removal of SET1A/B results in a larger reduction in H3K4me3 than removal of SET1A alone 92
4.2.3 Gene expression changes cannot be explained by changes in H3K4me3 96
4.2.4 Generating a SET1A conditional catalytic point mutant 98
4.2.5 SET1A is a transcriptional activator that can function independently of H3K4me3 102
4.3 Summary and Discussion 103
5. SET1A/B regulate RNAPII-mediated transcription 110
5.1 Introduction 110
5.2 Results 111
5.2.1 WDR82 is essential for transcription activation by SET1A 111
5.2.2 SET1A/B regulate RNAPII at gene promoters 116
5.2.3 The depletion of SET1A/B affects gene body-associated RNAPII 121
5.2.4 SET1A/B regulate gene transcription 126
5.3 Summary and discussion 132
6. Conclusions and future directions 140
A Appendices 146
Bibliography 155

List of Figures

Figure 1.1 Schematic of average ChIP profiles of phosphorylated RNAPII 4
Figure 1.2 Genomic distribution of histone H3 lysine 4 methylation 12
Figure 1.3 CpG islands and H3K4me3 13
Figure 1.4 Evolution of the H3K4 HMTs 22
Figure 1.5 Domain architecture of the mammalian H3K4 trimethyltransferases 24
Figure 1.6 Distribution of the H3K4 HMTs on chromatin 25
Figure 3.1 Generating a SET1A auxin-inducible degron line 62
Figure 3.2 Applying the dTAG system to SET1A 64
Figure 3.3 Addition of dTAG-13 results in the rapid depletion of SET1A 66
Figure 3.4 Loss of SET1A results in rapid changes in gene expression 67
Figure 3.5 SET1A positively regulates gene expression genome-wide 69
Figure 3.6 SET1A targets are enriched for chromatin features associated with active transcription 71
Figure 3.7 SET1A preferentially regulates the expression of CGI-associated genes 72
Figure 3.8 Generating a dTAG-SET1B cell line 73
Figure 3.9 SET1B contributes little to gene expression in ESCs 76
Figure 3.10 The expression of SET1B increases following depletion of SET1A 78
Figure 3.11 Charactering the dTAG-SET1A/B cell line 78
Figure 3.12 Depletion of SET1A/B results in widespread decreases in gene expression 80
Figure 3.13 Simultaneous removal of SET1A and SET1B increases the magnitude of gene expression changes 81
Figure 3.14 Removal of SET1A/B has a larger effect on cell growth than removal of SET1A alone 82
Figure 4.1 Removal of SET1A does not affect global H3K4me3 88
Figure 4.2 Optimising H3K4me3 ChIP conditions with synthetic modified nucleosomes 89
Figure 4.3 Removing SET1A has a subtle but pervasive effect on promoter-associated H3K4me3 90
Figure 4.4 Removal of SET1A results in the largest change in H3K4me3 at lowly expressed genes 92
Figure 4.5 Removal of both SET1A and SET1B does not affect bulk H3K4me3 93
Figure 4.6 Removal of both SET1A and SET1B has a larger effect on H3K4me3 than removal of SET1A alone 94
Figure 4.7 Removal of SET1A/B results in the largest reduction in promoter-associated H3K4me3 at highly expressed genes 96
Figure 4.8 Changes in H3K4me3 and gene expression were not strongly correlated at either time point 97
Figure 4.9 Identifying SET1A catalytic mutations 99
Figure 4.10 Schematic of a conditional SET1A catalytic point mutant 100
Figure 4.11 Characterising the SET1A conditional catalytic point mutant 101
Figure 4.12 SET1A functions as a transcriptional activator independently of its catalytic activity 103
Figure 5.1 The SET domain, RRM and linker are dispensable for transcription activation by SET1A 112
Figure 5.2 The interaction with WDR82 is essential for transcription activation by SET1A 114
Figure 5.3 Removal of WDR82 does not phenocopy SET1A/B depletion 115
Figure 5.4 Levels of RNAPII were not significantly affected following SET1A/B depletion 116

Figure 5.5 SET1A/B depletion results in a subtle but pervasive reduction in promoter-associated RNAPII 117
Figure 5.6 Overall levels of S5P RNAPII were not significantly affected by SET1A/B depletion 118
Figure 5.7 Removal of SET1A/B results in a widespread reduction in S5P RNAPII 120
Figure 5.8 SET1A/B depletion results in a larger reduction in RNAPII from gene bodies in comparison to promoters 122
Figure 5.9 Global levels of S2P RNAPII were not significantly affected by SET1A/B depletion 123
Figure 5.10 Removal of SET1A/B results in a pervasive reduction in S2P RNAPII 125
Figure 5.11 cTT-seq profiles nascent transcription 127
Figure 5.12 SET1A/B support RNAPII-mediated transcription 128
Figure 5.13 Gene body-associated transcription was reduced following SET1A/B depletion 130
Figure 5.14 Removal of SET1A/B results in pervasive reductions in anti-sense transcription from bidirectional promoters 132
Figure A1 SET1B localises to the nucleus in ESCs 146
Figure A2 SET1A/B regulate the expression of expressed, CGI-associated genes 146
Figure A3 SET1A/B targets are enriched for chromatin features associated with active transcription 147
Figure A4 SET1A occupancy does not explain gene expression outcomes 148
Figure A5 Levels of H3K4me3 scale with transcriptional activity in gene expression quartiles 148
Figure A6 Levels of H3K4me1 and H3K4me2 remain unchanged following depletion of SET1A/B 149
Figure A7 Genomic snapshots of H3K4me3 cChIP-seq and cRNA-seq following 24 hr SET1A/B depletion 149
Figure A8 Changes in H3K4me3 and gene expression were not strongly correlated at either time point in dTAG-SET1A cells 149
Figure A9 The conditional knockout of SET1A results in subtle changes in H3K4me3 150
Figure A10 The SET domain, RRM and linker region of SET1A are not sufficient to activate luciferase expression 151
Figure A11 SET1A DPR/AAA mutation inhibits its association with WDR82 (Courtesy of Jessica Kelley) 152
Figure A12 Changes in S5P RNAPII and promoter-associated S2P RNAPII are not correlated with gene expression 152
Figure A13 Removal of SET1A/B results in more pervasive effects on RNAPII than transcription 153
Figure A14 SET1A/B depletion does not result in transcription read-through at TESs 153
Figure A15 There is a moderate correlation between change in H3K4me3 and promoter-associated S5P RNAPII 154

List of Tables

Table 2.1 Name, genotype and origin of ESC lines used in this study 38
Table 2.2 Cas9 guide sequences used in this study 39
Table 2.3 Primers used to clone HDR targeting constructs 41
Table 2.4 Primers used to screen CRISPR targetings for homozygous insertion 42
Table 2.5 Antibodies used during this study 48
Table 2.6 Published datasets used in this study 59

List of common abbreviations

4sU	4-thiouridine
AID	Auxin-inducible degron
ATP	Adenosine triphosphate
BioCAP	Biotinylated CXXC affinity purification
bp	Base pair
Cas9	CRISPR-associated protein 9
CGI	CpG island
ChIP	Chromatin immunoprecipitation
ChIP-seq	ChIP followed by massively parallel sequencing
cChIP-seq	Spike-in calibrated ChIP-seq
CKO	Conditional knockout
CPA	Cleavage and polyadenylation
CpG	Cytosine-phosphate-guanine
CPM	Conditional point mutant
CRBN	Cerebelon
Cre-ERT2	Cre recombinase fused to the estrogen receptor
CRISPR	Clustered regularly interspaced short palindromic repeats
CTD	C-terminal domain
dTAG	$FKBP12^{F36V}$
DNMT	DNA methyltransferase
DSG	Disuccinimidyl glutarate
E	Embryonic day
ESC	mouse embryonic stem cell
FBS	Foetal bovine serum
FL	Full length
gDNA	Genomic DNA
GTF	General transcription factor
H2AK119	Histone H2A lysine 119
H2BK120	Histone H2B lysine 120
H3K4	Histone H3 lysine 4
H3K9	Histone H3 lysine 9
H3K27	Histone H3 lysine 27
H3K36	Histone H3 lysine 36
H3K79	Histone H3 lysine 79
HA	Homology arm
HDR	Homology directed repair
HRP	Horseradish peroxidase
HMT	Histone methyltransferase
IP	Immunoprecipitation
mAID	mini auxin-inducible degron (tag)
me1	Monomethylation
me2	Dimethylation
me3	Trimethylation
MNase	Micrococcal nuclease
NTD	N-terminal domain

OHT	4-hydroxytamoxifen
PIC	Pre-initiation complex
PBS	Phosphate buffered saline
PHD	Plant homeodomain
qPCR	Quantitative PCR
RNAPII	RNA Polymerase II
RNase	Ribonuclease
RNA-seq	RNA isolation followed by massively parallel sequencing
cRNA-seq	Spike-in calibrated RNA-seq
RPKM	Reads per kilobase per million
RRM	RNA recognition motif
RT	Room temperature
rTetR	Reverse tet repressor DNA binding domain
RT-qPCR	Reverse transcriptase-quantitative PCR
S2P	The RNAPII CTD phosphorylated on serine 2
S5P	The RNAPII CTD phosphorylated on serine 5
S7P	The RNAPII CTD phosphorylated on serine 7
T4P	The RNAPII CTD phosphorylated on threonine 4
Y1P	The RNAPII CTD phosphorylated on tyrosine 1
SAM	S-adenosyl methionine
SDS-PAGE	SDS-polyacrylamide gel electrophoresis
SET	Su(var)3-9, Enhancer-of-zeste and Trithorax
sgRNA	Single guide RNA
TAF	TBP-associated factor
TBP	TATA-binding protein
TetO	Tet repressor DNA binding sites
TF	Transcription factor
TES	Transcription end site
TSS	Transcription start site
TT-seq	Isolation of the transient transcriptome followed by massively parallel sequencing
TT(chem)-seq	Isolation of the transient transcriptome (using chemical fragmentation) followed by massively parallel sequencing
cTT-seq	Spike-in calibrated TT-seq
Ub	Ubiquitin
WRAD	WDR5, RBBP5, ASH2L, DPY30
WT	Wild type
ZF-CXXC	Zinc finger-CXXC

1. Introduction

1.1 The regulation of eukaryotic gene expression

Complex multicellular organisms are made up of a huge number of specialised cell types that, despite being phenotypically different, all possess the same DNA. During development, genes are differentially expressed in individual cell types in order to achieve cellular specialisation during differentiation and to enable cellular responses to environmental stimuli. The precise control of gene expression is therefore essential for the development of complex multicellular organisms. As the first stage in this process, transcription represents an important step at which gene expression can be regulated.

1.1.1 Transcription initiation

The transcription of protein-coding genes by RNA Polymerase II (RNAPII) initiates at transcription start sites (TSSs) embedded within core promoters (Haberle and Stark, 2018). On its own, RNAPII has a low affinity for core promoters and lacks the ability to efficiently initiate transcription (Schier and Taatjes, 2020). Eukaryotic promoters contain DNA motifs that are recognised by the general transcription factors (GTFs) (Juven-Gershon *et al.*, 2008). Binding of the GTFs to the promoter enables RNAPII to associate with DNA and initiate transcription (reviewed in Schier and Taatjes, 2020).

In the classical model of transcription initiation, proposed on the basis of biochemical and structural studies, a stepwise recruitment of the six main GTFs results in the formation of a transcriptionally-competent pre-initiation complex (PIC) (Van Dyke, Roeder and Sawadogo, 1988; Buratowski *et al.*, 1989; He *et al.*, 2013). The first GTF to associate with the core promoter, and therefore often described as the GTF responsible for 'nucleating' PIC formation, is TFIID (reviewed in Patel, Greber and Nogales, 2020). TFIID is a 1.3 mDa complex consisting of the DNA-binding protein TBP (TATA-binding protein) and 13 TBP-associated factors (TAFs). The association of TBP

and several of the TAF proteins with their corresponding DNA motifs in promoter elements is central to PIC formation and likely functions to position RNAPII with respect to the TSS (Louder *et al.*, 2016; Patel *et al.*, 2018). Binding of TBP 25-30 bp upstream of the TSS induces a 90° kink in the DNA, releasing repressive interactions between TBP and TFIID and allowing the formation of a network of interactions that would not be possible on linear DNA (Horikoshi *et al.*, 1992; Patel *et al.*, 2018; Schier and Taatjes, 2020). This supports further assembly of the PIC. TFIIA and TFIIB stabilise this bent DNA conformation and allow the association of TFIIF and RNAPII (Imbalzano, Zaret and Kingston, 1994; Hieb *et al.*, 2007). Assembly of the GTFs is complete after the association of TFIIE and TFIIH. The DNA translocase, XPB, a subunit of TFIIH, uses the energy derived from ATP hydrolysis to unwind promoter DNA and allow its passage into the polymerase active site (Tirode *et al.*, 1999; Kim, Ebright and Reinberg, 2000). Once the nascent RNA reaches a length of 12-15 nucleotides, structural changes in the PIC enable RNAPII to dissociate from the initiation factors and transition into elongation (Kostrewa *et al.*, 2009; X. Liu *et al.*, 2010; Bernecky *et al.*, 2016; He *et al.*, 2016).

Despite being sufficient to stimulate PIC assembly, core promoters support only a low basal level of transcription (Kadonaga, 2012; Haberle and Stark, 2018). Binding of sequence-specific transcription factors (TFs) to cis-regulatory elements in proximity to gene promoters can stimulate the activity and/or stability of RNAPII (Lambert *et al.*, 2018). Some TFs achieve this via low-complexity transactivation domains that mediate protein-protein interactions that potentiate transcription (reviewed in Frietze and Farnham, 2011). For example, TFs can stabilise binding of the GTFs to core promoters to directly stimulate PIC assembly (Horikoshi *et al.*, 1988). Alternatively, TFs can recruit transcription coactivator complexes that remodel the environment around gene promoters to make them more accessible to the transcription machinery (Trouche, Cook and Kouzarides, 1996).

In addition to promoter-proximal cis-regulatory elements, distal regulatory elements, often termed enhancers, can regulate promoters that may be tens, or even hundreds, of kilobases away (Spitz and Furlong, 2012; Shlyueva, Stampfel and Stark, 2014). For example, TFs have been proposed to recruit the mediator, a transcriptional coactivator, to enhancers (Fournier *et al.*, 2016; Quevedo *et al.*, 2019). The enhancers then communicate with gene promoters to support transcription initiation by stabilising the PIC and enabling the transcription machinery to escape from the promoter and enter into productive elongation (Allen and Taatjes, 2015; Soutourina, 2018). The coordinated activity of the transcription machinery at gene promoters and coactivator complexes at cis-regulatory elements is central to normal gene expression.

1.1.2 Dynamic post-translational modification of RNAPII

RNAPII is a multi-subunit complex, the largest protein of which, RPB1, contains a low-complexity, unstructured C-terminal domain (CTD). The mammalian RNAPII CTD is comprised of 52 repeats of the heptapeptide consensus sequence YSPTSPS (Corden *et al.*, 1985; P. Liu *et al.*, 2010). Although not required for RNAPII catalysis, the CTD is of central importance for the regulation of multiple steps of transcription and the coupling of transcription to various co-transcriptional processes, such as mRNA capping and splicing (reviewed in Hsin and Manley, 2012; Harlen and Churchman, 2017).

A key feature of the RNAPII CTD is its propensity to be dynamically post-translationally modified, most notably by phosphorylation of the non-proline residues within each heptapeptide repeat (reviewed in Harlen and Churchman, 2017). The phosphorylation states of the RNAPII CTD form a distinctive pattern across genes, with each of these modification states implicated in regulating various aspects of transcription and transcript processing (Figure 1.1). When RNAPII initially associates with the PIC, the CTD is in an unmodified state (Lu *et al.*, 1991). Upon the recruitment of TFIIH to the PIC, its CDK7 kinase subunit phosphorylates the RNAPII CTD on serine residues 5 and 7 of the heptapeptide repeats (S5P and S7P) (Glover-Cutter *et al.*, 2009; Boeing *et al.*, 2010).

These modifications contribute to the release of RNAPII from the promoter and the progression from transcription initiation into transcription elongation (Payne, Laybourn and Dahmus, 1989; Søgaard and Svejstrup, 2007; Abdella *et al.*, 2021). Additionally, S5P functions to recruit the capping enzyme to the nascent transcript, to begin the process of transcript maturation (McCracken *et al.*, 1997; Ho and Shuman, 1999).

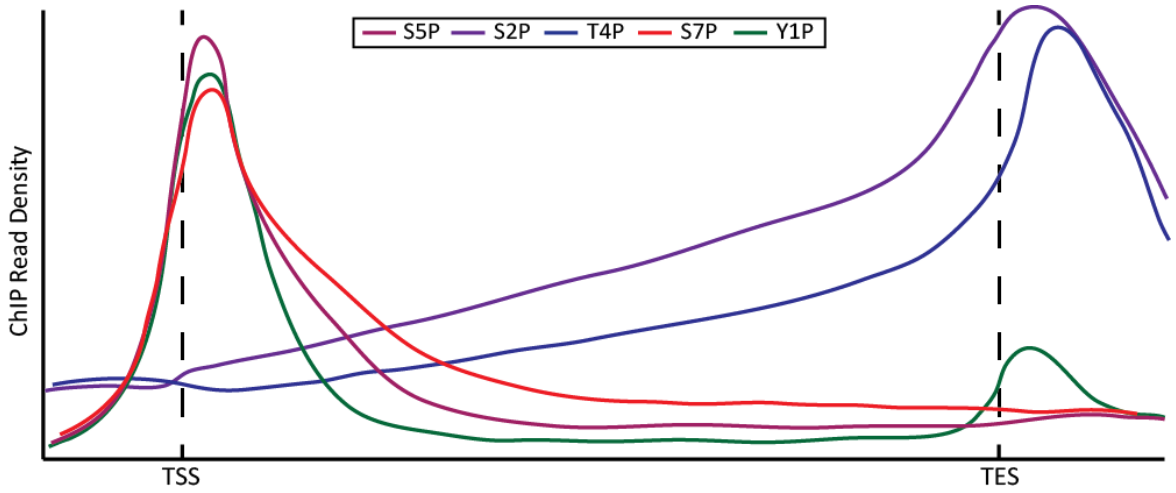


Figure 1.1 Schematic of average ChIP profiles of phosphorylated RNAPII
Schematic metaplots showing typical chromatin immunoprecipitation (ChIP) densities for the mammalian RNAPII CTD phosphorylated on tyrosine 1 (Y1P), serine 2 (S2P), threonine 4 (T4P), serine 5 (S5P) and serine 7 (S7P). Figure adapted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

After the release of RNAPII from the promoter and the formation of a short nascent RNA, the negative elongation factors, DSIF and NELF, associate with RNAPII (Wada *et al.*, 1998; Yamaguchi *et al.*, 1999; Vos *et al.*, 2018). The outcome of DSIF and NELF association is the pausing of RNAPII 20-60 bp downstream of the TSS, before it transitions into fully productive elongation (Fraser, Sehgal and Darnell, 1978; Gariglio, Bellard and Chambon, 1981; Gilmour and Lis, 1986; Rougvie and Lis, 1988; Giardina, Perez-Riba and Lis, 1992; Rasmussen and Lis, 1993). Such 'paused' polymerase was initially detected at the promoters of *Drosophila* heat shock genes in the inactive state and was proposed to enable the rapid release of RNAPII into productive elongation upon heat shock. Pausing is now considered to be a pervasive feature of eukaryotic protein-coding genes (Chao and Price, 2001; Muse *et al.*, 2007; Zeitlinger *et al.*, 2007; Rahl *et al.*, 2010; Jonkers, Kwak and Lis, 2014).

While initially suggested to allow rapid gene induction, alternative roles for RNAPII pausing, such as acting as a checkpoint to ensure proper coupling of transcription elongation and RNA processing, controlling transcription levels by preventing reinitiation by another RNAPII, and maintaining the promoter in a permissive state have also been proposed (Adelman and Lis, 2012; Shao and Zeitlinger, 2017).

The release of RNAPII from pausing into productive elongation depends on the recruitment of the elongation factor P-TEFb, and its associated kinase, CDK9 (Adelman and Lis, 2012). CDK9 phosphorylates the pausing factors NELF and DSIF, as well as serine 2 of the RNAPII CTD (Marshall *et al.*, 1996; Fujinaga *et al.*, 2004; Yamada *et al.*, 2006). This activity results in the dissociation of NELF, the conversion of DSIF into a positive elongation factor, and the release of RNAPII from pausing into an active elongation complex. During the process of productive elongation, the RNAPII CTD becomes more heavily phosphorylated on S2 (S2P) and threonine 4 (T4P), while S5P and S7P are removed (Figure 1.1). Changes in the phosphorylation state of RNAPII allow the association of a host of elongation factors that facilitate its passage through obstacles and ensure the proper maturation and splicing of the nascent transcript (reviewed in Hsin and Manley, 2012; Harlen and Churchman, 2017).

Transcription is terminated once RNAPII transcribes the polyA signal, which leads to the recruitment of the cleavage and polyadenylation (CPA) complex and termination factors to RNAPII. The CPA complex then cleaves the nascent transcript to release the mRNA (Colgan and Manley, 1997). Two main models have been proposed for the mechanism by which the polyA signal triggers transcription termination (reviewed in Proudfoot, 2016). In the allosteric model, transcription of the polyA signal induces a conformational change in RNAPII, which causes pausing and then dissociation (Logan *et al.*, 1987; Zhang, Rigo and Martinson, 2015). In the torpedo model of termination, RNAPII continues transcribing after the polyA signal and the 5'-3' exonuclease, XRN2, degrades the downstream transcript in kinetic competition with RNAPII transcription (West,

Gromak and Proudfoot, 2004; Fong *et al.*, 2015; Cortazar *et al.*, 2019). Once XRN2 'catches up' with RNAPII, this triggers its release from DNA. More recently, models have been proposed that incorporate aspects of both the allosteric and torpedo models of termination (Eaton *et al.*, 2020). Nonetheless, it is clear that transcription termination involves the concerted action of a host of termination factors that together contribute to the dissociation of RNAPII from DNA (Shi *et al.*, 2009).

1.1.3 DNA is packaged into chromatin

Although DNA sequence is important for the regulation of transcription, eukaryotic DNA is not present within nuclei as a naked template, but is rather found in a nucleic acid-protein complex known as chromatin (Kornberg and Lorch, 1999). The nucleosome, the basic repeating unit of chromatin, is composed of approximately 147 bp of DNA wound in 1.5 turns around a histone octamer comprising two copies each of histones H2A, H2B, H3 and H4 (Luger *et al.*, 1997). Histone H1 associates with the linker DNA between each nucleosome and likely functions to promote inter-nucleosomal interactions and chromatin compaction (Sato *et al.*, 1999).

It has been known for some time that chromatin functions to organise the genome into the comparatively small volume of the nucleus (Woodcock and Ghosh, 2010). However, it has become clear in recent years that chromatin can also profoundly affect DNA-templated processes, including replication and transcription (Groth *et al.*, 2007; Li, Carey and Workman, 2007). At the most basic level, the packaging of genomic DNA into nucleosomes can occlude TF binding motifs and inhibit the assembly of the basal transcription machinery through the compaction of nucleosome arrays (Knezetic and Luse, 1986; Lorch, LaPointe and Kornberg, 1987; Workman and Kingston, 1998; Cockerill, 2011). Therefore, a key role of transcription coactivator complexes is to 'remodel' the nucleosome landscape around gene promoters (Clapier and Cairns, 2009; Struhl and Segal, 2013). Chromatin remodellers use the energy derived from ATP hydrolysis to alter nucleosome positioning via the sliding or eviction of histone octamers (Hota and Bruneau, 2016). Therefore, nucleosome

remodelling complexes can increase the accessibility of promoters and cis-regulatory elements to the transcription machinery (Morris *et al.*, 2014; De Dieuleveult *et al.*, 2016). However, it is important to note that while many chromatin remodellers facilitate transcription, some can also contribute to gene repression through the formation of stable nucleosome arrays at gene promoters, restricting DNA accessibility (Bornelö *et al.*, 2018).

1.1.4 DNA methylation

The mammalian genome is pervasively methylated at the C5 position of the cytosine bases in the context of CG dinucleotides (CpG) (Lister *et al.*, 2009). A family of DNA methyltransferases (DNMTs) catalyses the methylation of CG dinucleotides throughout the mammalian genome. DNMT3A/B, aided during early development by their non-catalytic cofactor DNMT3L, primarily catalyse *de novo* DNA methylation (Okano *et al.*, 1999; Chédin, Lieber and Hsieh, 2002; Suetake *et al.*, 2004). DNMT1 is a maintenance methyltransferase that ensures that DNA methylation patterns are re-established after cell division (Vertino *et al.*, 2002). DNA methylation is therefore a heritable modification that can be stably maintained (Holliday and Pugh, 1975; Riggs, 1975). Pervasive CpG methylation throughout the genome has a widespread role in stable transcription repression and is important for silencing parasitic DNA elements, regulating imprinted gene expression and supporting X chromosome-inactivation (Mohandas, Sparkes and Shapiro, 1981; Venolia *et al.*, 1982; Frantz *et al.*, 2001).

Cytosine methylation is mutagenic; spontaneous deamination of methylated cytosine results in its conversion to thymine (Coulondre *et al.*, 1978; Duncan and Miller, 1980). Such a mutation is inherited when it occurs in the germline, resulting in the gradual depletion of CpG dinucleotides throughout evolution (Sved and Bird, 1990; Antequera, 2003). The CpG dinucleotide is therefore under-represented in mammalian genomes and is found at only ~20% of its expected frequency (Josse, Kaiser and Kornberg, 1961; Swartz, Trautner and Kornberg, 1962). However, short (~1-2 kb) regions of the genome, termed CpG islands (CGIs), remain non-methylated and have a CpG density

approaching its expected frequency (Bird *et al.*, 1985; Gardiner-Garden and Frommer, 1987). CGIs are associated with ~50-70% of vertebrate gene promoters, making them the most common class of promoter in mammalian cells (Tykocinski and Max, 1984; Bird *et al.*, 1985; Saxonov, Berg and Brutlag, 2006). In line with this, CGIs typically have chromatin features associated with transcriptional permissiveness; CGIs contain hyper-acetylated histones, are depleted of linker histone H1 and are more 'accessible' in nuclease digestion assays in comparison to CpG-poor gene promoters (Stalder *et al.*, 1980; Bird *et al.*, 1985; Antequera, Macleod and Bird, 1989; Tazi and Bird, 1990; Sabo *et al.*, 2004; Choi, 2010; Z. Li *et al.*, 2011; Fenouil *et al.*, 2012; Teif *et al.*, 2012). Furthermore, many CGIs that are not associated with annotated TSSs also exhibit promoter-like features, such as RNAPII occupancy (Illingworth *et al.*, 2010; Maunakea *et al.*, 2010). Importantly, methylation of CGIs is associated with transcriptional repression, suggesting that maintaining CGIs in a non-methylated state is essential for gene expression (Bird and Wolffe, 1999; Bird, 2002; Jones, 2012).

CpG islands typically remain free of DNA methylation regardless of the cell type or transcriptional state of the underlying gene, suggesting that the non-methylated state itself does not instruct gene expression (Weber *et al.*, 2007; Meissner *et al.*, 2008). How then do CGIs influence transcription? A breakthrough in our understanding about the mechanisms by which CGIs influence gene expression came with the identification of the zinc finger-CXXC (ZF-CXXC) protein domain, which binds to CpG dinucleotides only when they are non-methylated (Voo *et al.*, 2000; Lee, Voo and Skalnik, 2001). Bioinformatic analysis has now led to the identification of twelve ZF-CXXC domain-containing proteins that are typically found in complexes associated with chromatin modifying activities, implicated in either gene activation or repression (reviewed in Long, Blackledge and Klose, 2013). CGIs are therefore considered to be a regulatory platform that integrate the activities of chromatin-associated complexes to influence the transcriptional output at associated gene promoters (Blackledge and Klose, 2011; Deaton and Bird, 2011).

1.1.5 The post-translational modification of histones

The histone proteins around which DNA is wrapped to form nucleosomes can also be extensively post-translationally modified. Known histone modifications include, but are not limited to, methylation, acetylation, phosphorylation, and ubiquitination (Kouzarides, 2007; Rothbart and Strahl, 2014). While these modifications occur on residues throughout histone proteins, the N- or C-terminal tails that extend from the nucleosome core tend to be more heavily modified and are thought to be particularly important in gene regulation (Tropberger and Schneider, 2010; Rothbart and Strahl, 2014). Genomic approaches used to map the distribution of different histone modifications have revealed that they tend to occupy certain genomic regions (Dunham *et al.*, 2012). Some histone modifications, including methylation on H3K4, H3K79 and H3K36, as well as H2BK120 monoubiquitination and H3 and H4 lysine acetylation are associated with active transcription (Hebbes, Thorne and Crane-Robinson, 1988; Krogan *et al.*, 2003; Bannister *et al.*, 2005; Pokholok *et al.*, 2005; Zhu *et al.*, 2005; Pavri *et al.*, 2006; Barski *et al.*, 2007; Guenther *et al.*, 2007; Steger *et al.*, 2008; Karmodiya *et al.*, 2012). Conversely, methylation of H3K9 and H3K27 and monoubiquitination of H2AK119 are associated with transcription repression (Nakayama *et al.*, 2001; Cao *et al.*, 2002; Wang *et al.*, 2004).

A subset of histone modifications appears to directly regulate chromatin structure by influencing histone-DNA or nucleosome-nucleosome interactions (Wade, Pruss and Wolffe, 1997; Anderson, Lowary and Widom, 2001; Görisch *et al.*, 2005; Shogren-Knaak *et al.*, 2006). One such histone modification is acetylation, which can be found throughout the core histone proteins. Histone acetylation within the globular fold of histone proteins weakens histone-DNA interactions, while acetylation on the histone N-terminal tails weakens interactions between neighbouring nucleosomes (Anderson, Lowary and Widom, 2001; Shogren-Knaak *et al.*, 2006; Tropberger and Schneider, 2010). Therefore, histone acetylation has been implicated in reducing nucleosome stability and promoting a more open chromatin configuration. Indeed, the acetylation of histone

proteins has been extensively linked to transcriptional activity and chromatin accessibility of actively transcribed genomic regions (Hebbes, Thorne and Crane-Robinson, 1988; Wade, Pruss and Wolffe, 1997; Turner, 2000; Barnes, English and Cowley, 2019). Conversely, the removal of acetylation by histone deacetylases is often coincident with transcription repression.

Many histone modifications do not directly regulate chromatin architecture and instead function through the activity of so-called 'reader' proteins. Many chromatin-regulating complexes have been shown to either bind to, or be specifically excluded by, certain histone modifications. One such example of a histone modification with an extensive repertoire of 'reader' proteins is histone lysine methylation. Histone methylation is placed by SET (for Su(var)3-9, Enhancer-of-zeste and Trithorax) domain-containing histone methyltransferases (HMTs) that transfer a methyl group from the S-adenosylmethionine (SAM) donor onto several lysine residues throughout histone tails (Tschiersch *et al.*, 1994; Rea *et al.*, 2000; Zhang and Reinberg, 2001; Dillon *et al.*, 2005). Unlike acetylation, methylation preserves the histone charge, and therefore does not directly regulate nucleosome structure. Instead, proteins possessing domains such as the tudor domain or plant homeodomain (PHD), can bind to histones in a methylation-sensitive manner (Kim *et al.*, 2006; Greer and Shi, 2012). Interestingly, depending on the modified lysine residue, histone methylation has been implicated in both transcription activation and repression (Hyun *et al.*, 2017).

Frequently, the complexes that are proposed to 'read' a certain histone modification are also associated with chromatin-modifying activities themselves (Zhang, Cooper and Brockdorff, 2015). For example, histone-modifying complexes often contain protein domains capable of binding to specific histone modifications. The propensity of such complexes to 'read' a particular histone modification, and then 'write' their own histone modification, leads to the formation of chromatin domains containing histones that are combinatorially modified (Zhang, Cooper and Brockdorff, 2015). Certain proteins associated with reading the histone modification landscape are also capable of chromatin remodelling. For example, the SAGA chromatin remodelling complex

recognises histone modifications typically associated with active transcription, such as H3K4 methylation (Bian *et al.*, 2011). Furthermore, many chromatin regulators comprise multifunctional complexes that display several chromatin-modifying activities. For example, SAGA is associated with histone acetylation activity, in addition to chromatin remodelling (Ben-Shem, Papai and Schultz, 2020). This has led to a model in which specific histone modifications associated with either transcriptional repression or activation allow the recruitment of additional complexes which can further modify the chromatin landscape to support a specific transcriptional outcome (reviewed in Zhang, Cooper and Brockdorff, 2015; Allis and Jenuwein, 2016; Hyun *et al.*, 2017).

The misregulation of histone modifications is linked to developmental defects and diseases such as cancer, highlighting their importance in the regulation of transcription (Chi, Allis and Wang, 2010; Vallianatos and Iwase, 2015; Audia and Campbell, 2016; Jambhekar, Dhall and Shi, 2019). Therefore, a considerable research effort has been placed on understanding the mechanisms that govern histone modification patterns, and the role of these histone modifications in regulating gene expression.

1.2 H3K4 methylation in mammalian genomes

1.2.1 Distribution of H3K4 methylation

One of the most extensively studied histone modifications is the methylation of histone H3 on lysine 4. H3K4 can be mono-, di- or tri-methylated (H3K4me1/2/3). In simple eukaryotes, there is a gradient of H3K4 methylation states across genes: H3K4me3 is found at gene promoters, H3K4me2 within gene bodies, and H3K4me1 predominates towards the 3' end of genes (Santos-Rosa *et al.*, 2002; Chih *et al.*, 2005; Pokholok *et al.*, 2005; Soares *et al.*, 2017; Figure 1.2 A). Genome-wide analysis of H3K4 methylation in multicellular eukaryotes, which have larger and more complex genomes, has identified a more distinct pattern of H3K4 methylation. While the enrichment of H3K4me3 at gene promoters is conserved across all eukaryotes, in vertebrates

H3K4me1 is more broadly distributed and enriched at enhancers, while H3K4me2 tends to be elevated in regions flanking peaks of H3K4me3 (Liang *et al.*, 2004; Schneider *et al.*, 2004; Bernstein *et al.*, 2005; Barski *et al.*, 2007; Heintzman *et al.*, 2007, 2009; Figure 1.2 B).

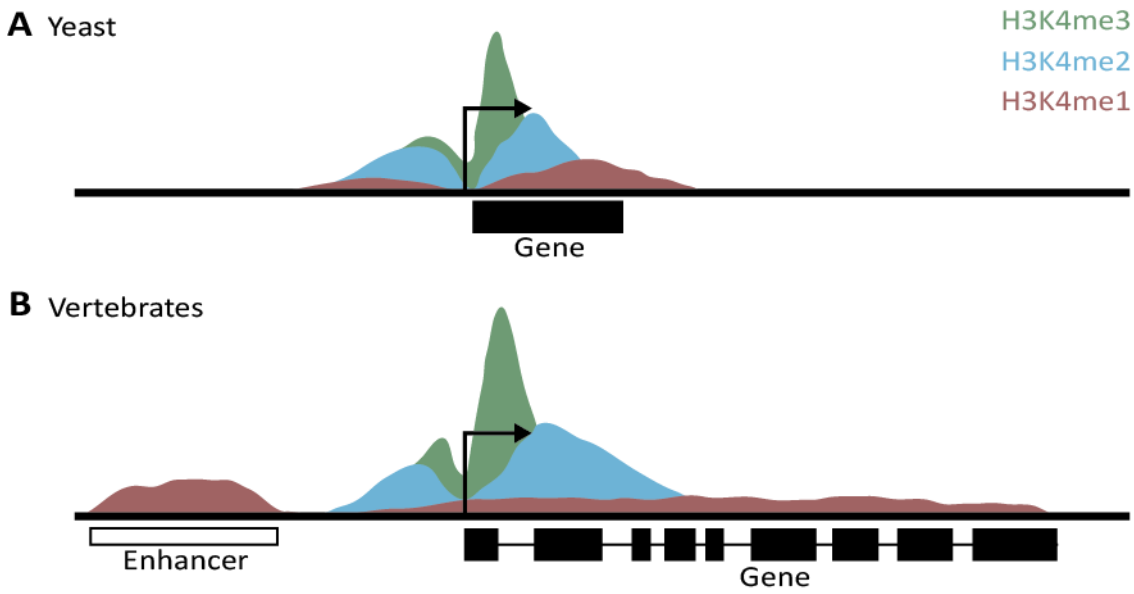


Figure 1.2 Genomic distribution of histone H3 lysine 4 methylation
Schematic depictions of the distribution of H3K4 methylation across **(A)** yeast and **(B)** vertebrate genes. Arrows indicate transcription start sites. Gene bodies are shown with black boxes representing exons. Figure adapted from Hughes, Kelley and Klose, 2020.

1.2.2 H3K4me3 and CpG islands

In vertebrate genomes the presence of H3K4me3 correlates closely with the underlying boundaries of promoter-associated CGIs and H3K4me3 is present at the majority of CGIs, irrespective of whether the associated gene is actively transcribed (Figure 1.3; Guenther *et al.*, 2007; Mikkelsen *et al.*, 2007; Illingworth *et al.*, 2008). Additionally, H3K4me3 is present at CGIs that are not associated with annotated TSSs and when exogenous CG-rich non-methylated DNA lacking a gene promoter is inserted into the mouse genome, it acquires H3K4me3 (Illingworth *et al.*, 2008, 2010; Maunakea *et al.*, 2010; Mendenhall *et al.*, 2010; Thomson *et al.*, 2010; Wachter *et al.*, 2014). This suggests that H3K4me3 deposition is intrinsically linked to the presence of a CGI.

1.2.3 H3K4me3 correlates with transcriptional activity

Although the presence of non-methylated, CG-rich DNA appears to be sufficient to create H3K4me3-modified chromatin, it is clear that it alone cannot fully explain the observed distribution of H3K4me3. Rather than being evenly distributed across all CGIs, levels of promoter-associated H3K4me3 typically correlate with the expression level of the associated gene (Barski *et al.*, 2007; Mikkelsen *et al.*, 2007; Guenther *et al.*, 2007; Figure 1.3). The relationship between H3K4me3 and transcriptional activity has been observed from yeast through to mammals and has led to the designation of H3K4me3 as an 'active' histone modification (Zhao and Garcia, 2015). Indeed, H3K4me3 has been proposed to regulate numerous aspects of chromatin architecture, which could implicate it in the positive regulation of gene expression (discussed in Chapter 1.3).

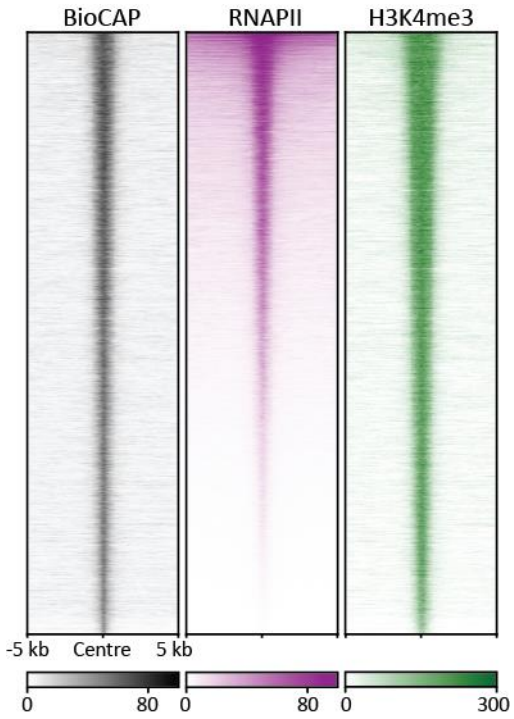


Figure 1.3 CpG islands and H3K4me3
Heatmaps illustrating non-methylated CpG-rich DNA (BioCAP-seq; Illingworth *et al.*, 2008) and RNAPII (transcription) and H3K4me3 ChIP-seq signal across non-methylated islands in mouse embryonic stem cells. The heatmaps are sorted by the RNAPII ChIP-seq signal (Datasets: Long *et al.*, 2013; Brown *et al.*, 2017; Turberfield *et al.*, 2019).

1.3 The influence of H3K4me3 on chromatin architecture

1.3.1 Does H3K4me3 prevent methylation of CpG islands?

The methylation of CGIs is associated with stable transcriptional repression (Bird and Wolffe, 1999; Bird, 2002; Jones, 2012). However, CGI-associated gene promoters tend to remain free of DNA methylation regardless of the transcriptional state of the associated gene (Weber *et al.*, 2007; Meissner *et al.*, 2008). It has been proposed that CGI-associated gene promoters must therefore remain non-methylated to render them sensitive to gene regulatory cues (Blackledge and Klose, 2011; Deaton and Bird, 2011). How is this made possible given that the rest of the genome is pervasively methylated? One explanation could be that certain features of CGIs are refractory to DNA methylation. Given that DNA methylation and H3K4me3 are almost entirely mutually exclusive, this has led to the suggestion that H3K4me3 may counteract DNA methylation (Okitsu and Hsieh, 2007; Weber *et al.*, 2007; Meissner *et al.*, 2008; Zhang *et al.*, 2009; Smallwood *et al.*, 2011; Singh *et al.*, 2013; Gahurova *et al.*, 2017; Greenfield *et al.*, 2018). Indeed, alterations in H3K4 methylation have been observed to influence DNA methylation in a number of studies (Milne *et al.*, 2002; Terranova *et al.*, 2006; Ciccone *et al.*, 2009; Stewart *et al.*, 2015). Yeast offer an interesting model system to study this relationship, since yeast genomes possess promoter-associated H3K4me3 but are devoid of DNA methylation. Expressing the mammalian DNMTs in yeast results in aberrant DNA methylation, but only at loci where H3K4me3 is absent (Hu *et al.*, 2009; Morselli *et al.*, 2015). These observations suggest that the presence of H3K4me3 could block DNMT activity. Indeed, biochemical and structural studies have revealed that H3K4me3 could inhibit the chromatin-binding of DNMT3A/B and DNMT3L, while also regulating the enzymatic activity of DNMT3A (Ooi *et al.*, 2007; Otani *et al.*, 2009; Zhang *et al.*, 2010; B. Z. Li *et al.*, 2011; Guo *et al.*, 2015; Noh *et al.*, 2015). The relationship between H3K4me3 and the inhibition of DNA methylation genome-wide remains unclear, however the presence of H3K4me3 clearly provides a potential mechanism by which the deposition of DNA methylation could be regulated.

1.3.2 Does H3K4me3 inhibit transcriptional repression?

CGI-associated gene promoters can be bound by numerous different chromatin-modifying complexes which possess transcription-activating or repressive activities (Blackledge and Klose, 2011; Deaton and Bird, 2011). It is therefore essential that the promoters of actively transcribed genes are resistant to the activities of the repressive complexes that bind them. Since there is extensive interplay between histone modifications, one could postulate that the histone modification status at actively transcribed genes may function to inhibit repressive chromatin states.

The polycomb group proteins, which form two catalytic complexes known as PRC1 and PRC2, have histone modifying activity; PRC1 and PRC2 place H2AK119ub and H3K27me3, respectively (Blackledge, Rose and Klose, 2015). Transcriptionally repressed CGI-associated gene promoters are typically found modified with these repressive histone modifications, creating a polycomb chromatin domain. The trithorax complexes, some of which deposit H3K4me3, were originally identified in *Drosophila* as inhibitors of polycomb-mediated silencing of the *Hox* genes during development (Ingham and Whittle, 1980; Kassis, Kennison and Tamkun, 2017). In mammals, there is evidence that the polycomb and H3K4 methylation systems function to oppose each other (Nakagawa *et al.*, 2008; D. H. Kim *et al.*, 2013; Mas *et al.*, 2018; Douillet *et al.*, 2020). From *in vitro* experiments, it is clear that H3K4me3 inhibits deposition of H3K27me3 by PRC2 on the same histone H3 tail (Schmitges *et al.*, 2011; Voigt *et al.*, 2012; Lechner, Agashe and Fierz, 2016; Chen *et al.*, 2018). However, the relationship between H3K4me3 and H3K27me3 may be more complicated than this would suggest. Somewhat counterintuitively, H3K4me3 and H3K27me3 can co-exist within a single nucleosome (Voigt *et al.*, 2012; Lechner, Agashe and Fierz, 2016). Moreover, histone modification profiling by Chromatin Immunoprecipitation (ChIP)-sequencing in mammalian cells has shown that certain CGI-associated promoters can be modified by both H3K27me3 and H3K4me3 in the same cell type (Azuara *et al.*, 2006; Bernstein *et al.*, 2006; Mikkelsen *et al.*, 2007).

Such a situation is referred to as a 'bivalent' state and has been proposed to maintain chromatin in a 'poised' configuration, ready for rapid gene induction (Voigt, Tee and Reinberg, 2013). However, whether such a chromatin configuration can be stably maintained remains unclear (Sneppen and Ringrose, 2019).

H3K9 methylation is typically associated with stably-repressed regions of the genome, such as telomeric or centromeric repeat regions (Becker, Nicetto and Zaret, 2016). However, H3K9me3 can also be found away from such regions and have a role in gene regulation (Bilodeau *et al.*, 2009; Karimi *et al.*, 2011; Ninova, Tóth and Aravin, 2019). H3K4me3 has been proposed to regulate H3K9 methylation by inhibiting binding of the H3K9 HMTs, preventing its deposition, and also by stabilising binding of H3K9 demethylases, thereby promoting its removal from gene promoters (Nishioka *et al.*, 2002; Huang *et al.*, 2006; Binda *et al.*, 2010; Feng *et al.*, 2010; Wu *et al.*, 2010; Fortschegger *et al.*, 2010; Horton *et al.*, 2010; Nikolov *et al.*, 2011; Wan *et al.*, 2013; Lohse *et al.*, 2013; Pedersen *et al.*, 2014; Pack, Yamamoto and Fujimori, 2016). Therefore, through regulating the deposition of two repressive histone modifications, H3K9me3 and H3K27me3, H3K4me3 may function to indirectly promote transcription by inhibiting chromatin-based transcriptional repression.

1.3.3 H3K4me3 supports the formation of a permissive chromatin state

Since H3K4me3 is associated with actively transcribed gene promoters, it has been proposed that its presence could directly contribute to a chromatin landscape that is permissive to gene transcription (Vermeulen and Timmers, 2010; Hughes, Kelley and Klose, 2020; Park, Kim and Kim, 2020). Through a number of biochemical screens, H3K4me3 has been linked to factors that mediate histone acetylation and chromatin remodelling, and has therefore been implicated in regulating promoter accessibility (Bartke *et al.*, 2010; Vermeulen *et al.*, 2010; Nikolov *et al.*, 2011; Eberl *et al.*, 2013).

The perturbation of H3K4me3 has been reported to result in reductions in histone acetylation *in vivo* (Z. Wang *et al.*, 2009; Clouaire, Webb and Bird, 2014). Furthermore, H3K4me3 has been implicated in supporting the chromatin association of several histone acetyltransferase complexes. For example, the SGF29 subunit of the SAGA complex is important for acetylation of its target loci and associates with H3K4me3 via its tandem tudor domains (Vermeulen *et al.*, 2010; Bian *et al.*, 2011). Furthermore, the mammalian ING proteins contain a PHD domain that preferentially binds to H3K4me3 (Peña *et al.*, 2006, 2008; Shi *et al.*, 2006; Champagne *et al.*, 2008; Palacios *et al.*, 2008; Hung *et al.*, 2009; Saksouk *et al.*, 2009; Kim *et al.*, 2016; Ormaza *et al.*, 2019). Interestingly, while ING3-5 associate with histone acetyltransferase complexes, ING1 and ING2 associate with the SIN3 histone deacetylase complex, which removes histone acetylation and can therefore counteract gene expression (Doyon *et al.*, 2006; Champagne and Kutateladze, 2009). It is possible that, by recruiting both histone acetyltransferases and deacetylases to gene promoters, H3K4me3 contributes to the dynamics and/or distribution of histone acetylation (Hazzalin and Mahadevan, 2005; Crump *et al.*, 2011). Further work is required to better define how histone acetylation is regulated in mammals, although H3K4me3 appears to contribute to this process.

Alterations in H3K4me3 have been associated with changes in nucleosome occupancy and one explanation for such an observation could be the association of chromatin remodelling complexes with H3K4me3 (Kurup, Campeanu and Kidder, 2019). The PHD domain of BPTF, a subunit of the NURF chromatin-remodelling complex, associates with H3K4me3, an interaction that appears to be required for appropriate NURF complex binding *in vivo* (Li *et al.*, 2006; Wysocka *et al.*, 2006; Ying Li *et al.*, 2016). BPTF also has a bromodomain that binds acetylated histone H4 and enables it to engage in multivalent interactions with chromatin carrying both H3K4me3 and H4 acetylation (Ruthenburg *et al.*, 2011). These modifications at actively transcribed genes may therefore reinforce a dynamic and accessible chromatin state to support ongoing gene transcription. In addition to BPTF/NURF, the ATP-dependent chromatin remodelling enzyme CHD1, which is thought to contribute to the maintenance of a transcriptionally-permissive chromatin state at gene

promoters, also binds to H3K4me3 via its double chromodomains (Flanagan *et al.*, 2005; Sims *et al.*, 2005; Gaspar-Maia *et al.*, 2009). This interaction has been proposed to enhance the nucleosome remodelling activity of CHD1 (Lin *et al.*, 2011). Together these observations suggest that H3K4me3 and histone acetylation may promote gene expression, at least in part, by recruiting chromatin remodelling enzymes to actively transcribed gene promoters.

1.3.4 H3K4me3 and gene transcription

H3K4me3 correlates with gene transcription and appears to contribute to a chromatin environment that is permissive to transcriptional activity. There is also evidence that H3K4me3 promotes gene transcription *in vitro* (Lin *et al.*, 2011; Jiang *et al.*, 2013; Lauberth *et al.*, 2013; Tang *et al.*, 2013). The TAF3 subunit of the TFIID complex contains a PHD domain that associates with H3K4me3 and stabilises PIC formation on *in vitro* transcription templates (Vermeulen *et al.*, 2007; van Ingen *et al.*, 2008; Nikolov *et al.*, 2011; Eberl *et al.*, 2013; Lauberth *et al.*, 2013). *In vivo,* TAF3 binding to H3K4me3 is essential for the appropriate induction of many TAF3 target genes (Lauberth *et al.*, 2013; van Nuland, Schram, *et al.*, 2013). Since the association of TFIID with chromatin is the initial step in transcription initiation (Patel *et al.*, 2018; Bhuiyan and Timmers, 2019), this implicates H3K4me3 as a key regulator of this process. It is therefore possible that, in addition to contributing to a transcriptionally permissive chromatin environment, H3K4me3 directly affects RNAPII function by mediating PIC formation.

Despite all of the evidence suggesting that there could be a direct relationship between the presence of H3K4me3 at gene promoters and active gene expression, whether H3K4me3 regulates gene expression is a subject of extensive debate within the field (Howe *et al.*, 2017). *In vivo* evidence that H3K4me3 directly contributes to gene expression genome-wide is lacking. From the earliest studies of H3K4me3 in yeast, it was apparent that its complete removal did not result in dramatic transcriptional outcomes (Miller *et al.*, 2001; Santos-Rosa *et al.*, 2002; Boa, Coert and Patterton, 2003; Venkatasubrahmanyam *et al.*, 2007; Guillemette *et al.*, 2011; Lenstra *et al.*, 2011;

Margaritis *et al.*, 2012; Ramakrishnan *et al.*, 2016; Murray *et al.*, 2019). Subsequent studies in *Drosophila* showed that the mutation of H3K4 to a non-methylatable substrate does not abolish the expression of a subset of developmental genes (Hödl and Basler, 2009, 2012). Determining the requirement for H3K4me3 in mammalian cells is complicated by the diversification in the H3K4 HMT machinery, which likely have overlapping functions in H3K4me3 deposition (discussed further in Chapter 1.4). To date, it has been impossible to generate mammalian cells completely devoid of H3K4me3. While the depletion of H3K4me3 through the perturbation of the H3K4 HMTs is associated clear defects in gene expression, there is frequently limited correlation between gene expression changes and alterations in H3K4me3 (P. Wang *et al.*, 2009; Clouaire *et al.*, 2012; Hu, Garruss, *et al.*, 2013; Denissov *et al.*, 2014; Austenaa *et al.*, 2015; Fang *et al.*, 2016; Brown *et al.*, 2017; Franks *et al.*, 2017; Hu *et al.*, 2017; Mas *et al.*, 2018; Sze *et al.*, 2020). Additionally, since the H3K4 HMTs likely have functions that are independent of H3K4me3, unpicking the contribution of the HMT itself from the effects of H3K4me3 is not straightforward. Nonetheless, the current evidence suggests that, at least in the steady state, alterations in H3K4me3 do not have a predictable effect on gene expression. Rather, it has been proposed that the contribution of H3K4me3 to gene expression is context-dependent and relies on other chromatin features, such as different histone modifications, present at gene promoters (Cano-Rodriguez *et al.*, 2016).

It has been proposed that H3K4me3 may be important for gene induction, for example during differentiation (Jiang *et al.*, 2011; Ying Li *et al.*, 2016; Cao *et al.*, 2017; Sze *et al.*, 2017). However, the reduction in H3K4me3 at genes induced by lipopolysaccharide in macrophages or DNA damage in mouse embryonic stem cells (ESCs), does not result in widespread transcriptional defects (Clouaire, Webb and Bird, 2014; Austenaa *et al.*, 2015). Furthermore, the reduction of H3K4me3 at lowly transcribed or inactive CGI-associated genes does not affect the induction of the majority of these genes during retinoic acid-induced differentiation, or differentiation into embryoid bodies (Hu, Garruss, *et al.*, 2013; Denissov *et al.*, 2014; Hu *et al.*, 2017; Mas *et al.*, 2018). Therefore, there is currently little evidence that the presence of H3K4me3 is essential for gene induction.

1.4 The H3K4 methylation system

1.4.1 Evolution of the H3K4 HMTs from yeast and *Drosophila*

The association of H3K4me3 with actively transcribed gene promoters is conserved from yeast through to mammals and throughout all domains of life, H3K4 methylation is deposited by proteins containing a highly conserved C-terminal SET domain (Meeks and Shilatifard, 2017; Cenik and Shilatifard, 2021). However, during evolution and the development of increasingly complex genomes, the mechanisms by which H3K4 methylation is deposited have become increasingly diverse (Figure 1.4).

In yeast, a single histone methyltransferase, ySet1, is responsible for placing all H3K4 methylation (Briggs *et al.*, 2001; Krogan *et al.*, 2002; Nagy *et al.*, 2002; Noma and Grewal, 2002; Santos-Rosa *et al.*, 2002). On its own ySet1 has minimal catalytic activity and must be present within a multi-protein complex, termed ySet1C, in order to efficiently methylate nucleosome substrates (Miller *et al.*, 2001; Roguev *et al.*, 2001, 2003; Krogan *et al.*, 2002; Nagy *et al.*, 2002; J. Kim *et al.*, 2013; Figure 1.4). Swd3, Swd1, Bre2 and Sdc1 are essential for ySet1 catalytic activity *in vitro.* ySet1C also associates with additional proteins, including Spp1 and Swd2, that further regulate its function and have been implicated in its recruitment to chromatin, as well as regulating its catalytic activity (Miller *et al.*, 2001; Acquaviva *et al.*, 2013; Soares and Buratowski, 2013; Howe *et al.*, 2017). The removal of ySet1 results in the complete loss of H3K4 methylation from yeast genomes (Briggs *et al.*, 2001; Krogan *et al.*, 2002; Nagy *et al.*, 2002; Noma and Grewal, 2002; Santos-Rosa *et al.*, 2002). However, yeast cells lacking ySet1 and H3K4 methylation are viable with mild transcriptional defects, albeit with slow growth phenotypes (Nislow, Ray and Pillus, 1997; Briggs *et al.*, 2001; Miller *et al.*, 2001; Noma and Grewal, 2002; Santos-Rosa *et al.*, 2002; Krogan *et al.*, 2002; Nagy *et al.*, 2002; Boa, Coert and Patterton, 2003; Venkatasubrahmanyam *et al.*, 2007; Guillemette *et al.*, 2011; Lenstra *et al.*, 2011; Margaritis *et al.*, 2012; Ramakrishnan *et al.*, 2016; Murray *et al.*, 2019). The only one of the ySet1C components that is essential for yeast growth is Swd2, and this is likely due

to its role within the termination complex, APT (Associated with Pta1), rather than its function within ySet1C (Roguev *et al.*, 2001; Cheng, He and Moore, 2004). Therefore, ySet1C and H3K4me3 are dispensable for yeast transcription and viability.

The H3K4 system has evolved in *Drosophila* to produce three proteins homologous to ySet1 capable of methylating H3K4: dSet1, dTrx and dTrr (Figure 1.4). Similarly to ySet1, dSet1, dTrx and dTrr each form a catalytic core through an association with Swd3, Swd1, Bre2 and Sdc1 homologues, known in *Drosophila* as the WRAD complex (for Wds, Rbbp5, Ash2 and Dpy30 – homologous to Swd3, Swd1, Bre2 and Sdc1, respectively; Mohan *et al.*, 2011). dSet1, dTrr and dTrx also each associate with unique accessory factors that perform a variety of functions within the individual complexes and contribute to their functional specialisation (Figure 1.4). Unlike ySet1, each of the *Drosophila* H3K4 HMTs is essential for normal development (Sedkov *et al.*, 2003; Smith *et al.*, 2004; Hallson *et al.*, 2012). This suggests that the three enzymes that place H3K4 methylation in *Drosophila* have non-redundant, essential functions.

dSet1 shares the closest evolutionary relationship with ySet1 and also associates with two accessory proteins that are homologous to Swd2 and Spp1: Wdr82 and Cfp1, respectively (Ardehali *et al.*, 2011; Mohan *et al.*, 2011; Hallson *et al.*, 2012). dSet1 is responsible for placing the majority of H3K4me2 and H3K4me3 in *Drosophila* and has been implicated in the regulation of gene transcription, at least at some specific loci (Ardehali *et al.*, 2011; Mohan *et al.*, 2011; Hallson *et al.*, 2012). dTrx and dTrr likely evolved through gene duplication events. dTrx appears to regulate H3K4me3 and gene expression at a unique subset of specific loci, primarily the *Hox* genes (Smith *et al.*, 2004; Ardehali *et al.*, 2011). As such, dTrx knockouts display homeotic transformations (Ingham and Whittle, 1980; Breen and Harte, 1991). In contrast, dTrr has been implicated in the deposition of H3K4me1 at enhancers (Ardehali *et al.*, 2011; Herz *et al.*, 2012; Kanda *et al.*, 2013; Rickels *et al.*, 2017). The genome-wide roles of dSet1, dTrx and dTrr remain to be properly defined,

however it seems likely that each complex regulates an independent aspect of the distribution of H3K4 methylation, with at least some function in gene expression.

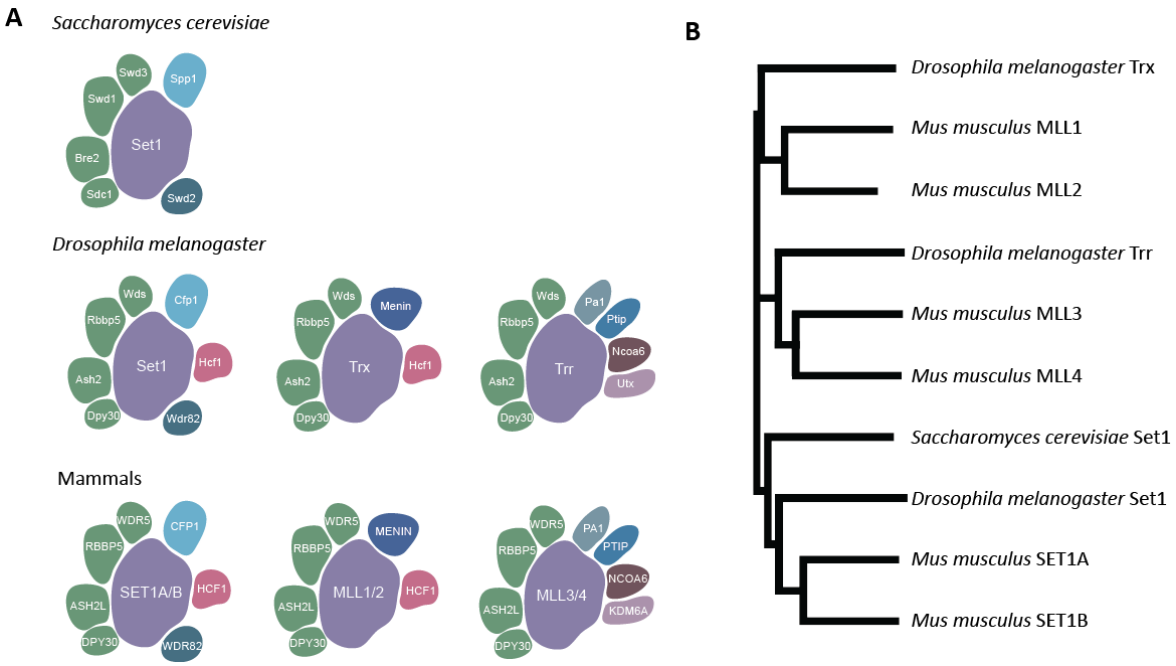


Figure 1.4 Evolution of the H3K4 HMTs
(A) The subunit composition of H3K4 HMTs from yeast, *Drosophila* and mammals. The mammalian H3K4 HMTs are split into three groups based on the homology of their catalytic subunit: SET1A/B, MLL1/2 and MLL3/4. Shared subunits are depicted in the same colour and position. Figure adapted from Howe *et al.*, 2017 and Hughes, Kelley and Klose, 2020.
(B) Sequence conservation of the H3K4 HMTs. Evolutionary distances were calculated using ClustalW (Madeira *et al.*, 2019).

1.4.2 Introducing the mammalian H3K4 HMTs

The H3K4 methylation system has further diversified in mammalian cells, to produce six proteins that are capable of placing H3K4 methylation: MLL1-4 and SET1A/B (Figure 1.4). The mammalian H3K4 HMTs are categorised into three pairs of structurally related proteins: MLL1/2, MLL3/4 and SET1A/B. As with all members of this protein family, MLL1-4 and SET1A/B share a highly homologous C-terminal SET domain and require an association with the WRAD complex for full catalytic activity (Dou *et al.*, 2006; Patel *et al.*, 2008, 2009; Zhang *et al.*, 2012; Shinsky *et al.*, 2015; Yanjing Li *et al.*, 2016). However, the three pairs of homologous proteins possess different additional protein domains and also associate with unique accessory factors that exert specific

functionality to each complex (Shilatifard, 2012; van Nuland, Smits, *et al.*, 2013). Each of the six mammalian H3K4 HMTs is essential for normal development, suggesting that they must perform at least partially non-redundant functions (Yu *et al.*, 1995; Yagi *et al.*, 1998; Glaser *et al.*, 2006; Lee and Skalnik, 2008; Lee *et al.*, 2013; Bledau *et al.*, 2014; Tusi *et al.*, 2015; Fang *et al.*, 2016; Ashokkumar *et al.*, 2020).

Understanding the evolution of the H3K4 HMTs from yeast and *Drosophila* has been informative for understanding their contribution to mammalian biology. SET1A/B are most closely related to dSet1 and ySet1, whereas MLL1-2 and MLL3-4 are most similar to dTrx and dTrr, respectively (Figure 1.4). As with dTrr, MLL3/4 have been shown to place H3K4me1 *in vitro* and *in vivo* (Hu, Gao, *et al.*, 2013; Kaikkonen *et al.*, 2013; Lee *et al.*, 2013; Tang *et al.*, 2013; Shinsky *et al.*, 2015; Kwon *et al.*, 2021). While SET1A/B and MLL1/2 are all capable of placing H3K4me3 *in vitro*, much like their *Drosophila* orthologues, they have been proposed to contribute to H3K4me3 deposition at distinct genomic loci *in vivo* (Shinsky *et al.*, 2015; Meeks and Shilatifard, 2017; Cenik and Shilatifard, 2021; Kwon *et al.*, 2021). The contribution of the mammalian H3K4 trimethyltransferases to H3K4me3 deposition and gene expression is the focus of this thesis.

1.4.3 MLL1/2

1.4.3.1 The association of MLL1/2 with chromatin

MLL1 was originally identified due to the presence of transforming MLL-fusion proteins in leukaemia. In such instances, the MLL1 N-terminus is translocated onto a chromatin-regulating protein, such as ENL or AF9 (Basu, Nandy and Biswas, 2020; Slany, 2020; Takahashi and Yokoyama, 2020). The N-termini of MLL1/2 contain all of their known DNA binding domains, including a ZF-CXXC domain (Figure 1.5). The retention of the ZF-CXXC domain is a universal feature of MLL1-fusion proteins and results in the aberrant targeting of the chromatin-regulating fusion partner to CGI-associated gene promoters (Okuda *et al.*, 2014; Slany, 2020). This typically results in widespread transcriptional defects. It was subsequently shown that the ZF-CXXC domain is also

essential for the correct genomic targeting of wild type (WT) MLL1/2 (Milne *et al.*, 2010; Hu *et al.*, 2017). In fact, MLL1/2 broadly occupy CGIs throughout mammalian genomes, a distribution most likely mediated by their ZF-CXXC domains (Guenther *et al.*, 2005; Denissov *et al.*, 2014; Hu *et al.*, 2017; Figure 1.6).

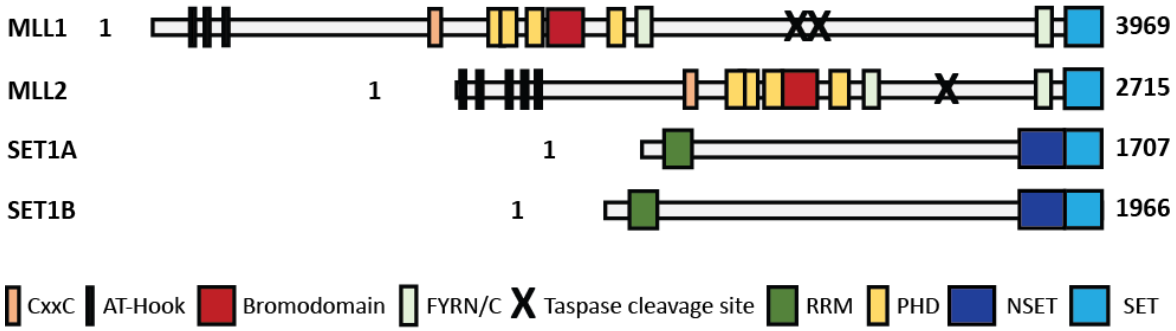


Figure 1.5 Domain architecture of the mammalian H3K4 trimethyltransferases
Protein domains present within each of the four mammalian H3K4 trimethyltransferases are shown. Numbers indicate the length of each protein. RRM: RNA Recognition Motif; FYRN/C: FY-rich domain, N-/C-terminal; PHD: Plant HomeoDomain; SET: Su(var)3-9/E(z)/Trithorax domain. Figure adapted from Crump and Milne, 2019.

ZF-CXXC domains have widely documented roles in recruiting chromatin-modifying proteins to CGIs. However, the ZF-CXXC domain of MLL1 has also been proposed to associate with PAF1, a component of the PAF1C complex, which has important roles in transcription elongation (Milne *et al.*, 2010; Muntean *et al.*, 2010; Hou *et al.*, 2019). This was proposed to be a mechanism for the recruitment of MLL1 to sites of active transcription, at least at a small number of loci. In addition to their ZF-CXXC domains, MLL1/2 also contain numerous other domains typically implicated in chromatin binding (Figure 1.5). For example, MLL1/2 each contain multiple PHD domains and the third PHD finger of MLL1 binds to H3K4me3 (Chang *et al.*, 2010; Milne *et al.*, 2010; Park *et al.*, 2010; Wang *et al.*, 2010). This interaction is important for MLL1 binding to a subset of target genes and has been proposed to drive the propagation of H3K4me3. MLL1/2 also each contain an atypical bromodomain and AT hook motifs which could potentially stabilise binding at acetylated histones or AT-rich DNA, respectively (Zeleznik-Le, Harden and Rowley, 1994; Wang *et al.*, 2010; Filippakopoulos *et al.*, 2012; Rodríguez *et al.*, 2015). Furthermore, the N-terminus of MLL1

associates with MENIN and LEDGF, and the interaction with the former is also conserved for MLL2 (Hughes *et al.*, 2004; Yokoyama *et al.*, 2004, 2005; Milne, Hughes, *et al.*, 2005; Caslini *et al.*, 2007; Yokoyama and Cleary, 2008; Murai *et al.*, 2011; Huang *et al.*, 2012; van Nuland, Smits, *et al.*, 2013). In certain contexts, MENIN and LEDGF are important for the expression of some MLL1 target genes and fusion protein-mediated leukaemogenesis (Hughes *et al.*, 2004; Yokoyama *et al.*, 2004, 2005; Milne, Hughes, *et al.*, 2005; Chen *et al.*, 2006; Caslini *et al.*, 2007; Brown-Bryan *et al.*, 2008; Yokoyama and Cleary, 2008; Huang *et al.*, 2012; Okuda *et al.*, 2014). Since the PWWP domain of LEDGF associates with H3K36me2, this interaction could influence MLL1 chromatin occupancy (Pradeepa *et al.*, 2012; Eidahl *et al.*, 2013; Okuda *et al.*, 2014; Zhu *et al.*, 2016). However, since WT MLL1/2 associate broadly with CGIs regardless of their transcriptional state, the contribution of these interactions for their genome-wide occupancy remains unclear (Figure 1.6). It is possible that these additional domains stabilise MLL1/2 binding at a subset of loci.

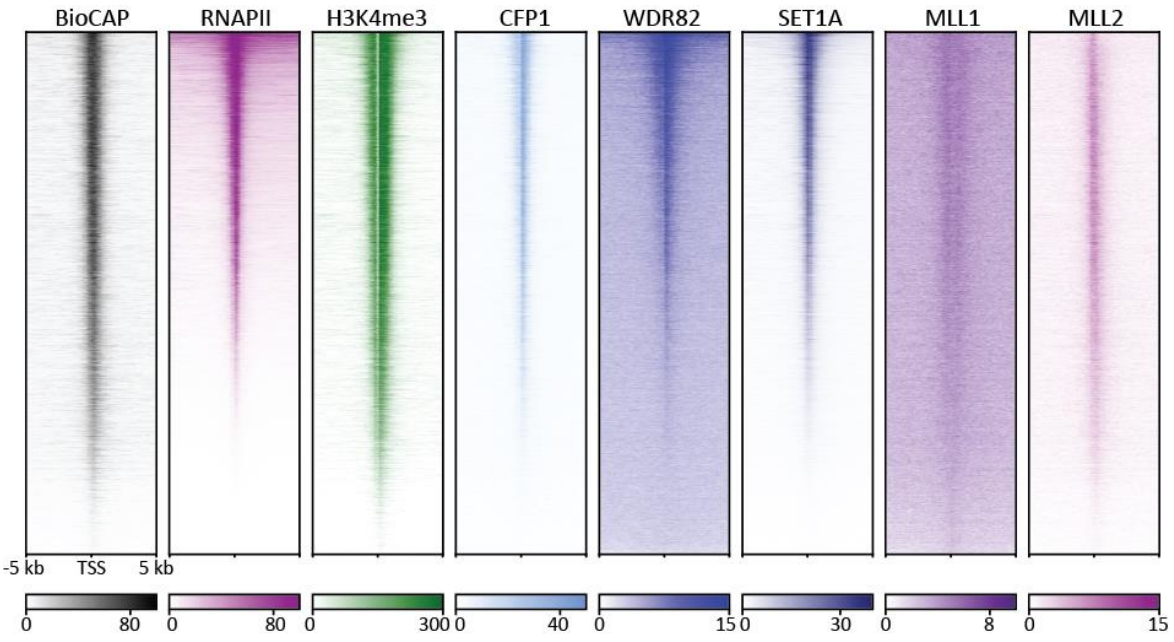


Figure 1.6 Distribution of the H3K4 HMTs on chromatin
Heatmaps illustrating non-methylated DNA (BioCAP), RNAPII (transcription), H3K4me3, CFP1, WDR82, SET1A, MLL1 and MLL2 ChIP-seq signal across all TSSs in mouse embryonic stem cells. The heatmaps are sorted by RNAPII signal (Datasets: Long *et al.*, 2013; Denissov *et al.*, 2014; Brown *et al.*, 2017; Turberfield *et al.*, 2019; Zhang *et al.*, 2019 and unpublished data from the Klose Lab).

1.4.3.2 The role of MLL1/2 in mammalian cells

Both MLL1 and MLL2 are essential for normal embryonic development. MLL1 knockout mice die around embryonic day E10, with haematopoietic and skeletal defects caused by altered *Hox* gene expression (Yu *et al.*, 1995; Yagi *et al.*, 1998). MLL2 knockout results in a much more severe phenotype than that of MLL1; MLL2 knockout mice experience delayed growth from E6.5 (Glaser *et al.*, 2006). Interestingly, the knockout of MLL1 or MLL2 in ESCs does not result in a genome-wide loss of H3K4me3 or genome-wide changes in gene expression, despite both complexes broadly occupying CGIs (Guenther *et al.*, 2005; Wu *et al.*, 2008; Glaser *et al.*, 2009; P. Wang *et al.*, 2009; Hu, Garruss, *et al.*, 2013; Denissov *et al.*, 2014; Hu *et al.*, 2017). Rather, much like *Drosophila* dTrx, MLL1/2 have been implicated in the deposition of H3K4me3 at, and the activation of, distinct subsets of developmental genes (Yagi *et al.*, 1998; Milne *et al.*, 2002; Nakamura *et al.*, 2002; Dou *et al.*, 2005; Glaser *et al.*, 2006; P. Wang *et al.*, 2009; Hu, Garruss, *et al.*, 2013; Denissov *et al.*, 2014). MLL2 has been specifically implicated in placing H3K4me3 at the promoters of lowly expressed, or 'bivalent', genes in ESCs (Hu, Garruss, *et al.*, 2013; Denissov *et al.*, 2014; Mas *et al.*, 2018; Sze *et al.*, 2020). The deposition of H3K4me3 by MLL2 at these loci has been proposed to enable their rapid transcriptional induction during differentiation (Voigt, Tee and Reinberg, 2013). However, the relevance of MLL2 and H3K4me3 for the expression of these genes remains unclear, since the loss of MLL2 and reductions in H3K4me3 typically do not result in dramatic gene expression perturbations of such genes, either in the steady state or during differentiation (Glaser *et al.*, 2009; Hu, Garruss, *et al.*, 2013; Denissov *et al.*, 2014; Hu *et al.*, 2017; Mas *et al.*, 2018). There is therefore currently little evidence that the presence of MLL2 and H3K4me3 is necessary to allow the rapid induction of lowly expressed genes during differentiation.

1.4.4 SET1A/B

1.4.4.1 The association of SET1A/B with chromatin

SET1A and SET1B, collectively known as the SET1 proteins, are highly homologous proteins that are the mostly closely related of the mammalian H3K4 HMTs to ySet1 and dSet1 (Lee and Skalnik, 2005; Lee *et al.*, 2007; Figure 1.4). As such, in addition to the WRAD complex, SET1A/B also associate with two accessory factors, CFP1 and WDR82, that are homologous to Spp1/dCfp1 and Swd2/dWdr82, respectively (Figure 1.4; Lee and Skalnik, 2005; Lee *et al.*, 2007; Lee and Skalnik, 2008; Wu *et al.*, 2008).

Much like MLL1/2, the genomic distribution of SET1A broadly follows the presence of CGIs (Thomson *et al.*, 2010; Denissov *et al.*, 2014; Brown *et al.*, 2017; Figure 1.6). However, SET1A is not uniformly distributed across CGIs. Rather, it preferentially occupies the CGI-associated gene promoters of actively transcribed genes (Brown *et al.*, 2017; Sze *et al.*, 2017, 2020). While the genome-wide distribution of SET1B remains to be determined, it would be predicted to associate with similar genomic loci to SET1A due to their high homology and presence within largely similar multi-subunit complexes. In contrast to the MLL proteins, SET1A/B do not possess DNA-binding domains and the association of SET1A/B with chromatin is thought to be primarily defined by CFP1 and WDR82 (Lee and Skalnik, 2008; Tate, Lee and Skalnik, 2010; Brown *et al.*, 2017; Franks *et al.*, 2017; Figure 1.5). The binding of the WRAD complex components to nucleosomes likely functions to support the binding of SET1A/B to target sites in chromatin (Chen *et al.*, 2011; Sarvan *et al.*, 2011; Bochyńska, Lüscher-Firzlaff and Lüscher, 2018; Mittal *et al.*, 2018).

CFP1 is the founding member of the ZF-CXXC domain-containing protein family and is an essential component of the SET1A/B complexes (Voo *et al.*, 2000; Carlone and Skalnik, 2001; Lee, Voo and Skalnik, 2001; Lee and Skalnik, 2005). *In vitro* binding studies have shown that the ZF-CXXC domain of CFP1 associates specifically with non-methylated CpG dinucleotides, an interaction which is essential for the association of CFP1 with chromatin (Voo *et al.*, 2000; Lee, Voo and Skalnik, 2001;

Xu *et al.*, 2011; Denissov *et al.*, 2014; Brown *et al.*, 2017). CFP1 is essential for the association of SET1A with CGIs, suggesting that SET1A/B associate with CGIs via the ZF-CXXC domain of CFP1 (Brown *et al.*, 2017).

Binding of the ZF-CXXC domain of CFP1 to CGIs cannot completely explain the genomic distribution of SET1A/B as SET1A is preferentially enriched at the promoters of actively transcribed genes (Brown *et al.*, 2017; Sze *et al.*, 2017, 2020). How is this enrichment achieved? A key feature of actively transcribed genes is the presence of promoter-associated H3K4me3 (Barski *et al.*, 2007; Mikkelsen *et al.*, 2007; Guenther *et al.*, 2007). In addition to its ZF-CXXC domain, CFP1 also contains a PHD finger that can bind to H3K4me3 (Eberl *et al.*, 2013; Mahadevan and Skalnik, 2016). This interaction provides the SET1A/B complexes with a preferential affinity for gene promoters already modified with H3K4me3. Mutations that disrupt H3K4me3 binding by CFP1 result in residual CFP1 occupancy that is distributed across CGIs (Brown *et al.*, 2017). In contrast, if the ZF-CXXC domain of CFP1 is mutated, there remains little, if any, CFP1 occupancy at CGIs (Brown *et al.*, 2017). This suggests that the SET1A/B complexes primarily associate with chromatin via the ZF-CXXC domain of CFP1, while multivalent engagement via the PHD domain functions to stabilise the occupancy of the SET1A/B complexes at gene promoters with pre-existing H3K4me3. This has been proposed to create a simple feedback mechanism whereby H3K4me3 is amplified at the promoters of actively transcribed genes through the stabilisation of SET1A/B to H3K4me3-containing chromatin (Brown *et al.*, 2017).

If SET1A/B enrichment at actively transcribed genes relies on pre-existing H3K4me3, how is the H3K4me3-elevated state initiated at such loci? One possibility is that MLL1/2 could initiate H3K4me3 deposition preferentially at actively transcribed genes. However, since MLL1/2 broadly occupy CGIs and appear to contribute to H3K4me3 at specific developmental genes, it seems unlikely that MLL1/2 are responsible for this initial enrichment of H3K4me3 (Chapter 1.4.3.2). Interestingly, in the absence of CFP1, SET1A is partially retained at the promoters of more highly

expressed genes (Brown *et al.*, 2017). It is therefore possible that there is another feature of actively transcribed genes that results in the enrichment of SET1A/B. Indeed, it has been proposed that the process of active transcription itself could be responsible for initiating this H3K4me3-amplifying cascade and the second unique accessory factor of SET1A/B, WDR82, has been implicated in this process. Through biochemical assays, WDR82 has been shown to associate with the initiating S5P form of RNAPII and has been suggested to mediate SET1A recruitment, at least in certain contexts (Lee and Skalnik, 2008; Lee *et al.*, 2010; Ebmeier *et al.*, 2017; Franks *et al.*, 2017). It has been proposed that the association of WDR82 with initiating RNAPII could preferentially enrich SET1A/B at the promoters of actively transcribed genes. Such an association would allow for the deposition of H3K4me3 at actively transcribed genes and further stabilisation of SET1A/B binding via the PHD domain of CFP1. However, interpreting the role of WDR82 in the context of the SET1A/B complexes is challenging because WDR82 is essential for SET1A stability (Lee and Skalnik, 2008; Wu *et al.*, 2008; Austenaa *et al.*, 2015; Franks *et al.*, 2017). Additionally, WDR82 also functions within the PNUTS-PP1 phosphatase complex with a role in transcription termination, confounding observations about the effects of its removal (Trinkle-Mulcahy *et al.*, 2006; Lee *et al.*, 2010; van Nuland, Smits, *et al.*, 2013; Austenaa *et al.*, 2015; Ebmeier *et al.*, 2017; Cortazar *et al.*, 2019). Therefore, the role of active transcription and the association of WDR82 with RNAPII in the function of SET1A/B remains enigmatic.

1.4.4.2 The role of SET1A/B in mammalian cells

Among the mammalian H3K4 HMTs, SET1A/B share the closest evolutionary relationship to yeast ySet1 and *Drosophila* dSet1 (Figure 1.4). Additionally, SET1A/B are capable of placing H3K4me1, me2 and me3 *in vitro* (Lee and Skalnik, 2005; Lee *et al.*, 2007; Wu *et al.*, 2008; Shinsky *et al.*, 2015; Kwon *et al.*, 2021). This has led to the suggestion that SET1A/B could be responsible for the majority of H3K4me3 in mammalian cells, similar to dSet1 in *Drosophila* (Meeks and Shilatifard, 2017; Cenik and Shilatifard, 2021).

Much of our understanding about the role of SET1A/B in mammalian cells has come from studies in which their chromatin binding has been perturbed by the removal of CFP1 or WDR82. In such studies, disrupting the association of SET1A/B with chromatin has been reported to result in a genome-wide reduction in H3K4me3, supporting the proposed role of SET1A/B as the major H3K4 trimethyltransferases in mammals (Lee and Skalnik, 2005; Wu *et al.*, 2008; Clouaire *et al.*, 2012; Vethantham *et al.*, 2012; Tang *et al.*, 2013; Clouaire, Webb and Bird, 2014; Brown *et al.*, 2017; Franks *et al.*, 2017; Yang *et al.*, 2020). ChIP-sequencing has revealed that the removal of SET1A/B results in the largest reductions in H3K4me3 at the most highly expressed genes, in keeping with the preferential enrichment of SET1A at these loci (Clouaire *et al.*, 2012; Clouaire, Webb and Bird, 2014; Brown *et al.*, 2017; Sze *et al.*, 2017, 2020). In contrast, H3K4me3 at lowly expressed genes, which appears to rely on MLL1/2, remained largely unchanged. This has led to a model whereby MLL1/2 place H3K4me3 at specific, lowly-expressed developmental genes, while SET1A/B are responsible for elevated H3K4me3 at actively transcribed genes (Meeks and Shilatifard, 2017; Cenik and Shilatifard, 2021).

Despite the observed reduction in H3K4me3 when SET1A/B are perturbed, and the association of H3K4me3 with active gene expression, the relationship between SET1A/B and the regulation of gene expression remains unclear. Typically, removal of CFP1 or WDR82 does result in changes in gene expression (Clouaire *et al.*, 2012; Clouaire, Webb and Bird, 2014; Austenaa *et al.*, 2015; Brown *et al.*, 2017; Franks *et al.*, 2017). However, these changes often manifest as increases in expression at some genes and decreases at others. For example, the conditional removal of CFP1 in ESCs resulted in the decreased expression of 1108 CFP1 target genes and the increased expression of 584 CFP1 target genes (Brown *et al.*, 2017). Similarly, the knockdown of WDR82 in haematopoietic progenitor cells resulted in 1831 genes with increased expression and 1451 genes with decreased expression (Franks *et al.*, 2017). Given that these perturbations resulted in both increases and decreases in gene expression, it has remained unclear whether SET1A/B are primarily involved in potentiating or counteracting steady state gene expression. Alternatively, it has been proposed

that SET1A/B and H3K4me3 deposition may contribute more specifically to the kinetics of transcription during gene induction, for example during differentiation or in response to other cellular stimuli (Carlone *et al.*, 2005; Young and Skalnik, 2007; Deng *et al.*, 2013; Tusi *et al.*, 2015; Mahadevan and Skalnik, 2016; Ying Li *et al.*, 2016; Cao *et al.*, 2017; Sze *et al.*, 2017). However, while SET1A/B removal has been reported to affect the induction of specific genes during differentiation, it remains unclear whether SET1A/B have a widespread role in gene induction (Deng *et al.*, 2013; Clouaire, Webb and Bird, 2014; Austenaa *et al.*, 2015; Ying Li *et al.*, 2016; Cao *et al.*, 2017; Sze *et al.*, 2017).

Interfering with the chromatin binding of SET1A/B through removal of CFP1 or WDR82 does not allow for the unique functions of SET1A and SET1B in H3K4me3 deposition and gene expression to be determined. Although SET1A and SET1B are highly homologous, their individual knockout leads to markedly different developmental phenotypes (Lee *et al.*, 2007; Bledau *et al.*, 2014). Loss of SET1A during mouse development results in the most extreme phenotype of any of the MLL/SET1 proteins; mice deficient for SET1A during development implant but fail to gastrulate and do not survive beyond embryonic day E7.5 (Bledau *et al.*, 2014). While SET1B is essential for development, SET1B knockout embryos can gastrulate and survive until day E11.5, albeit with an extremely growth retarded phenotype (Bledau *et al.*, 2014). Additionally, SET1A is essential for ESC proliferation, while SET1B is dispensable and cannot compensate for SET1A loss (Bledau *et al.*, 2014; Fang *et al.*, 2016). There is also evidence that SET1A and SET1B can associate with unique accessory factors not shared by the other protein. For example, SET1A specifically has been implicated in the repair of DNA damage, partly through a direct interaction with BOD1L, RAD18 or Cyclin K (van Nuland, Smits, *et al.*, 2013; Wang *et al.*, 2017; Higgs *et al.*, 2018; Hoshii *et al.*, 2018; Alsulami *et al.*, 2019). SET1A also directly interacts with HCF1, while SET1B has been shown to associate with BOD1 and the RNA-binding protein RBM15 (Wysocka *et al.*, 2003; Lee *et al.*, 2007; Lee and Skalnik, 2012; van Nuland, Smits, *et al.*, 2013; Wang *et al.*, 2017; Coker *et al.*, 2020).

Therefore, while both complexes clearly perform an essential function during development, their functions are likely to be at least partially non-redundant.

The individual contributions of SET1A and SET1B to the deposition of H3K4me3 and gene expression have begun to be dissected. The removal of SET1A by knockdown or knockout approaches has been reported to result in a global decrease in H3K4me3, though the extent of H3K4me3 loss at specific loci genome-wide remains unclear (Wu *et al.*, 2008; Bledau *et al.*, 2014; Fang *et al.*, 2016; Cao *et al.*, 2017; Sze *et al.*, 2020). While SET1B knockout alone appears to have a subtle effect on genome-wide H3K4me3, the combined removal of SET1A and SET1B has been reported to result in a more dramatic H3K4me3 reduction in comparison to the single knockouts alone (Wu *et al.*, 2008; Bledau *et al.*, 2014; Wang *et al.*, 2017; Sze *et al.*, 2020). This suggests that while SET1A may be responsible for placing the majority of H3K4me3 in mammalian cells, SET1B is able to place H3K4me3 to compensate for SET1A loss (Sze *et al.*, 2020). Additionally, while the depletion of SET1A results in widespread gene expression defects, these changes lack directionality, as observed with CFP1 or WDR82 knockouts (Fang *et al.*, 2016). The relationship between SET1B and genome-wide gene expression in ESCs remains untested, though its depletion was reported to have extremely subtle effects on gene expression in human cancer cells, but a more dramatic effect in mouse haematopoietic stem and progenitor cells (Wang *et al.*, 2017; Schmidt *et al.*, 2018). Therefore, while SET1A and SET1B clearly have a role in placing H3K4me3 and regulating gene expression, the individual contribution of each of these proteins remains largely unclear.

1.4.5 Can the H3K4 HMTs function independently of H3K4me3?

The SET1/MLL proteins together regulate the deposition H3K4me3, a histone modification implicated in the regulation of gene expression (Chapter 1.3). However, there is currently conflicting evidence about the contribution of H3K4me3 to gene expression (Chapter 1.3.4). Additionally, changes in gene expression upon perturbation of the MLL/SET1 complexes are not

necessarily associated with a change in H3K4me3. It is therefore possible that the SET1/MLL complexes regulate gene expression independently of H3K4me3. The SET1/MLL complexes could methylate non-histone substrates, or could even have functions that are completely independent of catalysis.

Experiments in yeast initially suggested that the SET1/MLL proteins could have substrates other than histone H3. ySet1 was shown to methylate the Dam1 kinetochore protein, which is involved in kinetochore assembly and chromosome segregation (Zhang *et al.*, 2005; Latham *et al.*, 2011). This activity implicates ySet1 in mitosis and highlights the possibility that MLL/SET1 proteins have non-histone substrates in mammals. Indeed, additional substrates for each of the mammalian MLL/SET1 HMTs have since been identified. In cancer cells, SET1A has been proposed to methylate the protein chaperone HSP70 and the TF YAP, implicating SET1A in the regulation of cancer cell proliferation (Cho *et al.*, 2012; Fang *et al.*, 2018). Each of the MLL proteins, as well as SET1B, have also been proposed to methylate p53 (Li *et al.*, 2020). While the function of these modifications remains to be determined, it is possible that by methylating other substrates, such as the transcription machinery, the SET1/MLL proteins regulate transcription independently of H3K4.

Another possibility is that the SET1/MLL complexes regulate gene expression completely independently of their methyltransferase activity. As already mentioned, MLL1 knockout mice display embryonic lethality (Yu *et al.*, 1995; Yagi *et al.*, 1998). However, the specific deletion of the MLL1 SET domain in mice does not phenocopy the embryonic lethality of full MLL1 knockout (Terranova *et al.*, 2006; Mishra *et al.*, 2014). This would suggest that the essential role of MLL1 in development is independent of catalysis. However, a point mutation in the SET domain of dTrx is sufficient to produce dTrx-dependent developmental phenotypes in *Drosophila* (Breen, 1999). Additionally, MLL1 HMT activity is essential for optimal transcription activation *in vitro* and *Hox* gene expression *in vivo* (Milne *et al.*, 2002; Dou *et al.*, 2005; Terranova *et al.*, 2006). This would argue that MLL1 catalytic activity is, in fact, essential for its role in gene expression. Likewise, the

activation of *in vitro* transcription by MLL2 relies on the presence of SAM and a methylation-competent H3K4 substrate (Jiang *et al.*, 2013). This suggests that MLL2 also depends on its catalysis, and specifically H3K4 methylation, to activate gene expression.

There is evidence that SET1A/B could possess both catalysis-dependent and -independent functions. Firstly, SET1B contributes to cancer cell viability independently of its SET domain (Wang *et al.*, 2017). Furthermore, whereas SET1A is essential for ESC growth, the constitutive removal of the SET1A SET domain has little effect on ESC proliferation (Bledau *et al.*, 2014; Fang *et al.*, 2016; Sze *et al.*, 2017). Instead, ESCs lacking the SET1A SET domain undergo abnormal differentiation (Sze *et al.*, 2017). These findings led to the proposal that SET1A catalytic activity has a role in the differentiation of ESCs, while its non-catalytic function contributes to self-renewal. In keeping with this, there is evidence that SET1A is required for growth of MLL-AF9 transformed leukaemia cells (Hoshii *et al.*, 2018). However, the SET1A SET domain is dispensable for leukaemia growth, again suggesting that the catalytic activity of SET1A is not required for cell proliferation. Interestingly, however, SET1A is unable to activate *in vitro* transcription on a H3K4 template that can't be methylated (Tang *et al.*, 2013). This suggests that in *in vitro* transcription systems, methylation of H3K4 is essential for the ability of SET1A to activate gene expression.

One way in which the SET1/MLL proteins have been proposed to regulate gene expression independently of their catalytic activity is through directly interacting with transcriptional coactivators. For example, MLL1 has been shown to associate with several histone acetyltransferases (Ernst *et al.*, 2001; Dou *et al.*, 2005; Paggetti *et al.*, 2010). In some instances, these interactions are important for the combined deposition of histone acetylation and H3K4me3 and the activation of target gene expression (Ernst *et al.*, 2001; Dou *et al.*, 2005; Paggetti *et al.*, 2010; Mishra *et al.*, 2014; Valerio *et al.*, 2017). SET1A has also been implicated in binding to histone acetyltransferases, such as p300, in order to concomitantly place H3K4me3 and histone acetylation at gene promoters (Tang *et al.*, 2013). While this was proposed to be a mechanism for SET1A

recruitment, such experiments raise the possibility that the MLL/SET1 complexes could contribute to transcription activation by directly influencing coactivator function and histone acetylation at their target gene promoters.

Another possibility is that the SET1/MLL proteins could directly influence RNAPII. Early work indicated that MLL1 interacts with RNAPII and it was subsequently shown to associate with the PAF1 complex, which has important roles in transcription elongation (Milne, Dou, *et al.*, 2005; Milne *et al.*, 2010; Muntean *et al.*, 2010). Such interactions could provide a mechanism for MLL1 to regulate RNAPII. Additionally, WDR82 of the SET1A/B complexes associates with the S5P modified form of RNAPII (Lee and Skalnik, 2008; Lee *et al.*, 2010; Ebmeier *et al.*, 2017). While this interaction has primarily been implicated in the recruitment of SET1A/B to chromatin, it is also possible that this is a means by which SET1A/B influence RNAPII. Nonetheless, the H3K4 HMTs are associated with activities that could result in changes in the chromatin architecture around gene promoters, or directly affect gene transcription, independently of H3K4me3.

1.5 Aims of thesis

SET1A/B are H3K4 HMTs that associate with gene promoters. Since SET1A/B are indispensable for normal mammalian development and often mutated in human disease, understanding their contribution to the regulation of gene expression is essential. Much of our current understanding about the role of SET1A/B in the deposition of H3K4me3 and the regulation of gene expression has come from studies in simpler eukaryotes or from mammalian systems in which their chromatin binding has been perturbed by the use of slow protein depletion systems. Such studies have led to a model whereby SET1A/B mediate the majority of H3K4me3 deposition in mammalian cells, with no predictable effect on gene expression. However, such experiments could be confounded by the use of slow protein depletion systems that are unable to capture the most primary effects of SET1A/B removal. Therefore, the central aim of this thesis is to examine the primary contribution of SET1A and SET1B to the regulation of mammalian gene expression.

To address this question, I employ genome engineering to generate ESC lines that enable the controlled and rapid degradation of SET1A and SET1B, both individually and in combination. By using quantitative genome-wide analysis, I first examine the contribution of SET1A and SET1B to gene expression in ESCs (Chapter 3). Since SET1A/B are H3K4 HMTs, I next examine their contribution to genome-wide H3K4me3 in ESCs, and the relationship between their roles in placing H3K4me3 and regulating gene expression (Chapter 4). I then use a synthetic reporter system to investigate the mechanisms by which SET1A/B could regulate gene expression and also examine their contribution to RNAPII function (Chapter 5). Together, this work provides evidence that SET1A/B are global transcriptional activators that can regulate gene transcription independently of their catalytic activity.

2. Materials and Methods

2.1 Cell lines and cell culture

2.1.1 Cell culture conditions

Mouse embryonic stem cells were grown in Dulbecco's Modified Eagle Medium (Gibco) supplemented with 15% fetal bovine serum (FBS, Biosera), 1x non-essential amino acids (Gibco), 2 mM L-glutamine (Gibco), 1x penicillin/streptomycin (Gibco), 0.5 mM beta-mercaptoethanol (Gibco), and 10 $ngml^{-1}$ leukemia inhibitory factor. ESCs were grown on gelatinised plates at 37°C and 5% CO_2. Cells were passaged using trypsin-EDTA (0.25%, Gibco) with 2% chicken serum or TrypLE Express Enzyme (1X, Gibco).

Drosophila melanogaster SG4 cells were grown adhesively at 25°C in Schneider's *Drosophila* Medium (Gibco) supplemented with 1x penicillin/streptomycin and 10% heat-inactivated FBS.

Human HEK293T cells were grown at 37°C and 5% CO_2 in Dulbecco's Modified Eagle Medium, supplemented with 10% FBS, 1x penicillin/streptomycin, 2 mM L-glutamine and 0.5 mM beta-mercaptoethanol.

2.1.2 Cell counting

A Countess II cell counter (Life Technologies) was used to count cells. Cell diameter > 12 µm gating was used for mammalian cells, while cell diameter > 3 µm and cell elongation < 70 settings were used for *Drosophila* cells.

2.1.3 Tamoxifen, dTAG-13 and Doxycycline treatments

Tamoxifen treatments of conditional knockout or conditional catalytic point mutant cells were performed using 800 nM 4-hydroxytamoxifen (OHT) for 96 hr to induce Cre-mediated recombination.

Cell lines expressing FKBP12^{F36V} (herein referred to as the dTAG) fusion proteins were treated with 100 nM dTAG-13 (Carole Bataille, Chemistry Research Laboratory, University of Oxford) to induce protein depletion. Media was pre-warmed for at least 1 hr before treatment.

Cells expressing rTetR-fusion proteins were treated with 1 µgml^{-1} doxycycline for 6 hr to induce binding to the TetO array.

2.1.4 Embryonic stem cell lines

All ESC lines used in this study are listed in Table 2.1.

Name	Genotype	Origin
TIR1	*Rosa26^{TIR1}*	Rhodes *et al.*, 2020
mAID-SET1A	*mAID-Setd1A; Rosa26^{TIR1}*	This study
mAID-SET1A, no TIR1	*mAID-Setd1A*	This study
dTAG-SET1A	*FKBP12^{F36V}-Setd1A*	This study
T7-dTAG-SET1A	*Triple T7-FKBP12^{F36V}-Setd1A*	This study (with Anna Lastuvkova)
dTAG-SET1B	*FKBP12^{F36V}-Setd1B*	This study
dTAG-SET1A/B	*FKBP12^{F36V}-Setd1A; FKBP12^{F36V}-Setd1B*	This study
ERT2-Cre recombinase	*Rosa26^{CreERT2}*	Blackledge *et al.*, 2020
SET1A conditional knock out	*Setd1A$^{fl/fl}$; Rosa26 ^{CreERT2}*	This study
SET1A conditional point mutant	*Setd1A$^{(WT->N1655S/Y1693F)fl/fl}$; Rosa26 ^{CreERT2}*	This study (with Anna Lastuvkova)
TOT2N	hBAC RP11 419E6 (chr8); 14x TetO	Neil P Blackledge *et al.*, 2014
Luciferase reporter cells	hBAC RP11 419E6 (chr8); 7xTetO-Ef1α-*Firefly Luciferase*	Anne Turberfield, unpublished
WDR82-dTAG	*Wdr82-FKBP12^{F36V}*	This study

Table 2.1 Name, genotype and origin of ESC lines used in this study

2.1.5 CRISPR-mediated genome editing

Cell lines containing tags or LoxP sites were engineered using CRISPR/Cas9-mediated genome editing. sgRNAs were designed using the CRISPOR online tool (http://crispor.tefor.net) and oligonucleotides encoding sgRNAs were cloned into the pSpCas9(BB)-2A-Puro plasmid as previously described (Ran *et al.*, 2013). All guide sequences used are listed in Table 2.2.

Targeting	**gRNA sequence**
Setd1A 5'	TAAATGAGCAAAGATGGACC
Setd1B 5'	TTCAGGTTGGGTTAACGGCA
Setd1A 5' LoxP site	CTTCGGACTGTCTCGAAAAT
Setd1A 3' LoxP site	GCTTGTAAGTCAATGAGGCC
Setd1A stop codon (for SET1A conditional point mutant downstream LoxP sites)	CGTGGCTCCCTCAACTGAGG
Setd1A conditional point mutant cassette insertion	GGCATCCCACTCTCTTATGG
Wdr82 3'	GCATCAGGATCAGTCATCAA

Table 2.2 Cas9 guide sequences used in this study

Targeting constructs for homology directed repair (HDR) were assembled by Gibson Assembly, as per the manufacturer's instructions (New England Biolabs, NEB). In brief, homology arms (HAs) of roughly equivalent length to the sequence to be inserted were amplified from ESC genomic DNA, $FKBP12^{F36V}$ tag sequences from pLEX_305-N-dTAG (pLEX_305-N-dTAG was a gift from James Bradner & Behnam Nabet (Addgene plasmid#91797;http://n2t.net/addgene:91797; RRID:Addgene_91797)) and mAID tag from AID-RING1B plasmid (Rhodes *et al.*, 2020). To generate the dTAG-SET1B targeting construct, the dTAG was amplified from the dTAG-SET1A plasmid. Targeting constructs for LoxP site insertion were synthesised by GeneArt (Invitrogen). To generate the SET1A conditional point mutant (CPM), the cassette containing the upstream LoxP sites, mutated exons and splice acceptor was synthesised by GeneArt and Gibson assembled with HAs as described above. The T7-dTAG-SET1A cell line was generated by Anna Lastuvkova, using the

same protocol as outlined above for targeting construct assembly. HA1-Triple T7 tag was amplified from the T7-SET1A targeting construct used previously (Brown *et al.*, 2017) and FKBP12^{F36V} tag-HA2 was amplified from the dTAG-SET1A plasmid. In all instances, targeting constructs were designed such that Cas9 recognition sites were disrupted after targeting. Primers used to clone homology arms and tag sequences are listed in Table 2.3.

Purpose	**Sequence**
SET1A HA1 F	GAAGGAAGGCCGTCAAGGCCGCATGAAGGTAAGGGGGCGCGAG
mAID-SET1A HA1 R	CTTTGCTCATTTACACTGTCAGG
mAID-SET1A mAID tag F	TTCCTGACAGTGTAAATGAGCAAAGATGAAGGAGAAGAGTGCTTGTCC
mAID-SET1A mAID tag R	TCCCGAACCTGAGCCTCCACTTCCACCTTTATACATCCTCAAATCGATTTTCC TC
mAID-SET1A HA2 F	TGGAAGTGGAGGCTCAGGTTCGGGAATGGACCAGGAAGGTGGG
SET1A HA2 R	GAAAGGAAGGCCCATGAGGCCCAGACCCACTGTCACAAGGCTTC
dTAG-SET1A HA1 R	CCTGCACTCCCATCTTTGCTCATTTACACTGTCAGGAA
dTAG-SET1A dTAG F	AAATGAGCAAAGATGGGAGTGCAGGTGGAAAC
dTAG-SET1A dTAG R	TCCCGAACCCGAGGATAACTTCGTATAATGTATGCTATACGAAGTTATGCC TCCACTTCCACCTTCCAGTTTTAGAAGCTCCACAT
dTAG-SET1A HA2 F	TACGAAGTTATCCTCGGGTTCGGGAATGGACCAGGAAGGTGGG
T7-dTAG-SET1A HA1-T7 F	ACTATAGGGCGAATTGAAGGAAGGCCGTCAAGGCCGCATGATCTGAGGG AGCCACGGGTCATGC
T7-dTAG-SET1A HA1-T7 R	ATGGTTTCCACCTGCACTCCCCCGGGTGATCTAGCTCCTTTCTCGA
T7-dTAG-SET1A dTAG-HA2 F	AAGGAGCTAGATCACCCGGGGGAGTGCAGGTGGAAAC
T7-dTAG-SET1A dTAG-HA2 R	TGGAAAGCGGGCAGTGAAAGGAAGGCCCATGAGGCCCAGACCCACTGTC ACA
dTAG-SET1B HA1 F	GAAGGAAGGCCGTCAAGGCCGCATGCGTAAACCCTCCGGAAAGG
dTAG-SET1B HA1 R (with a 2 bp mutation in Cas9 cut site)	CCTGCACTCCCATGCCGTTAACCCTTCCTGAAACCC
dTAG-SET1B dTAG F	GGGTTAACGGCATGGGAGTGCAGGTGGAAA
dTAG-SET1B dTAG R	GGTGACTGTTCTCTCCCGAACCCGAGGATAACTT
dTAG-SET1B HA2 F	TCGGGTTCGGGAGAGAACAGTCACCCCCAC
dTAG-SET1B HA2 R	GAAAGGAAGGCCCATGAGGCCCAGATAAGAATCCGGCTGGGCTTG

SET1A CPM Cassette HA1 F	GAAGGAAGGCCGTCAAGGCCGCATGGGGTAATGAGTTCCCAGGGC
SET1A CPM Cassette HA1 R	TATACGAAGTTATGATGCCATTCCTACATCCCTAGG
SET1A CPM Cassette F	AGGAATGGCATCATAACTTCGTATAGGATACTTTATACGAAGTTATGCAGA
SET1A CPM Cassette R	CATAAGAGAGTGGGTCGACTGGAGAGTGGCC
SET1A CPM Cassette HA2 F	TCTCCAGTCGACCCACTCTCTTATGGAGGTAGATTTG
SET1A CPM Cassette HA2 R	GAAAGGAAGGCCCATGAGGCCCAGAACACGCCCTAAGACTTCAGC
WDR82-dTAG HA1 F	ACTATAGGGCGAATTGAAGGAAGGCCGTCAAGGCCGCATGTGGGCTCAATCCCTCCAGTA
WDR82-dTAG HA1 R	GCCACCACCGCTTCCACCACCTCCTCCCGAACCTGAGCCTCCACTTCCACCGTCATCAATGGTGGGCAACCA
WDR82-dTAG dTAG F	GGAGGTGGTGGAAGCGGTGGTGGCGGCAGTGGCGGAGGCATGGGAGTGCAGGTGGAAAC
WDR82-dTAG dTAG R	GCCCAGCAGCATCAGGATCATTCCAGTTTTAGAAGCTCCACATCGAAG
WDR82-dTAG HA2 F	TGGAGCTTCTAAAACTGGAATGATCCTGATGCTGCTGGG
WDR82-dTAG HA2 R	TGGAAAGCGGGCAGTGAAAGGAAGGCCCATGAGGCCCAGAAGCCTCCTGACAAACAAGGT

Table 2.3 Primers used to clone HDR targeting constructs

ESCs were transfected at approximately 70% confluency. For CRISPR, 0.5 µg of guide plasmid and 2 µg of targeting construct were transfected in a 6 well plate using Lipofectamine 3000, according to the manufacturer's protocol (Invitrogen). The following morning, transfected cells were passaged to new plates at low density. After 4 hr, 1 $ugml^{-1}$ puromycin selection was applied. Selection was maintained for 48 hr. Individual colonies were picked into 96-well plates and positive clones identified by PCR screens of genomic DNA. PCR screens were performed using primers flanking the inserted sequence, and homozygous clones identified by an increase in product length corresponding to the desired insertion, with loss of the WT product. Screening primers are listed in Table 2.4. Clones were validated by western blotting.

To generate the SET1A conditional knockout cell line, the 5' LoxP site was first targeted into the ERT2-Cre recombinase cell line and homozygous clones carried forward for insertion of the 3' LoxP site.

To generate the SET1A conditional point mutant cell line, the 3' LoxP sites were first targeted into the ERT2-Cre recombinase cell line. Clones homozygote for LoxP site insertion were then targeted with the SET1A CPM cassette.

Insertion	**Sequence**
mAID-SET1A	F: GGATTCAGGGACCGAAGCT R: GAGCTTGTAGTTCCGCCACT
dTAG-SET1A / T7-dTAG-SET1A	F: CAAGATCTGAGGGAGCCACG R: GAGCTTGTAGTTCCGCCACT
dTAG-SET1B	F: CGTCGGAGAGGACAACAACT R: AGCCGGGTCAATCATCAACT
Setd1A 5' LoxP site	F: TTGTGTGTGCTCCTCTGTCC R: GCTGGCAGGAAGGAGGTATG
Setd1A 3' LoxP site	F: TCCAGTTCCTCCCCTCACTA R: TGCCAAACTGATGACCCGAG
SET1A conditional point mutant cassette	F: CGGTGATGTGATGTTCAGCC R: TTTCCCGCATGTCCGCTAC
SET1A conditional point mutant 3' LoxP sites	F: CTGGGCAGGGGTGAGGTG R: ACATTTACAGCTTCTTCTTCGCA
WDR82-dTAG	F: TGACTGTGGTGGGTGCAAGGG R: CGCAGTGAGAGCTGGAAAGGC

Table 2.4 Primers used to screen CRISPR targetings for homozygous insertion

2.1.6 Transient transfections

To express rTetR-fusion proteins, fragments comprising full SET1A and its truncated versions were Gibson assembled into a plasmid backbone containing a CAG promoter and either N- or C-terminal FLAG-StrepII tag, nuclear localisation signal and rTetR, as described previously (Blackledge *et al.*,

2014). Due to the smaller SET1A fragments giving leaky binding to the TetO, the rTetR was replaced with that from TetON-3G plasmid for all SET1A minimal domain fusions (pCAG-TetON-3G was a gift from Elena Cattaneo (Addgene plasmid # 96963; http://n2t.net/addgene:96963; RRID:Addgene_96963); Zhou *et al.*, 2006; Faedo *et al.*, 2017). ESCs were transfected at approximately 70% confluency with Lipofectamine 2000, according to the manufacturer's protocol (Invitrogen). For ChIP, cells were transfected in 10 cm dishes with 12 µg plasmid. For preparation of nuclear extract, cells were transfected in 6 well plates with 2.5 µg plasmid. For luciferase assays, cells were transfected in 96 well plates with 98 ng rTetR plasmid and 2 ng pRL *Renilla* Luciferase control reporter plasmid (Promega) to control for transfection efficiency. Each transfection was performed in six wells of a 96 well plate to obtain three technical replicates for both untreated and treated with doxycycline.

2.1.7 Luciferase reporter assays

Following overnight transfection of plasmids expressing rTetR-fusion proteins and *Renilla* luciferase, three wells for each transfection were treated with 1 µgml^{-1} doxycycline for 6 hr. The remaining three wells for each transfection were media changed. Luciferase reporter assays were performed with the Dual-Luciferase Reporter Assay System (Promega). In brief, cells were lysed in 20 µl 1x passive lysis buffer by shaking at room temperature (RT) for 20 min. 10 µl cell lysate was added to 50 µl Luciferase Assay Reagent II and Firefly luciferase measured using a 10 second measurement in a Luminometer. 50 µl Stop & Glo Reagent was added and *Renilla* luciferase measured using a 10 second measurement. Each Firefly reading was normalised to the respective *Renilla* reading. Technical replicates were averaged and normalised to the readings obtained in the absence of doxycycline.

2.1.8 Cell growth analysis

To generate growth curves, 3x10^5 cells were seeded into 2x wells of a 6 well plate for each time point to generate both a treated and untreated growth curve. Cells were counted using a Countess

Cell Counter at the relevant time points. Cells were labelled with 0.4 % trypan blue to stain dead cells and only alive cells were counted. All counts were performed in technical duplicate and the average count taken forward for growth analysis.

2.2 Protein Methods

2.2.1 Acid extraction of histones

Unless otherwise stated, all centrifugation steps involving protein samples were performed at 4°C. To prepare histone extracts, pelleted cells were washed in RSB (10 mM Tris HCl pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$), centrifuged at 240xg for 5 min and resuspended in RSB supplemented with 0.5% NP-40. Following incubation on ice for 10 min, cells were centrifuged at 500xg for 5 minutes. The nuclear pellet was resuspended in 5 mM $MgCl_2$, an equal volume of 0.8 M HCl added, and then incubated on ice for 20 min to extract histones. After centrifugation for 20 min at 18,000xg, the supernatant was taken and histones precipitated by adding TCA to 25% v/v and incubating on ice for 30 min. Histones were pelleted by centrifugation at 18,000xg for 15 min, and the pellet was washed twice in cold acetone. The histone pellet was resuspended by gentle vortexing in 1x SDS loading buffer (2% SDS, 100 mM Tris pH 6.8, 100 mM DTT, 10% glycerol, 0.1% bromophenol blue) and boiling at 95°C for 5 min. Any insoluble precipitate was pelleted by centrifugation at 18000xg for 15 min and the soluble fraction taken as the histone extract. Histone concentrations were compared across samples by SDS-polyacrylamide gel electrophoresis (SDS-PAGE) and Coomassie Blue staining.

2.2.2 Preparation of nuclear extract

Cell pellets were resuspended in 10 cell volumes of buffer A (10 mM Hepes pH 7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, 0.5 mM PMSF, 1x cOmplete protease inhibitor cocktail (Roche)) and incubated for 10 min on ice. After centrifugation at 500xg for 5 min, the cell pellet was resuspended in 3 cell volumes of buffer A supplemented with 0.1% NP-40 and incubated on ice for 10 min. Nuclei

were pelleted at 1500xg for 5 min. Where necessary, the supernatant was taken as the cytoplasmic extract. Nuclei were resuspended in 1 cell volume of buffer C (250 mM NaCl, 5 mM Hepes pH 7.9, 26% glycerol, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 0.5 mM DTT, 1x protease inhibitor cocktail). The volume of the nuclear suspension was measured and the NaCl concentration increased to 400 mM by dropwise addition of 5 M NaCl, assuming a 250 mM starting concentration. Nuclei were incubated at 4°C for 1 hr with gentle inversion to extract nuclear proteins. After centrifugation at 18,000xg for 20 min, nuclear proteins were recovered in the supernatant. Protein concentration was determined by Bradford assay (BioRad).

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

2.2.3 Immunoprecipitation

Immunoprecipitations (IPs) were performed using 550 µg of nuclear extract. Extracts were diluted to 500 µl with BC150 (150 mM KCl, 10% glycerol, 50 mM HEPES pH 7.9, 0.5 mM EDTA, supplemented with 1x protease inhibitor cocktail), with 250 units Benzonase Nuclease (Sigma). Protein A beads (Repligen) were blocked in BC150 supplemented with 1% Fish gelatin and 0.2 $mgml^{-1}$ BSA for 1 hr at 4°C. Extracts were pre-cleared with 50 µl slurry blocked beads for 1 hr at 4°C and then incubated rotating with antibody against the protein of interest overnight at 4°C. 50 µl slurry blocked Protein A beads were used to precipitate antibody-bound protein at 4°C for 3 hr. Beads were pelleted at 1000xg, and washed 3 times with BC150 + 0.02% NP-40, with one final wash in BC150. For subsequent western blotting, beads were resuspended in 2x SDS loading buffer, boiled at 95°C for 5 min, and the supernatant taken as the immunoprecipitate. An appropriate amount of nuclear extract was taken as the input sample and inputs were incubated in 1x SDS loading buffer at 95°C for 5 min.

2.2.4 SDS-PAGE

Where necessary, protein extracts were mixed with SDS loading buffer to 1x and boiled at 95°C for 5 min. Gels (0.1% SDS, 0.1% ammonium persulphate (Sigma), 0.1% TEMED (Sigma), 400 mM Tris HCl pH 8.8) were cast using the Mini-Protean Tetra Cell system (BioRad). Depending on the size of the protein of interest, resolving gel between 6% and 15% acrylamide/bis-acrylamide (BioRad) was used. Stacking gel contained 5% acrylamide/bis-acrylamide and 125 mM Tris HCl pH 6.8. Gels were run at 200 V in 1x SDS-PAGE running buffer (25 mM Tris, 192 mM glycine, 0.1% SDS). Alternatively, when analysing proteins running >180 kDa, 3-8% pre-cast NuPAGE Tris-Acetate gels were used (Invitrogen). Proteins were either visualised with Coomassie Blue staining or transferred to nitrocellulose membranes for western blotting.

2.2.5 Western blotting

Protein extracts were resolved using SDS-PAGE. Typically, proteins were transferred to nitrocellulose membrane by semi-dry transfer using the Trans-Blot Turbo Transfer System (BioRad). Transfer was performed as per the manufacturer's guidelines, depending on the size of the proteins being transferred. All blocking, antibody incubations and washing steps were performed in 1x PBS with 0.1% Tween 20 (Fisher) and 5% milk. Membranes were blocked for 30 min at RT and primary antibody incubations were then carried out overnight at 4°C (See Table 2.5 for antibody dilutions). Membranes were washed for 3x 10 min at RT, and then incubated with infrared-dye conjugated secondary antibody (1/15000) for 1 hr. After 3x 10 min washes and two PBS rinses, membranes were imaged using an Odyssey Fc system (LI-COR).

For IP samples where the protein of interest was smaller than 50 kDa, HRP (horseradish peroxidase) -conjugated VeriBlot secondary antibodies (Abcam) were used to avoid cross-reactivity with denatured IgG. Membranes were then washed 3 x 10 min and developed by chemiluminescence (Solution 1 – 100 mM Tris HCl, pH 8.5, 2.5 mM Luminal in DMSO, 396 µM p-Coumaric acid in DMSO;

Solution 2 – 100 mM Tris HCl, pH 8.5, 5.6 mM H_20_2). Solutions were mixed in a 1:1 ratio, applied to membranes for 1 min and membranes were then exposed to X-ray film (GE Healthcare).

For SET1B western blots, proteins were transferred to 0.45 µm nitrocellulose membrane by wet transfer. Transfer was performed in 1x wet transfer buffer (25 mM Tris, 192 mM glycine, 10% Methanol, 1% SDS) at 100 V for 90 min at 4°C. After blocking at RT and incubating with primary antibody overnight at 4°C, membranes were washed 3x 10 min and incubated with HRP-conjugated secondary antibodies at RT for 1 hr. Membranes were washed 3 x 10 min and developed by chemiluminescence, as described above.

2.2.6 Antibodies

Antibodies used in this study are listed in table 2.5.

Antigen	**Source**	**Dilution for western blotting**	**Amount used for IP/ChIP**
SET1A	Bethyl (A300-289A)	1:500	-
SET1A	Klose Lab	-	20 µl/IP
T7	Cell Signalling (D9E1X)	-	10 µl/ChIP
SUZ12	Cell Signalling (3737)	1:1000	-
WDR5	Cell Signalling (D9E1I)	1:1000	-
RBBP5	Cell Signalling (D3I6P)	1:1000	-
ASH2L	Cell Signalling (D93F6)	1:1000	-
WDR82	Cell Signalling (D2I3B)	1:1000	-
CFP1	Klose Lab	1:2000	-
SET1B	Cell Signalling (D1U5D)	1:500	-
Vinculin	Sigma (V9131)	1:1000	-
H3K4me3	Klose Lab	1:2000	1.5 µl/ChIP
H3K4me2	Life Technologies (710796)	1:2000	-
H3K4me1	Cell Signalling (D1A9)	1:2000	-
H3	Cell Signalling (96C10)	1:2000	-

H4	Cell Signalling (L64C1)	1:2000	-
FLAG-StrepII	Klose Lab	1:500	-
FLAG	Sigma (F1804)	-	5 µl/IP
CDK8	Santa Cruz (SC1521)	1:1000	-
TBP	Abcam (ab818)	1:2000	-
RNAPII NTD (total)	Cell Signalling (D8L4Y)	1:1000	15 µl/ChIP
RNAPII S5P	Cell Signalling (D9N5I)	1:1000	12.5 µl/ChIP
RNAPII S2P	Cell Signalling (E1Z3G)	1:1000	12.5 µl/ChIP
RING1B	Klose Lab	1:1000	-
Anti-mouse secondary (IRDye680)	LI-COR 68070	1:15000	-
Anti-rabbit secondary (IRDye800)	LI-COR 32211	1:15000	-
Anti-goat secondary (IRDye680)	LI-COR 68074	1:15000	-
Anti-rabbit secondary (HRP)	GE Healthcare	1:2500	-
Anti-mouse secondary (HRP)	GE Healthcare	1:2500	-
VeriBlot IP Detection Reagent (HRP)	Abcam (ab131366)	1:500	-

Table 2.5 Antibodies used during this study

2.3 Calibrated chromatin immunoprecipitation (cChIP)

2.3.1 Native cChIP-sequencing

Unless otherwise stated, ChIP was performed on ice and all centrifugations performed at 4°C.

For Native histone cChIP, 5 x 10^7 ESCs were mixed with 2 x 10^7 *Drosophila* SG4 cells and nuclei were released by resuspending in RSB (10 mM Tris HCl pH 8, 10 mM NaCl, 3 mM $MgCl_2$) with 0.1% NP40. Nuclei were pelleted at 1500xg for 5 min and then washed and resuspended in 1 ml MNase

digestion buffer (RSB with 0.25 M Sucrose and 3 mM $CaCl_2$, 1x protease inhibitor cocktail). Each sample was incubated with 200 units of MNase (Fermentas) at 37°C for 5 minutes, with gentle inversion. Digestion was stopped by addition of 4 mM EDTA. Following centrifugation at 1500xg for 5 min, the supernatant (S1 fraction) was transferred to a protein LoBind tube (Eppendorf). The remaining pellet was resuspended in 300 µl nucleosome release buffer (10 mM Tris HCl pH 7.5, 10 mM NaCl, 0.2 mM EDTA, 1x protease inhibitor cocktail), rotated at 4°C for 1 hr and then passed through a 27G needle five times using 1 ml syringe. Following centrifugation at 1500xg for 5 min, the supernatant (S2) was combined with S1 fraction, aliquoted and snap frozen. A small amount of S1/S2 DNA was purified using Qiagen's PCR purification kit and 1 µg visualised on a 1% agarose gel to check efficiency of MNase digestion.

For each IP, 100 µl S1/S2 nucleosomes were diluted to 1 ml total volume in native ChIP incubation buffer (70 mM NaCl, 10 mM Tris HCl, pH 7.5, 2 mM $MgCl_2$, 2 mM EDTA, 0.1 % Triton-X100, 1x protease inhibitor cocktail) and immunoprecipitated with the appropriate amount of antibody overnight at 4°C (see table 2.5). IPs were all set up in duplicate for each sample. 100 µl diluted chromatin was also set aside as an input sample. Protein A agarose beads (Repligen) were blocked with 1 $mgml^{-1}$ BSA and 1 $mgml^{-1}$ yeast tRNA in native ChIP incubation buffer, overnight at 4°C. 40 µl slurry of pre-blocked agarose beads were used to capture antibody-bound nucleosomes at 4°C for 1 hr. Beads were then washed 4x with Native ChIP wash buffer (20 mM Tris HCl, pH 7.5, 2 mM EDTA, 125 mM NaCl, 0.1 % Triton-X100, 1x protease inhibitor cocktail) and 1x Tris-EDTA (TE) buffer (10 mM Tris pH 8.0, 1 mM EDTA, supplemented with 1x protease inhibitor cocktail). DNA was eluted by vortexing for 30 min in elution buffer (1% SDS and 0.1 M $NaHCO_3$) and DNA was purified using a ChIP DNA Clean & Concentrator kit (Zymo Research). For each ChIP, DNA from the matched input control (10% of the IP) was also purified. Purified DNA was analysed using ChIP-qPCR and cChIP-seq libraries for both ChIP and input samples were prepared using NEBNext Ultra II DNA Library Prep Kit for Illumina following the manufacturer's guidelines (NEB).

In order to examine H3K4me3 antibody specificity, 0.5 µl of the SNAP-ChIP K-MetStat Panel of differentially modified nucleosomes (Epicypher) were spiked into native chromatin and ChIP performed as described above. The amount of each modified species pulled down during the ChIP was quantified by using barcode-specific qPCR primers, as per the manufacturer's protocol and as described in Shah *et al.*, 2018.

2.3.2 Crosslinked cChIP-sequencing

For double cross-linked SET1A ChIP, 5 x 10^7 ESCs were with fixed with 2 mM DSG (disuccinimidyl glutarate, ThermoFisher) for 50 min at 25°C and then 1% formaldehyde (methanol-free, Life Technologies) for 10 min. Alternatively, for single cross-linked RNAPII ChIP, 5 x 10^7 ESCs were fixed with 1% formaldehyde for 10 min at 25°C. Fixation was quenched using glycine added to 125 mM. Cells were then pelleted at 1000xg for 5 min and washed with PBS. Crosslinked ESCs were mixed with 1 x 10^5 crosslinked HEK293T/T7-SCC1 cells (1% formaldehyde, 15 min for SET1A ChIP; from Martin Houlard, Nasmyth lab) or 2 x 10^6 crosslinked HEK293T cells (1% formaldehyde, 10 min for RNAPII ChIP). Chromatin was prepared by incubation in 1 ml FA-lysis buffer (50 mM HEPES pH 7.9, 150 mM NaCl, 1 mM EDTA (2 mM EDTA for RNAPII ChIP), 0.5 mM EGTA, 0.5% NP-40, 0.1% Na-deoxycholate, 0.1% SDS, 1x protease inhibitor cocktail, 1 mM AEBSF, 10 mM NaF was also added fresh for RNAPII ChIP) on ice for 10 min. Chromatin was sonicated using a BioRuptor Pico sonicator (Diagenode) at 4°C. Sonication was performed using 23-30 cycles of 30s on/30s off at full power, shearing genomic DNA to an average size of 0.5 kb. The sonicated material was pelleted at 20,000xg for 20 min, and the supernatant taken as sonicated chromatin. A ten-fold dilution of chromatin in 0.1 M NaOH was quantified by NanoDrop. 50 µl chromatin was reverse crosslinked overnight at 65°C with 500 mM NaCl and 3 µl RNase A (Sigma). The following morning, 20 µg proteinase K was added and chromatin incubated at 45°C for 1 hr. DNA was then purified using Qiagen's PCR purification kit and 1 µg run on a 1% agarose gel to check sonication efficiency.

300 µg chromatin was used per IP. Chromatin was diluted to 1 ml total volume per IP in FA lysis buffer (unless otherwise stated, 1x protease inhibitor cocktail, 1 mM AEBSF and 10 mM NaF for RNAPII ChIP were added to all buffers). An additional volume of diluted chromatin was taken to use as an input sample. Protein A agarose beads (Repligen) were blocked with 1 $mgml^{-1}$ BSA and 1 $mgml^{-1}$ yeast tRNA in 1x TE buffer at 4°C for 1 hr. Chromatin was pre-cleared with agarose beads (40 µl slurry beads per ChIP) at 4°C for 1 - 2 hr. The input sample was taken from the pre-cleared chromatin, and the remainder was immunoprecipitated overnight at 4°C with the appropriate amount of antibody (see Table 2.5). Antibody-bound chromatin was precipitated for 3 hr at 4°C using 40 µl slurry of blocked Protein A agarose beads. Washes were carried out for 5 min each at 4°C, using FA lysis buffer, FA lysis buffer with 500 mM NaCl, 1x DOC buffer (10 mM Tris HCl, pH 8, 250 mM LiCl, 1 mM EDTA (2 mM EDTA for RNAPII ChIP), 0.5% NP-40, 0.5% Na-deoxycholate), and 2 washes with TE buffer. DNA was eluted by vortexing for 30 min in elution buffer (1% SDS and 0.1 M $NaHCO_3$). Cross-links were reversed for ChIPs and inputs at 65°C overnight with 200 mM NaCl and 2 µl RNase A (Sigma). Samples were then incubated with 20 µg Proteinase K for 1 hr at 45°C. DNA for ChIPs and inputs was purified using a ChIP DNA Clean & Concentrator kit (Zymo Research). Purified DNA was analysed using ChIP-qPCR and cChIP-seq libraries for both ChIP and input samples were prepared using NEBNext Ultra II DNA Library Prep Kit for Illumina following manufacturer's guidelines (NEB).

2.4 RNA methods

2.4.1 RNA extraction, reverse transcription and quantitative PCR

RNA was extracted using TRIzol reagent (Invitrogen), according to the manufacturer's guidelines. 400 ng purified RNA was reverse transcribed using the imProm-II system (Promega) using random primers and RNasin ribonuclease inhibitor. A control lacking reverse transcriptase was included for each experiment. cDNA was diluted to 200 µl in nuclease-free water and analysed by quantitative

real-time PCR (qPCR) using SensiMix SYBR Green mastermix (Bioline). Each qPCR reaction was run in technical duplicate.

qPCR primers were designed using Primer-BLAST, limiting the product size range to 80-120 bp, melting temperatures to 59-61°C and temperature difference to 1.0°C (Ye *et al.*, 2012). cDNA primers were designed to span large introns, so that cDNA and gDNA products could be clearly distinguished. Primer pairs were chosen that had limited off-target binding and their efficiency was tested using a serial dilution curve of DNA.

2.4.2 Calibrated total RNA-sequencing (cRNA-seq)

ESCs were counted and mixed with 1/4 the number of SG4 cells and 4:5 of the cell mixture taken to extract RNA using TRIzol reagent, according to the manufacturer's protocol (Invitrogen). RNA was quantified with the Qubit RNA BR assay kit (Life technologies) and 9 µg purified RNA was resuspended in 100 µl nuclease-free water and treated with the TURBO DNA-free kit, using 6 units TURBO DNase (Thermo Scientific). RNA was quantified using the Qubit RNA HS assay kit (Life technologies) and 900 ng RNA was depleted of rRNA using the NEBNext rRNA depletion kit (NEB). RNA was then checked for efficient rRNA depletion on an RNA Pico Bioanalyser chip (Agilent). RNA-seq libraries were prepared from an equal amount of ribo-depleted RNA using the NEBNext Ultra II directional RNA library prep kit (NEB), including 2.5-4 min fragmentation at 94°C.

1:5 of the mouse:fly cell mixture was taken to extract genomic DNA to sequence as a matched input. gDNA was extracted using the Zymo Quick DNA miniprep kit (Zymo Research) and A260:A280 recorded by Nanodrop. If necessary, gDNA was further cleaned up using the Genomic DNA Clean and Concentrator kit (Zymo Research). 10-90 ng genomic DNA was used to make libraries with the NEBNext Ultra II FS DNA library prep kit, as per the manufacturer's instructions (NEB).

2.4.3 Transient-Transcriptome (TT)-sequencing

To enable quantification of the TTchem-seq protocol, 40 million ESCs were seeded into 2 x 15cm dishes. 4 hr later, 100 nM dTAG-13 was added to 1 x dish for 105 minutes, while the second dish was media changed. Media was replaced on both dishes to pre-warmed media containing 500 µM 4-thiouridine (4sU, Glentham Life Sciences), with or without 100 nM dTAG-13. Cells were incubated with 4sU for 15 min to label newly-transcribed RNA. Cells were harvested directly into TRIzol reagent to immediately stop 4sU-labelling. 4sU labelled ESCs were snap-frozen in TRIzol reagent. The number of harvested cells was assumed to be equal between the untreated and dTAG-13-treated plates.

100 million *Drosophila* SG4 cells were seeded into a 15 cm dish. Once attached, SG4 cells were treated with 500 µM 4sU for 15 min, harvested into TRIzol Reagent and snap-frozen in 10 million aliquots.

To perform TT-seq I utilised a recently developed protocol, termed TTchem-seq, which relies on chemical fragmentation of the RNA. TTchem-seq was performed as previously described, with minor modifications (Gregersen, Mitter and Svejstrup, 2020). In brief, the TRIzol containing 4sU-labelled ESCs and SG4 cells was thawed and the equivalent of 10 million 4sU-labelled SG4 cells were mixed with 40 million 4sU-labelled ESCs. RNA was extracted using the Direct-zol DNA/RNA Miniprep kit (Zymo Research) as per the manufacturer's protocol and treated with the TURBO DNA-free kit, using 6 units TURBO DNase. The RNA concentration was recorded with the Qubit RNA BR assay kit and 60 µg taken into 100 µl nuclease-free water and fragmented on ice with 20 µl 1 M NaOH for 20 min. Fragmentation was stopped with 80 µl 1 M Tris, pH 6.8 and the RNA was cleaned up with Micro Bio-Spin P-30 gel columns (Biorad). RNA was biotin-labelled with 50 µl 0.1 $mgml^{-1}$ MTSEA biotin-XX linker (Biotium) with 3 µl biotin buffer (833 mM Tris HCl, pH 7.4, 83.3 mM EDTA) for 30 min at RT. Biotin-labelled RNA was purified with Phenol/Chloroform/Isoamyl alcohol (Invitrogen). Streptavidin pull down was performed with the µMACS Streptavidin Kit (Miltenyi

Biotec), washing the columns three times with pre-warmed pull-down wash buffer (100 mM Tris HCl, pH 7.4, 10 mM EDTA, 1 M NaCl and 0.1% Tween 20) and 3x RT pull down wash buffer. Biotin-labelled RNA was eluted with 100 µl elution buffer (100 mM DTT in nuclease-free water) and cleaned up with the RNeasy MinElute Cleanup kit (QIAGEN), adjusting the amount of ethanol to capture RNA < 200 nucleotides in length. RNA was quantified using the Qubit RNA HS assay kit and RNA libraries were prepared from 20 ng RNA with the Ultra II directional RNA library prep kit, as per the manufacturer's guidelines for rRNA depleted and FFPE RNA (NEB).

2.5 Massively parallel sequencing and data processing

2.5.1 Massively parallel sequencing

All sequencing experiments were carried out in at least biological triplicate, unless otherwise stated. Sequencing samples were indexed using NEBNext Multiplex Oligos (NEB). The average size and concentration of all sequencing libraries was analysed using a Bioanalyser High Sensitivity DNA Kit (Agilent) followed by qPCR using SensiMix SYBR Green mastermix (Bioline) and KAPA Illumina DNA standards (Roche). Libraries were sequenced using a NextSeq500 (Illumina). ChIP and TT-seq libraries were sequenced with 40 bp paired-end reads and RNA-sequencing libraries with 80 bp paired-end reads, with the help of Amanda Williams (Department of Zoology, University of Oxford).

2.5.2 Read alignment

The alignment pipelines used in this study were developed by Hamish King and Nadya Fursova. For cChIP-seq, paired-end reads were aligned to the concatenated mouse mm10 and spike-in genomes (mm10+dm6 for Native cChIP and mm10+hg19 for cross-linked cChIP) using Bowtie 2 with the '-no-mixed' and '-no-discordant' options (Langmead and Salzberg, 2012). Reads that mapped more than once were discarded and PCR duplicates were removed using Sambamba (Tarasov *et al.*, 2015).

For cRNA-seq and cTT-seq, reads that aligned to the mm10 and dm6 rDNA genomic sequences (GenBank: BK000964.3 and M21017.1) were first identified using Bowtie 2 with '-very-fast', '-no-mixed' and '-no-discordant' options) and discarded (Langmead and Salzberg, 2012). Unmapped reads were then aligned to the concatenated mm10 and dm6 genomes using STAR (Dobin *et al.*, 2013). To improve mapping of intronic sequences, reads that failed to map using STAR were aligned using Bowtie 2 with '-sensitive-local', '-no-mixed' and '-no-discordant' options (Langmead and Salzberg, 2012). Uniquely aligned reads from the last two steps were combined for further analysis and PCR duplicates were removed using Sambamba (Tarasov *et al.*, 2015). For the corresponding genomic DNA-seq input experiments, paired-end read alignment and processing was carried out as described above for cChIP-seq.

2.5.3 Data normalisation

Sequencing datasets were calibrated to the spike-in *Drosophila* or human genomes, as described previously (Fursova *et al.*, 2019; Turberfield *et al.*, 2019; Blackledge *et al.*, 2020). Uniquely aligned mm10 reads were randomly subsampled using factors that reflected the total number of dm6 or hg19 reads in each sample. For cRNA-seq and cChIP-seq, the downsampling factors were additionally corrected using the ratio of dm6 or hg19/mm10 total read counts in corresponding input samples to account for any small variation in spike-in cell mixing. For cRNA-seq, corresponding genomic DNA-seq samples were used as inputs.

After normalisation, read coverages for individual biological replicates were compared across regions of interest using the multiBamSummary and plotCorrelation functions from deepTools (Ramírez *et al.*, 2014). Biological replicates correlated well with each other (Pearson correlation coefficient>0.9) and were merged for subsequent analysis. Genome coverage tracks were generated using the pileup function from MACS2 for cChIP-seq and genomeCoverageBed from BEDTools for cRNA-seq and visualised using the UCSC genome browser (Kent *et al.*, 2002; Zhang *et al.*, 2008; Quinlan, 2014; Casper *et al.*, 2018). Differential genome coverage tracks (log2 fold change

of two conditions) were obtained using the bigwigCompare function from deepTools (Ramírez *et al.*, 2014).

2.5.4 Peak calling and annotation

H3K4me3 peak sets were generated from each ChIP replicate using MACS2 ('BAMPE' and 'broad' options specified), with merged input samples from each biological replicate used for background normalisation (Zhang *et al.*, 2008). A high stringency H3K4me3 peak set comprising peaks called in three biological replicates in untreated samples in both dTAG-SET1A and dTAG-SET1A/B cells were used for further analysis. TSS-associated H3K4me3 peaks were identified using BEDTools Intersect (Quinlan, 2014), only carrying forward peaks which overlapped directly with a TSS from a custom-built, non-redundant mm10 set (n=20633, Rose *et al.*, 2016; TSS-associated H3K4me3 peaks n=14546). For the H3K4me3 datasets, reads were quantified within peaks due to the high variability of peak widths.

T7-SET1A peak sets were generated from each biological ChIP replicate using MACS2 ("BAMPE" option specified), with merged input samples used for background normalisation. Only peaks present in each replicate of untreated cells were carried forward. TSS-associated T7-SET1A peaks were identified using BEDTools Intersect (Quinlan, 2014), only carrying forwards peaks which overlapped directly with a TSS (n=10480). Reads were quantified within peaks.

To annotate genes as CGI or non-CGI associated, non-methylated islands were identified using MACS2 peak calling from BioCAP-seq (Biotinylated CXXC Affinity Purification-seq) data using a matched input control (n=27047; Long *et al.*, 2013). CGI-associated genes were annotated using BEDTools Intersect (Quinlan, 2014) to identify genes with a TSS which was within 2.5 kb of a non-methylated island (n=14687). Remaining TSSs were defined as 'non-CGI-associated' (n=5946).

For the purposes of RNAPII analysis gene promoters were defined as TSS ± 0.5 kb in order to reduce any contribution from gene body signal. Gene bodies were defined as TSS to TES (transcription end site) and the termination window as TES + 3 kb.

2.5.5 Differential gene expression analysis

Differential gene expression analysis was performed as described previously (Fursova *et al.*, 2019; Turberfield *et al.*, 2019; Blackledge *et al.*, 2020). In brief, mm10 read counts were obtained from individual biological replicates prior to spike-in normalisation using a SAMtools-based custom Perl script within a custom-built, non-redundant mm10 gene set (n=20633, Rose *et al.*, 2016). In contrast, dm6 reads were pre-normalised to the dm6/mm10 ratio in the matched input library and read counts obtained from a set of unique dm6 refGene genes. For cTT-seq, in which there was no matched input, dm6 counts were obtained from the raw, non-normalised reads. These dm6 counts were used to calculate normalisation Size Factors using the DESeq2 package (Love, Huber and Anders, 2014). These size factors were then applied to DESeq2 analysis of the mm10 read counts. A change was considered significant based on a threshold of p-adj < 0.05 and fold change > 1.5. For visualization purposes, DESeq2-normalized read counts were averaged across the replicates and used to calculate RPKM (Reads per kilobase per million).

2.5.6 Readcount quantification and enrichment analysis for cChIP-seq

For cChIP-seq, genomic regions of interest were annotated with read counts using BEDTools multicov (Quinlan, 2014). For comparative analysis of untreated cChIP-seq levels, read counts were obtained from spike-in normalised and merged biological replicates. Differential enrichment analysis was performed using DESeq2, as described above for cRNA-seq (Love, Huber and Anders, 2014). mm10 read counts from individual biological replicates prior to spike-in normalisation were obtained using relevant regions of interest: peaks for H3K4me3 and T7-SET1A and promoters, gene bodies or termination windows for RNAPII. Spike-in reads were pre-normalised using the dm6/mm10 ratio in the corresponding input sample and reads counted in appropriate control regions: dm6 TSS ± 2 kb of a custom unique set for H3K4me3 and 20 kb bins for RNAPII. Counting in bins was performed using multiBamSummary ('-outRawCounts') from deepTools (Ramírez *et al.*, 2014).

2.5.7 Box plots, correlation and statistical analysis

Boxplots and scatterplots were made using R and ggplot2. Boxes for the boxplots show interquartile range (IQR), centre line represents median, and the whiskers extend by 1.5x IQR. R was used to perform regression analyses using standard linear models and to calculate Pearson correlation coefficients. Correlation matrices were generated with the ggcor function from the GGally R package and scatterplots were coloured by density with stat_density2d. Student's T tests were also performed in R with samples considered to be independent and two-sided alternative hypothesis, unless otherwise stated.

2.5.8 Metaplots and Heatmaps

computeMatrix and plotProfile/plotHeatmap from deepTools were used to perform metaplot and heatmap analysis of read density at regions of interest (Ramírez *et al.*, 2014). Metaplot profiles represent the mean read density over a set of genomic regions. To ensure strandedness for cTT-seq, computeMatrix was used to obtain sense and antisense read density within required regions on each strand. Stranded matrices were then merged using computeMatrixOperations.

2.5.9 Published datasets used in this study

Published sequencing datasets were used in this study. These are listed in table 2.6. WDR82 ChIP-seq dataset (unpublished) was generated by myself and Chloe Tubman.

Study	**Accession code**	**Dataset(s)**
Long *et al.*, 2013	GSE43512	BioCAP-seq
Turberfield *et al.*, 2019	GSE126864	RNAPII ChIP-seq
Brown *et al*. 2017	GSE93538	H3K4me3, CFP1 and SET1A ChIP-seq
Zhang *et al*. 2019	GSE107406	MLL1 ChIP-seq
Denissov *et al.*, 2014	GSE52071	MLL2 ChIP-seq
Fursova et al. 2021	GSE161996	H3K27ac ChIP-seq and ATAC-seq

Dobrinić, Szczurek and Klose, 2021	GSE159400	RING1B and SUZ12 ChIP-seq

Table 2.6 Published datasets used in this study

3. SET1A/B have a widespread role in gene expression

3.1 Introduction

The mammalian SET1 proteins, SET1A and SET1B, are two highly homologous histone methyltransferases shown to trimethylate histone H3 at lysine 4 *in vitro* (Lee and Skalnik, 2005; Lee *et al.*, 2007; Lee and Skalnik, 2008; Wu *et al.*, 2008; Shinsky *et al.*, 2015; Kwon *et al.*, 2021). Genome-wide chromatin binding analysis has established that SET1A associates with CGI-associated gene promoters and preferentially occupies the promoters of more highly expressed genes (Brown *et al.*, 2017; Sze *et al.*, 2017, 2020). Similarly, the levels of promoter-associated H3K4me3 correlate with gene transcription and numerous H3K4me3 'reader' proteins identified so far have documented roles in regulating gene expression (Barski *et al.*, 2007; Guenther *et al.*, 2007; Mikkelsen *et al.*, 2007; Chapter 1.3). Together, these observations have led to the proposal that SET1A/B deposit H3K4me3 at gene promoters to enable the recruitment of H3K4me3-binding proteins that regulate transcription (Vermeulen and Timmers, 2010; Meeks and Shilatifard, 2017; Park, Kim and Kim, 2020; Cenik and Shilatifard, 2021). However, despite this view being widely held, there is very little previous work directly examining the role of SET1A and SET1B in the regulation of gene expression.

Previous attempts to examine the function of the SET1 complexes in mammalian cells have used RNAi-based depletion or Cre-LoxP-mediated gene knockout of *Setd1A/B* or, more commonly, the SET1A/B complex components, *Cfp1* and *Wdr82* (Wu *et al.*, 2008; Lee and Skalnik, 2008; Thomson *et al.*, 2010; Vethantham *et al.*, 2012; Bledau *et al.*, 2014; Austenaa *et al.*, 2015; Fang *et al.*, 2016; Brown *et al.*, 2017; Franks *et al.*, 2017; Wang *et al.*, 2017; Schmidt *et al.*, 2018; Yang *et al.*, 2020). Somewhat surprisingly, given the occupancy of SET1A at gene promoters, and the proposed role of H3K4me3 in gene expression, the effects on gene expression in these contexts have been subtle and lacked directionality. However, these experiments have inherent limitations. Firstly, the

removal of CFP1 or WDR82 does not result in the complete removal of SET1A/B from chromatin (Lee and Skalnik, 2008; Brown *et al.*, 2017). Additionally, RNAi fails to completely remove the target protein and the kinetics of protein loss in Cre-LoxP systems is a function of mRNA and protein turnover, resulting in protein depletion on a timescale of several days. Given that the systems that regulate gene expression function on far more rapid timescales, these experiments do not measure the primary effects of SET1 depletion. More recently, new systems have been developed that specifically degrade a protein of interest, rather than targeting its corresponding RNA or gene (Nishimura *et al.*, 2009; Nabet *et al.*, 2018). These systems typically induce complete protein degradation within hours. Therefore, to understand the function of SET1A/B in mammalian cells, I set out to use a degron system to induce their rapid depletion and directly examine the role of each protein in regulating gene expression.

3.2 Results

3.2.1 Generating a SET1A auxin-inducible degron line

In order to examine how SET1A and SET1B contribute to mammalian gene expression, I set out to develop cell lines in which I could rapidly remove both proteins. I chose to do this in ESCs due to their ease of genetic manipulation and the wealth of other information regarding gene expression and chromatin modifications that is available in this cell type. I focussed initially on SET1A as it is essential for ESC viability, whereas SET1B is dispensable (Bledau *et al.*, 2014; Fang *et al.*, 2016; Cao *et al.*, 2017). To enable the rapid depletion of SET1A, I initially employed the auxin-inducible degron (AID) system (Nishimura *et al.*, 2009). The AID system relies on tagging a protein of interest with the AID tag. Upon the addition of auxin, an exogenous E3 ubiquitin ligase, TIR1, targets AID-tagged proteins for proteasomal degradation. This results in rapid protein removal (Figure 3.1 A).

To create a SET1A AID cell line, I began with an ESC line in which the *Oryza Sativa* TIR1 E3 ubiquitin ligase (*os*TIR1) is stably expressed from the *Rosa26* locus (Rhodes *et al.*, 2020). I then used

CRISPR/Cas9 genome editing (Garneau *et al.*, 2010; Gasiunas *et al.*, 2012; Jinek *et al.*, 2012; Ran *et al.*, 2013) and homology directed repair (HDR) to add the mini AID tag (mAID; Kubota *et al.*, 2013), followed by a short glycine-serine linker, to both copies of the *Setd1A* gene (Figure 3.1 B) . I chose to tag SET1A at its N-terminus, since the C-terminus contains its catalytic SET domain and previous experiments have shown that tagging ySet1 at its C-terminus causes a reduction in catalytic activity (Roguev *et al.*, 2001; Krogan *et al.*, 2002). I identified clones homozygous for mAID-SET1A by genomic DNA PCR screening and confirmed successful tagging by western blotting (Figure 3.1 C). Unfortunately, in all homozygous clones tested, the levels of mAID-SET1A were reduced by approximately 50% compared to WT SET1A. This suggests that the mAID tag was affecting either expression of SET1A or SET1A protein stability.

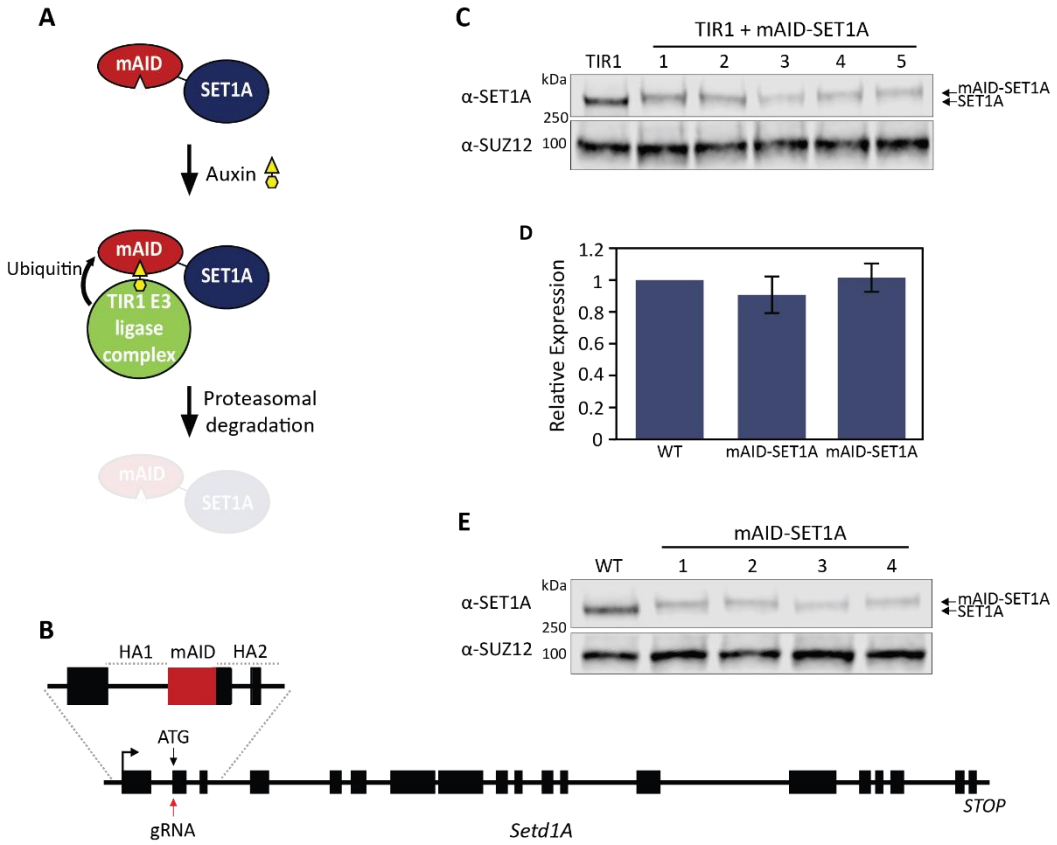


Figure 3.1 Generating a SET1A auxin-inducible degron line
(A) A schematic of the auxin inducible degron system applied to SET1A. Treatment of cells expressing TIR1 and mAID-SET1A with auxin results in the recruitment of the TIR1 ubiquitin ligase complex to mAID-SET1A. This leads to the polyubiquitination and proteasomal degradation of mAID-SET1A.
(B) A CRISPR/Cas9 strategy to endogenously tag SET1A with the mAID tag. Exons are indicated by black boxes and the black arrow depicts the transcription start site. The red arrow indicates the Cas9 cut-site,

defined by the specific gRNA used. Tag insertion was achieved using homology directed repair with a repair template containing the mAID tag and glycine-serine linker flanked by two homology arms.
(C) Western blot of nuclear extracts showing SET1A protein levels in five homozygous mAID-SET1A cell lines expressing TIR1. 'TIR1' represents an ESC line expressing TIR1 from the *Rosa26* locus, but with untagged SET1A. SUZ12 is a loading control.
(D) RT-qPCR for *Setd1A* in TIR1 cells and two clones expressing TIR1 and SET1A endogenously tagged with the mAID tag. Expression is normalised to the RNAPIII-transcribed tRNA-Lys and computed relative to levels in TIR1 cells. Error bars represent SEM for three biological replicates.
(E) Western blot of nuclear extracts showing expression of SET1A in four homozygous mAID-SET1A cell lines that do not express TIR1. WT represents wild type ESCs, which express untagged SET1A and do not express TIR1.

To determine whether the reduced levels of SET1A protein were a result of reduced expression, I examined *Setd1A* transcript levels by RT-qPCR (Figure 3.1 D). The levels of *Setd1A* transcript remained unchanged, suggesting that the reductions in SET1A were occurring at the protein level. I therefore wondered whether the location or nature of the tag might be interfering in some way with protein stability. I tried a number of different strategies to alleviate this problem, including using different linkers and tagging SET1A internally. However, the levels of mAID-SET1A were always reduced in comparison to WT SET1A (data not shown).

Another explanation for reduced mAID-SET1A levels could be 'leaky' degradation by *os*TIR1 in the absence of auxin, as has previously been observed (Kubota *et al.*, 2013; Natsume *et al.*, 2016). To examine this possibility, I tagged SET1A with the mAID tag in ESCs that do not express *os*TIR1. However, I still observed a reduction in the levels of mAID-SET1A in this cell line (Figure 3.1 E). It is therefore likely that tagging SET1A with the mAID tag makes it inherently unstable and subject to proteolysis.

3.2.2 A system to induce controlled and rapid degradation of SET1A

Given that the mAID tag was deleterious to SET1A levels, I decided to try an alternative degron system which uses a different degron tag. The dTAG system works in a similar way to the AID system, but takes advantage of a natively expressed E3 ubiquitin ligase, Cerebelon (CRBN; Nabet *et al.*, 2018). A modified FKBP12 tag ($FKBP12^{F36V}$; herein referred to as the dTAG) is endogenously

fused to the protein of interest. The addition of a bifunctional molecule (dTAG-13), which simultaneously binds to CRBN and the dTAG, recruits CRBN to the protein of interest, resulting in its ubiquitination and proteasomal degradation (Figure 3.2 A).

I used CRISPR/Cas9-mediated HDR to add the dTAG to the N-terminus of SET1A in WT ESCs. Importantly, the levels of dTAG-SET1A were comparable to those of WT SET1A, indicating that the dTAG does not affect SET1A expression or protein stability (Figure 3.2 B). To examine whether dTAG-SET1A incorporates normally into the SET1A complex, I immunoprecipitated SET1A from nuclear extract and examined the levels of WDR5, RBBP5, ASH2L, CFP1 and WDR82 (Figure 3.2 C). This showed that dTAG-SET1A interacts with the SET1A complex components similarly to WT SET1A. Therefore, the presence of the dTAG does not affect either expression of SET1A or SET1A complex formation.

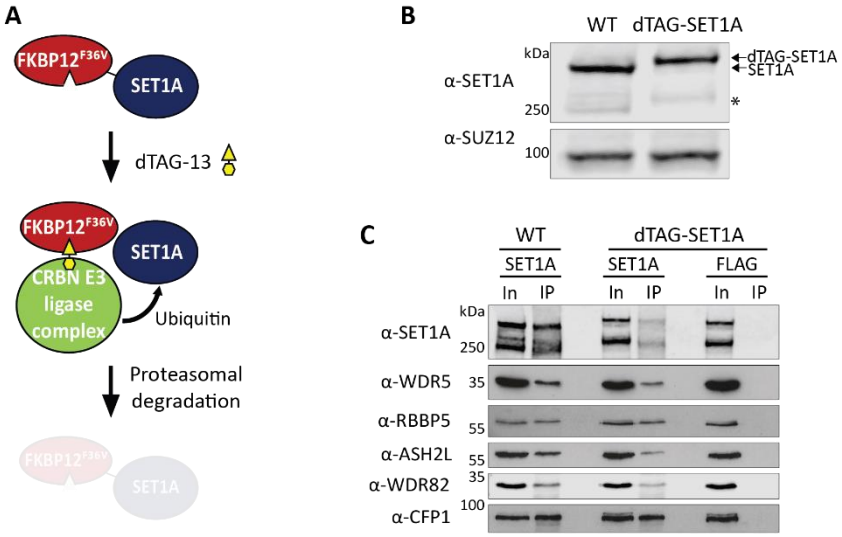


Figure 3.2 Applying the dTAG system to SET1A
(A) A schematic of the dTAG system applied to SET1A. Treatment of cells expressing FKBP12^{F36V}(dTAG)-SET1A with dTAG-13 brings the CRBN E3 ligase complex into close proximity to SET1A, resulting in its polyubiquitination and proteasomal degradation.
(B) Western blot of nuclear extracts showing expression of dTAG-SET1A in comparison to WT SET1A. SUZ12 is a loading control. * denotes a common SET1A degradation product.
(C) Immunoprecipitation of SET1A from nuclear extracts from WT and dTAG-SET1A ESCs. Levels of interacting WDR5, RBBP5, ASH2L, WDR82 and CFP1 are shown. 12% input was loaded ('In'). Nuclear extracts from dTAG-SET1A cells were also immunoprecipitated with an α-FLAG-tag antibody as a control.

To test whether dTAG-SET1A could be inducibly degraded, I treated cells with dTAG-13 over a 24 hour time course. This revealed that addition of dTAG-13 results in the complete depletion of dTAG-SET1A within just 2 hours (Figure 3.3 A). Since SET1A functions as part of a large multi-protein complex (Lee and Skalnik, 2005; Lee *et al.*, 2007; Lee and Skalnik, 2008; Wu *et al.*, 2008), I next wanted to understand whether the levels of other SET1A complex components were affected following proteasomal degradation of SET1A. The levels of WDR5, RBBP5, ASH2L and WDR82 remained largely unchanged following SET1A depletion, whereas the levels of CFP1 were significantly decreased (Figure 3.3 A). This is unsurprising since CFP1 is the only protein examined that is exclusively part of the SET1 complexes (van Nuland, Smits, *et al.*, 2013) and suggests that an inability to interact with SET1A may render CFP1 unstable.

Although western blot analysis suggested that dTAG-SET1A was completely degraded within 2 hours, I wanted to confirm that it was also completely removed from chromatin. I attempted to carry out SET1A ChIP using polyclonal antibodies raised against a fragment of the protein. However, this failed to yield reproducible ChIP signal. Previously, SET1A ChIP was successfully performed using a triple-T7 epitope tag knocked into the endogenous gene (Brown *et al.*, 2017). I therefore utilised a new cell line whereby SET1A was N-terminally tagged with both the T7 tag and dTAG (Courtesy of Anna Lastuvkova). The addition of the T7 tag did not affect the expression of dTAG-SET1A or the kinetics of proteasomal degradation (Figure 3.3 B and C). Using this cell line, I then performed calibrated ChIP-sequencing (cChIP-seq) using an antibody specific for the T7 epitope tag. To enable the ChIP signal to be quantified, I spiked the experiment with a small number of HEK293T cells expressing T7-tagged SCC1 (Martin Houlard, Nasymth Lab). Inspection of individual SET1A target genes, as well as plotting T7-dTAG-SET1A cChIP-seq signal over all TSSs, confirmed that SET1A was lost from chromatin following a 2 hour treatment with dTAG-13 (Figure 3.3 D and E).

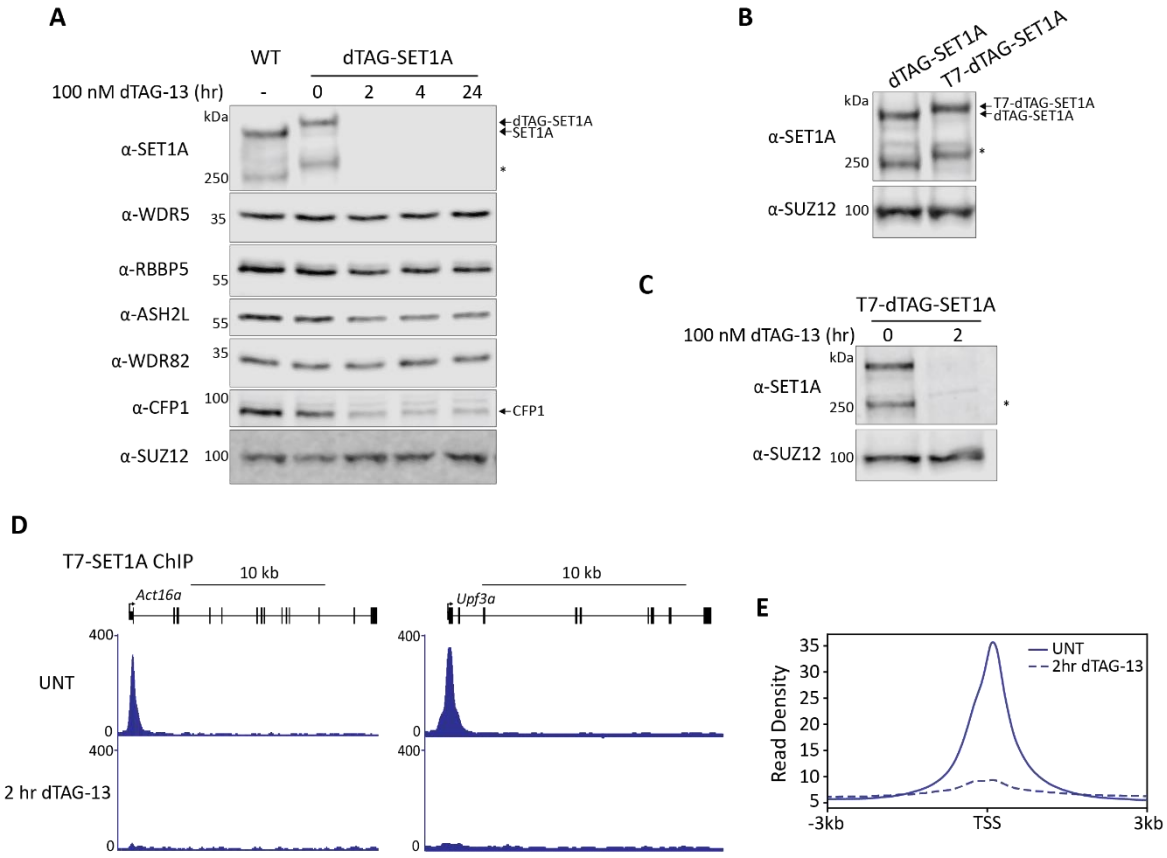


Figure 3.3 Addition of dTAG-13 results in the rapid depletion of SET1A

(A) Western blot of nuclear extracts from WT ESCs and dTAG-SET1A cells treated with dTAG-13 over a 24 hr time course. Levels of each of the SET1A complex components, besides DPY30, are shown. WT represents untargeted ESCs and SUZ12 is a loading control. * denotes a common SET1A degradation product.

(B) Western blot of nuclear extract showing levels of SET1A from untreated dTAG-SET1A and T7-dTAG-SET1A cells.

(C) Western blot of nuclear extract showing levels of SET1A from T7-dTAG-SET1A cells treated with dTAG-13 for 2 hr.

(D) Genomic snapshots showing T7-dTAG-SET1A cChIP-seq signal at two genes with a promoter-associated SET1A peak before (UNT) and after 2 hr treatment with dTAG-13.

(E) A metaplot of T7-dTAG-SET1A cChIP-seq signal in untreated cells (UNT) and cells treated with dTAG-13 for 2 hr. The metaplot is centred at the TSS of all genes (n=20633).

3.2.3 SET1A positively regulates gene expression genome-wide

Previous experiments perturbing the function of SET1A, or its orthologues in simpler eukaryotes, have suggested that SET1A has subtle, non-directional effects on gene expression, despite being broadly located at the promoters of actively transcribed genes (Santos-Rosa *et al.*, 2002;

Venkatasubrahmanyam *et al.*, 2007; Guillemette *et al.*, 2011; Lenstra *et al.*, 2011; Margaritis *et al.*, 2012; Fang *et al.*, 2016; Ramakrishnan *et al.*, 2016; Brown *et al.*, 2017; Sze *et al.*, 2017, 2020; Murray *et al.*, 2019). However, these gene expression changes could be confounded by the use of slow protein depletion systems that are unable to capture the most primary effects on gene expression. Having established a system with which I could rapidly remove SET1A, I set out to examine how its depletion affected gene expression in a time course experiment. I initially focussed on a subset of genes where SET1A has previously been proposed to contribute to gene expression (Brown *et al.*, 2017). Remarkably, after only 4 hours of SET1A depletion, I observed major effects on gene expression (Figure 3.4). On average, at the earliest time points of SET1A removal, there was a decrease in the expression of the SET1A target genes. Interestingly, this effect was not apparent at later time points, when expression had typically increased beyond initial levels. This experiment demonstrates the advantage of using rapid protein depletion systems to enable the primary effects of SET1A on gene expression to be examined.

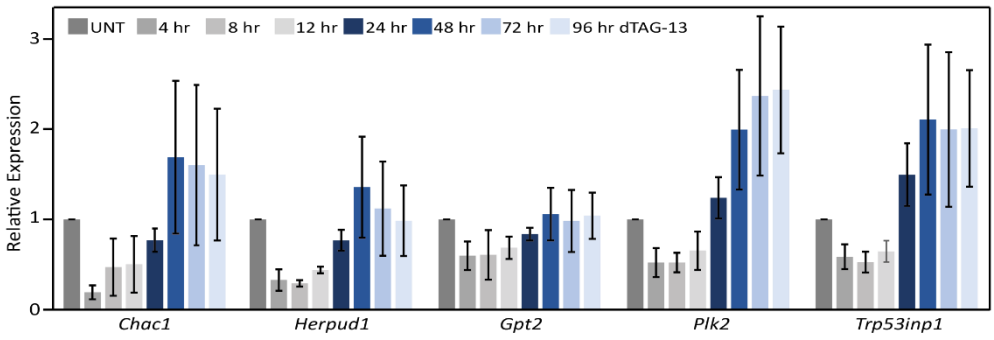


Figure 3.4 Loss of SET1A results in rapid changes in gene expression
RT-qPCR analysis to examine gene expression following a time course of SET1A removal. Expression levels of five SET1A target genes were normalised to the RNAPIII-transcribed *U6 snRNA* and are shown relative to levels in untreated cells. Error bars represent SEM of three biological replicates.

To examine the effect of SET1A depletion on gene expression throughout the genome, I next performed calibrated total RNA-sequencing (cRNA-seq). Using a calibrated sequencing approach allows gene expression changes to be quantified by using reads obtained from a spike-in genome as an internal normalisation point (Bonhoure *et al.*, 2014; Orlando *et al.*, 2014; Hu *et al.*, 2015). This approach is particularly important if the transcript levels of most genes were to change in one

direction, as such outcomes would be lost when using classical differential gene expression analyses normalising to total library size. Based on my gene-specific analysis, it was clear that the primary effects on gene expression manifested rapidly after SET1A removal. Therefore, I chose to examine gene expression after 2 hours of dTAG-13 treatment, a time point at which SET1A was completely depleted from chromatin (Figure 3.3). Interestingly, after only 2 hours of SET1A removal, substantial gene expression changes were observed (Figure 3.5 A). There was a large number of significantly decreased genes (2157), and a much smaller number of significantly increased genes (493). This suggests that SET1A typically functions to support gene expression, in agreement with its enrichment at actively transcribed genes (Brown *et al.*, 2017; Sze *et al.*, 2017, 2020). Importantly, I also performed cRNA-seq in WT ESCs treated with dTAG-13 for 2 hours as a control (Figure 3.5 B). There were no significant gene expression changes, confirming that the changes observed following SET1A depletion were due to SET1A loss.

In order to understand the effect of more prolonged culture without SET1A on gene expression, I also performed cRNA-seq at a later time point (24 hr) after SET1A removal (Figure 3.5 C). As at 2 hours, the expression of the majority of significantly changed genes was reduced following 24 hours of SET1A depletion. Interestingly, plotting the density of log2 fold change of all genes revealed that there was a subtle global decrease in gene expression following 24 hours of SET1A depletion. Therefore, the removal of SET1A has a widespread effect on gene expression in ESCs. However, there were fewer significantly changed genes at 24 hours than at 2 hours. In fact, over half of the significantly decreased genes at 2 hours were not significantly decreased at 24 hours (Figure 3.5 D). In contrast, there were 427 genes that were only significantly decreased at the 24 hr time point. This suggests that there is a subset of genes significantly differentially expressed only in the earliest time points of SET1A removal and a subset of genes that become differentially expressed only following longer-term culture without SET1A. It is possible that the expression changes observed following 24 hours of SET1A depletion are confounded by non-primary effects that arise due to SET1A loss.

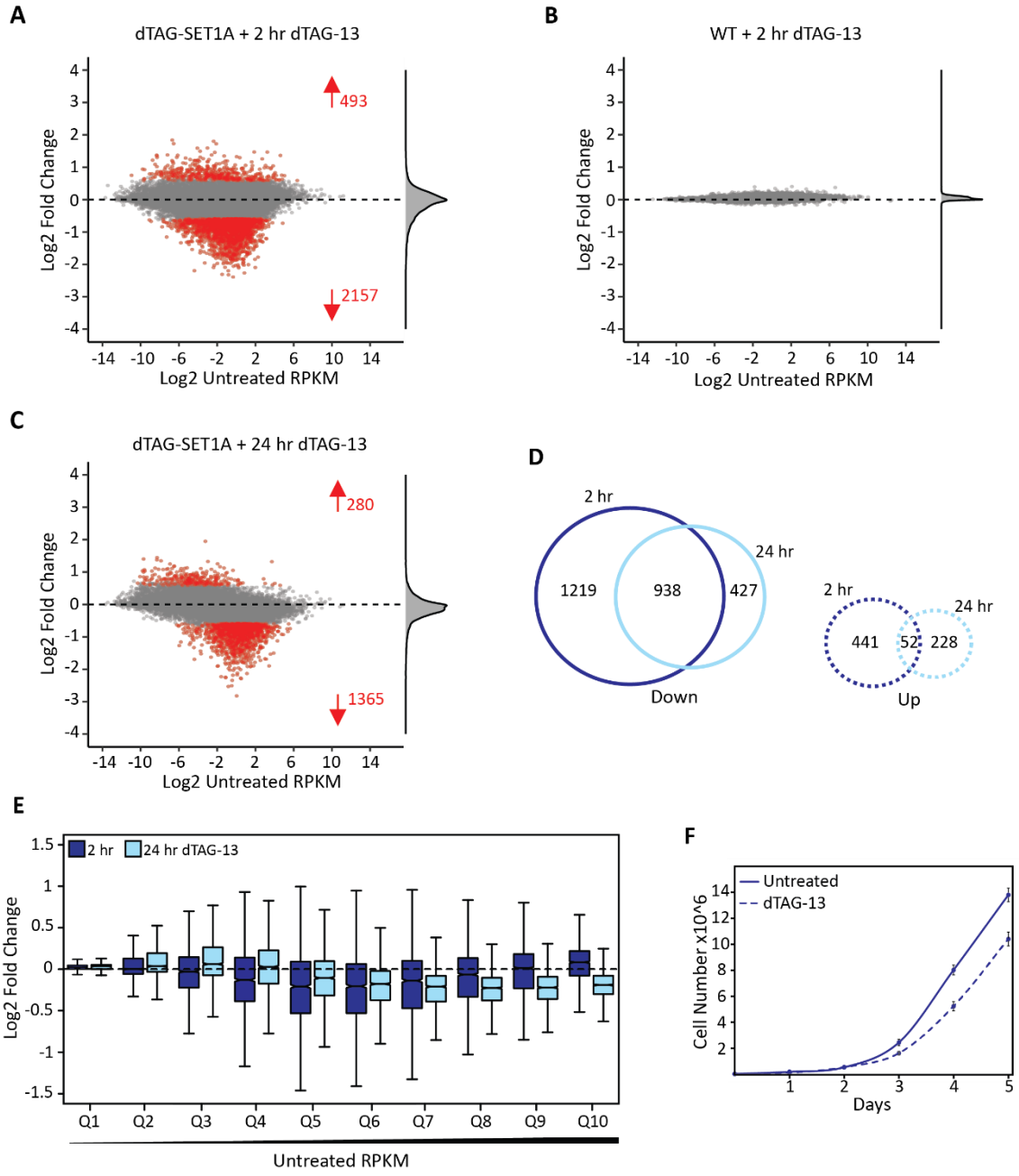


Figure 3.5 SET1A positively regulates gene expression genome-wide

(A) MA-plot of log2 fold change in gene expression (cRNA-seq) in dTAG-SET1A cells following 2 hr dTAG-13 treatment plotted against log2 mean RPKM for a non-redundant mm10 gene set (n=20633). Significantly changed (p-adj<0.05, fold change>1.5) genes are coloured red and the numbers of significant changes are shown. A density plot of log2 fold change in cRNA-seq signal is shown to the right of the plot.

(B) As in (A), but WT ESCs treated with dTAG-13 for 2 hr (two biological replicates).

(C) As in (A), but dTAG-SET1A cells treated with dTAG-13 for 24 hr.

(D) A Venn diagram showing the overlap of genes significantly changed after 2 hr and 24 hr SET1A depletion.

(E) Boxplots of log2 fold change in gene expression of all genes (n=20633) following 2 hr and 24 hr SET1A

depletion, split into deciles of equal size based on gene expression level in untreated cells (Untreated RPKM).
(F) Growth curve for untreated and dTAG-13-treated dTAG-SET1A cells cultured for up to 5 days. Error bars represent SEM of three biological replicates.

In order to understand the differences between the gene expression changes at the two time points, I divided all genes into deciles of equal size based on their initial expression level and compared the fold changes in gene expression of the genes in each decile (Figure 3.5 E). On average, after 2 hours of SET1A depletion, genes with moderate initial expression were reduced. In contrast, after 24 hours the decreased genes were more highly expressed in untreated cells. This is in keeping with the localisation of SET1A to the promoters of actively transcribed genes and indicates that moderately expressed genes are more rapidly affected by SET1A depletion, whereas prolonged culture without SET1A results in decreased expression of more highly expressed genes. It is possible that the reason for this difference is that 2 hours is too short a time for turnover of the RNA of more highly expressed genes, given that the median half-life of mouse transcripts is 4 hours (Herzog *et al.*, 2017). Interestingly, the expression of lowly expressed genes was typically slightly increased at 24 hours compared to 2 hours. This could indicate that there was a general recovery in gene expression levels of more lowly expressed genes following longer-term depletion of SET1A. This again highlights the importance of examining the acute effects of SET1A removal. Nonetheless, it is clear that SET1A has a widespread role in supporting active gene expression in ESCs. Consistent with the major changes in gene expression following SET1A removal, cells grew more slowly in the absence of SET1A (Figure 3.5 F). This is in agreement with previous observations that SET1A is required for normal ESC growth and viability (Bledau *et al.*, 2014; Fang *et al.*, 2016; Cao *et al.*, 2017).

3.2.4 Genes regulated by SET1A are actively transcribed and associated with CpG islands

Having established that SET1A has a widespread role in regulating gene expression, I next wanted to examine promoter-associated chromatin features at the genes that are regulated by SET1A

(Figure 3.6). As would be predicted, at both time points the promoters of the genes with significantly decreased expression were enriched for SET1A. Additionally, in agreement with the positive contribution of SET1A to gene expression, such genes were also enriched for chromatin features typically associated with active gene expression, including H3K4me3, H3K27ac and increased accessibility (ATAC-seq; Haberle and Stark, 2018). In contrast, genes that significantly increased in expression following SET1A removal were typically depleted of SET1A and chromatin features associated with active gene expression and enriched for components of the polycomb repressive complexes (RING1B and SUZ12; Blackledge, Rose and Klose, 2015). Therefore, the expression of actively transcribed genes that are enriched for SET1A was typically decreased after SET1A depletion, while the expression of a small number of polycomb target genes was increased.

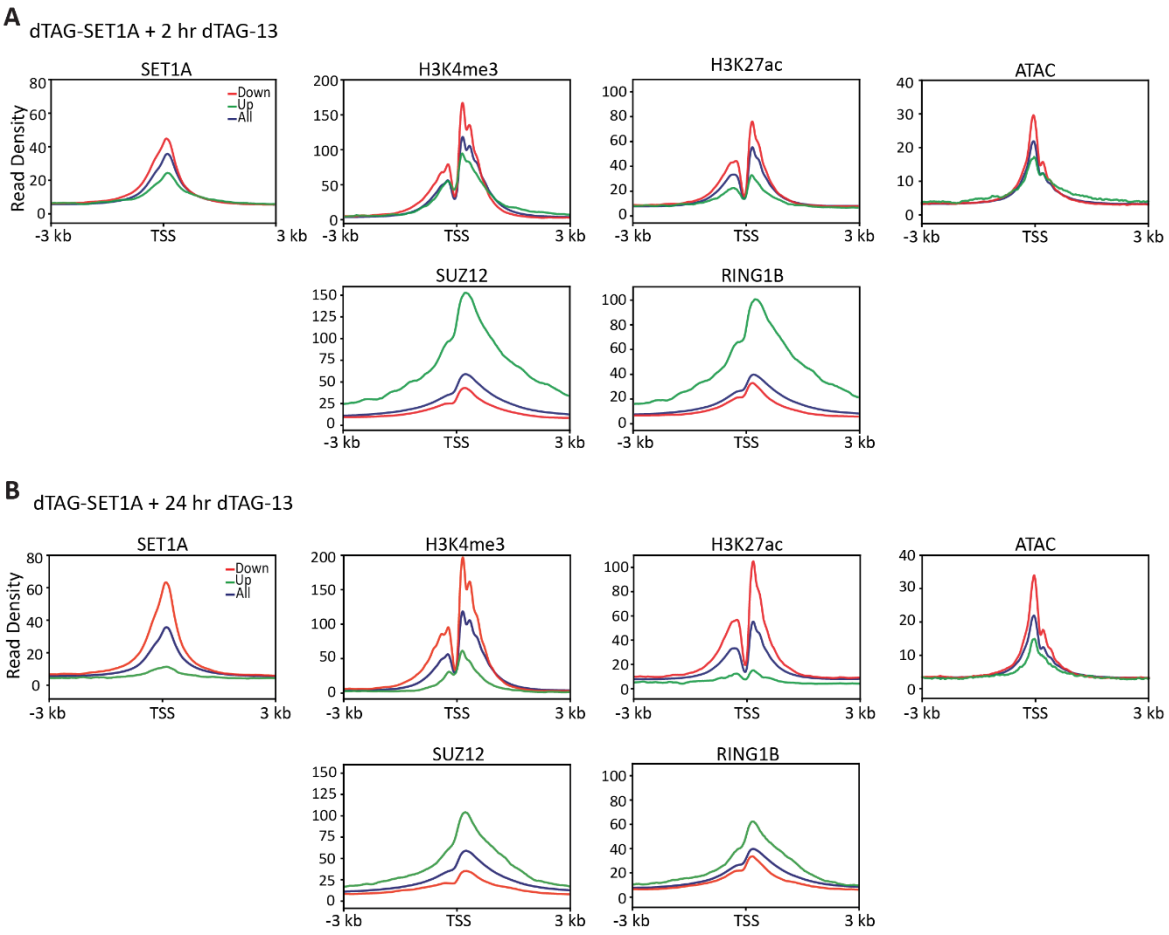


Figure 3.6 SET1A targets are enriched for chromatin features associated with active transcription
Metaplots of SET1A, H3K4me3, H3K27ac, SUZ12 and RING1B ChIP-seq and ATAC-seq signal at the TSSs of all genes (n=20633) or genes significantly increased ('Up') or decreased ('Down') at 2 hr ((A), Up

n=493, Down n=2157) or 24 hr ((B), Up n=280, Down n=1365) SET1A depletion (Datasets: Brown *et al.*, 2017; Dobrinić, Szczurek and Klose, 2021; Fursova *et al.*, 2020).

CpG islands are present at ~65% of gene promoters in ESCs and SET1A preferentially binds to CGI-associated gene promoters via CFP1 (Thomson *et al.*, 2010; Denissov *et al.*, 2014; Brown *et al.*, 2017). Therefore, I wanted to examine whether SET1A preferentially regulates the expression of CGI-associated genes. Indeed, while the expression of CGI-associated genes was reduced on average following the depletion of SET1A, the expression of non-CGI-associated genes typically remained unchanged (Figure 3.7 A). Additionally, genes that displayed significant reductions in expression following SET1A depletion were enriched for CGI-associated genes (Figure 3.7 B). Interestingly, while the genes with increased expression following 2 hours SET1A depletion were enriched for CGI-associated genes, those genes with increased expression following 24 hours SET1A depletion were typically depleted of CGI-associated genes. This again suggests that the gene expression outcomes at 24 hours may not represent the most primary effects of SET1A depletion. Nonetheless, it is clear that SET1A preferentially supports the expression of CGI-associated genes, an observation likely explained by the close relationship between SET1A occupancy and the presence of a CGI.

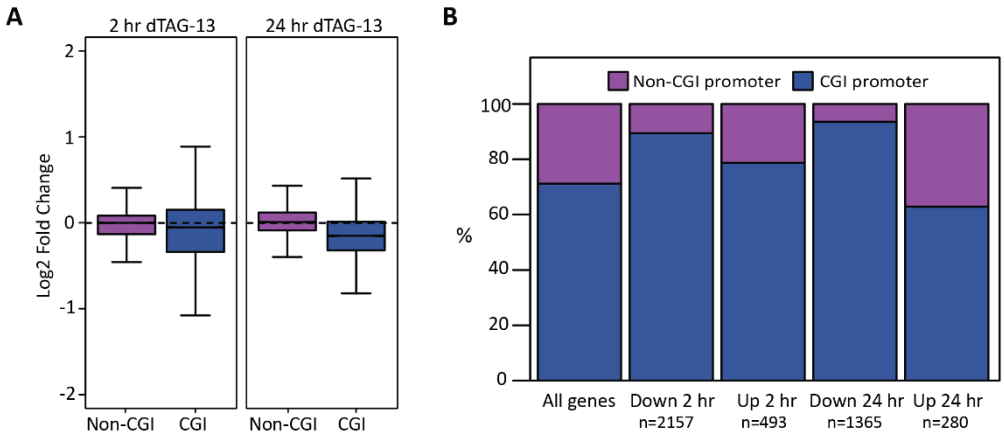


Figure 3.7 SET1A preferentially regulates the expression of CGI-associated genes
(A) Box plots of log2 fold change in gene expression (cRNA-seq) of CGI-associated (n=14687) and non-CGI-associated (n=5946) genes following 2 hr and 24 hr dTAG-13 treatment in dTAG-SET1A cells.
(B) Bar chart showing the percentage of genes that have a CGI-associated gene promoter for all genes (n=20633) and genes that were significantly changed at 2 hr and 24 hr SET1A depletion.

3.2.5 SET1B regulates the expression of a small number of genes

SET1B is a paralogue of SET1A and is present in cells in a similar multi-subunit complex (Lee and Skalnik, 2005; Lee *et al.*, 2007; Lee and Skalnik, 2008; Wu *et al.*, 2008). SET1B is expressed in ESCs and, although it has previously been proposed to reside predominantly in the cytoplasm (Wang *et al.*, 2017; Sze *et al.*, 2020), I found that SET1B is localised to the nucleus (Figure A1). Having established that SET1A has a widespread role in regulating gene expression in ESCs, I wanted to examine whether SET1B has a similar function. To achieve this, I used the same strategy as for SET1A to generate a dTAG-SET1B cell line (Figure 3.8 A). dTAG-SET1B was expressed at similar levels to WT SET1B and the addition of dTAG-13 induced rapid dTAG-SET1B degradation (Figure 3.8 B).

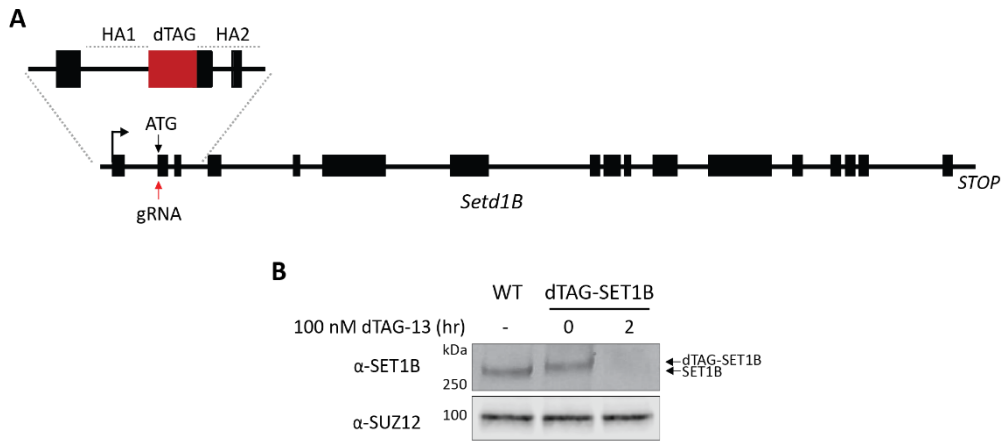


Figure 3.8 Generating a dTAG-SET1B cell line
(A) CRISPR-Cas9 strategy to endogenously tag SET1B with the dTAG. Exons are shown as black boxes and the black arrow depicts the transcription start site. The red arrow indicates the Cas9 cut-site, defined by the specific gRNA used. Tag insertion was achieved using HDR with a repair template containing the tag and glycine-serine linker flanked by two homology arms.
(B) Western blot of nuclear extracts showing levels of SET1B from WT ESCs and dTAG-SET1B cells treated with dTAG-13 for 2 hr. SUZ12 is a loading control.

To examine how the depletion of SET1B affects gene expression, I performed cRNA-seq in the dTAG-SET1B cell line at the same time points as for SET1A (Figure 3.9). Unlike SET1A, SET1B removal had minimal effects on gene expression. After 2 hours, a time point at which the removal of SET1A affected the expression of more than 2500 genes, only 404 genes were significantly affected by

SET1B depletion (Figure 3.9 A). The magnitude of change in expression following SET1B removal was also smaller than following SET1A removal. Additionally, while the overwhelming majority of differentially expressed genes were reduced following SET1A depletion, depletion of SET1B resulted in an approximately equal number of increased and decreased genes. After 24 hours of SET1B removal, only seven genes were differentially expressed, indicating that there was an almost complete recovery in gene expression at this time point (Figure 3.9 B). Together, this suggests that, compared to SET1A, SET1B contributes very little to gene expression in ESCs. This is consistent with removal of SET1B having no effect on cell growth (Figure 3.9 C).

Despite the moderate effects of SET1B removal on gene expression, I wanted to understand whether the genes affected by SET1B depletion were also affected when SET1A was removed. Of the 404 genes significantly changed following 2 hours of SET1B depletion, 231 were also significantly affected by loss of SET1A at the same time point (Figure 3.9 D). Interestingly, SET1A depletion typically had a slightly more dramatic effect on the expression of these genes compared to SET1B depletion (Figure 3.9 E, the 'Shared' gene set). Therefore, this small subset of genes are regulated by both SET1A and SET1B, but SET1A has a slightly more prominent role in regulating their expression.

There were 173 genes with significantly changed expression following SET1B depletion that were not significantly affected by SET1A depletion (Figure 3.9 D). Although these genes were not captured as significantly affected by differential gene expression analysis criteria, when I examined their expression levels in cells depleted of SET1A, they did have altered expression on average, just to a lesser magnitude than was observed for SET1B depletion (Figure 3.9 E, 'SET1B-dependent' genes). In contrast, there were 2419 genes only significantly affected when SET1A was removed, and the expression of these SET1A-dependent genes was only very subtly affected by SET1B depletion (Figure 3.9 E, 'SET1A-dependent' genes). Indeed, by looking at individual genes, I could identify just a very small number of truly SET1B-specific genes, a number of genes regulated by

both SET1A and SET1B, and a much larger number of genes regulated by SET1A but that were only very subtly affected by SET1B removal (Figure 3.9 F). Therefore, this comparative analysis shows that SET1B makes a minimal contribution to the expression of just a small subset of genes, the majority of which are also regulated by SET1A. The regulation of SET1B target genes by SET1A could explain why there was a general recovery in gene expression following prolonged culture without SET1B (Figure 3.9 B).

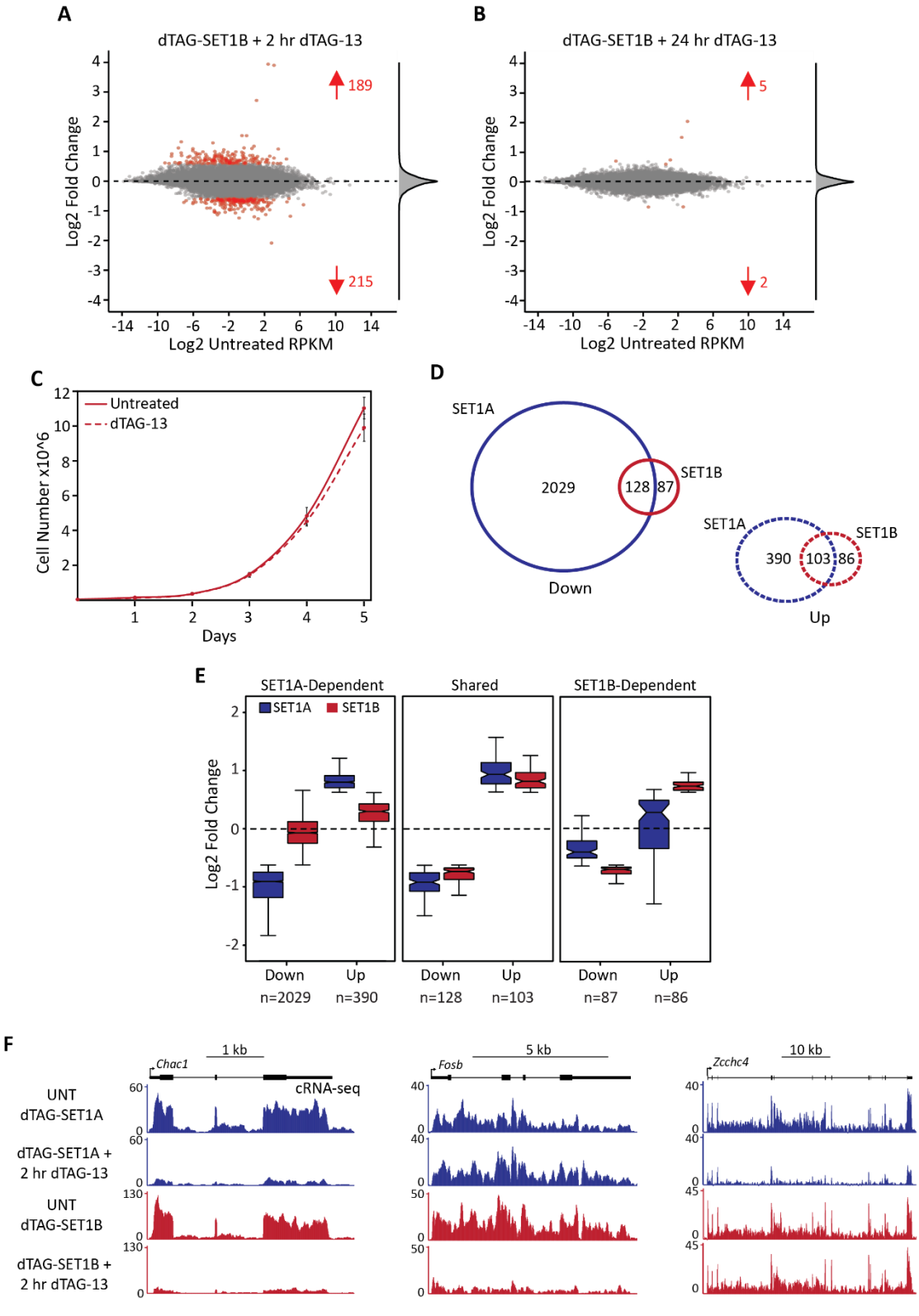


Figure 3.9 SET1B contributes little to gene expression in ESCs

(A) MA-plot of log2 fold change in gene expression (cRNA-seq) in dTAG-SET1B cells following 2 hr dTAG-13 treatment plotted against log2 mean RPKM for a non-redundant mm10 gene set (n=20633). Genes experiencing significant changes are coloured red and numbers of significant changes are shown (p-adj<0.05, fold change>1.5). A density plot of the log2 fold change in cRNA-seq signal is shown to the right of the plot.

(B) As in (A), but following 24 hr treatment with dTAG-13.
(C) Growth curve for untreated and dTAG-13-treated dTAG-SET1B cells cultured for 5 days. Error bars represent SEM of three biological replicates.
(D) A Venn diagram showing the overlap between genes significantly decreased ('Down') or increased ('Up') in expression following 2 hr SET1A depletion and 2 hr SET1B depletion.
(E) Box plots of log2 fold change in gene expression of all significantly changed genes at 2 hr SET1A or SET1B depletion. These genes are split into three groups, depending on whether they were significantly changed following depletion of both SET1A and SET1B (Shared), or only one of the two proteins (SET1A- or SET1B-Dependent). The number of genes in each subset are as per the Venn diagrams in (D).
(F) Representative genome shapshots of cRNA-seq signal following 2 hr dTAG-13 treatment in dTAG-SET1A and dTAG-SET1B cells. Snapshots show one gene regulated by both SET1A and SET1B (*Chac1*), one gene regulated by SET1B only (*Fosb*) and one gene regulated by SET1A only (*Zcchc4*).

3.2.6 The removal of SET1A and SET1B results in a widespread reduction in gene expression

SET1A has a widespread role in the regulation of gene expression in ESCs, whereas SET1B has a minor role at a limited number of genes. Interestingly, following the removal of SET1A, levels of SET1B were increased (Figure 3.10). This was apparent at both the RNA and protein level and raised the possibility that, although SET1B has a small effect on gene expression, it could compensate for the loss of SET1A. If this were the case, it would be expected that the combined removal of both SET1A and SET1B would result in more dramatic gene expression changes than following removal of SET1A alone. To investigate this possibility, I generated a cell line in which both SET1A and SET1B were engineered to have an N-terminal dTAG. In this cell line, treatment with dTAG-13 induced rapid removal of both SET1A and SET1B (Figure 3.11). Similarly to the depletion of SET1A, the levels of WDR5, RBBP5, ASH2L and WDR82 remained largely unchanged following SET1A/B depletion. However, the levels of CFP1 were reduced and this was slightly more pronounced than after removal of SET1A alone. This again suggests that CFP1 may be unstable when not associated with the SET1 complexes.

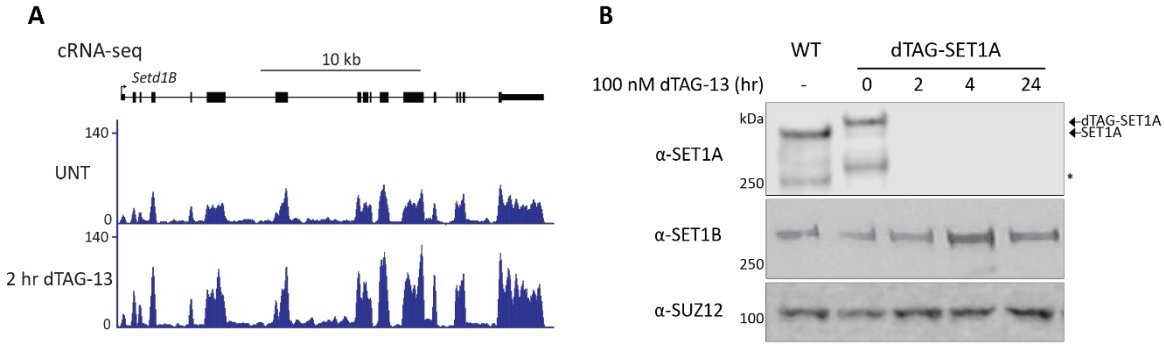


Figure 3.10 The expression of SET1B increases following depletion of SET1A
(A) cRNA-seq signal at *Setd1B* in dTAG-SET1A cells before (UNT) and after treatment with dTAG-13 for 2 hr.
(B) Western blot of nuclear extract from dTAG-SET1A cells showing the expression of SET1A and SET1B during a SET1A depletion timecourse. SUZ12 is a loading control. * denotes a common degradation product of SET1A.

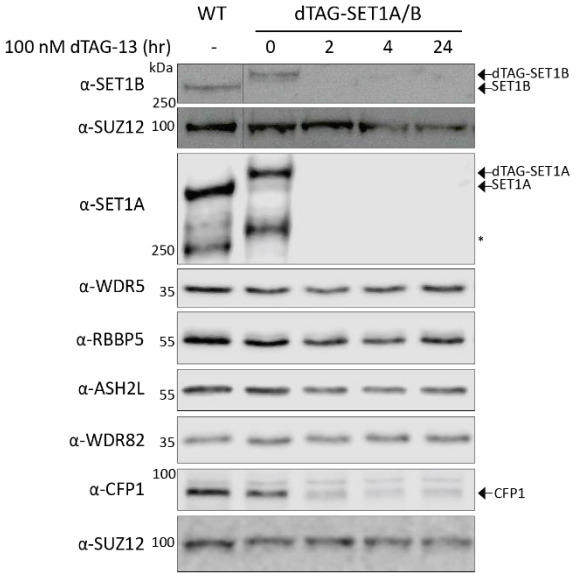


Figure 3.11 Charactering the dTAG-SET1A/B cell line
Western blot of nuclear extracts from WT ESCs and dTAG-SET1A/B cells treated with dTAG-13 over a 24 hr time course. Levels of each of the SET1 complex components, besides DPY30, are shown. SUZ12 is a loading control and * denotes a common degradation product of SET1A.

To examine how gene expression was affected after the combined removal of SET1A/B, I performed cRNA-seq after 2 hours and 24 hours of dTAG-13 treatment in the dTAG-SET1A/B cell line. As expected, the combined removal of SET1A/B predominantly resulted in decreases in gene expression at both time points (Figure 3.12 A and B). Qualitatively, the expression changes observed upon loss of both SET1A/B largely recapitulated those observed following the depletion of SET1A alone. Firstly, there were fewer significantly changed genes at 24 hours than at 2 hours.

Additionally, the genes decreased at 2 hours were initially moderately expressed, while those decreased at 24 hours were more highly expressed, and the differentially expressed genes were enriched for CGI-associated genes (Figure A2). However, from the density plot of the log2 fold changes, it is clear that in SET1A/B-depleted cells there was a pervasive reduction in gene expression, which became more apparent at 24 hours (Figure 3.12 A and B). This indicates that the combined removal of SET1A/B caused more pronounced effects on gene expression than the individual removal of SET1A. Indeed, following removal of both SET1A/B there were far more significantly changed genes than when SET1A was removed on its own (4487 at 2 hr and 4071 at 24 hr). Overlapping the significantly changed genes in the dTAG-SET1A/B line with those in the dTAG-SET1A line uncovered a group of genes that only became differentially expressed when both SET1A and SET1B were removed simultaneously, but not when SET1A was removed on its own (Figure 3.12 C-E). Since I have already established that some genes are regulated by both SET1A and SET1B (Figure 3.9), I wanted to understand whether these genes were sites of combined regulation. Plotting the expression change of these genes in each cell line revealed that they were modestly affected by both SET1A and SET1B depletion but more dramatically affected when both SET1A/B were removed together (Figure 3.13, 'SET1A/B-dependent' gene set). Interestingly, SET1A and SET1B actually appeared to have a largely additive effect on their expression. For these genes, it is likely that when either SET1A or SET1B is removed, the presence of the other protein is sufficient to at least partially support gene expression. This additive effect was not evident at 24 hours, however this was expected given the almost complete recovery in gene expression following prolonged removal of SET1B only (Figure 3.9 B).

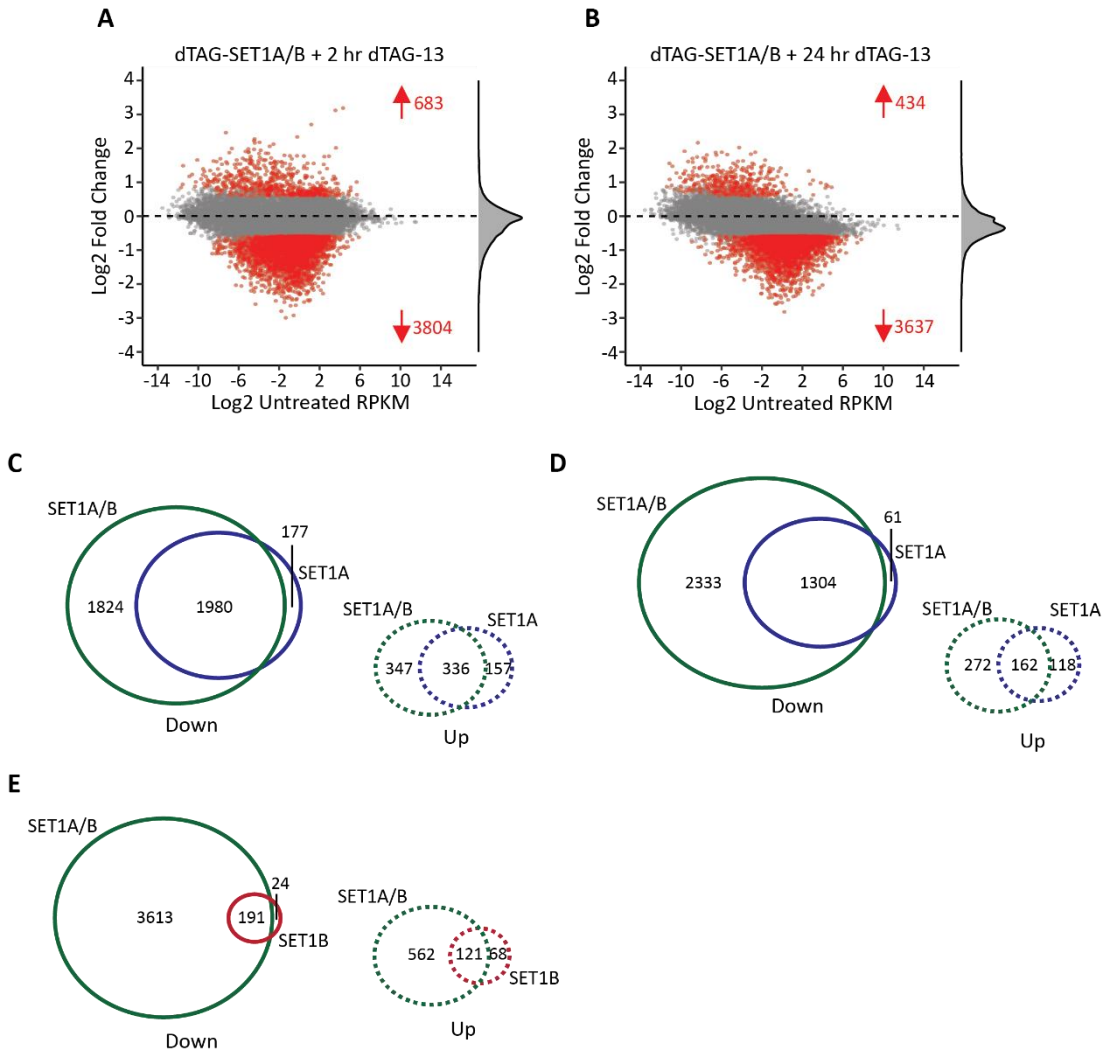


Figure 3.12 Depletion of SET1A/B results in widespread decreases in gene expression
(A) MA-plot of log2 fold change in gene expression (cRNA-seq) in dTAG-SET1A/B cells following 2 hr dTAG-13 treatment plotted against log2 mean RPKM for a non-redundant mm10 gene set (n=20633). Significant changes are coloured red and numbers of significant changes are shown (p-adj<0.05, fold change>1.5). A density plot of the log2 fold change in cRNA-seq signal is shown to the right.
(B) As in (A), but dTAG-SET1A/B cells treated with dTAG-13 for 24 hr.
(C) and (D) Venn diagrams illustrating the overlap between genes significantly changed following depletion of SET1A and SET1A/B at 2 hr (C) or 24 hr (D).
(E) Venn diagrams illustrating the overlap between genes significantly changed following depletion of SET1B and SET1A/B for 2 hr.

I have established that there is a subset of genes that were only significantly differentially expressed when both SET1A and SET1B were removed. This is due to the relatively minor but additive roles of both SET1A and SET1B at these loci. However, I have also previously established that SET1B makes a minor contribution to gene expression at the genes that are significantly

decreased when SET1A is removed on its own (Figure 3.9). I therefore wanted to determine whether the additive regulation by SET1A and SET1B was also apparent at these SET1A-dependent genes. Indeed, when both SET1A and SET1B were simultaneously removed, the magnitude of change of these genes was larger than following removal of SET1A alone (Figure 3.13, 'SET1A-dependent' genes). This suggests that while SET1A clearly has a much greater role in the regulation of gene expression, SET1A and SET1B actually largely co-regulate an overlapping set of target genes. Additionally, in the dTAG-SET1A/B cell line, the magnitude of change of the SET1A-dependent genes was larger than the SET1A/B-dependent subset (Figure 3.13, green bars). Therefore, the genes most dependent on the SET1 proteins for their expression are mostly regulated by SET1A. The contribution of SET1B is largest at the genes where the contribution from SET1A and the additive contribution by SET1A/B is smaller. The greater gene expression outcome following depletion of SET1A and SET1B is supported by growth analysis, which indicates that ESCs grew more slowly when both SET1A/B were removed, in comparison to SET1A alone (Figure 3.14).

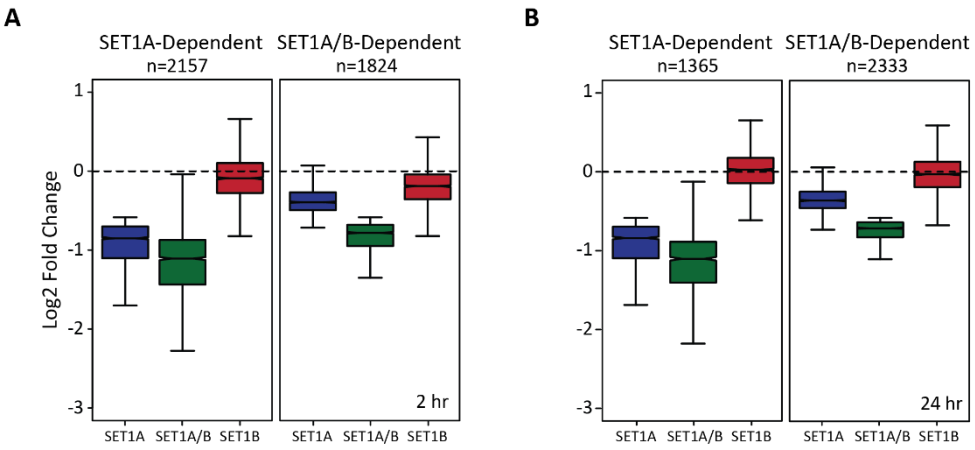


Figure 3.13 Simultaneous removal of SET1A and SET1B increases the magnitude of gene expression changes

Box plots of log2 fold change in gene expression in the dTAG-SET1A, dTAG-SET1B or dTAG-SET1A/B cell lines. Genes are split into those that were significantly decreased in the dTAG-SET1A line ('SET1A-Dependent') or those genes only significantly decreased in the dTAG-SET1A/B line but not the dTAG-SET1A line ('SET1A/B-Dependent'). Box plots are at 2 hr (A) or 24 hr (B).

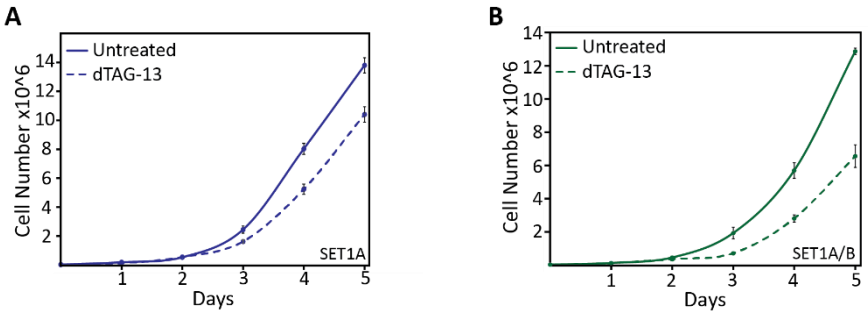


Figure 3.14 Removal of SET1A/B has a larger effect on cell growth than removal of SET1A alone
Growth curves for untreated and dTAG-13-treated dTAG-SET1A (A) or dTAG-SET1A/B (B) cells cultured for up to 5 days. Error bars represent SEM of three biological replicates.

3.3 Summary and Discussion

The SET1 complexes are highly conserved throughout eukaryotic evolution. Mammalian SET1A/B, as well as their orthologues in lower eukaryotes, preferentially associate with active gene promoters and deposit H3K4me3, a histone modification associated with active transcription (Meeks and Shilatifard, 2017; Cenik and Shilatifard, 2021). However, previous studies have suggested that the SET1 complexes make a relatively minor contribution to the regulation of gene expression. In *S. cerevisiae,* the sole orthologue of SET1A/B, ySet1, is dispensable for viability and its removal has only modest effects on gene expression (Nislow, Ray and Pillus, 1997; Briggs *et al.*, 2001; Miller *et al.*, 2001; Noma and Grewal, 2002; Santos-Rosa *et al.*, 2002; Krogan *et al.*, 2002; Nagy *et al.*, 2002; Boa, Coert and Patterton, 2003; Venkatasubrahmanyam *et al.*, 2007; Guillemette *et al.*, 2011; Lenstra *et al.*, 2011; Margaritis *et al.*, 2012; Ramakrishnan *et al.*, 2016; Murray *et al.*, 2019). Similarly, when the murine SET1A/B complexes were perturbed previously, gene expression outcomes were typically subtle and lacked directionality (Clouaire *et al.*, 2012; Clouaire, Webb and Bird, 2014; Austenaa *et al.*, 2015; Fang *et al.*, 2016; Brown *et al.*, 2017; Franks *et al.*, 2017). However, both SET1A and SET1B are essential for mouse development, though it has remained unclear to what extent they regulate gene expression (Bledau *et al.*, 2014; Tusi *et al.*, 2015; Fang *et al.*, 2016).

Here I set out to examine whether SET1A and SET1B contribute to the regulation of gene expression in ESCs. By using rapid protein depletion systems coupled with quantitative gene expression analysis, I found that within just 2 hours of SET1A/B depletion, there was a widespread reduction in gene expression (Figure 3.12). This demonstrates that SET1A/B have a central role in supporting gene expression in ESCs. Interestingly, these gene expression outcomes were much more dramatic and directional than would be predicted from previous studies. Based on these findings, I propose that the true effects of SET1A/B removal on gene expression were overlooked previously due to two central limitations in the experimental systems that were employed. Firstly, in previous studies of mammalian SET1A/B, conditional genetic deletion or RNAi approaches were employed. Such systems are slow due to their reliance on natural protein turnover. Here, using a rapid degron system, I now discover that removal of SET1A/B causes an almost immediate and profound effect on gene expression. These effects are confounded by non-primary outcomes, and possibly compensated for to some extent, at later time points. Therefore, degron approaches are key to uncovering the primary effects of SET1A/B on gene expression, and this can partially explain why these effects were overlooked in previous studies. Secondly, previous genome-wide analyses of SET1A/B perturbation typically did not employ spike-in calibration. Such a normalisation method is essential for identifying highly directional effects on gene expression. Here, by including spike-in calibration, I uncovered a clear directionality in the gene expression changes following SET1A/B removal and provided clear evidence that these proteins play a fundamental role in supporting mammalian gene expression. This is consistent with their essential role during mouse development (Bledau *et al.*, 2014; Tusi *et al.*, 2015; Fang *et al.*, 2016). Based on the finding that degron approaches and calibrated gene expression analyses are essential for capturing the effects of the mammalian SET1 proteins on gene expression, it will be important to revisit whether the SET1 proteins also play a conserved role in supporting gene expression in lower eukaryotes.

SET1A and SET1B are highly homologous and found within largely similar multi-protein complexes (Lee and Skalnik, 2005; Lee *et al.*, 2007; Lee and Skalnik, 2008; Wu *et al.*, 2008). I presented

evidence that SET1A and SET1B make a shared contribution to the expression of a number of target genes (Figure 3.13). However, SET1A clearly has a much more extensive role in supporting gene expression than SET1B (Figure 3.5 and 3.9). This is consistent with the essential role of SET1A in ESC growth and the earlier embryonic lethality caused by SET1A knockout in comparison to SET1B (Bledau *et al.*, 2014; Fang *et al.*, 2016; Cao *et al.*, 2017). However, given the close homology between SET1A and SET1B, it is unclear why SET1A has a much greater role in the regulation of gene expression. One possibility that could explain these findings is simply that SET1A is more highly expressed than SET1B, as has been proposed based on proteomics analysis in HeLa cells (van Nuland, Smits, *et al.*, 2013). Based on my cRNA-seq analysis, levels of the *Setd1B* mRNA (4.4 RPKM) were only slightly lower than those of *Setd1A* (6.7 RPKM). It remains possible that SET1B protein levels could be much lower in comparison to SET1A than the levels of their respective mRNAs would suggest and this possibility remains to be examined in detail in ESCs. However, it was previously shown that the over-expression of SET1B cannot compensate for growth defects caused by SET1A removal and I observed that the contribution of SET1B to gene expression was most pronounced at the genes that were only subtly affected by the loss of SET1A (Bledau *et al.*, 2014 and Figure 3.13). This suggests that SET1A and SET1B, regardless of their expression levels and some shared functions, must also have distinct roles in gene regulation.

Going forward it will be important to examine in more detail what enables the distinct functions of SET1A and SET1B in the regulation of gene expression. One explanation could come in the form of the accessory proteins that are unique to the SET1A and SET1B complexes; SET1A associates with HCF1, Cyclin K and BOD1L, while SET1B associates with BOD1 and RBM15 (Wysocka *et al.*, 2003; Lee *et al.*, 2007; Lee and Skalnik, 2012; van Nuland, Smits, *et al.*, 2013; Wang *et al.*, 2017; Higgs *et al.*, 2018; Hoshii *et al.*, 2018; Coker *et al.*, 2020). It is possible that these accessory factors contribute to regulating gene expression. Alternatively, the accessory factors might define complex-specific target site recognition. For example, HCF1 has been implicated in the recruitment of SET1A to a specific subset of cell-cycle regulated gene promoters (Tyagi *et al.*, 2007). It would

therefore be useful to determine the genome-wide binding of SET1B. To address this, I sought to characterise the binding of SET1B by ChIP-seq using commercially available antibodies, but without success. As an alternative, similar epitope tagging approaches to those used for SET1A would provide an avenue to directly compare SET1A and SET1B binding in the genome. Using these approaches, one could explore any unique binding properties of SET1A or SET1B and, if they do occupy distinct regions of the genome, determine whether this relies on the complex-specific proteins and whether it explains the distinct effects on gene expression.

Furthermore, it will be important to investigate the mechanisms by which SET1A/B regulate gene expression. Interestingly, following SET1A/B depletion, the expression of some polycomb target genes was increased (Figures 3.6 and A3). This could suggest that the function of SET1A/B is not limited to supporting the expression of actively transcribed genes. However, since SET1A is relatively depleted from these genes, this observation might not be a primary effect of SET1A/B removal. In contrast, SET1A is enriched at the genes whose expression was reduced in its absence (Figures 3.6 and A3). However, I observed that SET1A occupancy alone cannot explain the effects of SET1A/B removal on gene expression (Figure A4). Since SET1A/B place H3K4me3 and H3K4me3 is extensively linked to active gene expression (Lee and Skalnik, 2005; Lee *et al.*, 2007; Wu *et al.*, 2008; Chapter 1.3), one could predict that SET1A/B support gene expression via H3K4me3. However, there is evidence to suggest that SET1A/B can function independently of their catalysis (Chapter 1.4.5) and the extremely rapid nature of the gene expression changes observed following SET1A/B depletion might support a role in regulating RNAPII. I will explore potential mechanisms by which SET1A/B regulate gene expression throughout the remainder of my thesis.

4. SET1A/B regulate gene expression independently of H3K4me3

4.1 Introduction

H3K4me3 is a histone modification universally associated with CGI-associated gene promoters, where its levels scale with transcription; more highly expressed genes have more promoter-associated H3K4me3 (Barski *et al.*, 2007; Guenther *et al.*, 2007; Mikkelsen *et al.*, 2007; Illingworth *et al.*, 2008). Given the relationship between H3K4me3 and gene expression, a considerable effort has been placed on understanding the mechanisms that govern H3K4me3 deposition. This has been relatively straightforward in yeast, where ySet1 places all H3K4 methylation (Briggs *et al.*, 2001; Krogan *et al.*, 2002; Nagy *et al.*, 2002; Noma and Grewal, 2002; Santos-Rosa *et al.*, 2002). However, understanding the mechanisms that enable H3K4me3 deposition in higher eukaryotes is complicated by the diversification in the H3K4 HMT machinery during evolution (Chapter 1.4). Here, the potential for functional redundancy between the H3K4 HMTs makes examining their contribution to H3K4me3 difficult.

There are four HMTs that are thought to trimethylate H3K4 in mammalian cells: MLL1/2 and SET1A/B. Removal of MLL1/2 in ESCs has revealed that they place H3K4me3 primarily at the promoters of lowly expressed genes (P. Wang *et al.*, 2009; Hu, Garruss, *et al.*, 2013; Denissov *et al.*, 2014; Mas *et al.*, 2018; Sze *et al.*, 2020). This accounts for a small proportion of total H3K4me3 in mammalian genomes, since H3K4me3 is most enriched at highly expressed genes (Barski *et al.*, 2007; Guenther *et al.*, 2007; Mikkelsen *et al.*, 2007). SET1A on the other hand preferentially associates with the promoters of highly expressed genes and, along with SET1B, has been implicated in the deposition of H3K4me3 at these loci (Clouaire *et al.*, 2012; Clouaire, Webb and Bird, 2014; Brown *et al.*, 2017; Sze *et al.*, 2017, 2020). Furthermore, perturbations of SET1A/B or *Drosophila* dSet1 have appeared to cause dramatic reductions in H3K4me3 (Lee and Skalnik, 2005; Wu *et al.*, 2008; Ardehali *et al.*, 2011; Mohan *et al.*, 2011; Vethantham *et al.*, 2012; Clouaire *et al.*,

2012; Hallson *et al.*, 2012; Tang *et al.*, 2013; Clouaire, Webb and Bird, 2014; Fang *et al.*, 2016; Brown *et al.*, 2017; Cao *et al.*, 2017; Franks *et al.*, 2017; Sze *et al.*, 2020; Yang *et al.*, 2020). This has led to a model whereby the SET1 complexes are thought to be primarily responsible for the enrichment of H3K4me3 at actively transcribed genes (Meeks and Shilatifard, 2017; Cenik and Shilatifard, 2021). In fact, a conditional knockout of SET1A alone was reported to cause a roughly 50% reduction in H3K4me3 (Bledau *et al.*, 2014). However, depending on the system used to perturb SET1A/B function, the reported effects on overall levels of H3K4me3 are highly variable. Additionally, previous attempts to directly perturb SET1A/B were not accompanied by a genome-wide and quantitative analysis of the effects on H3K4me3. Therefore, the contribution of SET1A and SET1B to placing H3K4me3 at the genome scale in mammalian cells remains elusive.

In chapter 3 I established that SET1A/B are key regulators of gene expression in ESCs. Since H3K4me3 has previously been proposed to directly support gene transcription by stabilising PIC assembly (Vermeulen *et al.*, 2007; van Ingen *et al.*, 2008; Eberl *et al.*, 2013; Lauberth *et al.*, 2013; van Nuland, Schram, *et al.*, 2013), one could predict that the role of SET1A/B in regulating gene expression is mediated through H3K4me3 deposition. However, there is evidence that SET1A/B can function independently of H3K4me3; both complexes likely methylate additional substrates and have even been proposed to have functions completely independent of their catalytic activity (Chapter 1.4.5). I therefore set out to examine the contribution of SET1A/B to the levels of H3K4me3 in ESCs and to develop systems to investigate whether SET1A/B regulate gene expression through H3K4me3 deposition.

4.2 Results

4.2.1 Removal of SET1A results in a subtle but widespread reduction in H3K4me3

To begin to examine the contribution of SET1A to H3K4me3 in ESCs, I used the dTAG-SET1A cell line to quantify the levels of bulk H3K4me3 in histone extracts during a time course of SET1A

depletion (Figure 4.1). Surprisingly, levels of H3K4me3 remained unchanged, even after 5 days of treatment. This suggests that SET1A alone is not responsible for placing the majority of H3K4me3 in ESCs, as has been previously proposed based on observations made following depletion of SET1A in a classical conditional knockout cell line (Bledau *et al.*, 2014).

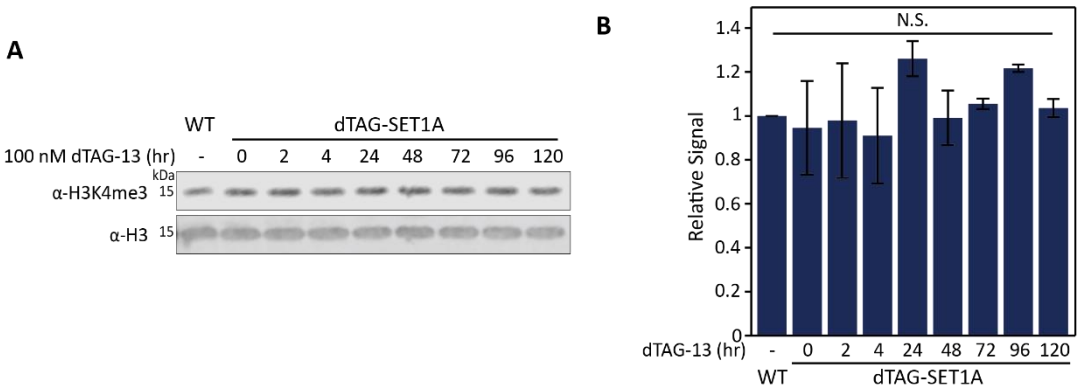


Figure 4.1 Removal of SET1A does not affect global H3K4me3
(A) Representative western blot of histone extracts showing the levels of H3K4me3 from WT ESCs and dTAG-SET1A cells treated with dTAG-13 over a 120 hr time course. H3 is a loading control.
(B) Quantification of H3K4me3 levels from western blots of histone extracts from dTAG-SET1A cells treated with dTAG-13 over a 120 hr time course. H3K4me3 signal is normalised to histone H3 and expressed relative to WT. Error bars show SEM from two biological replicates. N.S. represents not significant in a paired Student's T test ($p>0.05$).

Although overall levels of H3K4me3 were unaffected following SET1A depletion, I wanted to examine whether SET1A has an important role in placing H3K4me3 at specific gene promoters. Such a contribution cannot be assessed by examining bulk levels of H3K4me3 by western blotting. Therefore, I set out to use calibrated H3K4me3 ChIP-seq. Since H3K4me1 and H3K4me2 are common in mammalian genomes, using an antibody that is highly specific for H3K4me3 and does not cross-react with H3K4me1/2 is essential. The Klose group has previously used a polyclonal H3K4me3 antibody in cChIP-seq (Brown *et al.*, 2017). However, I wanted to further optimise the ChIP conditions in order to ensure that the H3K4me3 signal was highly specific. To do this, I spiked native chromatin with a panel of differentially modified synthetic nucleosomes that were individually barcoded before performing ChIP using different amounts of the H3K4me3 antibody (Shah *et al.*, 2018). I was then able to quantify how much of each nucleosome species was pulled

down during the ChIP experiment by interrogating the sample using qPCR with barcode-specific primers. This experiment revealed that the antibody was highly specific for H3K4me3 in all tested conditions (Figure 4.2). Using less antibody in the ChIP increased specificity for H3K4me3 over lower methylation states. However, after reducing the amount of antibody to less than 1.5 µl, any gain in specificity was negligible. As expected, with smaller volumes of antibody the recovery of H3K4me3-modified nucleosomes was reduced. I therefore used 1.5 µl of this antibody in all future ChIP experiments, since this amount showed an approximately 45-fold enrichment for H3K4me3 over H3K4me2, without compromising too much on the amount of H3K4me3-modified material isolated during the ChIP.

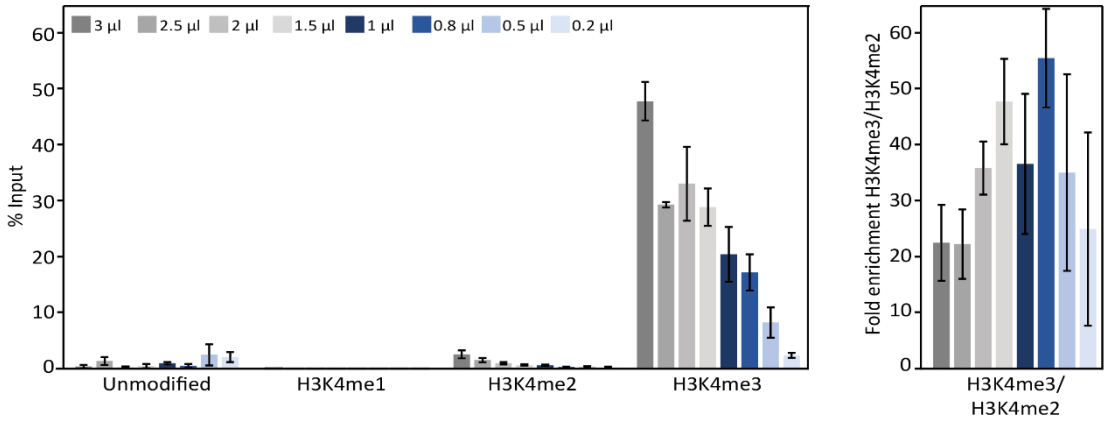


Figure 4.2 Optimising H3K4me3 ChIP conditions with synthetic modified nucleosomes
H3K4me3 ChIP-qPCR for unmodified, H3K4me1, H3K4me2 and H3K4me3 barcoded nucleosomes spiked into native ESC chromatin (Shah *et al.*, 2018). A decreasing amount of H3K4me3 antibody was used in the ChIP reactions. Mean H3K4me3 signal from three biological replicates is shown, as is the ratio of H3K4me3/H3K4me2. Error bars represent SEM.

Having optimised ChIP conditions to yield highly specific H3K4me3 signal, I next wanted to examine the effect of SET1A removal on TSS-associated H3K4me3 genome-wide. To this end, I performed H3K4me3 cChIP-seq in the dTAG-SET1A cell line treated with dTAG-13 for 2 hours and 24 hours, time points when extensive gene expression changes were observed (Chapter 3). If the effects on gene expression were related to alterations in promoter-associated H3K4me3, one might predict that any H3K4me3 changes would be observable within a similar timeframe. Additionally, since H3K4me3 has been proposed to be a highly dynamic histone modification, with an estimated half-

life of less than 7 hours in HeLa cells (Zheng *et al.*, 2014), I reasoned that a 24 hour treatment would be long enough to observe any contribution of SET1A to H3K4me3. As expected, H3K4me3 signal peaked at the TSSs of expressed genes and was primarily associated with CGI-associated gene promoters but largely absent from non-CGI-associated gene promoters (Figure 4.3 A and B). On average, depleting SET1A for 2 hours resulted in a very subtle decrease in H3K4me3 at TSSs (Figure 4.3 C). Plotting the log2 fold change in H3K4me3 across all TSSs highlighted a pervasive but subtle reduction in H3K4me3 following 2 hours of SET1A depletion (Figure 4.3 B). Based on a half-life prediction of approximately 7 hr, such a subtle decrease in H3K4me3 following 2 hour SET1A depletion might be expected. Interestingly, however, at the 24 hour time point, rather than observing a further decrease, H3K4me3 increased when compared to untreated cells (Figure 4.3 B and C). Together, this suggests that SET1A contributes rather modestly to H3K4me3, possibly due to the presence of a mechanism that compensates for H3K4me3 deposition when SET1A is removed.

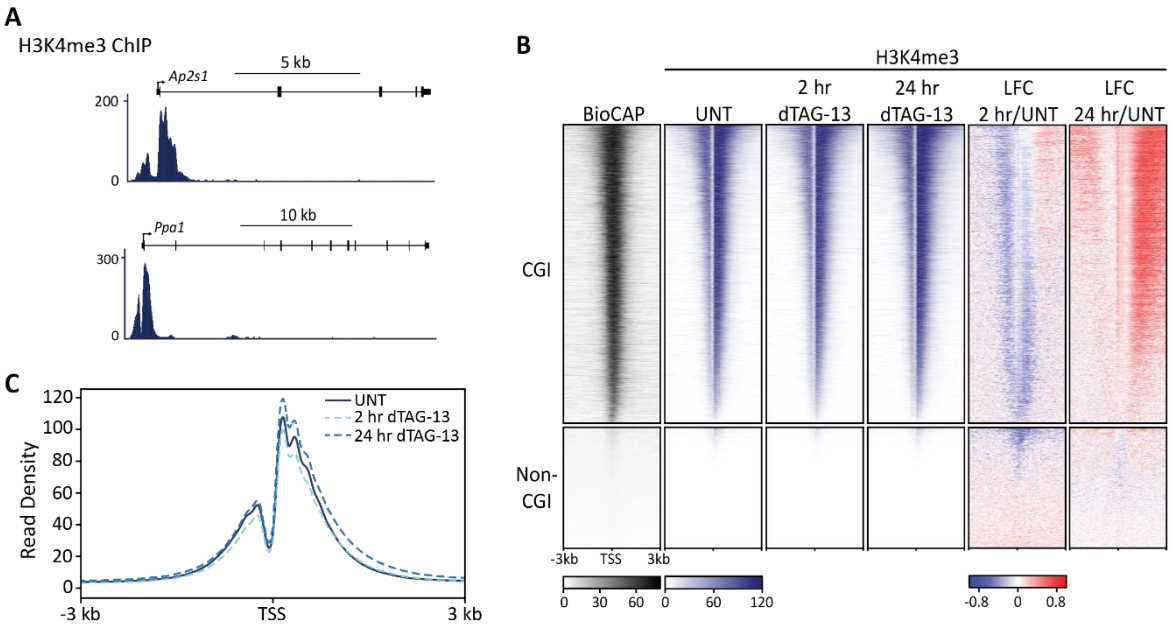


Figure 4.3 Removing SET1A has a subtle but pervasive effect on promoter-associated H3K4me3
(A) Genomic snapshots showing H3K4me3 cChIP-seq signal at the TSSs of two actively transcribed genes in untreated dTAG-SET1A cells.
(B) Heatmaps of H3K4me3 cChIP-seq signal in untreated, 2 hr and 24 hr dTAG-13-treated dTAG-SET1A cells, as well as the log2 fold change (LFC) between treated and untreated conditions. BioCAP-seq signal is also shown (Long *et al.*, 2013). Genes are split into 'CGI' or 'non-CGI' depending on the presence of a

CGI within 2.5 kb of the TSS ('CGI' n=14687, 'non-CGI' n=5946). Heatmaps are sorted by mean H3K4me3 signal in untreated cells.
(C) Metaplot of H3K4me3 signal in untreated, 2 hr and 24 hr dTAG-13-treated dTAG-SET1A cells. Signal is centred at all TSSs (n=20633).

The heatmap analysis of changes in H3K4me3 revealed that the promoters with lower levels of H3K4me3 in untreated cells experienced both the largest reduction in H3K4me3 at 2 hours and the smallest increase at 24 hours (Figure 4.3 B). Since H3K4me3 scales with transcriptional activity, this suggests that promoter-associated H3K4me3 is more dependent on SET1A at lowly expressed genes. To investigate this possibility, I divided all genes with a peak of promoter-associated H3K4me3 in untreated cells into quartiles based on their initial expression level and examined the effect of SET1A removal on promoter-associated H3K4me3 within each quartile (Figure 4.4). As would be predicted, the level of H3K4me3 in each quartile scaled with their transcriptional activity (Figure A5). Quantification of the change in promoter-associated H3K4me3 confirmed that removal of SET1A for 2 hours resulted in a greater reduction in H3K4me3 at lowly expressed genes (Figure 4.4 A). More highly expressed genes typically retained more H3K4me3 in the absence of SET1A. The finding that lowly expressed genes are more dependent on SET1A for H3K4me3 deposition is surprising, given that SET1A is enriched at the promoters of highly expressed genes (Brown *et al.*, 2017; Sze *et al.*, 2017, 2020). Following 24 hour SET1A depletion, H3K4me3 typically returned to levels similar to those in untreated cells at lowly expressed genes. In contrast, at highly expressed genes, H3K4me3 increased compared to the untreated state (Figure 4.4 A). This suggests that there is a mechanism that places H3K4me3 in the absence of SET1A, which contributes most prominently to H3K4me3 deposition at highly expressed genes.

Interestingly, the heatmaps illustrating the changes in H3K4me3 across TSSs revealed that H3K4me3 at highly expressed genes was decreased upstream of TSSs following 2 hour SET1A depletion (Figure 4.4 B). This same decrease was not apparent towards gene bodies. Additionally, H3K4me3 gain at 24 hours was more pronounced towards the gene bodies of highly expressed

genes. This suggests that there may be a compensatory mechanism acting primarily downstream of TSSs to deposit H3K4me3 at highly expressed genes in the absence of SET1A. The same mechanism could be responsible both for the maintenance of H3K4me3 following acute SET1A removal and the gain in H3K4me3 during longer-term culture without SET1A.

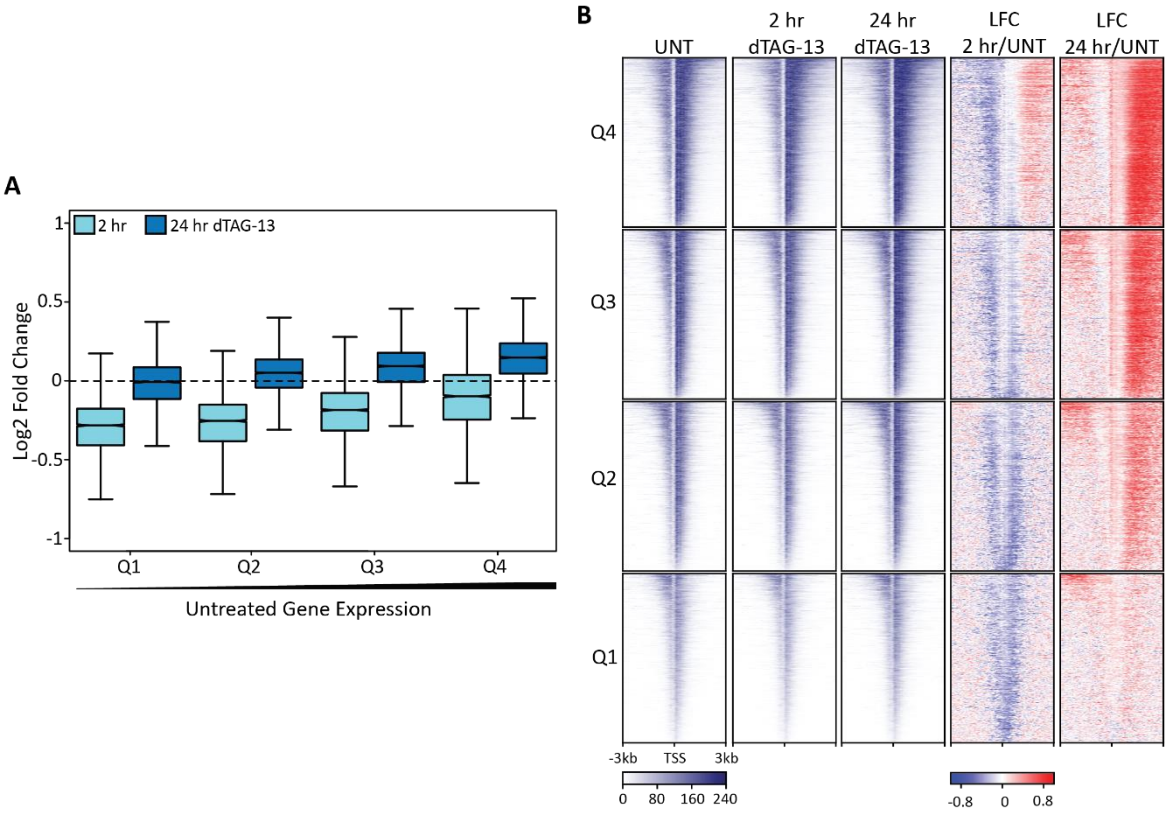


Figure 4.4 Removal of SET1A results in the largest change in H3K4me3 at lowly expressed genes
(A) Box plots of log2 fold change in H3K4me3 cChIP-seq signal at promoter-associated H3K4me3 peaks after 2 hr and 24 hr SET1A removal (n=14546). Genes are split into quartiles of equal size based on gene expression level in untreated dTAG-SET1A cells.
(B) Heatmaps of H3K4me3 cChIP-seq signal at each gene with a promoter-associated H3K4me3 peak in untreated cells (n=14536), split into the same expression quartiles as in (A). H3K4me3 cChIP-seq signal in untreated cells and dTAG-SET1A cells treated with dTAG-13 for 2 hr and 24 hr is shown. The log2 fold change (LFC) in H3K4me3 at 2 hr and 24 hr compared to untreated cells is also shown. Heatmaps are sorted by mean signal in untreated cells.

4.2.2 Removal of SET1A/B results in a larger reduction in H3K4me3 than removal of SET1A alone

I have established that SET1A makes a minimal but widespread contribution to promoter-associated H3K4me3 in ESCs. However, the removal of CFP1 and WDR82 have previously been reported to cause more pronounced reductions in H3K4me3 (Lee and Skalnik, 2005; Wu *et al.*,

2008; Clouaire *et al.*, 2012; Vethantham *et al.*, 2012; Tang *et al.*, 2013; Clouaire, Webb and Bird, 2014; Brown *et al.*, 2017; Franks *et al.*, 2017; Yang *et al.*, 2020). Following removal of SET1A, the reduction in H3K4me3 was smaller at the promoters of more highly expressed genes (Figure 4.4 A). These are the loci where SET1A is enriched and also where SET1B would be predicted to bind (Brown *et al.*, 2017; Sze *et al.*, 2017, 2020). Since the expression of SET1B increased following the depletion of SET1A (Figure 3.10), I was curious whether SET1B was compensating for the loss of SET1A at highly expressed genes. Therefore, I aimed to examine whether the combined removal of SET1A and SET1B results in a more pronounced effect on H3K4me3 than removal of SET1A alone.

To determine whether removal of SET1A/B affected bulk levels of H3K4me3, I analysed histone extracts from the dTAG-SET1A/B cell line during a time course of SET1A/B depletion. Much like depletion of SET1A, the removal of SET1A/B did not affect the bulk levels of H3K4me3 by western blotting (Figure 4.5). In keeping with this observation, levels of H3K4me1 and H3K4me2 also remained unchanged (Figure A6). This suggests that SET1A/B make a minor contribution to the overall levels of H3K4me3 in ESCs.

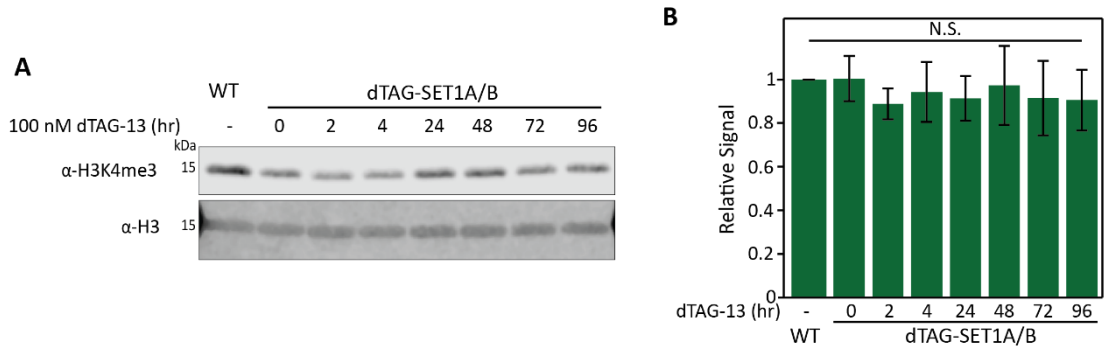


Figure 4.5 Removal of both SET1A and SET1B does not affect bulk H3K4me3
(A) Representative western blot showing levels of H3K4me3 in histone extracts from dTAG-SET1A/B cells treated with dTAG-13 over a 96 hr time course. H3 is a loading control.
(B) Quantification of H3K4me3 from western blots of histone extracts from dTAG-SET1A/B cells treated with dTAG-13 over a 96 hr time course. Signal is normalised to histone H3 and expressed relative to WT. Error bars show SEM of two biological replicates. N.S. represents not significant in a paired Student's T test ($p>0.05$).

In order to examine the effect of the combined removal of SET1A/B on promoter-associated H3K4me3, I next performed cChIP-seq in the dTAG-SET1A/B cell line at the same time points as for dTAG-SET1A. Removal of SET1A/B for 2 hours resulted in a relatively minor reduction in overall H3K4me3 at TSSs (Figure 4.6 A). However, this reduction was larger than when SET1A only was removed (Figure 4.6 B and C). Therefore, the acute depletion of SET1A/B has a larger overall effect on promoter-associated H3K4me3 than removing SET1A alone. As with the depletion of SET1A, H3K4me3 levels recovered following 24 hours SET1A/B removal (Figure 4.6 A). However, this increase in H3K4me3 was smaller than following longer-term depletion of SET1A (Figure 4.6 B and C). This suggests that SET1B is at least partially responsible for the increase in H3K4me3 observed at the 24 hour time point in the dTAG-SET1A line (Figure 4.3).

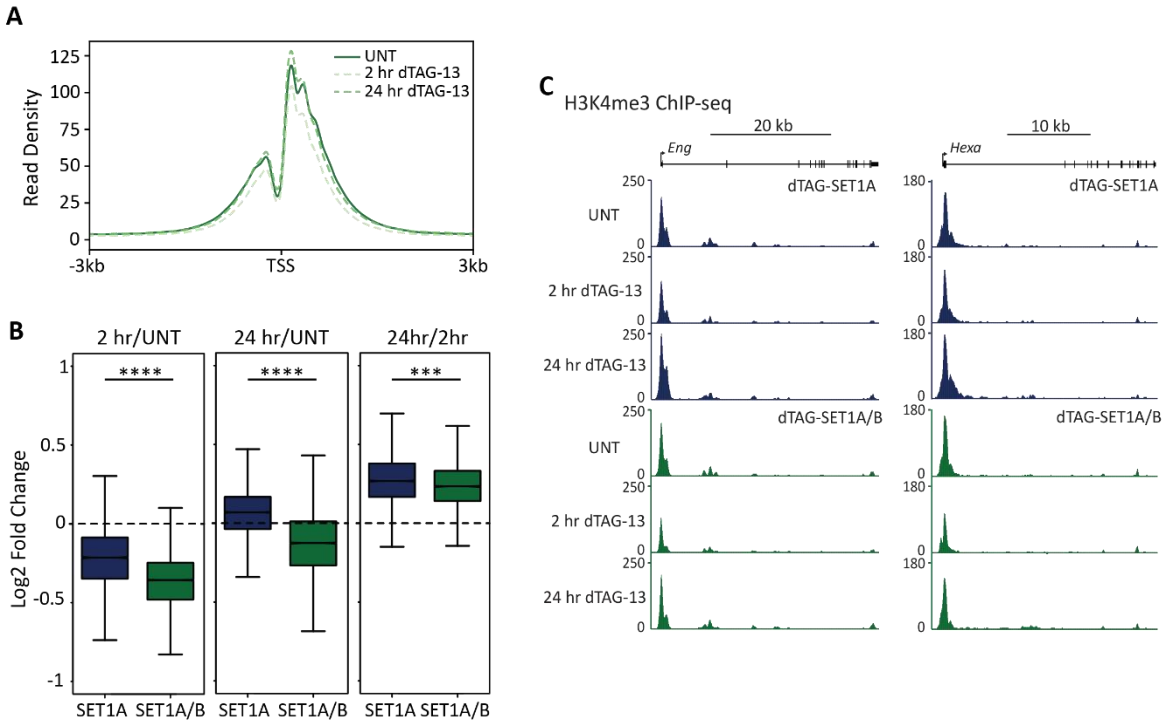


Figure 4.6 Removal of both SET1A and SET1B has a larger effect on H3K4me3 than removal of SET1A alone

(A) Metaplot of H3K4me3 cChIP-seq signal centred at all TSSs (n=20633) in untreated, 2 hr and 24 hr dTAG-13-treated dTAG-SET1A/B cells.

(B) Boxplots of the log2 fold change in H3K4me3 cChIP-seq signal at promoter-associated H3K4me3 peaks (n=14546) at each time point in both dTAG-SET1A and dTAG-SET1A/B cell lines. *** represents $p<10^{-94}$ and **** represents $p<10^{-100}$ as measured by Student's T test.

(C) Genomic snapshots showing H3K4me3 cChIP-seq signal at the TSSs of two genes (*Eng* and *Hexa*) with a larger depletion of H3K4me3 following removal of SET1A/B in comparison to SET1A alone.

I next wanted to examine how depletion of SET1A/B affects H3K4me3 at genes with different expression levels in untreated cells. Within 2 hours of SET1A/B removal there was a general decrease in promoter-associated H3K4me3 across all gene expression quartiles (Figure 4.7). However, in contrast to the SET1A-only depletion, the effect of SET1A/B removal was largest at more highly expressed genes. Strikingly, the reduction in H3K4me3 at highly expressed genes was also apparent both upstream and downstream of TSSs (Figure 4.7 B). This suggests that SET1B places H3K4me3 around the TSSs of highly expressed genes in the absence of SET1A. Simultaneous removal of SET1A/B therefore results in a subtle but pervasive reduction in promoter-associated H3K4me3 genome-wide, with a larger reduction at more highly expressed genes. This observation is consistent with previous observations made upon the removal of CFP1, which can target SET1A/B to CGIs (Clouaire *et al.*, 2012; Clouaire, Webb and Bird, 2014; Brown *et al.*, 2017).

I next wanted to understand whether SET1B was responsible for the observed increase in H3K4me3 following prolonged culture without SET1A, which was especially evident at highly expressed genes (Figure 4.4). Following a prolonged 24 hour treatment to remove SET1A/B, H3K4me3 at highly expressed genes remained reduced in comparison to untreated cells, suggesting that SET1B is at least partially responsible for the recovery in H3K4me3 at highly expressed genes observed following depletion of SET1A (Figure 4.7). This indicates that SET1B can compensate for SET1A loss in H3K4me3 deposition at highly expressed genes. In contrast, H3K4me3 at lowly expressed genes recovered to levels similar to those in untreated cells. Therefore, SET1B is not responsible for the deposition of H3K4me3 at lowly expressed genes in the absence of SET1A. There must therefore be an additional mechanism to place H3K4me3 at lowly expressed genes when SET1A/B are removed.

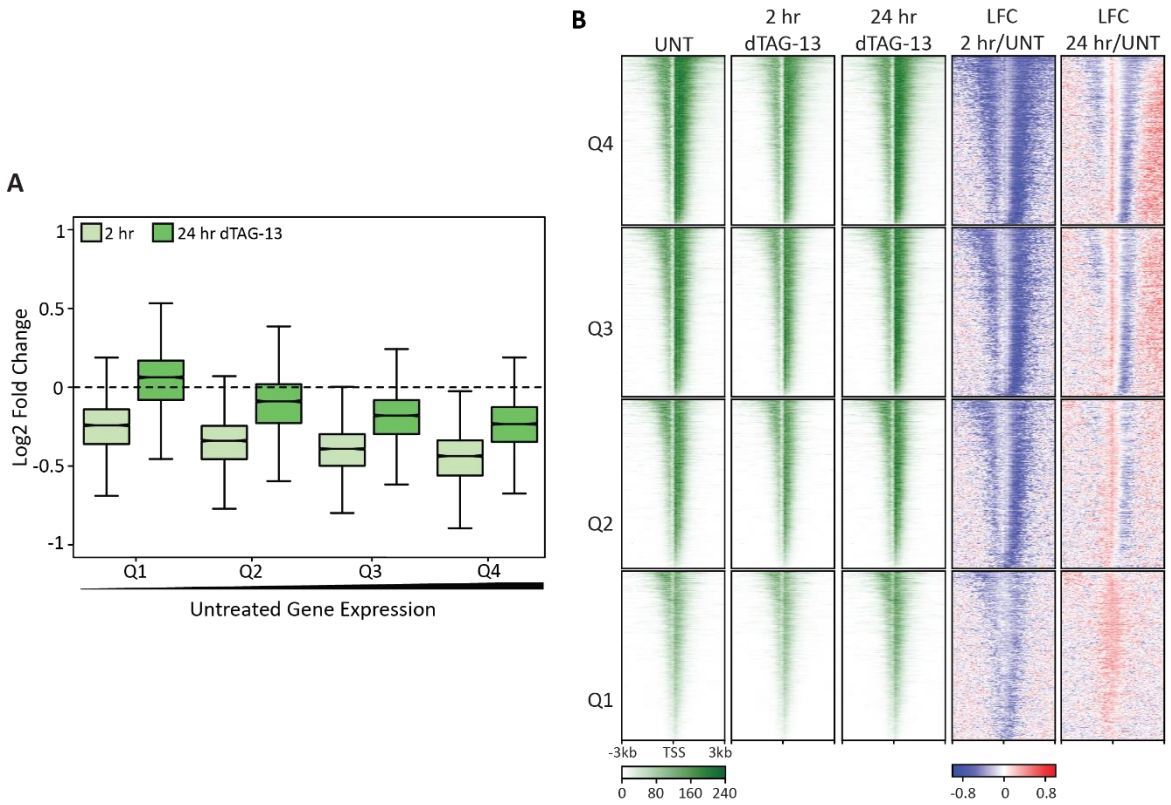


Figure 4.7 Removal of SET1A/B results in the largest reduction in promoter-associated H3K4me3 at highly expressed genes

(A) Box plots of log2 fold change in H3K4me3 cChIP-seq signal at promoter-associated H3K4me3 peaks after 2 hr and 24 hr of SET1A/B removal (n=14546). Genes are split into quartiles of equal size based on gene expression level in untreated dTAG-SET1A/B cells.

(B) Heatmaps of H3K4me3 cChIP-seq signal at each gene with a promoter-associated H3K4me3 peak in untreated cells (n=14536), split into the same expression quartiles as in (A). H3K4me3 cChIP-seq signal in untreated cells and dTAG-SET1A/B cells treated with dTAG-13 for 2 hr and 24 hr is shown. The log2 fold change (LFC) in H3K4me3 at 2 hr and 24 hr compared to untreated cells is also shown. Heatmaps are sorted by mean signal in untreated cells.

4.2.3 Gene expression changes cannot be explained by changes in H3K4me3

Having established that SET1A/B positively regulate gene expression genome-wide (Chapter 3) and make a small contribution to promoter-associated H3K4me3, I next wanted to examine whether the gene expression changes observed following removal of SET1A/B correlated with the changes in H3K4me3. Interestingly, these two features were very poorly correlated at 2 hours of SET1A/B removal, while by 24 hours a modest correlation was observed (Figure 4.8 A and B). By looking at individual genes, I could find examples whereby H3K4me3 was relatively depleted with little

change in gene expression, or where gene expression was greatly reduced but H3K4me3 unchanged (Figure 4.8 C and D and Figure A7). This was also the case when SET1A only was removed (Figure A8) and suggests that SET1A/B might regulate gene expression independently of H3K4me3, as has previously been proposed (Clouaire *et al.*, 2012; Brown *et al.*, 2017; Sze *et al.*, 2017; Murray *et al.*, 2019).

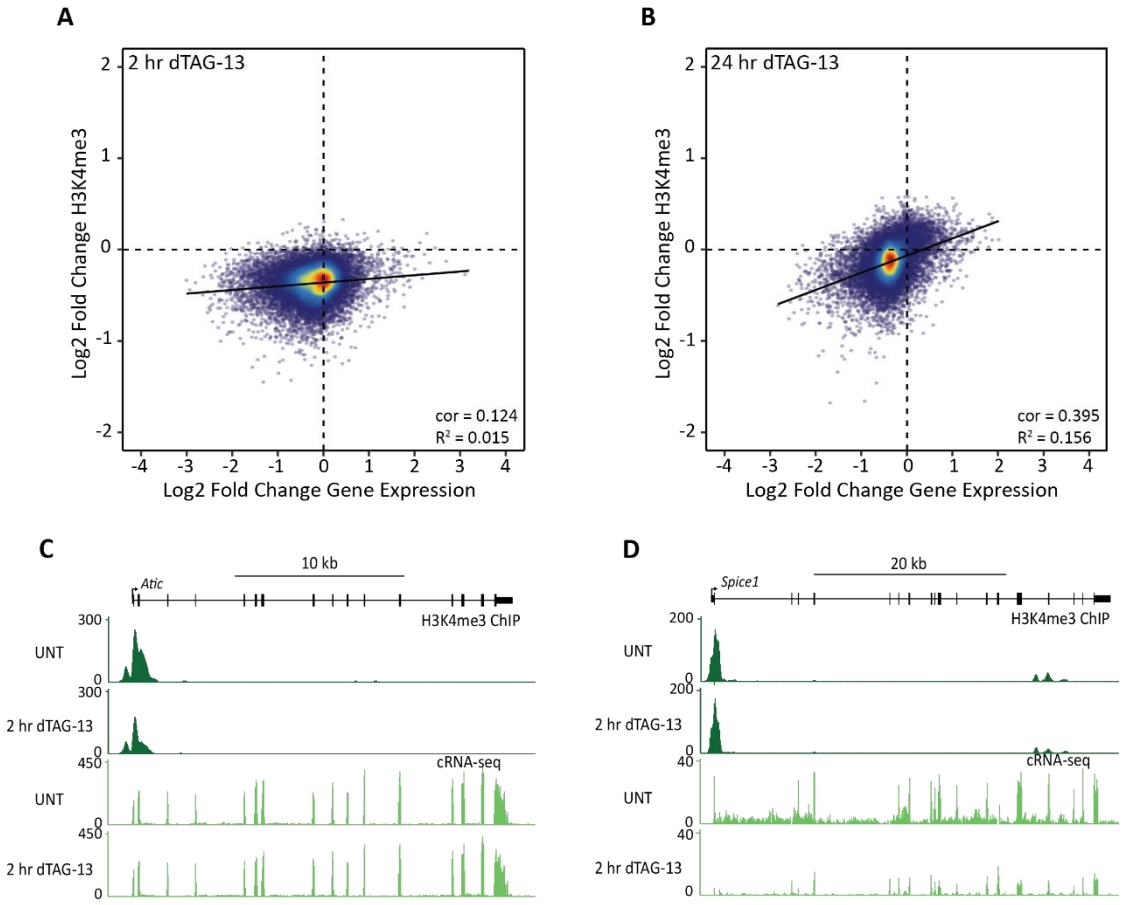


Figure 4.8 Changes in H3K4me3 and gene expression were not strongly correlated at either time point

(A) and (B) Scatterplots comparing the changes in gene expression (cRNA-seq) and H3K4me3 cChIP-seq signal at promoter-associated H3K4me3 peaks in dTAG-SET1A/B cells treated with dTAG-13 for 2 hr (A) or 24 hr (B). Only genes that have a promoter-associated H3K4me3 peak in untreated cells are included (n=14546). The dashed lines indicate log2 fold change=0, while the solid line represents linear regression. Cor denotes the Pearson correlation coefficient and R^2 the coefficient of determination for linear regression.

(C) and (D) Genomic snapshots showing H3K4me3 cChIP-seq signal and cRNA-seq signal for a gene with reduced H3K4me3 and unchanged expression (*Atic*, C) and a gene with reduced expression but unchanged H3K4me3 (*Spice1*, D) following depletion of SET1A/B for 2 hr.

4.2.4 Generating a SET1A conditional catalytic point mutant

The widespread reduction in gene expression after only 2 hours of SET1A/B depletion, accompanied by very subtle changes in H3K4me3, suggests that SET1A/B may regulate gene expression independently of H3K4me3. However, directly testing the contribution of H3K4me3 to these gene expression outcomes is technically challenging. One way to examine if SET1A/B regulate gene expression through H3K4me3 could be to ask whether their catalytic activity is required. To answer this question, I employed a Cre-LoxP system that enables the inducible conversion of WT SET1A into a catalytically inactive form of the protein (Blackledge *et al.*, 2020). Since there is not a high resolution structure of SET1A, there remains limited information about specific mutations that might abolish its catalytic activity. I therefore chose to use two mutations found within the proposed SAM co-ordinating pocket of the active site, N1655S and Y1693F (Zhang *et al.*, 2002; Weirich *et al.*, 2015; Yanjing Li *et al.*, 2016). Importantly, these two residues are highly conserved and have previously been shown to disrupt catalysis of other SET domain-containing proteins (Figure 4.9 A; Zhang *et al.*, 2002; Williamson *et al.*, 2013; Weirich *et al.*, 2015; Wang *et al.*, 2019). In order to confirm that these two mutations rendered SET1A catalytically inactive, I utilised a cell line (TOT2N) in which a cassette containing fourteen Tet repressor DNA binding sites (TetO) have been engineered into a single site on chromosome 8 in ESCs (Blackledge *et al.*, 2014; Dimitrova *et al.*, 2018). Fusing a protein of interest to the reverse Tet Repressor (rTetR) enables its recruitment to the TetO array upon the addition of doxycycline. By performing H3K4me3 ChIP-qPCR using primers across the TetO array, I found that tethering of rTetR-Full length (FL) SET1A to the TetO array resulted in the deposition of H3K4me3 (Figure 4.9 B). However, this was not the case when SET1A contained the N1655S/Y1693F mutations. Importantly, the expression of the rTetR-fusion proteins was similar, confirming that this was not a result of protein expression and that the N1655S and Y1693F mutations render SET1A catalytically inactive (Figure 4.9 C).

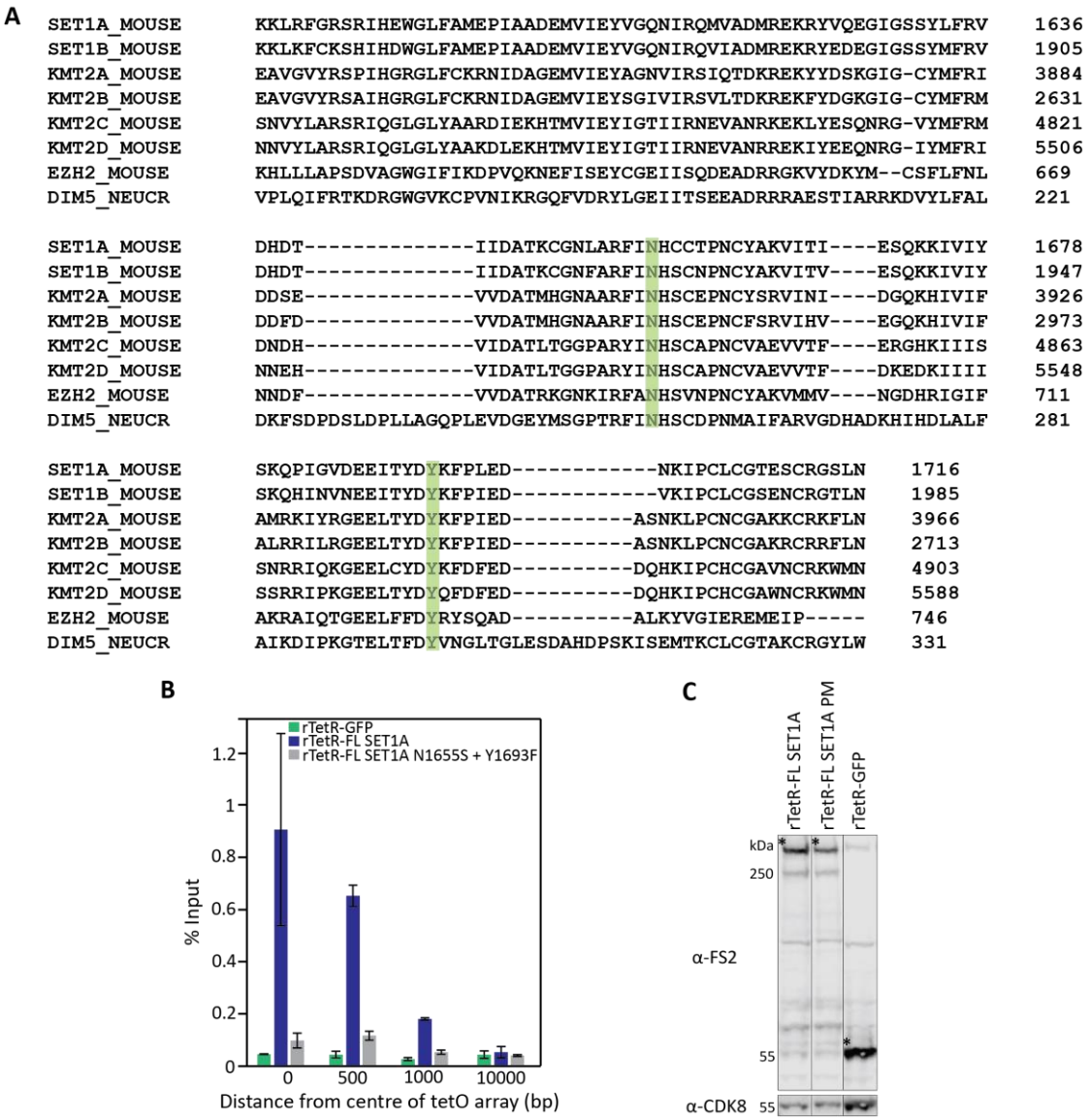


Figure 4.9 Identifying SET1A catalytic mutations

(A) Multiple sequence alignment of the SET domain of mouse SET1A against the five other mouse H3K4 HMTs, as well as mouse EZH2 and *Neurospora* DIM5, from which the catalytic mutations were identified. The two residues mutated in the SET1A catalytic point mutant are highlighted. Alignments were generated with ClustalW (Madeira *et al.*, 2019).

(B) H3K4me3 ChIP-qPCR cross the TetO array following recruitment of rTetR-GFP, rTetR-full length SET1A (FL) or rTetR-FL SET1A N1655S + Y1693F. Error bars represent SEM from two biological replicates.

(C) Western blot of nuclear extract from TOT2N cells transiently transfected with the rTetR-GFP, rTetR-FL SET1A or rTetR-FL SET1A N1655S + Y1693F (PM) fusion proteins. Blot was probed with FLAG-StrepII antibody, which recognises the tags that were fused to the rTetR. CDK8 is a loading control. * denotes the specific rTetR fusion from each transfection.

To set up the SET1A conditional point mutant (CPM) system, I used an ESC line in which ERT2-Cre recombinase is expressed from the *Rosa26* locus. Since the two catalytic mutations are in the final

two exons of SET1A, I fused the two mutated exons together and used CRISPR/Cas9-mediated HDR to insert them into the intron preceding the second-to-last exon in the antisense orientation (Figure 4.10). The whole cassette was then flanked by inverted double LoxP/Lox2272 sites. In the absence of tamoxifen, the mutated exons would be spliced out from the transcript, resulting in the expression of WT SET1A. Upon the addition of tamoxifen, ERT2-Cre recombinase would enter the nucleus and cause recombination between the compatible inverted LoxP sites, causing inversion of the cassette. The remaining pair of LoxP sites would then be in the same orientation, leading to recombination and deletion of the intervening sequence, locking the cassette into place. Insertion of a strong splice acceptor just upstream of the mutated exons should cause a splicing event that results in the expression of the catalytically inactive form of SET1A from the endogenous locus.

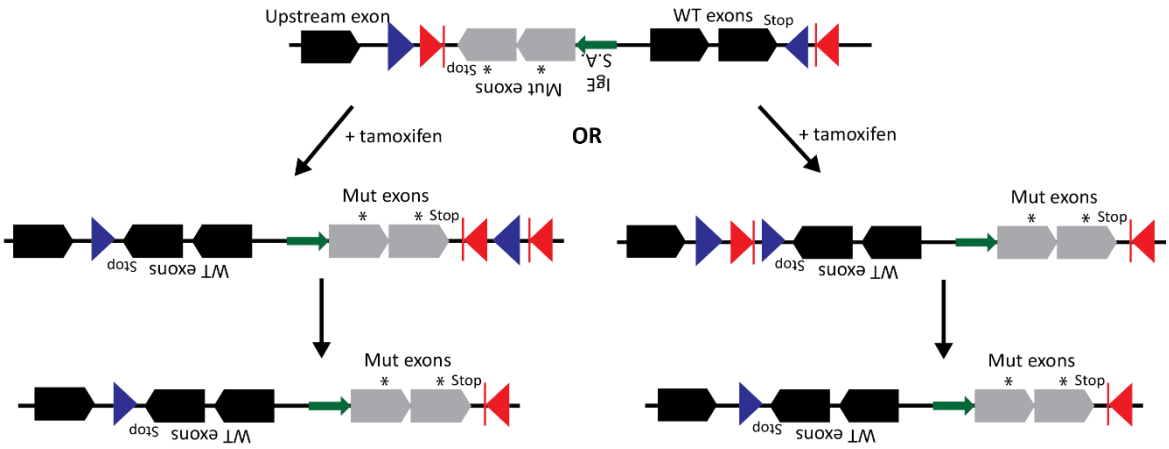


Figure 4.10 Schematic of a conditional SET1A catalytic point mutant
Schematic showing the second-to-last *Setd1A* intron targeted with the final two exons containing the N1655S and Y1693F catalytic mutations in the antisense orientation, before and after tamoxifen addition. Exons are shown by boxes and LoxP sites by triangles. Stop is the transcription termination site and IgE S.A. a strong splice acceptor.

After confirming that the conditional catalytic point mutant cassette and 3' LoxP sites were successfully targeted into ERT2-Cre recombinase-expressing ESCs by PCR screening of genomic DNA, I wanted to test whether addition of tamoxifen caused flipping of the cassette and inversion of the mutated exons into the sense orientation. To do this, I performed RT-qPCR using primers binding within the WT exons, which would only give a product in untreated cells, or primers binding

within the mutated exons, which would only give a product upon cassette inversion (Figure 4.11 A and B). This revealed that the cassette was inverting as predicted and that the mutated exons were being spliced onto the upstream WT exon. Therefore, the mutated *Setd1A* RNA was being transcribed from its endogenous locus. Unfortunately, western blot analysis of SET1A protein levels revealed that the levels of the mutant form of SET1A were dramatically reduced in comparison to WT SET1A (Figure 4.11 C). Importantly, RT-qPCR analysis using primers upstream of the CPM cassette indicated that the levels of mutated *Setd1A* transcript were comparable to WT *Setd1A* (Figure 4.11 D). Therefore, the reduction in the levels of SET1A was occurring at the protein level. I tried a number of different approaches to alleviate this problem, including using different combinations of catalytic mutations and different systems to induce conversion to the mutant protein, but levels of mutated SET1A were always reduced in comparison to WT. This suggests that the catalytically inactive form of SET1A is unstable.

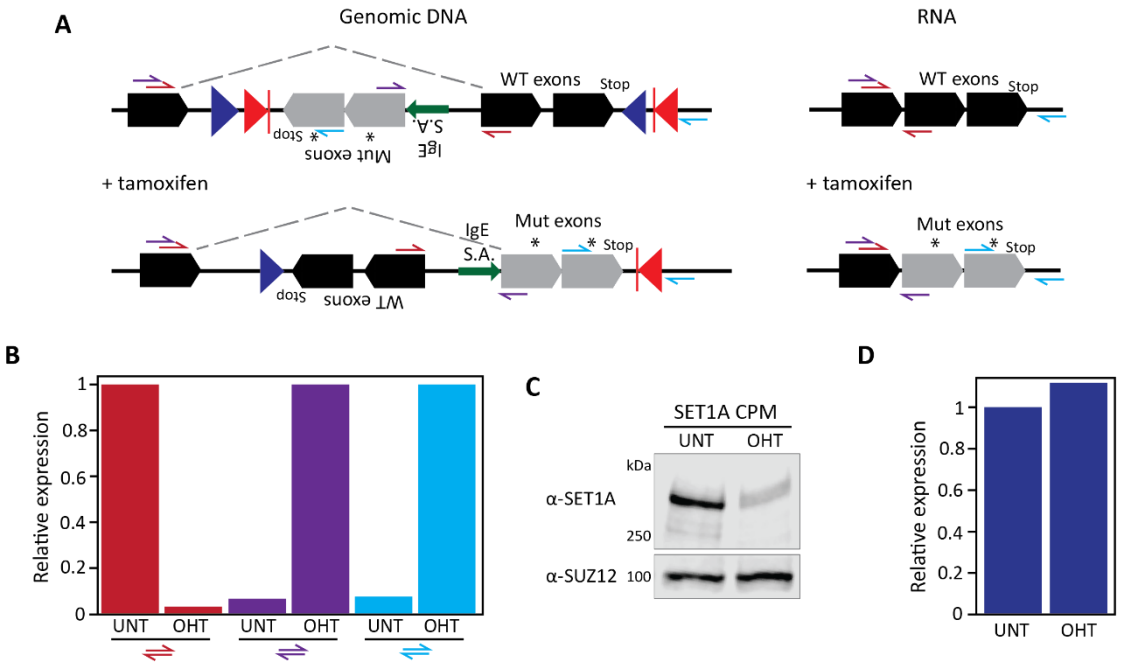


Figure 4.11 Characterising the SET1A conditional catalytic point mutant
(A) Schematic of the PCR screening strategy for the SET1A CPM line. The genomic DNA showing the final *Setd1A* intron and final three exons before and after tamoxifen treatment is shown on the left, and the corresponding mRNA on the right.
(B) RT-qPCR analysis of the SET1A CPM before (UNT) and after (OHT) tamoxifen treatment. The location of the primer pairs is shown in (A). One primer pair (red) matches to the WT allele and should not amplify upon tamoxifen treatment. The remaining two primer pairs (purple and blue) should only

amplify if the point mutant cassette flipped after tamoxifen treatment. Expression is normalised to the RNAPIII-transcribed tRNA-Lys and expressed relative to either untreated (WT exons) or tamoxifen treated (mutated exons).
(C) Western blot showing levels of SET1A in nuclear extract from SET1A CPM cells before (UNT) and after (OHT) tamoxifen treatment. SUZ12 is a loading control.
(D) RT-qPCR analysis of SET1A CPM using primers upstream of the CPM cassette, before (UNT) and after (OHT) tamoxifen treatment. Expression is normalised to the RNAPIII-transcribed tRNA-Lys and expressed relative to untreated levels.

4.2.5 SET1A is a transcriptional activator that can function independently of H3K4me3

SET1A/B clearly regulate gene expression and this activity appears to be uncoupled from H3K4me3. However, attempts to directly examine the requirement for SET1A catalytic activity using an endogenous catalytic point mutation were unsuccessful (Chapter 4.2.4). I therefore decided to use an alternative system to test the ability of catalytically active and inactive forms of SET1A to regulate gene expression. In this synthetic system, the TOT2N cell line was engineered to replace the TetO array with a CG-free cassette containing seven TetO binding sites, an Ef1α promoter and a Firefly luciferase reporter construct (Figure 4.12 A; Anne Turberfield, unpublished; Klug and Rehli, 2006). In this cell line, the ability of rTetR-fusion proteins to activate gene expression can be examined by measuring luciferase levels.

Binding of rTetR-Full length (FL) SET1A to the TetO array resulted in an approximately 18-fold induction of luciferase expression (Figure 4.12 B). This was not the case when rTetR-GFP was recruited to the TetO array as a control. This suggests that SET1A functions as a transcriptional activator and the recruitment of SET1A to an artificial, CG-free gene promoter is sufficient to promote gene expression. Interestingly, tethering of the catalytically inactive form of SET1A induced expression of luciferase similarly to WT SET1A (Figure 4.12 B). Importantly, expression of the rTetR-fusion proteins were all highly comparable, confirming that this result is not a function of rTetR-fusion protein expression (Figure 4.12 C). Therefore, in this synthetic system, SET1A is able to activate transcription independently of its catalytic activity and, therefore, its ability to deposit H3K4me3.

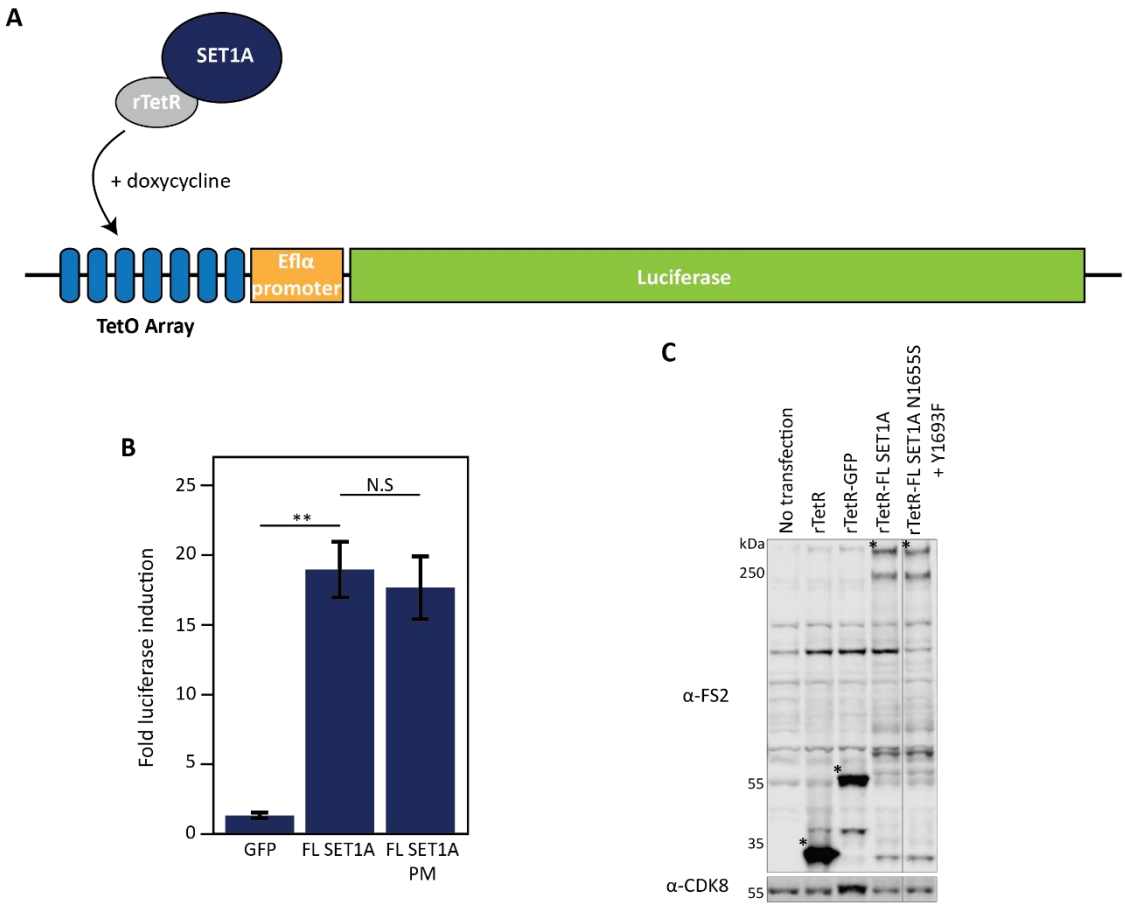


Figure 4.12 SET1A functions as a transcriptional activator independently of its catalytic activity
(A) Schematic of the luciferase reporter assay.
(B) Bar chart showing the fold induction in luciferase levels following tethering of GFP, full length (FL) SET1A or N1655S/Y1693F SET1A catalytic point mutant (PM) to the TetO array. Luciferase expression is normalised to *Renilla* luciferase which was transiently transfected along with the rTetR-fusions to account for transfection efficiency and expressed relative to luciferase levels in the absence of doxycycline. Mean luciferase induction is shown and error bars represent SEM from three biological replicates. N.S represents not significant in a Student's T test ($p>0.05$) and ** represents $p<0.015$.
(C) Western blot of nuclear extract from untransfected luciferase reporter cells or cells transiently transfected with the annotated rTetR-fusion proteins. The blot was probed with FLAG-StrepII antibody, which recognises the tags that were fused to the rTetR. CDK8 is a loading control and * denotes the specific rTetR fusion protein from each transfection.

4.3 Summary and Discussion

The association of H3K4me3 with gene promoters is conserved across eukaryotic evolution. However, the mechanisms that govern H3K4me3 deposition have become increasingly complex in higher eukaryotes (Chapter 1.4). Four mammalian HMTs have been shown to be capable of

trimethylating H3K4 *in vitro*: MLL1/2 and SET1A/B (Shinsky *et al.*, 2015; Meeks and Shilatifard, 2017; Cenik and Shilatifard, 2021; Kwon *et al.*, 2021). SET1A/B have been proposed to be responsible for elevated H3K4me3 at the promoters of highly expressed genes and for placing the majority of H3K4me3 in mammalian genomes (Lee and Skalnik, 2005; Wu *et al.*, 2008; Vethantham *et al.*, 2012; Clouaire *et al.*, 2012; Tang *et al.*, 2013; Bledau *et al.*, 2014; Clouaire, Webb and Bird, 2014; Fang *et al.*, 2016; Brown *et al.*, 2017; Cao *et al.*, 2017; Franks *et al.*, 2017; Yang *et al.*, 2020; Sze *et al.*, 2020). However, the contribution of these complexes to the deposition of H3K4me3 at the genome scale has remained unclear.

By using the rapid protein depletion systems developed in chapter 3, coupled with quantitative H3K4me3 ChIP-sequencing, I found that SET1A/B make a pervasive, but subtle, contribution to promoter-associated H3K4me3 in ESCs (Figure 4.7). The contribution of SET1A/B to H3K4me3 was largest at the promoters of highly expressed genes, in keeping with previous evidence that SET1A/B place H3K4me3 specifically at highly expressed genes and with the enrichment of SET1A at such loci (Clouaire *et al.*, 2012; Clouaire, Webb and Bird, 2014; Brown *et al.*, 2017; Sze *et al.*, 2017, 2020). However, the removal of SET1A/B also resulted in reductions in H3K4me3 at lowly expressed genes, revealing that they have a pervasive role in depositing promoter-associated H3K4me3. This is consistent with the localisation of SET1A across the majority of CGI-associated gene promoters and suggests that SET1A/B place H3K4me3 even at low-occupancy TSSs (Brown *et al.*, 2017).

SET1A/B have previously been proposed to place the majority of H3K4me3 in mammalian cells (Lee and Skalnik, 2005; Wu *et al.*, 2008; Vethantham *et al.*, 2012; Clouaire *et al.*, 2012; Tang *et al.*, 2013; Bledau *et al.*, 2014; Clouaire, Webb and Bird, 2014; Fang *et al.*, 2016; Brown *et al.*, 2017; Cao *et al.*, 2017; Franks *et al.*, 2017; Yang *et al.*, 2020; Sze *et al.*, 2020). However, this assumption is in part based on the observations that their orthologues in yeast and *Drosophila* contribute significantly to H3K4me3 deposition and by examining bulk H3K4me3 by western blotting. By using cChIP-seq, a much more sensitive and quantitative analysis, I observed that SET1A/B make a much more

subtle contribution to levels of H3K4me3 than previously proposed. It is possible that the reason for the modest decrease in H3K4me3 when SET1A/B were removed is simply that H3K4me3 was not turned over during the experiment. While a number of H3K4me3 demethylases exist in mammalian cells, they have predominantly been implicated in fine-tuning levels of H3K4me3 at gene promoters and their contribution to overall levels of H3K4me3 remains unclear (Schmitz *et al.*, 2011; Xie *et al.*, 2011; Gu and Lee, 2013; Kidder *et al.*, 2013; Kidder, Hu and Zhao, 2014; He and Kidder, 2017). Interestingly, extensive nucleosome remodelling and eviction of H3K4me3-modified nucleosomes at TSSs during transcription initiation has been implicated in the turnover of H3K4me3 (Zheng *et al.*, 2014). Given that I observed a widespread reduction in gene expression upon SET1A/B depletion, it is possible that levels of gene transcription were reduced, which could therefore reduce nucleosome eviction and result in slower H3K4me3 turnover. However, while slow removal of H3K4me3 could mask H3K4me3 changes at early time points of SET1A/B removal, such an explanation is not consistent with the observed subtle effect following longer-term culture without SET1A/B. Given that the estimated half-life of H3K4me3 in HeLa cells is 7 hours (Zheng *et al.*, 2014), one might predict that depleting SET1A/B for 24 hours in ESCs, which divide more rapidly in culture, would result in a reduction in SET1A/B-deposited H3K4me3 of approximately 85%. Additionally, I observed only subtle changes in H3K4me3 using a classical SET1A conditional knockout cell line, where SET1A was completely depleted after 96 hours of tamoxifen treatment and would, therefore, be a more similar system to those used in previous studies (Figure A9; Seibler *et al.*, 2003). There was also no major change in the levels of H3K4me3 measured by western blotting in the dTAG-SET1A/B line up to 96 hours after SET1A/B removal (Figure 4.5). If SET1A/B are responsible for depositing the majority of H3K4me3, as has been previously proposed, even without active demethylation and nucleosome exchange, the levels of H3K4me3 should be naturally diluted during cell division over such a timescale. While cell growth was affected by SET1A/B depletion, the cells were still actively proliferating. Additionally, I actually observed that levels of H3K4me3 recovered to some extent during prolonged culture without SET1A/B (Figure

4.6 and 4.7). This indicates that other H3K4 HMTs must actively contribute to H3K4me3 deposition in the absence of SET1A/B.

The contribution of SET1A/B to H3K4me3 deposition was found to be more modest than expected based on the prevailing views in the field. This indicates that other H3K4 HMTs must play more significant roles in the deposition of H3K4me3 than previously appreciated. MLL1/2 associate with CGIs genome-wide via their ZF-CXXC domains, and have been proposed to primarily contribute to H3K4me3 deposition at lowly expressed genes (Guenther *et al.*, 2005; P. Wang *et al.*, 2009; Hu, Garruss, *et al.*, 2013; Denissov *et al.*, 2014; Hu *et al.*, 2017; Mas *et al.*, 2018; Sze *et al.*, 2020). This could explain why I observed a less dramatic reduction in H3K4me3 at lowly expressed genes in the absence of SET1A/B (Figure 4.7). However, I also observed that the removal of SET1A/B did not abolish the enrichment of H3K4me3 at highly expressed genes. On the contrary, following longer-term depletion of SET1A/B, levels of H3K4me3 at least partially recovered at such loci. Therefore, it is possible that in the absence of SET1A/B, MLL1/2 could be responsible for maintaining H3K4me3 at highly expressed genes, as well as at lowly expressed genes (Sze *et al.*, 2020). Alternatively, other H3K4 HMTs, such as MLL3/4, might contribute to H3K4me3 deposition in mammals. While MLL3/4 predominantly monomethylate H3K4 *in vitro*, there is evidence that these enzymes retain some residual trimethyltransferase activity (Lee *et al.*, 2013; Tang *et al.*, 2013; Kwon *et al.*, 2021). Furthermore, the depletion of dTrr was previously observed to result in global reductions in H3K4me3 in *Drosophila* (Ardehali *et al.*, 2011). In the future, it will be important to determine whether MLL3/4 can place H3K4me3 *in vivo* and contribute to the deposition of H3K4me3 at gene promoters. Furthermore, additional mammalian H3K4 HMTs that fall outside of the SET1/MLL family have been identified (Wang *et al.*, 2001; Nishioka *et al.*, 2002; Hamamoto *et al.*, 2004). It will be important to determine whether these HMTs, or additional as-yet undiscovered H3K4 HMTs, make a substantial contribution to the levels of H3K4me3 in mammals, explaining the findings presented here.

In light of my findings, our understanding of the mechanisms that dictate H3K4me3 deposition in mammalian cells will need to be revisited. Clearly, the contribution of SET1A/B to H3K4me3 is much more limited than previously proposed and other H3K4 HMTs must significantly contribute or compensate in the absence of SET1A/B. To further understand the mechanisms that govern H3K4me3 deposition in mammalian cells, the contribution of MLL1/2 to H3K4me3 genome-wide will need to be examined using similar degron systems and calibrated H3K4me3 ChIP-seq. Additionally, creation of a SET1A/B/MLL1/2 degron line would help to determine whether these are the predominant H3K4 HMTs in mammals and, if not, provide a tool with which to identify other putative H3K4me3 HMTs. If there are as-yet unidentified H3K4me3 HMTs, this would suggest that this system has expanded and diversified in mammals even more extensively than previously thought.

The observation that the simultaneous removal of SET1A/B resulted in a larger reduction in H3K4me3 at highly expressed genes than SET1A removal alone suggests that SET1B can place H3K4me3 at these loci in the absence of SET1A (Figure 4.4 and 4.7). Indeed, although the contribution of SET1B to the deposition of H3K4me3 has been proposed to be relatively minor, it has previously been observed to compensate for SET1A loss in H3K4me3 deposition (Wu *et al.*, 2008; Bledau *et al.*, 2014; Wang *et al.*, 2017; Sze *et al.*, 2020). It will be important to determine whether the role of SET1B in H3K4me3 deposition at highly expressed genes is relevant in the steady state or is merely a consequence of SET1A removal and subsequent compensation. Interestingly, I found that the contribution of SET1B to H3K4me3 was greatest downstream from TSSs, towards gene bodies. The recruitment of SET1A/B to chromatin has previously been proposed to be coupled with transcription, possibly via a direct interaction between WDR82 and RNAPII (Lee and Skalnik, 2008; Lee *et al.*, 2010; Ebmeier *et al.*, 2017; Franks *et al.*, 2017). While the contribution of this interaction to genome-wide SET1A/B occupancy remains to be determined, such an interaction could explain the observed contribution of SET1B to H3K4me3 deposition downstream of TSSs.

The removal of SET1A/B results in dramatic and widespread reductions in gene expression, as well as a subtle, but pervasive reduction in H3K4me3. Given that H3K4me3 has been extensively linked to active gene expression, one could predict that the changes in gene expression were due to changes in H3K4me3. However, there was only a moderate correlation between gene expression change and change in H3K4me3 (Figure 4.8). This suggests that, in the context of SET1A/B removal, there is no simple relationship between changes in H3K4me3 and gene expression. This could be explained by a context-dependent role of H3K4me3, which might render some genes more sensitive to the loss of H3K4me3 than others and would, therefore, not result in a discernible correlation between these effects (Cano-Rodriguez *et al.*, 2016). It is also possible that H3K4me3 may contribute to some other feature of transcription that is not captured in the cRNA-seq analysis.

Despite the modest effect of SET1A/B on H3K4me3 and the lack of a clear correlation between changes in H3K4me3 and gene expression, I was nevertheless keen to examine whether there was any connection between the catalytic activity of SET1A/B and gene expression. To examine this, I tried to create a conditional catalytic point mutant in SET1A to determine whether loss of its catalytic activity recapitulated the gene expression changes observed upon complete protein removal. Unfortunately, mutated SET1A created in this way proved to be unstable (Figure 4.11). In order to revisit whether SET1A functions through H3K4me3 deposition in the future, one might try using different mutations to the ones used here to generate a more stable SET1A catalytic point mutant. However, if catalytically inactive SET1A is inherently unstable, such an experiment might not be possible. An important related question is whether H3K4me3 itself has a role in the regulation of gene expression. However, to directly examine the contribution of H3K4me3 to gene expression *in vivo* it would be necessary to develop systems where H3K4me3 can be completely removed, without disturbing the formation of the HMT complexes. For example, one could develop conditional catalytic point mutants for each of the H3K4 HMTs. However, given that I was unable to generate such a system for SET1A, and this approach does not directly examine the role of H3K4me3, further approaches might need to be explored. In *Drosophila*, this question was

addressed by generating H3K4-methylation resistant histones (Hödl and Basler, 2009, 2012). However, given that the histone genes have duplicated multiple times during eukaryotic evolution, performing a similar experiment in mammalian cells would be complex and could also disrupt other H3K4me3-independent functions of the N-terminal tail of H3. Alternatively, one could imagine a system whereby a H3K4me3 demethylase is targeted specifically to gene promoters by dCas9. This could enable the removal of H3K4me3 without disturbing the H3K4 HMT complexes. While H3K4me3 has been shown to stimulate transcription *in vitro* (Lin *et al.*, 2011; Jiang *et al.*, 2013; Lauberth *et al.*, 2013; Tang *et al.*, 2013), understanding the contribution of H3K4me3 to gene expression *in vivo* remains one of the key outstanding questions in the field.

Using a synthetic reporter system I observed that the recruitment of SET1A to a gene promoter is sufficient to promote gene expression (Figure 4.12). This suggests that SET1A may function as an instructive transcriptional activator. Given the high degree of homology between SET1A and SET1B, it will be important to determine if this activity is also conserved within SET1B. Importantly, tethering of GFP did not activate luciferase expression, suggesting that activation is not merely a consequence of tethering a protein, which might influence chromatin accessibility, to the promoter. Furthermore, I observed that the ability of SET1A to activate luciferase was not diminished when it was mutated to be catalytically inactive. Therefore, even if the deposition of H3K4me3 by SET1A does have a role in promoting transcriptional activity, its contribution is negligible in this system. This is important given that I could not definitively establish whether H3K4me3 contributes to the gene expression changes observed when SET1A/B were removed. Further, this experiment suggests that SET1A can activate gene transcription independently of its ability to methylate H3K4me3 or any other potential substrates. Indeed, both SET1A and SET1B have previously been proposed to have functions that are independent of their catalytic activity (Chapter 1.4.5; Sze *et al.*, 2017; Wang *et al.*, 2017). In the remainder of my thesis, I will examine potential mechanisms by which SET1A/B could activate gene expression independently of their catalysis.

5. SET1A/B regulate RNAPII-mediated transcription

5.1 Introduction

In chapter 3 I found that SET1A/B support gene expression in ESCs. However, this activity appears to be uncoupled from H3K4me3 (Chapter 4). Furthermore, using a synthetic reporter system I observed that SET1A can activate reporter gene expression independently of its methyltransferase activity (Figure 4.12). Given that SET1A/B are indispensable for mammalian development and often mutated in human disease, understanding the mechanisms by which they regulate gene expression is essential (Kranz and Anastassiadis, 2020).

SET1A/B are large proteins that share a highly similar domain architecture (Figure 1.5). The N-terminus of each protein mediates their association with WDR82 and also contains a RNA recognition motif (RRM) (Lee *et al.*, 2007; Lee and Skalnik, 2008). Yeast Set1 possesses two RRMs that have been implicated in chromatin occupancy, though the function of the RRM in the mammalian proteins remains unclear (Trésaugues *et al.*, 2006; Battaglia *et al.*, 2017; Luciano *et al.*, 2017; Sayou *et al.*, 2017). The RRM is followed by a large (approximately 1250 amino acid) central linker region that comprises predominantly low complexity sequences and lacks any annotated domain other than a HCF1 binding site in SET1A (Wysocka *et al.*, 2003; Lee *et al.*, 2007). Interestingly however, this central linker region is essential for the function of SET1A in supporting the expression of DNA damage response genes and leukaemia cell growth (Hoshii *et al.*, 2018). This suggests that the central region of SET1A could be an important domain for the activation of gene expression. The SET domain is present at the C-termini of SET1A/B and, along with the N-SET domain, interacts with CFP1 and the WRAD complex (Lee and Skalnik, 2008). However, given that SET1A can activate gene expression independently of its HMT activity (Figure 4.12), and that the SET1A/B complexes do not possess any other known catalytic activity, I hypothesised that they may regulate gene expression by interacting with other proteins that influence transcription.

One way in which SET1A/B could regulate gene expression independently of catalysis is through directly interacting with transcription coactivator complexes at gene promoters. For example, SET1A has been proposed to associate with p300 (Tang *et al.*, 2013). Alternatively, SET1A/B could affect gene expression by interacting with RNAPII. In support of this, the WDR82 subunit of the SET1A/B complexes associates with the S5P form of RNAPII (Lee and Skalnik, 2008; Lee *et al.*, 2010; Ebmeier *et al.*, 2017). It is possible that through binding to RNAPII at gene promoters, WDR82 could contribute to polymerase stability and/or activity at actively transcribed genes. Here I set out to identify the regions of SET1A/B that are essential for the activation of gene expression and to begin to dissect the mechanisms through which activation by SET1A/B is achieved.

5.2 Results

5.2.1 WDR82 is essential for transcription activation by SET1A

In chapter 4, I established that tethering full length SET1A to a gene promoter activates transcription of an associated reporter gene. I therefore decided to use this reporter system to investigate the mechanisms by which SET1A activates gene expression. I generated rTetR-fusion protein constructs containing various truncated forms of SET1A and examined the ability of these fusion proteins to activate reporter gene expression (Figure 5.1 A). Importantly, expression of these rTetR-fusion proteins was comparable to rTetR-FL SET1A (Figure 5.1 B). In chapter 4 I found that catalytically inactive SET1A activated luciferase expression comparatively to WT SET1A. In keeping with this observation, rTetR-SET1A truncated before the N-SET domain also activated luciferase expression, albeit to a slightly lower magnitude than full length SET1A (Figure 5.1 C). This is interesting since the SET1A ΔSET protein would be unable to interact with the WRAD complex or CFP1, in addition to being catalytically inactive (Lee and Skalnik, 2008). Therefore, in this system SET1A can activate gene expression independently of catalysis and its association with WDR5, RBBP5, ASH2L, DPY30 and CFP1. Furthermore, SET1A lacking the RRM or the central linker region also activated luciferase expression (Figure 5.1 C). Together, this suggests that the RRM, central

linker and the SET domain are all dispensable for reporter gene activation. In keeping with this observation, each of these domains alone was not sufficient to activate luciferase expression (Figure A10).

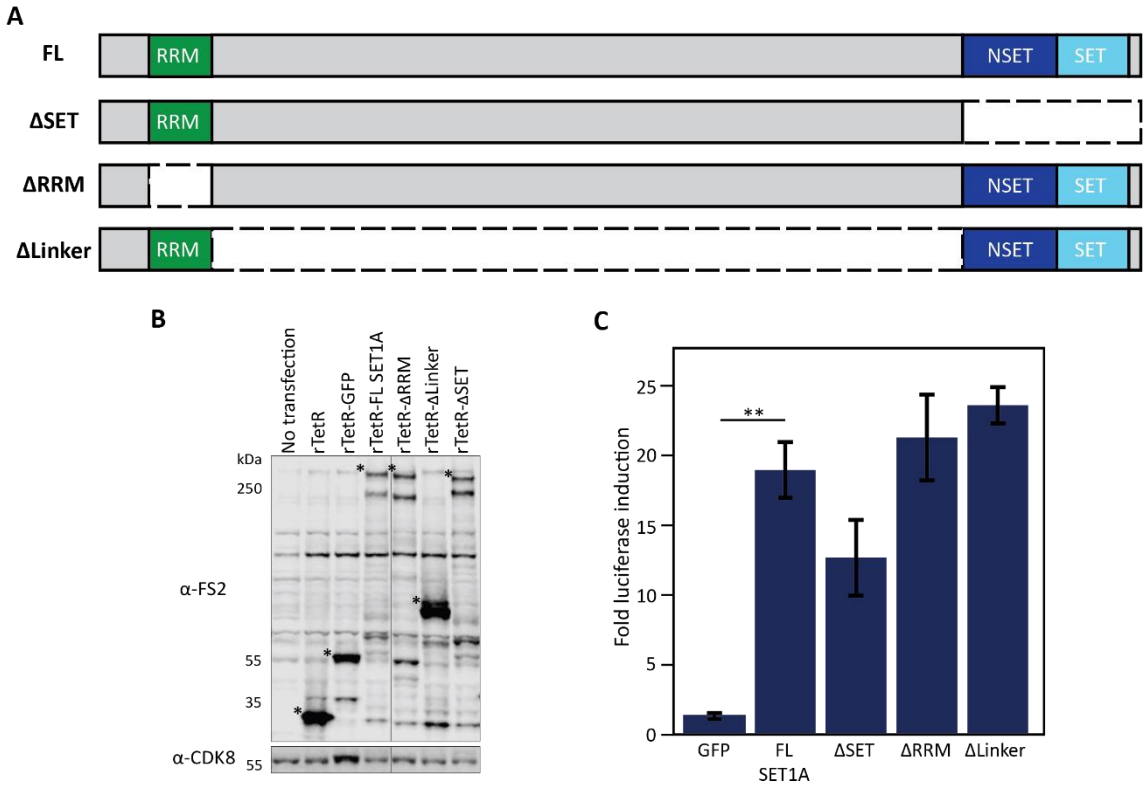
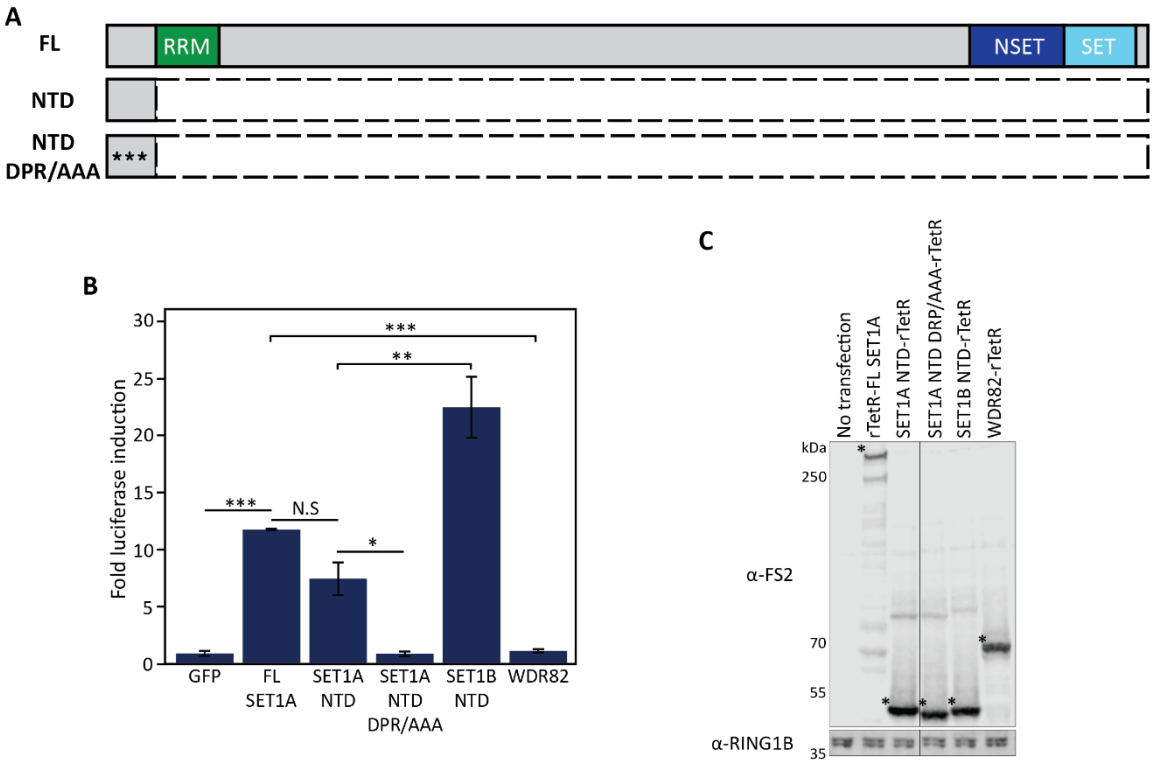


Figure 5.1 The SET domain, RRM and linker are dispensable for transcription activation by SET1A
(A) Schematic of the SET1A constructs tethered to the TetO array.
(B) Western blot of nuclear extract from untransfected luciferase reporter cells or cells transiently transfected with the annotated rTetR-fusion proteins. The blot was probed with FLAG-StrepII antibody, which recognises the tags that were fused to the rTetR. CDK8 is a loading control and * denotes the specific rTetR fusion protein from each transfection.
(C) Bar chart showing the fold induction in luciferase levels following tethering of GFP, full length (FL) SET1A, SET1A ΔSET, ΔRRM and ΔLinker to the TetO array. Luciferase expression is normalised to *Renilla* luciferase which was transiently transfected along with the rTetR-fusions to account for transfection efficiency and expressed relative to luciferase levels in the absence of doxycycline. Mean luciferase induction is shown and error bars represent SEM from three biological replicates. ** represents $p<0.015$ by Student's T test. If not indicated, changes were not significant ($p>0.05$).

The luciferase reporter assays indicated that there is no single domain of SET1A from the RRM to its very C-terminus that is necessary or sufficient to activate reporter gene expression. Therefore, I tested the ability of the N-terminal 84 amino acids of SET1A (N-terminal domain, NTD) to activate

luciferase expression. This region corresponds to the initial methionine up to the beginning of the RRM. Surprisingly, the SET1A NTD activated luciferase expression, suggesting that this region of SET1A is sufficient to activate transcription (Figure 5.2 A-C). The SET1A NTD does not possess any known catalytic function, but it does interact with WDR82 (Lee *et al.*, 2007; Lee and Skalnik, 2008). In order to test the significance of this interaction, I introduced three mutations (D62A, P63A, R64A) into the SET1A NTD that render it unable to associate with WDR82 (Jessica Kelley, Figure A11) and then tested its ability to activate luciferase expression (Figure 5.2 B). Interestingly, the SET1A NTD did not activate gene expression when mutated in this way. Therefore, the ability of the SET1A NTD to activate reporter gene expression is dependent on its ability to associate with WDR82.

SET1B also interacts with WDR82 and SET1A and SET1B share extensive sequence similarity within their N-terminal domains (Figure 5.2 D; Lee *et al.*, 2007). Therefore, I wanted to examine whether the ability of the SET1A NTD to activate reporter gene expression is conserved within SET1B. The SET1B NTD, again defined as the region from methionine 1 to the start of the RRM, activated gene expression (Figure 5.2 B). In fact, the activation by the NTD of SET1B was greater than that observed for SET1A. While this could be due to technical variability, such as differences in the expression of the SET1A and SET1B NTD proteins, this may also suggest that SET1B is a more potent transcriptional activator than SET1A. This would be surprising since the depletion of SET1B resulted in much more subtle effects on gene expression than the removal of SET1A (Chapter 3). Interestingly, tethering WDR82 alone did not activate luciferase expression, suggesting that WDR82 itself is not sufficient to activate transcription (Figure 5.2 B). Overall, these experiments suggest that the N-terminal domains of SET1A and SET1B are sufficient to activate transcription, and that an association with WDR82 is essential for this outcome, at least in the context of SET1A.



D

```
SET1A       -----------MDQEGGGDGQKAPSFQWRNYKLIVDPALDPALRRPSQKVYRYDGVHFSV   49
SET1B       MENSHPHHHHQQPPPQPGPSGERRNHHWRSYKLMIDPA----LKKGHHKLYRYDGQHFSL   56

SET1A       SDSKYTPVEDLQDPRCH-VRSKARDFSLPVPKFKLD     84
SET1B       AMSSNRPVEIVEDPRVVGIWTKNKELELSVPKFKID     92
```

Figure 5.2 The interaction with WDR82 is essential for transcription activation by SET1A

(A) Schematic of the SET1A constructs tethered to the TetO array.

(B) Bar chart showing the fold induction in luciferase levels following tethering of GFP, full length (FL) SET1A, SET1A NTD, SET1A WDR82 binding mutant NTD, SET1B NTD and WDR82 to the TetO array. Luciferase expression is normalised to *Renilla* luciferase which was transiently transfected along with the rTetR-fusions to account for transfection efficiency and expressed relative to luciferase levels in the absence of doxycycline. Mean luciferase induction is shown and error bars represent SEM from three biological replicates. N.S represents not significant in a Student's T test ($p>0.05$), *** represents $p<0.0002$, ** represents $p<0.015$ and * represents $p<0.05$.

(C) Western blot of nuclear extract from untransfected luciferase reporter cells or cells transiently transfected with the annotated rTetR-fusion proteins. The blot was probed with FLAG-StrepII antibody, which recognises the tags that were fused to the rTetR. RING1B is a loading control and * denotes the specific rTetR fusion protein from each transfection.

(D) Multiple sequence alignment of the SET1A and SET1B NTDs. Identical residues are coloured green and similar residues coloured blue. The three residues mutated in SET1A to render it unable to associate with WDR82 are highlighted. Alignments were generated with ClustalW (Madeira *et al.*, 2019).

Having established that the interaction with WDR82 is required for reporter gene activation by SET1A, I next wanted to examine whether WDR82 is required for gene expression in a native

context. I therefore generated a WDR82 degron cell line by using CRISPR/Cas9-mediated HDR to tag WDR82 with a C-terminal dTAG in WT ESCs. Expression of WDR82-dTAG was comparable to WT WDR82, and the addition of dTAG-13 induced complete WDR82 depletion within 2 hours (Figure 5.3 A). However, levels of SET1A were reduced following short-term depletion of WDR82, as observed previously in long-term perturbation systems (Lee and Skalnik, 2008; Wu *et al.*, 2008; Austenaa *et al.*, 2015; Franks *et al.*, 2017). In order to examine whether the removal of WDR82 resulted in gene expression outcomes similar to those observed following the depletion of SET1A/B, I analysed the expression of a panel of SET1A/B target genes by RT-qPCR (Figure 5.3 B). The depletion of WDR82 or SET1A/B resulted in a similar reduction in expression of approximately half of the analysed genes. Importantly, these gene expression changes could be confounded by the reductions in the levels of SET1A when WDR82 is removed. Interestingly, however, there were also several genes (*Fam175a*, *Ears2*, *Rfesd*) whose expression was decreased in the absence of SET1A/B, but not WDR82. Therefore, despite WDR82 being essential for transcriptional activation by SET1A/B in a synthetic reporter system, the removal of WDR82 does not completely phenocopy the removal of SET1A/B.

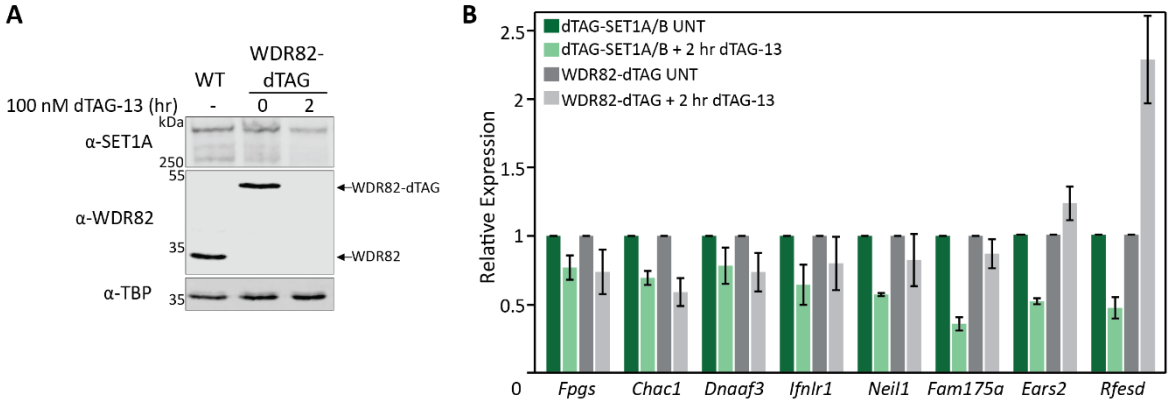


Figure 5.3 Removal of WDR82 does not phenocopy SET1A/B depletion
(A) Western blot showing levels of WDR82 and SET1A in nuclear extracts from WT ESCs and WDR82-dTAG cells untreated and treated with dTAG-13 for 2 hr. TBP is a loading control.
(B) RT-qPCR analysis to examine gene expression following 2 hr depletion of SET1A/B or WDR82. Expression levels of the SET1A/B target genes were normalised to the RNAPIII-transcribed *U6 snRNA* and shown relative to the levels in untreated cells. Error bars represent SEM from three biological replicates.

5.2.2 SET1A/B regulate RNAPII at gene promoters

I have established that SET1A/B have a widespread role in supporting gene expression in ESCs and that their ability to activate reporter gene expression is dependent on their interaction with WDR82. Since WDR82 directly interacts with RNAPII (Lee and Skalnik, 2008; Lee *et al.*, 2010; Ebmeier *et al.*, 2017), I wanted to examine whether the removal of SET1A/B affects the function of RNAPII. At first, I used the dTAG-SET1A/B cell line to examine the bulk levels of total RNAPII following SET1A/B depletion. Levels of total RNAPII remained largely unchanged, indicating that the overall stability of RNAPII was not dramatically affected by the removal of SET1A/B (Figure 5.4).

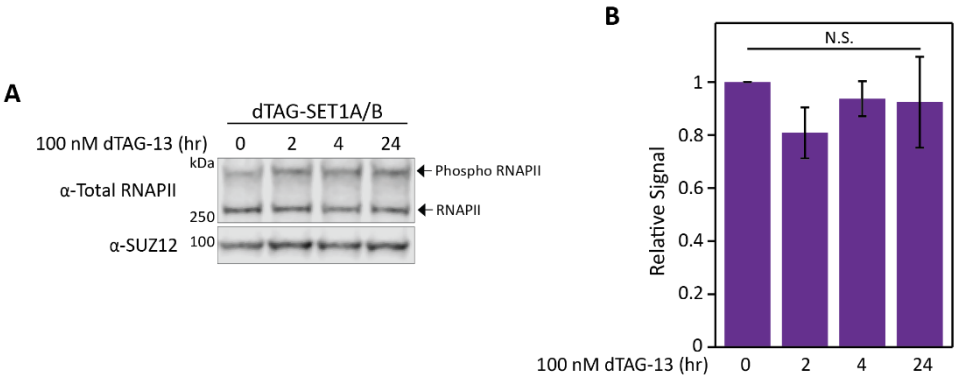


Figure 5.4 Levels of RNAPII were not significantly affected following SET1A/B depletion
(A) Representative western blot of nuclear extracts showing the levels of total RNAPII from dTAG-SET1A/B cells treated with dTAG-13 over a 24 hr time course. SUZ12 is a loading control.
(B) Quantification of western blots of total RNAPII signal from dTAG-SET1A/B cells treated with dTAG-13 over a 24 hr time course. Signal is normalised to SUZ12 and expressed relative to untreated. Error bars show SEM from two biological replicates. N.S. represents not significant in a paired Student's T test ($p>0.05$).

Although total levels of RNAPII were not significantly affected by the depletion of SET1A/B, I wanted to examine whether its occupancy on chromatin at individual genes was affected. To this end, I performed cChIP-seq for total RNAPII in the dTAG-SET1A/B cell line following 2 hours of SET1A/B depletion. As previously described, the RNAPII distribution in untreated cells was characterised by a strong enrichment at TSSs with lower levels throughout gene bodies (Figure 5.5 A). The removal of SET1A/B resulted in a subtle decrease in RNAPII occupancy across the majority of gene promoters, accompanied by a very subtle widening in the RNAPII peak (Figure 5.5 B and

C). Genes with significantly reduced expression typically experienced the greatest reduction in promoter-associated RNAPII (Figure 5.5 D). However, promoter-associated RNAPII was also slightly reduced on average even at genes that were not defined as significantly decreased by cRNA-seq. Therefore, there was a subtle, but widespread reduction in promoter-associated RNAPII following the depletion of SET1A/B. These observations suggest that SET1A/B make a small contribution to the regulation of RNAPII occupancy at gene promoters.

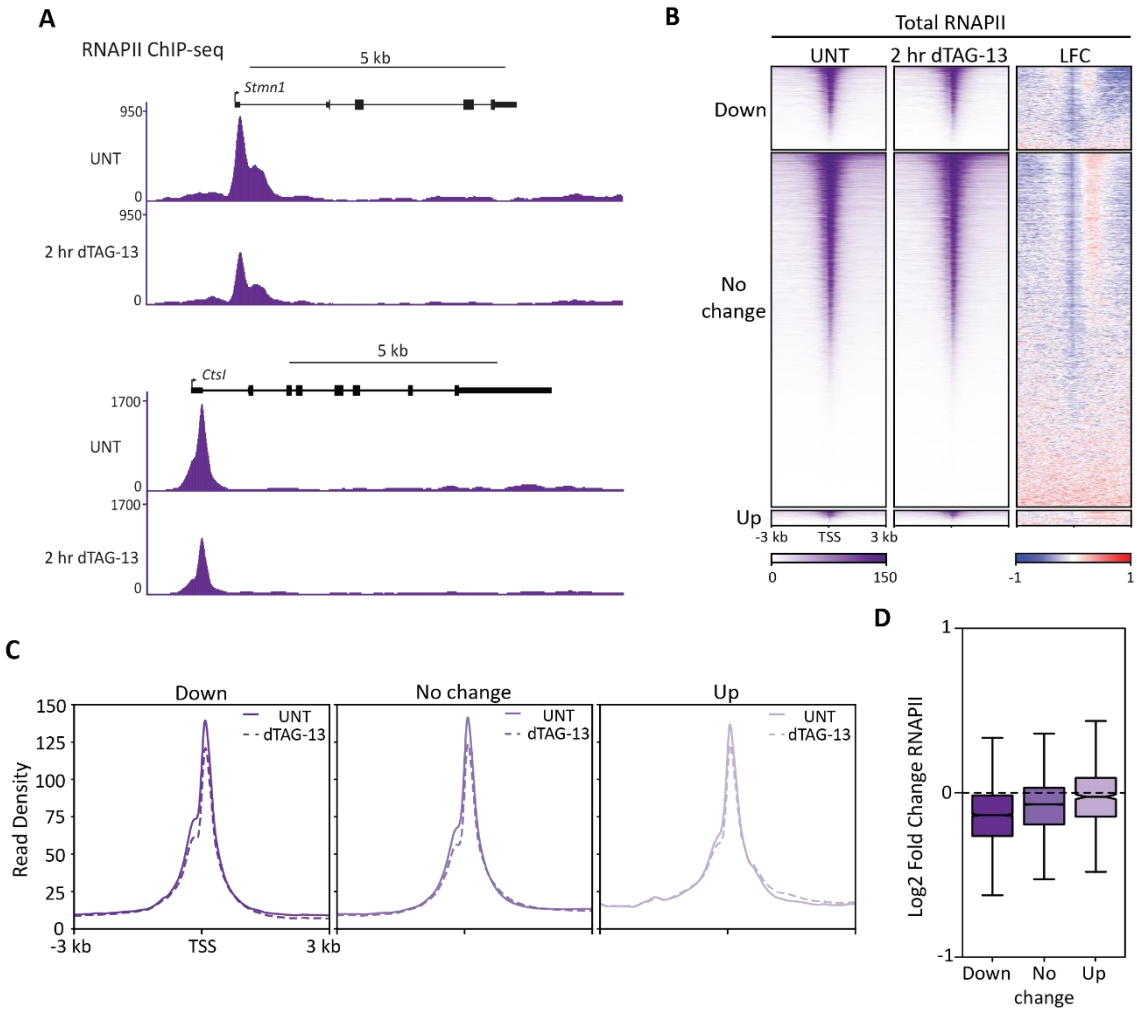


Figure 5.5 SET1A/B depletion results in a subtle but pervasive reduction in promoter-associated RNAPII

(A) Genomic snapshots showing total RNAPII cChIP-seq signal at two actively transcribed genes in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells.

(B) Heatmaps of total RNAPII cChIP-seq signal at all TSSs in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells. The log2 fold change (LFC) in total RNAPII is also shown. Genes are split into those that experienced a significant reduction (Down, n=3804), increase (Up, n=683), or no change (No change, n=16146) in expression based on cRNA-seq analysis. Intervals are sorted by total RNAPII cChIP-seq signal in untreated cells.

(C) Metaplots of total RNAPII cChIP-seq signal at TSSs in untreated cells (UNT) and 2 hr dTAG-13-treated dTAG-SET1A/B cells. Genes are split into the same groups as in (B).
(D) Box plots of the log2 fold change in total RNAPII cChIP-seq signal at gene promoters (TSS ± 0.5 kb) after 2 hr SET1A/B removal. Genes are split into the same groups as in (B).

Although the depletion of SET1A/B resulted in only subtle effects on promoter-associated RNAPII occupancy, this did suggest that the function of RNAPII was disrupted in some way. I therefore wanted to examine in more detail how the depletion of SET1A/B affects RNAPII. The initial stage of transcription culminates in PIC assembly and, during the final step of this process, the association of TFIIH results in RNAPII being phosphorylated on serine 5 (Harlen and Churchman, 2017). Given the pervasive effects on promoter-associated total RNAPII, and the previously reported interaction between WDR82 and S5P RNAPII (Lee and Skalnik, 2008; Lee *et al.*, 2010; Ebmeier *et al.*, 2017), I set out to examine how the depletion of SET1A/B affects S5P RNAPII. I therefore analysed the levels of S5P RNAPII by western blotting (Figure 5.6). Levels of S5P RNAPII were slightly reduced following the depletion of SET1A/B. However, this decrease was not significant, indicating that the removal of SET1A/B does not have a pronounced effect on the overall serine 5 phosphorylation status of RNAPII.

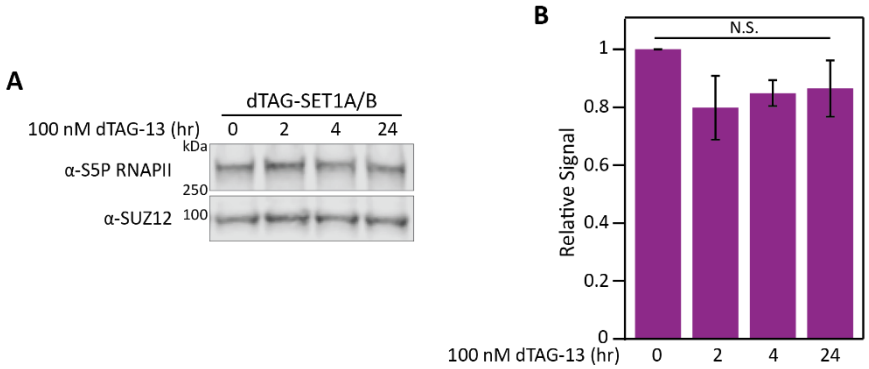


Figure 5.6 Overall levels of S5P RNAPII were not significantly affected by SET1A/B depletion
(A) Representative western blot of nuclear extracts showing the levels of S5P RNAPII from dTAG-SET1A/B cells treated with dTAG-13 over a 24 hr time course. SUZ12 is a loading control.
(B) Quantification of western blots of S5P RNAPII signal from dTAG-SET1A/B cells treated with dTAG-13 over a 24 hr time course. Signal is normalised to SUZ12 and expressed relative to untreated. Error bars show SEM from two biological replicates. N.S. represents not significant in a paired Student's T test ($p>0.05$).

Given that the depletion of SET1A/B resulted in a subtle, albeit not significant, reduction in bulk S5P RNAPII, I next examined the occupancy of S5P RNAPII across genes by cChIP-seq following 2 hours of SET1A/B depletion. As previously described, S5P RNAPII was enriched at gene promoters and was also present throughout gene bodies (Figure 5.7 A). Interestingly, levels of S5P RNAPII were dramatically reduced both at gene promoters and in gene bodies when SET1A/B were removed (Figure 5.7 B-D). This decrease was larger than the reduction in total RNAPII, indicating that the removal of SET1A/B specifically affects the phosphorylation of serine 5 of the RNAPII CTD. Interestingly, after normalising S5P signal for changes in total RNAPII, promoter- and gene body-associated S5P RNAPII levels were similarly reduced (Figure 5.7 C and D). Moreover, S5P RNAPII was similarly reduced across all genes, regardless of the gene expression outcomes. Therefore, the removal of SET1A/B results in a genome-wide reduction in the serine 5 phosphorylation of RNAPII. This could suggest that SET1A/B regulate gene expression by supporting the phosphorylation of RNAPII during initiation.

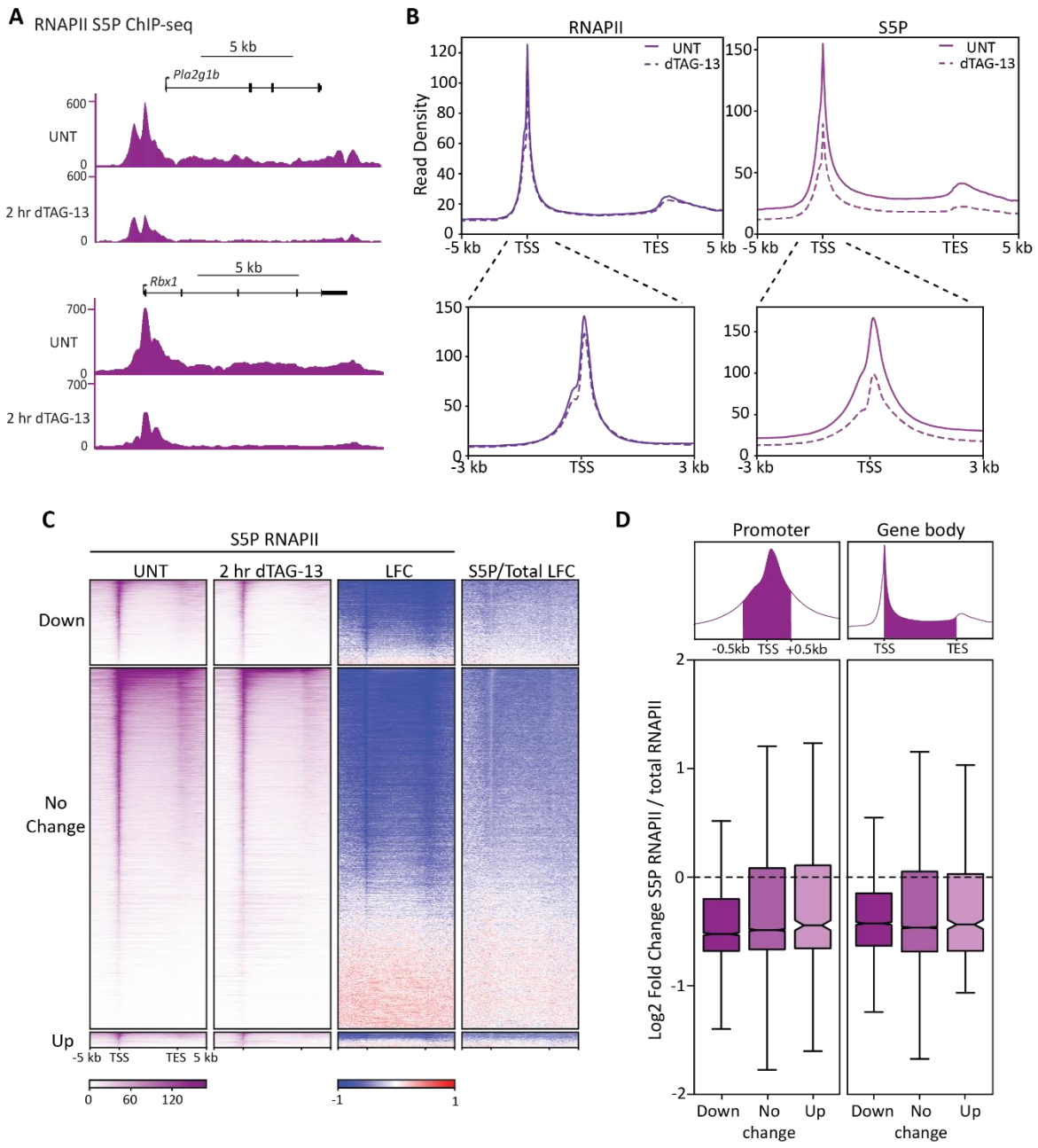


Figure 5.7 Removal of SET1A/B results in a widespread reduction in S5P RNAPII

(A) Genomic snapshots showing S5P RNAPII cChIP-seq signal at two actively transcribed genes in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells.

(B) Metaplots of total RNAPII and S5P RNAPII cChIP-seq signal across all gene bodies and TSSs (n=20633) in untreated cells (UNT) and 2 hr dTAG-13-treated dTAG-SET1A/B cells.

(C) Heatmaps of S5P RNAPII cChIP-seq signal at all gene bodies in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells. The log2 fold change (LFC) in S5P RNAPII following SET1A/B depletion and the log2 fold change in S5P RNAPII normalised for total RNAPII (S5P/Total) are also shown. Genes are split into those that experienced a significant reduction (Down, n=3804), increase (Up, n=683), or no change (No Change, n=16146) in expression based on cRNA-seq analysis. Intervals are sorted by total RNAPII cChIP-seq signal in untreated cells.

(D) Box plots of the log2 fold change in S5P RNAPII cChIP-seq signal normalised for total RNAPII cChIP-seq signal at gene promoters (TSS ± 0.5 kb) or gene bodies (TSS to TES) after 2 hr SET1A/B removal. Genes are split into the same groups as in (C).

5.2.3 The depletion of SET1A/B affects gene body-associated RNAPII

The depletion of SET1A/B resulted in clear defects to RNAPII at gene promoters; levels of promoter-associated RNAPII and S5P RNAPII were pervasively reduced. However, these changes in RNAPII were more pervasive than the effects on gene expression. I therefore wanted to examine whether RNAPII defects at promoters were associated with additional defects in the function of RNAPII during subsequent steps of transcription. I observed that at the genes with significantly reduced expression, the reduction in total RNAPII occupancy was typically larger in gene bodies than at gene promoters (Figure 5.8 A and B). In fact, at such genes, the largest reduction in total RNAPII occupancy actually occurred towards the end of gene bodies and beyond TESs. In contrast, genes with unchanged expression typically had levels of gene body-associated RNAPII similar to those in untreated cells, while genes with increased expression had on average slightly increased levels of total RNAPII in their gene bodies. Interestingly, while there was no correlation between the changes in promoter-associated total RNAPII and gene expression, changes in gene body-associated RNAPII were moderately correlated with the changes in gene expression (Figure 5.8 C and D). Therefore, while promoter-associated total RNAPII was pervasively reduced when SET1A/B were removed, changes in gene body-associated RNAPII were more indicative of effects on gene expression.

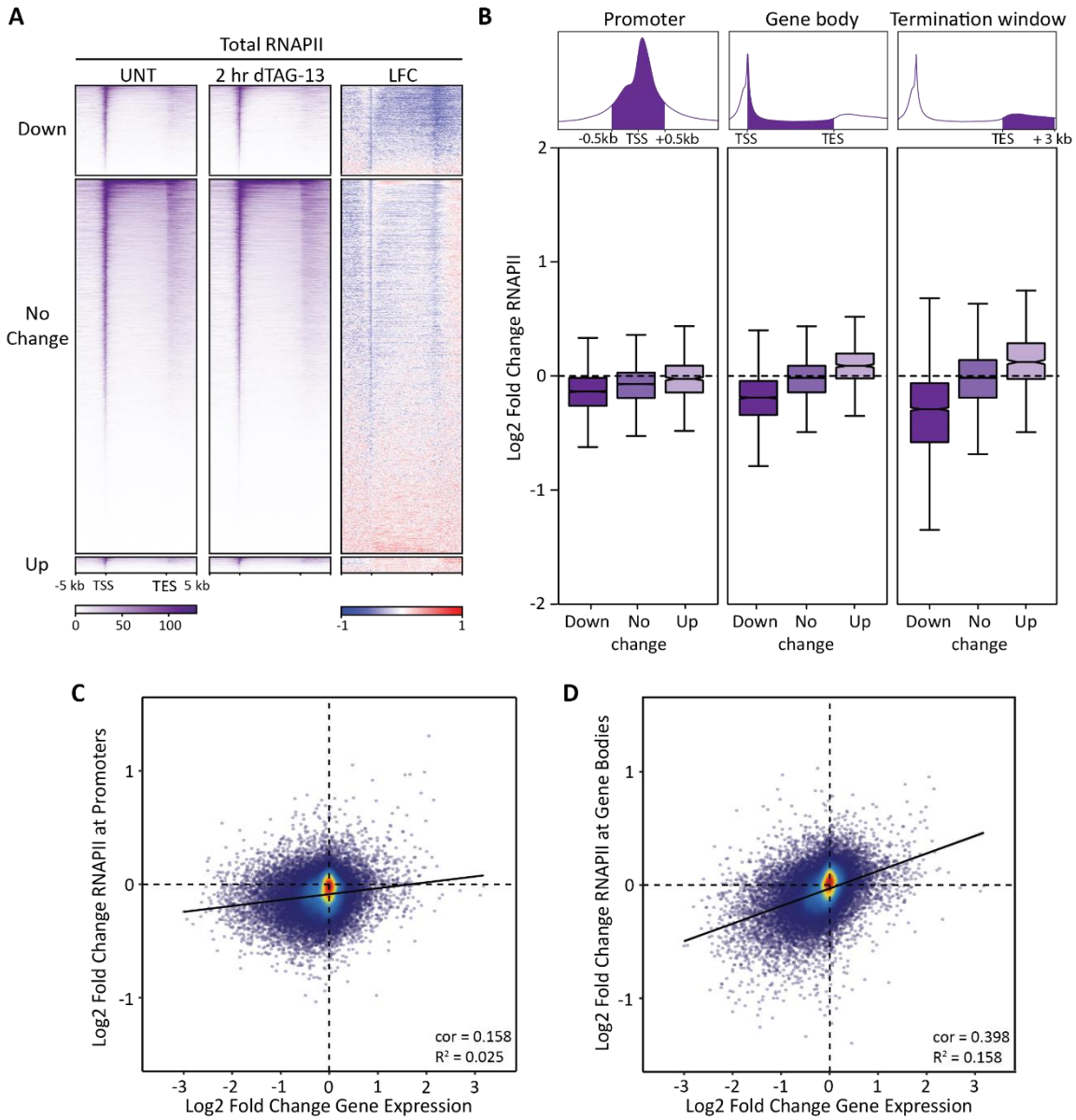


Figure 5.8 SET1A/B depletion results in a larger reduction in RNAPII from gene bodies in comparison to promoters

(A) Heatmaps of total RNAPII cChIP-seq signal over all gene bodies in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells. The log2 fold change (LFC) in RNAPII following SET1A/B depletion is also shown. Genes are split into those that experienced a significant reduction (Down, n=3804), increase (Up, n=683), or no change (No change, n=16146) in expression based on cRNA-seq analysis. Intervals are sorted by total RNAPII cChIP-seq signal in untreated cells.

(B) Box plots of the log2 fold change in total RNAPII cChIP-seq signal at gene promoters (TSS ± 0.5 kb), gene bodies (TSS to TES) or termination windows (TES + 3 kb) after 2 hr SET1A/B removal. Genes are split into the same groups as in (A).

(C) and (D) Scatterplots comparing the log2 fold change in cRNA-seq signal and promoter-associated (C) or gene body-associated (D) total RNAPII cChIP-seq signal in dTAG-SET1A/B cells treated with dTAG-13 for 2 hr. The dashed lines indicate log2 fold change=0, while the solid line represents linear regression. Cor denotes the Pearson correlation coefficient and R^2 the coefficient of determination for linear regression.

Serine 2 of the RNAPII CTD becomes phosphorylated shortly after its release from promoters and is therefore indicative of the elongating form of RNAPII (Harlen and Churchman, 2017). Since the genes with significantly reduced expression experienced a larger reduction in the occupancy of RNAPII within their gene bodies in comparison to their promoters, I wanted to examine how the depletion of SET1A/B affects S2P RNAPII. I therefore examined the levels of S2P RNAPII by western blotting. As with total RNAPII and S5P RNAPII, the depletion of SET1A/B resulted in a subtle, but not statistically significant, reduction in S2P RNAPII (Figure 5.9).

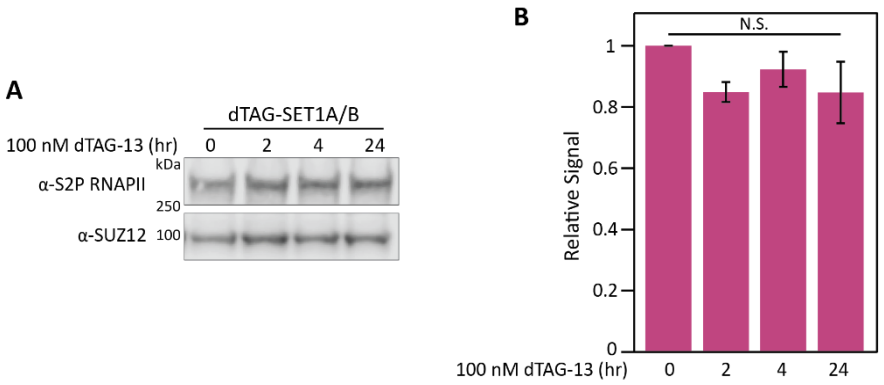


Figure 5.9 Global levels of S2P RNAPII were not significantly affected by SET1A/B depletion
(A) Representative western blot of nuclear extracts showing the levels of S2P RNAPII from dTAG-SET1A/B cells treated with dTAG-13 over a 24 hr time course. SUZ12 is a loading control.
(B) Quantification of western blots of S2P RNAPII signal from dTAG-SET1A/B cells treated with dTAG-13 over a 24 hr time course. Signal is normalised to SUZ12 and expressed relative to untreated. Error bars show SEM from two biological replicates. N.S. represents not significant in a paired Student's T test ($p>0.05$).

In order to further examine the effect of SET1A/B removal on S2P RNAPII, I next performed cChIP-seq. As previously described, S2P RNAPII was present throughout gene bodies and peaked towards transcription end sites (Figure 5.10 A). The removal of SET1A/B resulted in a pervasive reduction in S2P RNAPII and this reduction was larger than in total RNAPII (Figure 5.10 B-D). This suggests that the removal of SET1A/B also affects the phosphorylation of serine 2. As with S5P RNAPII, levels of S2P RNAPII were globally reduced, even at genes whose expression was not reduced following the depletion of SET1A/B. However, the genes with reduced expression typically lost more S2P RNAPII in comparison to genes whose expression was not reduced. In fact, the reduction in gene-body

associated S2P RNAPII was moderately correlated with the change in gene expression (Figure 5.10 E). Among the RNAPII changes that I examined, gene body-associated total RNAPII and S2P RNAPII were the most highly correlated with the gene expression outcomes (Figure 5.8 C and D, 5.10 E and F and A12). Therefore, while the removal of SET1A/B results in pervasive reductions in promoter-associated RNAPII and S5P RNAPII, reductions in S2P and total RNAPII occupancy within gene bodies are more closely related to the gene expression outcome. Overall, this suggests that SET1A/B influence the occupancy and phosphorylation of RNAPII to enable normal transcription.

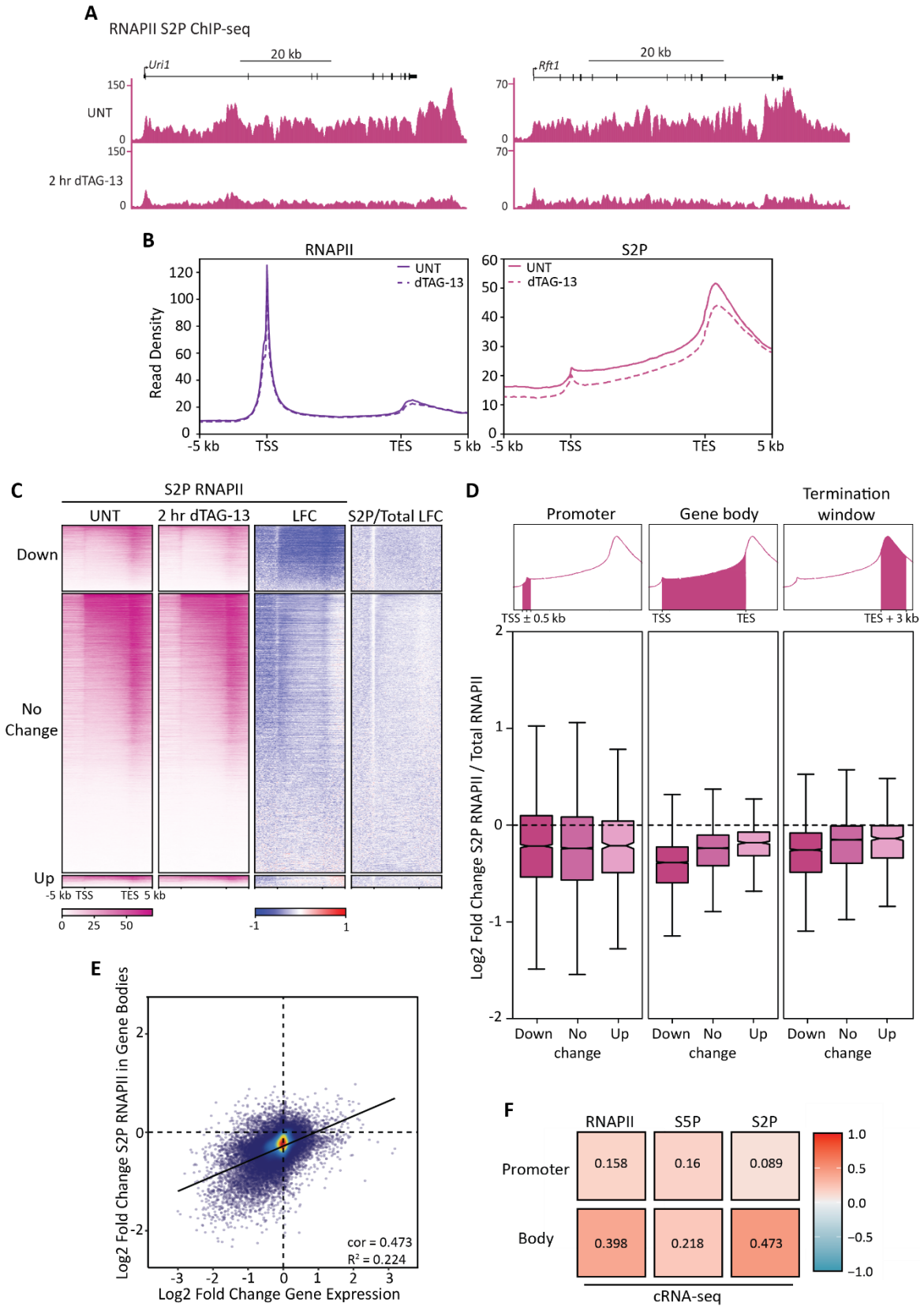


Figure 5.10 Removal of SET1A/B results in a pervasive reduction in S2P RNAPII

(A) Genomic snapshots showing S2P RNAPII cChIP-seq signal at two actively transcribed genes in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells.

(B) Metaplots of total RNAPII and S2P RNAPII cChIP-seq signal at all gene bodies (n=20633) in untreated cells (UNT) and 2 hr dTAG-13-treated dTAG-SET1A/B cells.
(C) Heatmaps of S2P RNAPII cChIP-seq signal at all gene bodies in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells. The log2 fold change (LFC) in S2P RNAPII following SET1A/B depletion and the log2 fold change in S2P RNAPII normalised for total RNAPII (S2P/Total) are also shown. Genes are split into those that experienced a significant reduction (Down, n=3804), increase (Up, n=683), or no change (No change, n=16146) in expression based on cRNA-seq analysis. Intervals are sorted by total RNAPII cChIP-seq signal in untreated cells.
(D) Box plots of the log2 fold change in S2P RNAPII cChIP-seq signal normalised for total RNAPII signal at gene promoters (TSS ± 0.5 kb), gene bodies (TSS to TES) or termination windows (TES + 3 kb) after 2 hr SET1A/B removal. Genes are split into the same groups as in (C).
(E) Scatterplot comparing the log2 fold changes in cRNA-seq signal and gene body-associated S2P RNAPII cChIP-seq signal in dTAG-SET1A/B cells treated with dTAG-13 for 2 hr. The dashed lines indicate log2 fold change=0, while the solid line represents linear regression. Cor denotes the Pearson correlation coefficient and R^2 the coefficient of determination for linear regression.
(F) Pearson correlation coefficients of log2-fold changes in gene expression (cRNA-seq) with log2-fold changes in cChIP-seq signal for total RNAPII, S5P RNAPII and S2P RNAPII at gene promoters or gene bodies in dTAG-SET1A/B cells following 2 hr dTAG-13 treatment.

5.2.4 SET1A/B regulate gene transcription

The depletion of SET1A/B resulted in a pervasive reduction in promoter-associated RNAPII and S5P RNAPII and the genes with reduced expression experienced the largest reduction in RNAPII occupancy and S2P RNAPII in their gene bodies. In chapter 3 I characterised the role of SET1A/B in gene expression by cRNA-seq, which measures RNAs being actively transcribed, as well as those that have been spliced and processed into mRNA. Therefore, such measurements of gene expression can be influenced by several aspects of RNA metabolism, including transcription rate, co-transcriptional processing, and RNA turnover. Given that the depletion of SET1A/B resulted in pronounced defects in the occupancy and phosphorylation of RNAPII, I reasoned that it would be necessary to examine the effect of SET1A/B depletion on ongoing transcription.

Measures of nascent transcription typically rely on 4sU labelling of newly synthesised RNA followed by biotin conjugation and streptavidin pull down to specifically capture nascent transcripts (Rabani *et al.*, 2011; Windhager *et al.*, 2012). A further development of this methodology, called Transient Transcriptome-sequencing (TT-seq) uses 4sU labelling coupled with RNA fragmentation to

specifically capture the transcripts that were synthesised during the labelling period (Schwalb *et al.*, 2016; Gregersen, Mitter and Svejstrup, 2020). This provides a high-resolution genome-wide view of ongoing transcription. Therefore, I utilised a spike-in calibrated TT-seq (cTT-seq) approach to examine nascent transcription following the depletion of SET1A/B for 2 hours. Importantly, in contrast to cRNA-seq, reads obtained in cTT-seq were not biased towards exons and were instead present throughout transcribed genes (Figure 5.11). This indicates that I was able to efficiently capture ongoing transcription.

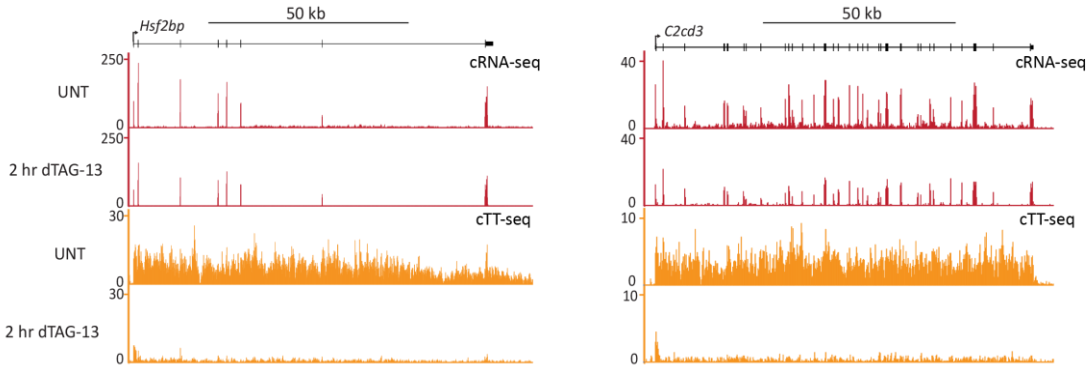


Figure 5.11 cTT-seq profiles nascent transcription
Genomic snapshots showing cRNA-seq and cTT-seq signal at two actively transcribed genes in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells.

In order to understand the effects of SET1A/B removal on nascent transcription, I performed differential gene expression analysis using the cTT-seq data. Qualitatively, the changes in nascent transcription as measured by cTT-seq were similar to the changes in steady-state RNA as measured by cRNA-seq (Figure 5.12). Furthermore, the majority of significantly reduced genes were identified using both steady-state RNA and nascent transcription measurements, and the genes that were only identified as significantly decreased by one measurement were decreased on average using the other (Figure 5.12 B). Therefore, changes in steady state gene expression and nascent transcription were largely similar. Interestingly, however, while cRNA-seq identified more than 600 genes with increased expression following SET1A/B depletion, only 75 genes were identified with increased transcription by cTT-seq. Moreover, the magnitude of change of these 75 genes was

relatively small. This cTT-seq experiment confirms that SET1A/B predominantly support RNAPII-mediated gene transcription in ESCs.

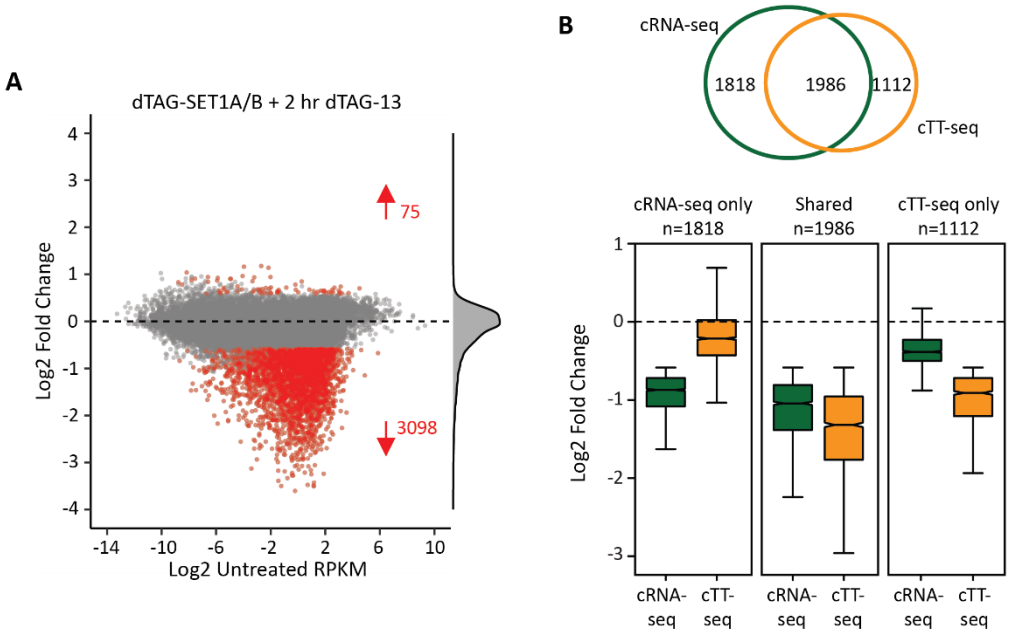


Figure 5.12 SET1A/B support RNAPII-mediated transcription
(A) MA-plot of log2 fold change in nascent transcription (cTT-seq) in dTAG-SET1A/B cells following 2 hr dTAG-13 treatment plotted against log2 mean RPKM in untreated cells for a non-redundant mm10 gene set (n=20633). Significantly changed (p-adj<0.05, fold change>1.5) genes are coloured red and the numbers of significant changes are shown. A density plot of log2 fold change in cTT-seq signal is shown to the right of the plot.
(B) Top panel: A Venn diagram showing the overlap between genes significantly decreased following 2 hr SET1A/B depletion as measured by cRNA-seq or cTT-seq. Bottom panel: Box plots of log2 fold change in gene expression, as measured using either cRNA-seq or cTT-seq, of all significantly decreased genes at 2 hr SET1A/B depletion. These genes are split into three groups, depending on whether they were significantly decreased in cRNA-seq only, cTT-seq only or in both cRNA-seq and cTT-seq (Shared). The number of genes in each subset are as per the Venn diagram.

I identified 3098 genes with significantly reduced transcription by differential gene expression analysis (Figure 5.12 A). Interestingly, nascent transcription of these genes was typically reduced both proximal to TSSs and within gene bodies (Figure 5.13 A). However, while levels of transcription close to TSSs were reduced, there was a very clear trend for a further decrease in transcription downstream of TSSs (Figure 5.13 A-C). Interestingly, the location of this transcription decrease was consistently 500-1000 bp downstream of the TSSs across these genes (Figure 5.13 C). This suggests that the ability of RNAPII to transcribe further than 500-1000 bp downstream of TSSs is reduced in

the absence of SET1A/B. As with cRNA-seq, the changes in transcription as measured by cTT-seq correlated well with the changes in gene body-associated polymerase, especially its S2P form (Figure 5.13 D and E). Altogether, this reveals that SET1A/B have important roles in supporting gene transcription in ESCs, and this is most likely mediated through influencing the occupancy and phosphorylation of RNAPII.

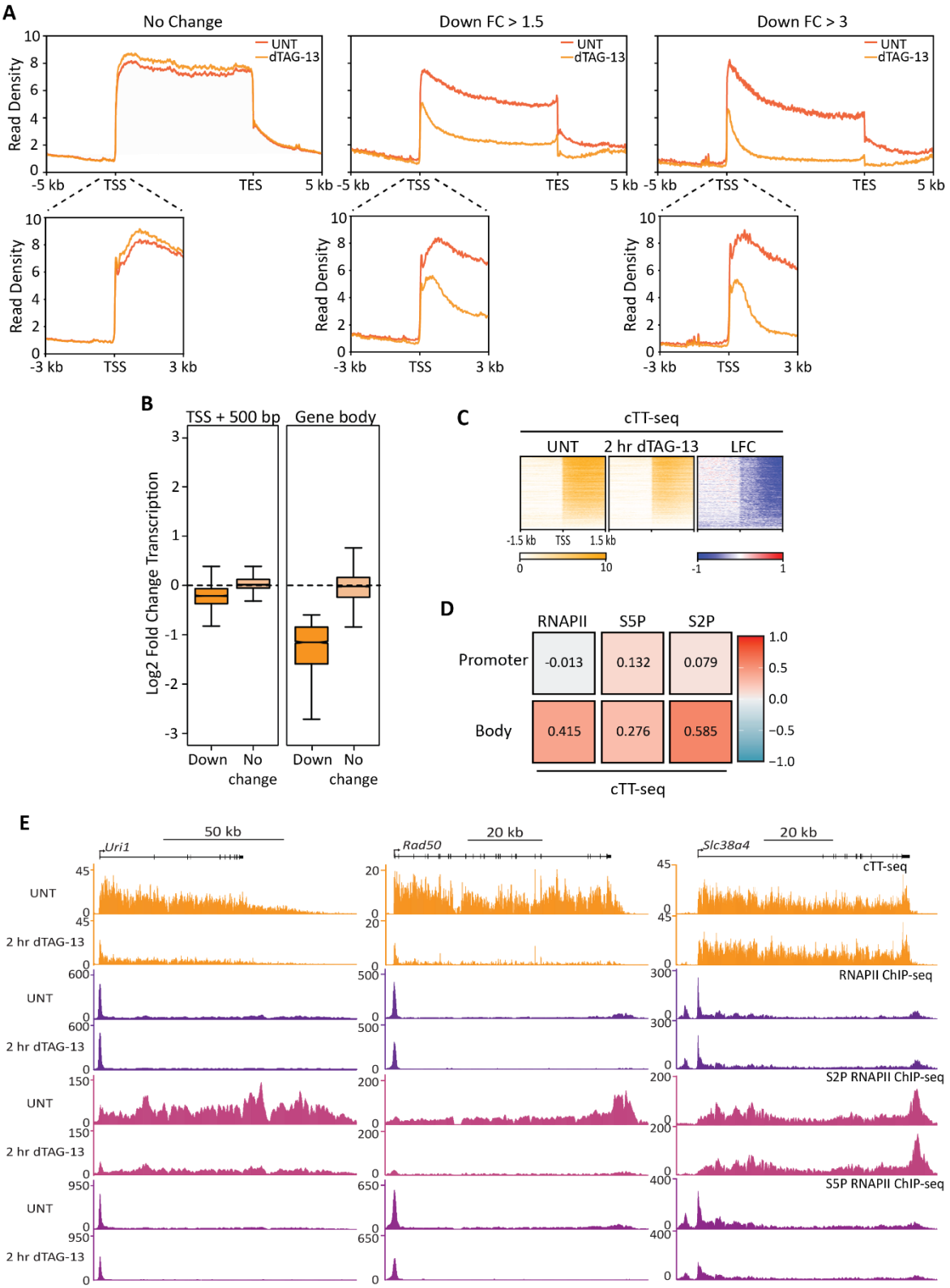


Figure 5.13 Gene body-associated transcription was reduced following SET1A/B depletion

(A) Metaplots of cTT-seq signal across gene bodies and TSSs in untreated (UNT) and 2 hr dTAG-13-treated dTAG-SET1A/B cells. Genes are split into those not significantly changed following SET1A/B depletion (No Change, n=17460), significantly decreased with a fold change greater than 1.5 (Down FC

> 1.5, n=3098), and significantly decreased with a fold change greater than 3 (Down FC > 3, n=765), as measured by cTT-seq.
(B) Box plots of the log2 fold change in cTT-seq signal at TSSs or gene bodies after 2 hr SET1A/B removal. Genes are split into those that experienced a significant reduction (Down, n=3098) or no change (No change, n=17460) in transcription based on cTT-seq analysis.
(C) Heatmaps of cTT-seq signal at TSSs in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells. The log2 fold change (LFC) in cTT-seq signal is also shown. Only genes that experienced a significant reduction in transcription based on cTT-seq analysis are shown (n=3098). Intervals are sorted by total RNAPII cChIP-seq signal in untreated cells.
(D) Pearson correlation coefficients of log2-fold changes in transcription (cTT-seq) with log2-fold changes in cChIP-seq signal for total RNAPII, S5P RNAPII and S2P RNAPII at gene promoters or gene bodies in dTAG-SET1A/B cells following 2 hr dTAG-13 treatment.
(E) Genomic snapshots showing cTT-seq, total RNAPII cChIP-seq, S5P RNAPII cChIP-seq and S2P RNAPII cChIP-seq in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells at two genes with significantly reduced transcription (*Uri1* and *Rad50*) and one gene with unchanged transcription (*Slc38a4*).

cTT-seq profiles ongoing transcription, allowing for the observation of 'transient' transcripts (Schwalb *et al.*, 2016). Anti-sense transcription from bidirectional promoters is now considered to be a pervasive feature of mammalian gene promoters (Core, Waterfall and Lis, 2008; Seila *et al.*, 2008). Given that I observed pervasive defects in the occupancy and phosphorylation of RNAPII at gene promoters, I was curious to examine whether the depletion of SET1A/B affects anti-sense transcription. Interestingly, the depletion of SET1A/B resulted in pervasive reductions in anti-sense transcription from bidirectional promoters, even at genes where sense transcription was unaffected (Figure 5.14). This suggests that perturbations of RNAPII function at bidirectional promoters could be associated with pervasive reductions in anti-sense transcription, while just a subset of these genes are sensitive to significantly reduced transcription in the sense direction. Nonetheless, it is clear that SET1A/B regulate key aspects of RNAPII function and RNAPII-mediated transcription.

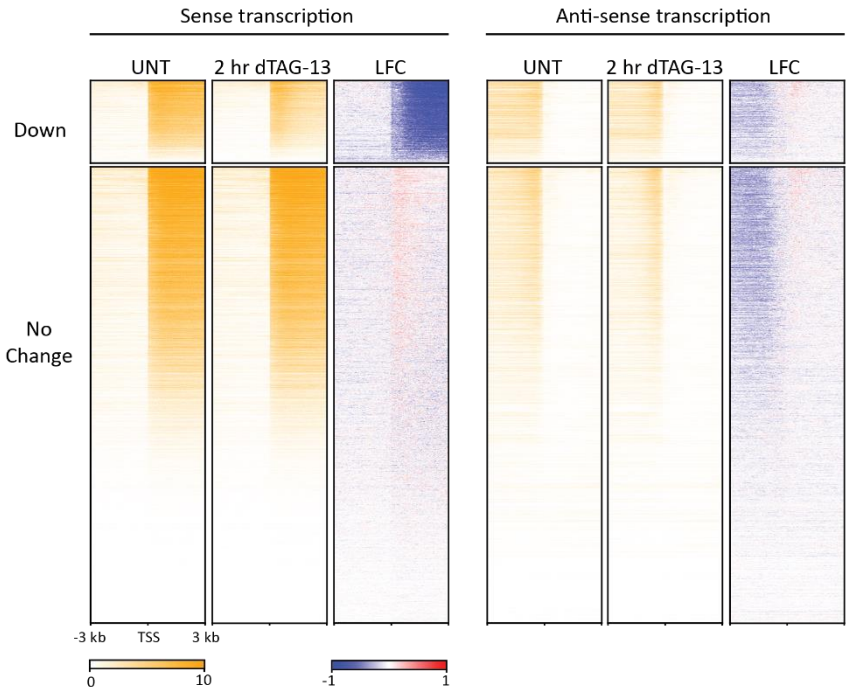


Figure 5.14 Removal of SET1A/B results in pervasive reductions in anti-sense transcription from bidirectional promoters

Heatmaps of cTT-seq signal in the sense and anti-sense orientation from TSSs in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells. Also shown is log2 fold change in the cTT-seq signal (LFC). Genes are split into those that experienced a significant reduction (Down, n=3098), or no change (No Change, n=17460) in sense transcription based on cTT-seq analysis. Intervals are sorted by total RNAPII cChIP-seq signal in untreated cells.

5.3 Summary and discussion

SET1A/B support gene transcription in ESCs. However, since this function appears to be largely independent of their catalytic activity, the mechanisms by which SET1A/B regulate gene expression remain enigmatic. Here I used the reporter system developed in chapter 4, RNAPII cChIP-seq and nascent transcription analysis to investigate these mechanisms. I found that the N-termini of SET1A/B are sufficient to activate reporter gene expression and that the interaction with WDR82 is essential for this outcome, at least in the context of SET1A. Furthermore, the removal of SET1A/B resulted in pervasive reductions in the occupancy and phosphorylation of RNAPII, as well as dramatic defects in transcription. This suggests that the role of SET1A/B in supporting transcription is dependent on WDR82 and involves influencing the occupancy and phosphorylation of RNAPII.

The depletion of SET1A/B resulted in pervasive reductions in the phosphorylation of the RNAPII CTD on serine 5 and serine 2, suggesting that SET1A/B could influence the function of the RNAPII CTD kinases and/or phosphatases. Furthermore, reductions in S5P RNAPII at gene promoters could actually perturb the deposition of S2P (Eick and Geyer, 2013; Ebmeier *et al.*, 2017; Joo *et al.*, 2019). Since S5P and S2P recruit various key regulatory factors required for different steps of ongoing transcription to RNAPII (Harlen *et al.*, 2016), this reduced phosphorylation could be a cause of further perturbations to RNAPII function. For example, reduced phosphorylation could impair the recruitment of important elongation factors to RNAPII and could also have implications for co-transcriptional processes, such as mRNA capping and splicing. Since defects in the phosphorylation of RNAPII could result in the formation of a polymerase that is less competent for transcription, an important role of SET1A/B in the regulation of gene expression could be to support the phosphorylation of RNAPII at gene promoters.

Changes in the occupancy and phosphorylation of RNAPII are often indicative of perturbations in the transcription cycle. The removal of SET1A/B resulted in reduced promoter-associated RNAPII and pervasive reductions in the initiation-associated phosphorylation of serine 5. Furthermore, genes with reduced transcription experienced reduced promoter-proximal transcription, as measured by cTT-seq. These observations suggest that SET1A/B could influence an aspect of transcription initiation prior to TFIIH and CDK7 engagement with the PIC. It is tempting to speculate that the removal of SET1A/B could therefore reduce transcription initiation. However, more high-resolution techniques that directly measure initiation frequency will be required to examine this possibility in the future (Gressel *et al.*, 2017; Lim, 2018; Wissink *et al.*, 2019). Nonetheless, as discussed above, widespread perturbations to the phosphorylation of RNAPII during initiation could be a cause of further defects to transcription. However, the contribution of these effects on promoter-associated RNAPII to gene expression remain unclear, since they were more pervasive than the effects on gene body-associated polymerase, gene expression and gene transcription (Figure 5.7, 5.8 and A13).

The removal of SET1A/B resulted in pervasive reductions in the phosphorylation of serine 2. Since serine 2 is phosphorylated upon the release of RNAPII from pausing and is then progressively phosphorylated during transcription (Harlen and Churchman, 2017), this could suggest that SET1A/B influence RNAPII pause-release. Indeed, genes with reduced expression experienced a larger reduction in total RNAPII occupancy within their gene bodies in comparison to their promoters (Figure 5.8). However, reduced release of RNAPII from pausing would be expected to result in increased RNAPII occupancy at gene promoters. I observed that the depletion of SET1A/B resulted in a very subtle widening of the promoter-associated RNAPII peak (Figure 5.5). However, overall promoter-associated RNAPII occupancy was pervasively reduced. While it is possible that such an effect could be masked by additional reductions in transcription initiation, I also observed in cTT-seq that the major reduction in gene body transcription occurred approximately 500-1000 bp downstream from TSSs (Figure 5.13 C). This indicates that transcription can proceed past the promoter-proximal pause site (20-60 bp downstream of TSSs) in the absence of SET1A/B. Together, this would argue that the removal of SET1A/B does not significantly affect promoter-proximal pausing. Alternatively, reductions in the phosphorylation of serine 2 could be associated with altered RNAPII elongation rate. However, reduced elongation rate might be expected to result in increased gene body-associated polymerase. The removal of SET1A/B resulted in reduced gene body-associated polymerase at the genes with reduced expression (Figure 5.8). Therefore, this would argue against the possibility that SET1A/B significantly affect RNAPII elongation rate. Overall, while the removal of SET1A/B resulted in pervasive defects to the occupancy and phosphorylation of RNAPII, there was no obvious effect on promoter-proximal pausing or RNAPII elongation rate.

Transcription was clearly reduced in the absence of SET1A/B. Interestingly, genes with significantly reduced transcription typically experienced a pronounced reduction in transcription 500-1000 bp downstream from TSSs, as well as dramatic reductions in gene body-associated RNAPII (Figure 5.8 and 5.13). These observations could be explained by increased premature transcription termination. In contrast to increased pausing or reduced elongation rate, premature termination

would not be predicted to result in increased RNAPII occupancy, provided that terminated RNAPII was turned over. It will therefore be important to address whether SET1A/B depletion results in increased premature termination in future work. Premature termination frequently arises due to the recognition of cryptic polyA signals by the CPA complex and this can result in exosome-mediated transcript degradation (Kamieniarz-Gdula and Proudfoot, 2019). Therefore, one could examine whether short transcripts accumulate upon inhibition of the exosome in the absence of SET1A/B. Alternatively, one might expect that the occupancy of termination factors might shift more 5' if premature termination was increased. However, if the removal of SET1A/B does result in premature transcription termination, it is unclear how this could align with the activation observed in reporter assays, or whether such an effect could be captured in the absence of a tethered functional SET1A. Therefore, it seems likely that SET1A/B could affect several aspects of RNAPII function.

I observed that the ability of SET1A to activate reporter gene expression is dependent on its interaction with WDR82 (Figure 5.2). Could the function of WDR82 therefore explain the effects of SET1A/B depletion on RNAPII? Mass spectrometry analysis has indicated that WDR82 interacts with several other proteins not found within the SET1A/B complexes (Lee *et al.*, 2010). For example, WDR82 associates with RPB1 when it is phosphorylated on serine 5 (Lee and Skalnik, 2008; Lee *et al.*, 2010; Ebmeier *et al.*, 2017). While this has been proposed to be a mechanism whereby active transcription results in the recruitment of the SET1A/B complexes to gene promoters, this could also be an interaction that affects the function of initiating RNAPII. For example, WDR82 binding to S5P RNAPII could stabilise RNAPII occupancy, contributing to transcription initiation. Interestingly, ySet1, Swd2 and RNAPII have been proposed to form a ternary complex in yeast, whereby the association of Swd2 with RNAPII is dependent on the presence of ySet1 (Lee and Skalnik, 2008; Bae *et al.*, 2020). I observed that WDR82 was not sufficient to activate reporter gene expression (Figure 5.2). It will be important to determine if tethering of WDR82 results in the recruitment of SET1A to the reporter gene promoter. If not, the

inability of WDR82 to activate reporter gene expression could be related to its requirement for SET1A to associate with RNAPII.

As well as having the ability to interact with additional proteins as part of the SET1A/B complexes, WDR82 can also form a stable association with the PNUTS-PP1 protein phosphatase complex, which dephosphorylates S5P RNAPII and contributes to transcription termination (Trinkle-Mulcahy *et al.*, 2006; Lee *et al.*, 2010; van Nuland, Smits, *et al.*, 2013; Austenaa *et al.*, 2015; Ebmeier *et al.*, 2017; Cortazar *et al.*, 2019). The details of the association between RNAPII and WDR82-SET1A/B or WDR82-PNUTS-PP1 are currently unclear. For example, it seems likely that WDR82-SET1A/B and WDR82-PNUTS-PP1 could compete for binding to RNAPII. In this instance, one could imagine that the removal of SET1A/B would result in the aberrant association of WDR82-PNUTS-PP1 with S5P RNAPII. If this were the case, this could explain why the removal of SET1A/B results in pervasive reductions in S5P RNAPII. Furthermore, since PNUTS-PP1 functions to limit RNAPII elongation rate within gene bodies (Cortazar *et al.*, 2019), increased binding of WDR82-PNUTS-PP1 to RNAPII could also reduce RNAPII elongation rate. However, as already discussed, it seems unlikely that SET1A/B depletion significantly affects RNAPII elongation rate. Another possibility is that increased binding of WDR82-PNUTS-PP1 to RNAPII promotes premature transcription termination in the absence of SET1A/B. The hypothesis that WDR82-SET1A/B and WDR82-PNUTS-PP1 could compete for binding to RNAPII, and that this could contribute to the effects of SET1A/B depletion on gene transcription, postulates that the removal of SET1A/B would be epistatic with that of the PNUTS-PP1 complex. In turn, since WDR82 would be proposed to mediate the interaction between the PNUTS-PP1 complex and S5P RNAPII, one could also predict that WDR82 depletion might not result in reduced expression of SET1A/B target genes. Indeed, I observed that the expression of some SET1A/B target genes was not affected by WDR82 depletion (Figure 5.3). This is despite the fact that WDR82 depletion results in reduced SET1A stability and has been proposed to reduce the chromatin association of SET1A/B (Lee and Skalnik, 2008; Wu *et al.*, 2008; Austenaa *et al.*, 2015; Franks *et al.*, 2017). It will therefore be important in future studies to determine whether epistasis with PNUTS-

PP1 contributes to the effect of SET1A/B on gene expression. Indeed, there is evidence from yeast to suggest that the function of Swd2 within ySet1C might be antagonistic with its role in transcription termination (Soares and Buratowski, 2012; Lee *et al.*, 2018).

The depletion of WDR82, SET1A/B and PNUTS-PP1 have previously been proposed to result in read-through transcription at TESs in macrophages and Swd2 has long been implicated in transcription termination in yeast (Cheng, He and Moore, 2004; Austenaa *et al.*, 2015). However, I observed no evidence of read-through transcription at TESs in the absence of SET1A/B (Figure A14). In fact, levels of total RNAPII were actually further reduced beyond the TES in comparison to gene bodies (Figure 5.8). This suggests that the removal of SET1A/B does not perturb transcription termination at TESs.

Whether SET1A/B regulate RNAPII and gene transcription through WDR82 genome-wide remains unclear and will require further examination in the future. Firstly, it will be important to examine how the depletion of SET1A/B affects promoter-associated WDR82. Removal of ySet1 in yeast leads to large reductions in Swd2 promoter-occupancy and residual Swd2 binding is mostly distributed throughout gene bodies, likely due to its role in the termination machinery (Soares and Buratowski, 2012). I attempted to examine WDR82 occupancy following SET1A/B depletion by performing cChIP-seq for WDR82 in the dTAG-SET1A/B cell line. However, this was unsuccessful and will require revisiting in the future. Secondly, it will be important to directly test in genome-wide experiments whether WDR82 is required for transcriptional activation by SET1A/B. One way of addressing this question could be to generate a conditional WDR82 binding mutant using a similar system to the one used to generate the conditional catalytic point mutant (Chapter 4.2.4). However, such an experiment could be confounded by reduced SET1A stability in the absence of WDR82 binding (Lee and Skalnik, 2008; Wu *et al.*, 2008; Austenaa *et al.*, 2015; Franks *et al.*, 2017; Figure 5.3). An alternative approach could be to perform rescue experiments in the dTAG-SET1A cell line with overexpressed SET1A mutants. Finally, the contribution of the WDR82-PNUTS-PP1

complex to the gene expression outcomes following SET1A/B depletion will need to be examined. For example, by combining SET1A/B and PNUTS-PP1 depletion, the possibility of epistasis could be examined by cTT-seq. Alternatively, such an effect could also be examined by performing cTT-seq in the WDR82-dTAG cell line. While WDR82 is clearly essential for the activation of reporter gene expression by SET1A, it remains unexplored whether this (i) also applies to SET1B, (ii) is required for genome-wide transcriptional activation, and (iii) whether WDR82-PNUTS-PP1 contributes to this outcome.

From the gene-specific RT-qPCR analysis, I observed that the effect of SET1A/B removal on gene expression was more dramatic than the effect of WDR82 depletion (Figure 5.3). As discussed, this could be due to epistasis with the WDR82-PNUTS-PP1 complex at certain target genes. However, this could also suggest that SET1A/B have additional roles in the regulation of gene expression that are independent of WDR82. One possibility is that SET1A/B support the formation of the PIC or stability of the transcription machinery at gene promoters. In chapter 4 I observed that H3K4me3 was modestly reduced when SET1A/B were removed. H3K4me3 has previously been proposed to stabilise PIC formation via direct TAF3 binding (Vermeulen *et al.*, 2007; van Ingen *et al.*, 2008; Nikolov *et al.*, 2011; Eberl *et al.*, 2013; Lauberth *et al.*, 2013; van Nuland, Schram, *et al.*, 2013). Furthermore, global reductions in promoter-associated RNAPII and pervasive reductions in S5P and S2P have previously been associated with reductions in promoter-associated H3K4me3 (Mas *et al.*, 2018; Sze *et al.*, 2020). I observed that there was no correlation between the changes in H3K4me3 and promoter-associated total RNAPII, and a mild correlation between the changes in H3K4me3 and promoter-associated S5P RNAPII (Figure A15). Nonetheless, it is possible that reductions in H3K4me3 could lead to a pervasive reduction in initiating RNAPII from gene promoters. However, the role that this might play in gene expression is unclear since the loss of H3K4me3 and promoter-associated RNAPII were not correlated with changes in gene expression. Additionally, SET1A can activate gene expression independently of its catalytic activity (Chapter 4.2.5). It is possible that the role of SET1A/B in gene expression may be simultaneously mediated through different

mechanisms, for example through the regulation of RNAPII activity by WDR82, PIC formation by H3K4me3, and maybe even additional, currently unidentified mechanisms.

The very N-termini of SET1A/B were sufficient to activate reporter gene expression (Figure 5.2). This raises some interesting questions about the function of the rest of the SET1A/B proteins. It is possible that the remaining domains are important for transcription activation in different contexts. For example, the central linker region of SET1A has been proposed to be important for transcriptional activation of the DNA damage response genes in leukaemia cells (Hoshii *et al.*, 2018). Alternatively, the SET domain of SET1A was shown to be important for ESC differentiation (Sze *et al.*, 2017). This suggests that SET1A/B could have several functions in the regulation of gene expression, dependent on cell type or developmental context.

The other domains of SET1A/B likely possess functions that are independent of direct transcription regulation. For example, the C-termini of SET1A/B comprising the SET domain, WRAD complex and CFP1 are all likely to be important for the recruitment of SET1A/B to chromatin. The identification of unique domains required for transcriptional activation and complex recruitment provides an interesting model for the activity of SET1A/B at gene promoters. In this model, the ZF-CXXC domain of CFP1 would mediate the association of SET1A/B with CGI-associated gene promoters (Voo *et al.*, 2000; Lee and Skalnik, 2005; Brown *et al.*, 2017). SET1A/B could then contribute to transcription activation via WDR82 and simultaneously deposit H3K4me3. H3K4me3 could subsequently be bound by CFP1, further amplifying SET1A/B occupancy and transcriptional activation at the promoters of actively transcribed genes (Eberl *et al.*, 2013; Mahadevan and Skalnik, 2016; Brown *et al.*, 2017). This would create a simple feedback loop to potentiate transcriptional activation by SET1A/B.

6. Conclusions and future directions

SET1A/B are two highly conserved HMTs that associate with gene promoters. Here they deposit H3K4me3, a histone modification that correlates with transcriptional activity (Barski *et al.*, 2007; Mikkelsen *et al.*, 2007; Guenther *et al.*, 2007). An abundance of H3K4me3 'reader' proteins have been identified and proposed to modify the surrounding chromatin to promote transcriptional activity or even directly support PIC formation (Chapter 1.3). However, despite the links between SET1A/B, H3K4me3 and active transcription, previous work has shown that perturbing the function of SET1A/B does not have predictable effects on gene expression (Clouaire *et al.*, 2012; Clouaire, Webb and Bird, 2014; Austenaa *et al.*, 2015; Brown *et al.*, 2017; Franks *et al.*, 2017). Furthermore, the sole H3K4 HMT in yeast, ySet1, is dispensable for yeast viability (Nislow, Ray and Pillus, 1997; Briggs *et al.*, 2001) and H3K4me3 is not essential for gene expression in *Drosophila* (Hödl and Basler, 2009, 2012). Therefore, the role of SET1A/B in the regulation of gene expression, either through the deposition of H3K4me3 or otherwise, has remained largely unclear. Here I have used quantitative gene expression analysis to show that SET1A/B broadly support gene expression in ESCs (Chapter 3). Additionally, I have shown that their role in regulating gene expression appears to be uncoupled from the deposition of H3K4me3 (Chapter 4). Furthermore, using a synthetic reporter system, I have revealed that SET1A/B can activate gene transcription through their highly conserved N-terminal domains and, at least when it comes to SET1A, HMT activity is dispensable for transcriptional activation (Chapter 5). Finally, I have shown that SET1A requires its association with WDR82 to activate gene expression and that SET1A/B influence the occupancy and phosphorylation of RNAPII to support gene transcription in ESCs (Chapter 5).

SET1A/B have previously been proposed to be responsible for placing the majority of H3K4me3 in mammalian genomes (Meeks and Shilatifard, 2017; Cenik and Shilatifard, 2021). However, this proposal is largely based on the function of their orthologues in simpler eukaryotes or western blot analysis of histone proteins in mammalian cells. In contrast to the prevailing views in the field, I

found that the removal of SET1A/B results in only modest reductions in H3K4me3 (Chapter 4). This argues against SET1A/B being primarily responsible for the deposition of H3K4me3, at least in ESCs. Therefore, future work will need to further address the division of labour responsible for H3K4me3 deposition. For example, it is possible that MLL1/2 have more pronounced roles in the deposition of H3K4me3, particularly at active gene promoters, than have previously been appreciated. Moreover, given that MLL3/4 retain residual trimethyltransferase activity *in vitro,* it may also be necessary to re-examine their contribution to H3K4me3 deposition *in vivo* (Lee *et al.*, 2013; Tang *et al.*, 2013; Kwon *et al.*, 2021). Alternatively, it is possible that there are as-yet undiscovered H3K4 HMTs in mammals. By combining degron systems to simultaneously remove SET1A/B and MLL1/2, these possibilities could be explored. Furthermore, it will be interesting to examine whether the other H3K4 HMTs also activate gene expression and, if so, to examine the requirement for their catalytic activity in this process. MLL1/2 do not associate with WDR82 (Shilatifard, 2012; van Nuland, Smits, *et al.*, 2013), suggesting that if they were to activate transcription independently of catalysis, this must occur through a mechanism that is distinct to that of SET1A/B. Given that MLL1/2 associate with various factors not found within the SET1A/B complexes, these proteins could provide interesting candidates for catalysis-independent functions.

Previous studies in which the function of SET1A/B has been perturbed have reported that they have rather unpredictable, non-directional effects on gene expression (Clouaire *et al.*, 2012; Clouaire, Webb and Bird, 2014; Austenaa *et al.*, 2015; Brown *et al.*, 2017; Franks *et al.*, 2017). However, such experiments have relied on slow protein depletion systems that fail to capture the primary effects of SET1A/B removal. Here I used the dTAG system to induce the controlled, rapid degradation of SET1A/B, allowing the acute effects of their removal to be examined (Nabet *et al.*, 2018). By coupling rapid protein depletion with quantitative gene expression analysis, I found that SET1A/B perform a previously unappreciated function in supporting gene expression in ESCs (Chapter 3). While the depletion of SET1A resulted in dramatic reductions in gene expression, the contribution of SET1B was rather more subtle. Nevertheless, I observed that SET1A and SET1B

make a largely additive contribution to gene expression, resulting in dramatic and widespread reductions in gene expression in their absence. This widespread and directional effect was surprising based on previous studies and implicates SET1A/B as important regulators of gene expression in mammalian cells. Based on these findings, it will be interesting to apply rapid depletion and quantitative analysis in simpler eukaryotes to examine if this functionality of SET1A/B is conserved and was simply overlooked previously due to the limitations of classical perturbation approaches and non-quantitative analyses.

The mechanism(s) by which SET1A/B regulate gene expression remain to be fully elucidated. However, I have made several interesting observations that have allowed me to form a speculative model for how SET1A/B could support gene transcription in ESCs. I observed in a reporter assay that the ability of SET1A to activate gene expression is dependent on its interaction with WDR82. Future work will be required to examine whether this interaction is also important for transcriptional activation by SET1B and whether SET1A/B support gene transcription through WDR82 in a genome-wide manner. Nonetheless, the function of WDR82 provides several possible mechanisms by which the SET1A/B complexes could support transcription. Firstly, WDR82 interacts with S5P RNAPII (Lee and Skalnik, 2008; Lee *et al.*, 2010; Ebmeier *et al.*, 2017). I observed that the removal of SET1A/B results in pervasive reductions in RNAPII occupancy and serine 5 phosphorylation at gene promoters (Chapter 5). It is therefore possible that SET1A/B support the function of RNAPII during initiation through WDR82 binding to polymerase. This interaction could stabilise RNAPII occupancy at gene promoters, for example. In order to examine this possibility in the future, it would be interesting to identify mutations within SET1A/B and/or WDR82 that would disrupt the ability of the SET1A/B complexes to associate with RNAPII. One could then examine whether binding of WDR82-SET1A/B to RNAPII contributes to its function at gene promoters. Furthermore, WDR82 forms a stable association with the PNUTS-PP1 complex, in which PP1 is a protein phosphatase with specificity for serine 5 of the RNAPII CTD (Trinkle-Mulcahy *et al.*, 2006; Lee *et al.*, 2010; van Nuland, Smits, *et al.*, 2013; Ebmeier *et al.*, 2017). I have hypothesised that the

association of WDR82-SET1A/B with S5P RNAPII could compete with that of the WDR82-PNUTS-PP1 complex. In this scenario, WDR82-PNUTS-PP1 binding to S5P RNAPII would be increased in the absence of SET1A/B. In turn, this could lead to the aberrant removal of serine 5 phosphorylation from RNAPII. Such a hypo-phosphorylated RNAPII could be less competent for ongoing transcription and also disrupt numerous activities required to process nascent transcripts into mature mRNA. Therefore, the role of WDR82-SET1A/B at gene promoters could be to (i) support RNAPII occupancy, (ii) promote the phosphorylation of serine 5, possibly by preventing its dephosphorylation by the PNUTS-PP1 complex. It may therefore be interesting in the future to use more high-resolution techniques to examine whether defects in promoter-associated RNAPII are associated with reduced transcription initiation events (Mahat *et al.*, 2016; Gressel *et al.*, 2017; Lim, 2018; Wissink *et al.*, 2019).

I observed that the genes with significantly reduced transcription in the absence of SET1A/B experienced a dramatic reduction in transcription 500-1000 bp downstream from the TSSs (Figure 5.13). I propose that this sudden decline in transcription could be due to premature transcription termination mediated by WDR82-PNUTS-PP1. To examine this possibility in the future, it will firstly be essential to determine whether the defects in nascent transcription 500-1000 bp downstream of TSSs are a consequence of premature termination. Since premature termination frequently results in exosome-mediated transcript degradation (Kamieniarz-Gdula and Proudfoot, 2019), one might predict that short nascent transcripts would accumulate if the exosome was inhibited in the absence of SET1A/B. Furthermore, it will be important in future work to examine whether the PNUTS-PP1 complex contributes to the transcription defects observed in the absence of SET1A/B. One could examine this possibility by combining the SET1A/B degron with a PNUTS-PP1 degron. cTT-seq could then be used to examine if the defects in gene body-associated transcription are dependent on the presence of a functioning PNUTS-PP1 complex. If so, this would implicate PNUTS-PP1 in mediating premature transcription termination in the absence of SET1A/B. Furthermore, one could also use this system to examine whether PNUTS-PP1 contributes to S5P

dephosphorylation when SET1A/B are removed. Together this could suggest that SET1A/B have an important role in antagonising the activities of the PNUTS-PP1 complex. Therefore, examining the role of the PNUTS-PP1 complex in the RNAPII and transcription defects observed upon the depletion of SET1A/B will be essential for future studies.

Understanding the contribution of H3K4me3 to gene expression remains one of the key outstanding questions in the field (Howe *et al.*, 2017). Examining the role of H3K4me3 was beyond the scope of this thesis, since the perturbation of SET1A/B was not associated with dramatic defects in H3K4me3 and I was unable to generate systems to specifically perturb the catalytic activity of SET1A. However, SET1A/B do contribute modestly to H3K4me3 deposition genome-wide (Chapter 4). Even though H3K4me3 appears to be dispensable for transcriptional activation by SET1A/B, it is still possible that the deposition of H3K4me3 plays an important role in active gene expression. H3K4me3 supports gene transcription on *in vitro* transcription templates and is bound by components of the pre-initiation complex (Vermeulen *et al.*, 2007; van Ingen *et al.*, 2008; Lin *et al.*, 2011; Nikolov *et al.*, 2011; Eberl *et al.*, 2013; Jiang *et al.*, 2013; Lauberth *et al.*, 2013; Tang *et al.*, 2013). This suggests that H3K4me3 can directly contribute to active transcription. It is possible that SET1A/B could contribute to the function of RNAPII at gene promoters via the pervasive deposition of H3K4me3. Furthermore, extensive previous work has indicated that H3K4me3 could promote the formation of a transcriptionally permissive chromatin state (Chapter 1.3). The presence of H3K4me3 has been implicated in inhibiting the deposition of DNA methylation and polycomb-repressive histone modifications and in promoting histone acetylation and chromatin remodelling (Hughes, Kelley and Klose, 2020). In addition, H3K4me3 is recognised by the CFP1 subunit of the SET1A/B complexes at the promoters of actively transcribed genes, an interaction proposed to amplify the occupancy of SET1A/B at such loci (Eberl *et al.*, 2013; Mahadevan and Skalnik, 2016; Brown *et al.*, 2017). One could predict that the association of SET1A/B with gene promoters would result in transcriptional activation via WDR82 and simultaneous deposition of H3K4me3. Therefore, the activation of gene expression could be coincident with the formation of

a feedback loop that would further amplify SET1A/B occupancy and, therefore, active transcription. Furthermore, H3K4me3 could contribute to PIC assembly and also function to create a chromatin landscape around CGI-associated gene promoters that would be permissive to transcriptional activity (Chapter 1.3). Since the removal of SET1A/B resulted in only subtle alterations in H3K4me3, it seems likely that the other H3K4 HMTs must also be involved in establishing such a system. For example, given that MLL1/2 broadly associate with CGI-associated gene promoters (Guenther *et al.*, 2005; Denissov *et al.*, 2014; Hu *et al.*, 2017a), deposition of H3K4me3 by MLL1/2 could be important for establishing this feedback loop. The Klose lab and others have previously proposed that CGIs may create bistable chromatin states at gene promoters to regulate gene expression (Deaton and Bird, 2011; Binder *et al.*, 2013; Klose *et al.*, 2013; Steffen and Ringrose, 2014). In such a system, the establishment of permissive chromatin states following transcription initiation could function to potentiate new rounds of transcription, provide a memory of recent transcriptional activity, and allow for transcriptional activation to be maintained when the levels of an activation signal drop below the threshold required for initial activation. Recent work has established that bistable chromatin states are likely to exist (Sneppen and Ringrose, 2019; Reinig *et al.*, 2020) and understanding the contribution of SET1A/B and H3K4me3 deposition in such a system will be an exciting avenue for future research.

A Appendices

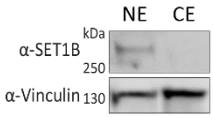


Figure A1 SET1B localises to the nucleus in ESCs
Western blot showing the levels of SET1B in 25 µg nuclear extract (NE) and 25 µg cytoplasmic extract (CE) from WT ESCs. Vinculin is a loading control.

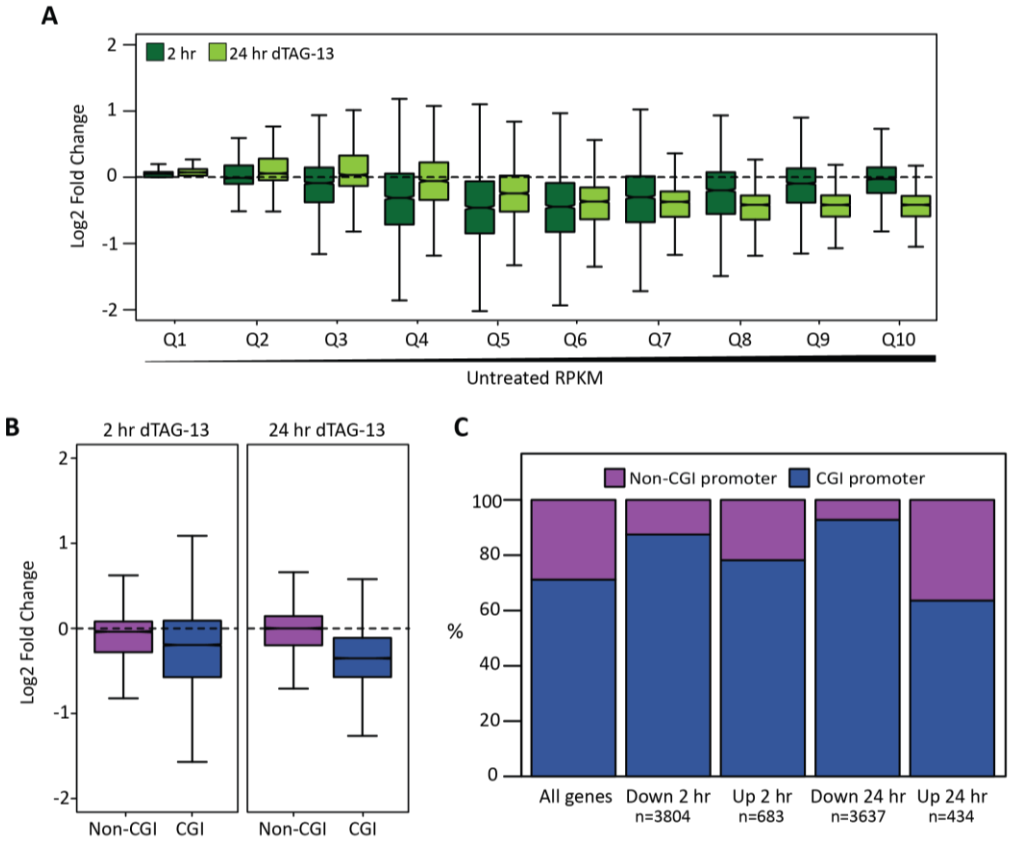


Figure A2 SET1A/B regulate the expression of expressed, CGI-associated genes
(A) Boxplots of log2 fold change in gene expression (cRNA-seq) of all transcripts (n=20633) following 2 hr and 24 hr SET1A/B depletion, split into deciles of equal size based on untreated gene expression levels (Untreated RPKM).
(B) Box plots of log2 fold change in gene expression of CGI-associated (n=14687) and non-CGI-associated (n=5946) genes following 2 hr and 24 hr dTAG-13 treatment in dTAG-SET1A/B cells.
(C) Bar chart showing the percentage of genes that have a CGI-gene promoter for all genes (n=20633) and genes that were significantly differentially expressed at 2 hr and 24 hr SET1A/B depletion.

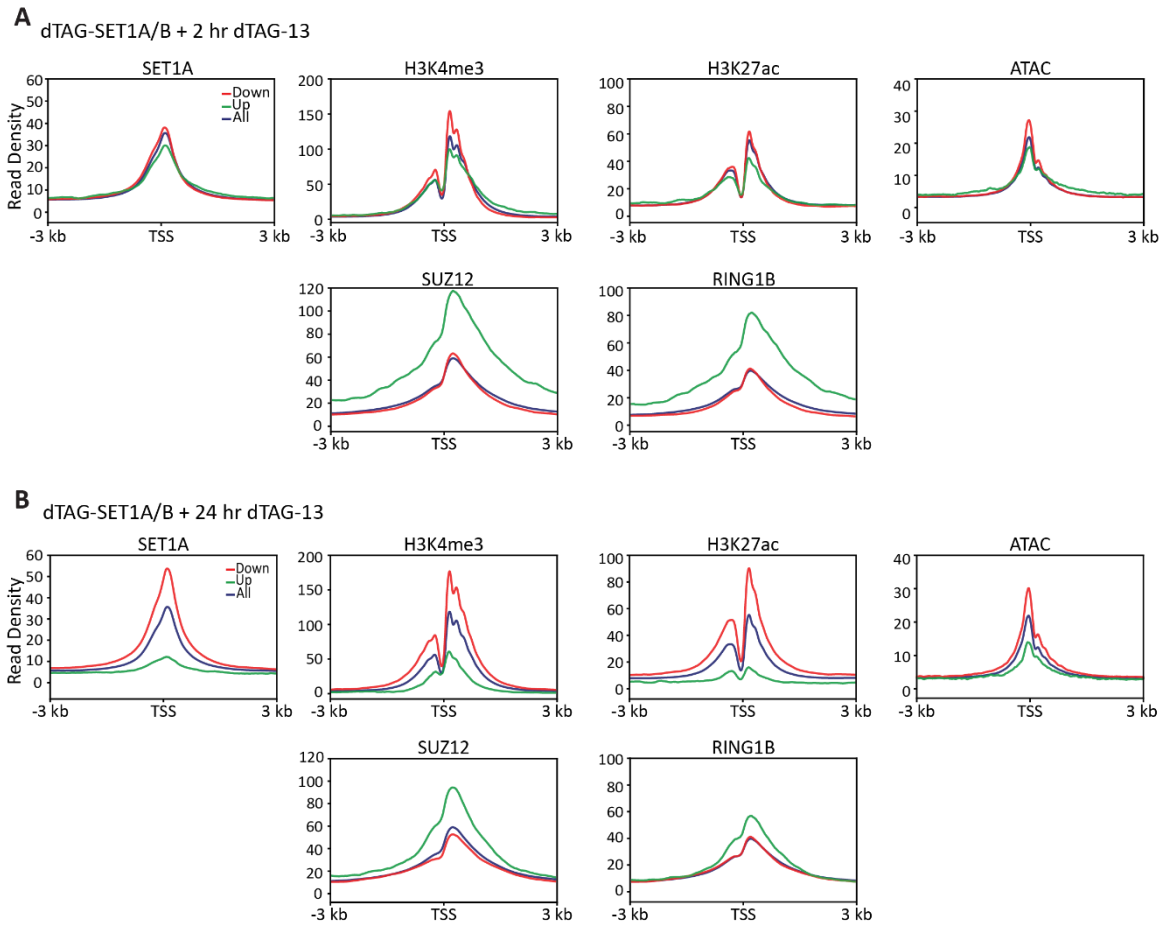


Figure A3 SET1A/B targets are enriched for chromatin features associated with active transcription
Metaplots of SET1A, H3K4me3, H3K27ac, SUZ12 and RING1B ChIP-seq and ATAC-seq signal at the TSS of all genes (n=20633) or genes significantly increased ('Up') or decreased ('Down') at 2 hr ((A), Up n=673, Down n=3804) or 24 hr ((B), Up n=434, Down n=3637) SET1A/B depletion (Datasets: Brown *et al.*, 2017; Dobrinić, Szczurek and Klose, 2021; Fursova *et al.*, 2020).

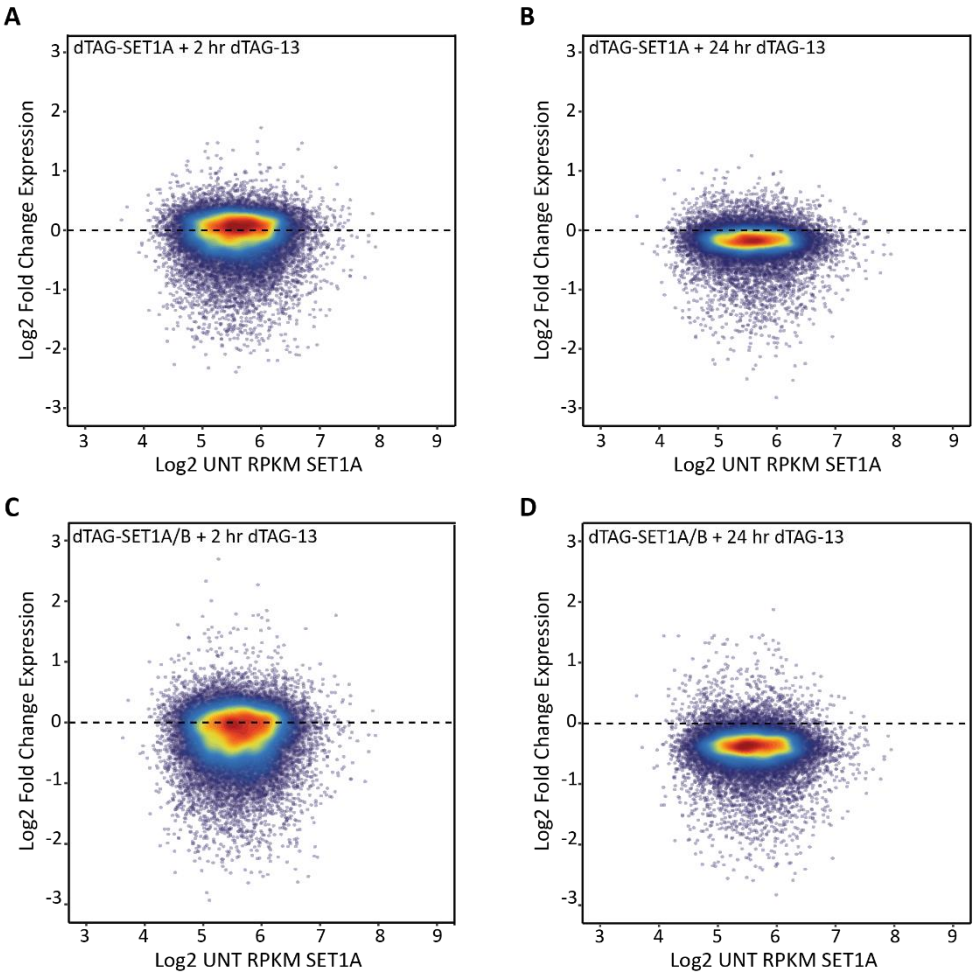


Figure A4 SET1A occupancy does not explain gene expression outcomes
Scatterplots comparing the levels of T7-dTAG-SET1A cChIP-seq signal in untreated cells (Log2 UNT RPKM) and the log2 fold change in gene expression (cRNA-seq) in dTAG-SET1A cells (A and B) and dTAG-SET1A/B cells (C and D) treated with dTAG-13 for 2 hr or 24 hr. Only genes that have a TSS-associated SET1A peak in untreated cells are included (n=10480). The dashed line indicates log2 fold change in cRNA-seq=0.

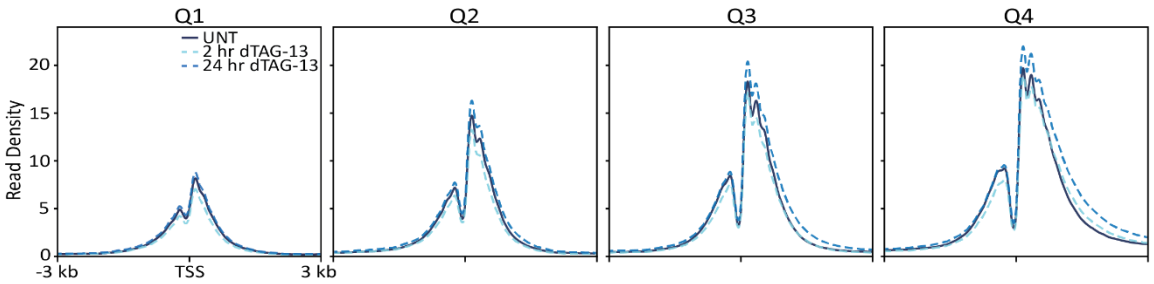


Figure A5 Levels of H3K4me3 scale with transcriptional activity in gene expression quartiles
Metaplots of H3K4me3 cChIP-seq signal at promoter-associated H3K4me3 peaks split into quartiles of equal size based on gene expression level in untreated dTAG-SET1A cells. Q1 corresponds to the most lowly expressed genes and Q4 to the most highly expressed genes. Only genes with a promoter-associated H3K4me3 peak in untreated cells are included (n=14546). Mean H3K4me3 signal in untreated dTAG-SET1A cells and cells treated with dTAG-13 for 2 hr and 24 hr and is shown.

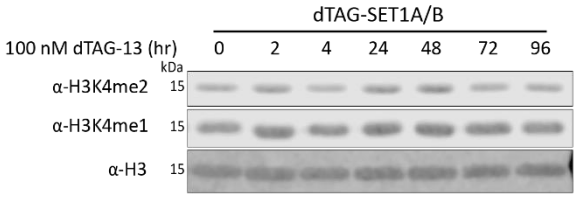


Figure A6 Levels of H3K4me1 and H3K4me2 remain unchanged following depletion of SET1A/B
Representative western blot showing levels of H3K4me1 and H3K4me2 in histone extracts from dTAG-SET1A/B cells treated with dTAG-13 over a 96 hr time course. H3 is a loading control.

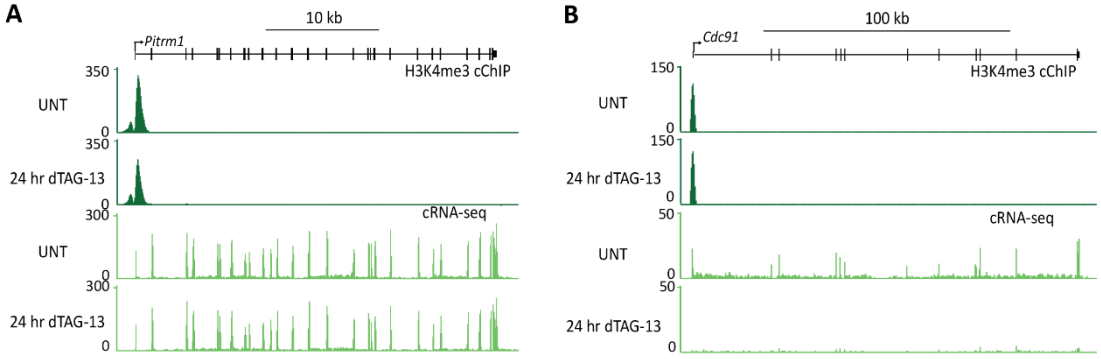


Figure A7 Genomic snapshots of H3K4me3 cChIP-seq and cRNA-seq following 24 hr SET1A/B depletion
Genomic snapshots showing H3K4me3 cChIP-seq signal and cRNA-seq signal at one gene with reduced H3K4me3 and unchanged expression (*Pitrm1,* A) and one gene with reduced expression but unchanged H3K4me3 (*Cdc91*, B) following depletion of SET1A/B for 24 hr.

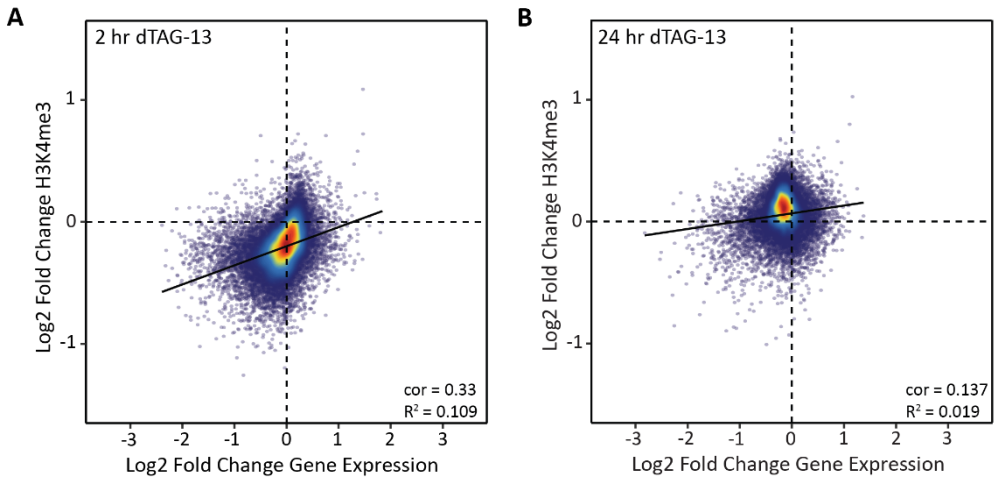


Figure A8 Changes in H3K4me3 and gene expression were not strongly correlated at either time point in dTAG-SET1A cells
Scatterplots comparing the log2 fold change in gene expression (cRNA-seq) and H3K4me3 cChIP-seq signal at promoter-associated H3K4me3 peaks in dTAG-SET1A cells treated with dTAG-13 for 2 hr (A) or 24 hr (B). Only genes that have a promoter-associated H3K4me3 peak in untreated cells are included (n=14546). The dashed lines indicate log2 fold change=0, while the solid line represents linear regression. Cor denotes the Pearson correlation coefficient and R^2 the coefficient of determination for linear regression.

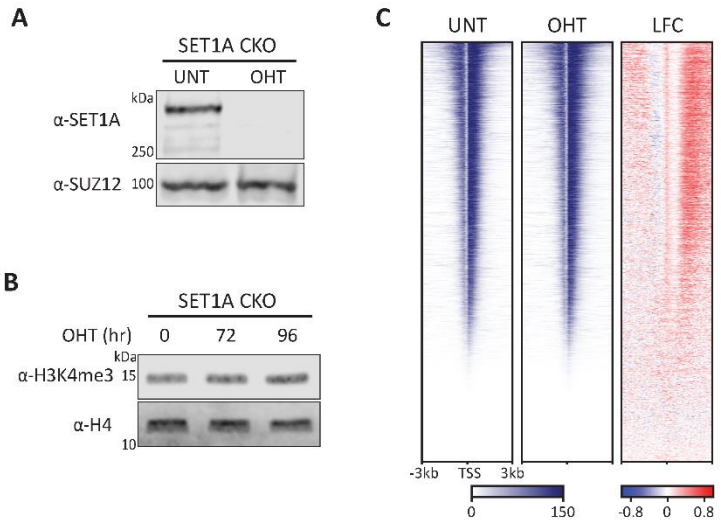


Figure A9 The conditional knockout of SET1A results in subtle changes in H3K4me3

(A) Western blot of nuclear extract showing levels of SET1A in the SET1A conditional knockout (CKO) cell line untreated (UNT) and treated with tamoxifen (OHT) for 96 hr. SUZ12 is a loading control.

(B) Western blot showing levels of H3K4me3 in histone extract from SET1A CKO cells untreated and treated with tamoxifen (OHT) for 72 and 96 hr. H4 is a loading control.

(C) Heatmaps of H3K4me3 cChIP-seq signal at all TSSs (n=20633) in untreated and 96 hr OHT-treated SET1A CKO cells. The log2 fold change (LFC) in H3K4me3 cChIP-seq signal is also shown. Heatmaps are sorted by mean signal in untreated cells.

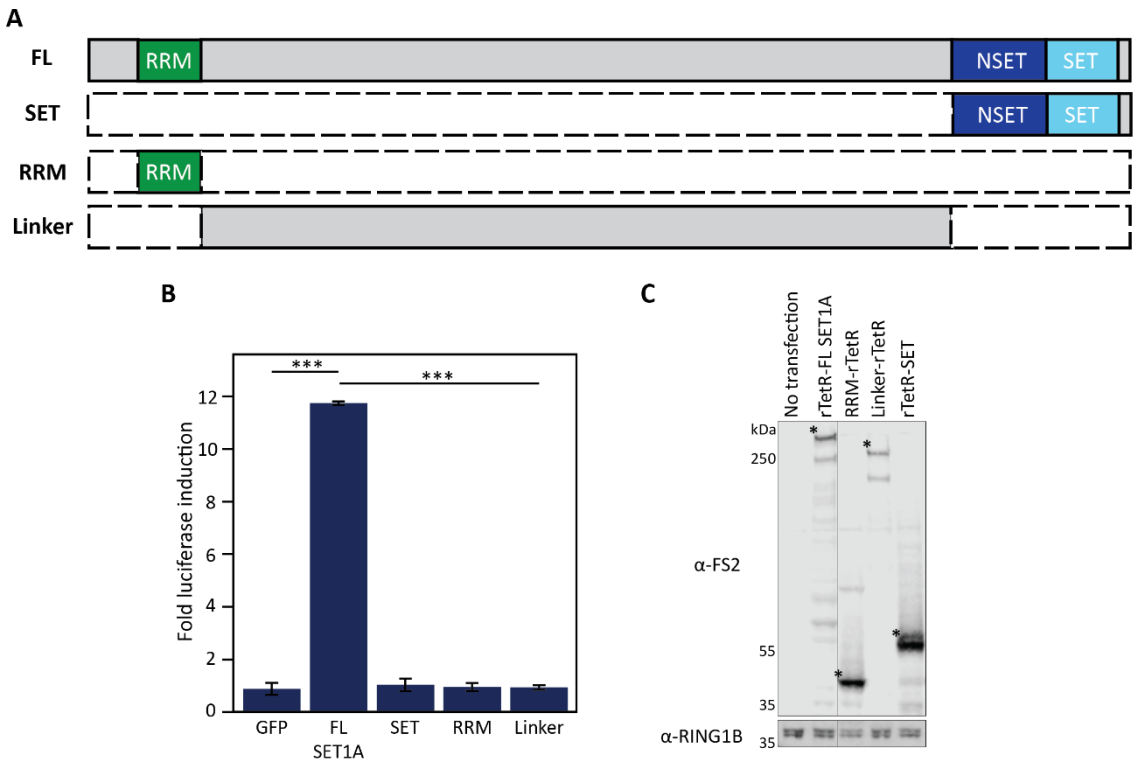


Figure A10 The SET domain, RRM and linker region of SET1A are not sufficient to activate luciferase expression

(A) Schematic of the SET1A constructs tethered to the TetO array.

(B) Bar chart showing the fold induction in luciferase levels following tethering of GFP, full length (FL) SET1A, SET1A SET domain, RRM or linker to the TetO array. Luciferase expression is normalised to *Renilla* luciferase which was transiently transfected along with the rTetR-fusions to account for transfection efficiency and expressed relative to luciferase levels in the absence of doxycycline. Mean luciferase induction is shown and error bars represent SEM from three biological replicates. *** represents $p<0.001$ by Student's T test.

(C) Western blot of nuclear extract from untransfected luciferase reporter cells or cells transiently transfected with the annotated rTetR-fusion proteins. The blot was probed with FLAG-StrepII antibody, which recognises the tags that were fused to the rTetR. RING1B is a loading control and * denotes the specific rTetR fusion from each transfection.

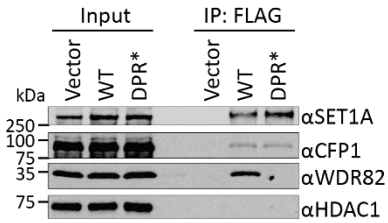


Figure A11 SET1A DPR/AAA mutation inhibits its association with WDR82 (Courtesy of Jessica Kelley)
FLAG IP from nuclear extracts of HEK 293T cells transiently transfected with vectors expressing either WT or DPR/AAA mutant SET1A (DPR*) with an N-terminal FLAG tag. IP from cells transfected with an empty vector was also performed as a control. Levels of interacting WDR82 and CFP1 are shown. 10% input was loaded and HDAC1 is a loading control.

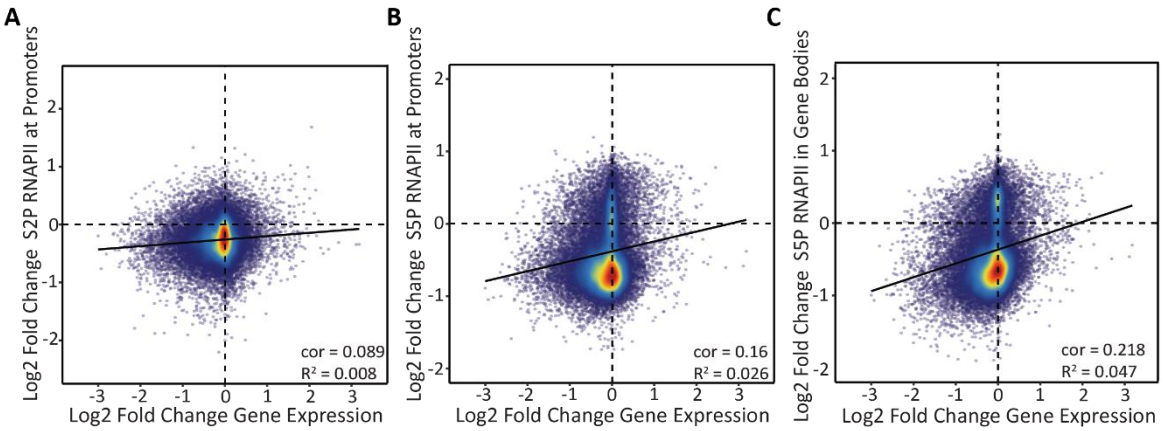


Figure A12 Changes in S5P RNAPII and promoter-associated S2P RNAPII are not correlated with gene expression
Scatterplots comparing the log2 fold change in cRNA-seq signal and promoter-associated S2P RNAPII (A), promoter-associated S5P RNAPII (B) or gene body-associated S5P RNAPII (C) cChIP-seq signal in dTAG-SET1A/B cells treated with dTAG-13 for 2 hr. The dashed lines indicate log2 fold change=0, while the solid line represents linear regression. Cor denotes the Pearson correlation coefficient and R^2 the coefficient of determination for linear regression.

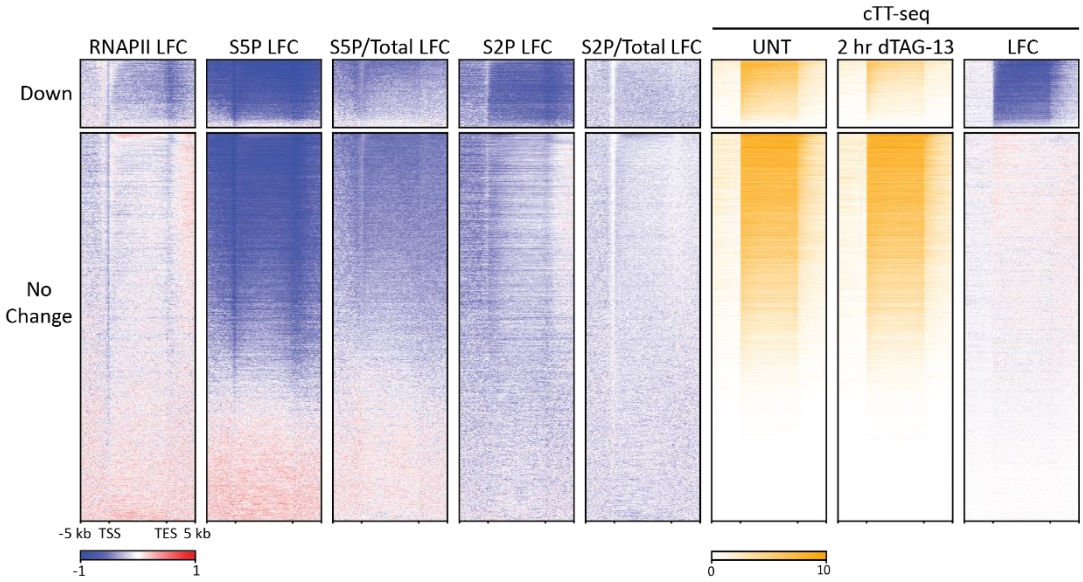


Figure A13 Removal of SET1A/B results in more pervasive effects on RNAPII than transcription
Heatmaps of the log2 fold change in cChIP-seq signal at gene bodies for total RNAPII (RNAPII LFC), S5P RNAPII (S5P LFC), S5P RNAPII normalised for total RNAPII (S5P/Total LFC), S2P RNAPII (S2P LFC) and S2P RNAPII normalised for total RNAPII (S2P/Total LFC) in dTAG-SET1A/B cells depleted of SET1A/B for 2 hr. Also shown is cTT-seq signal in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells and the log2 fold change in cTT-seq (cTT-seq LFC). Genes are split into those that experienced a significant reduction (Down, n=3098), or no change (No Change, n=17460) in transcription based on cTT-seq analysis. Intervals are sorted by total RNAPII cChIP-seq signal in untreated cells.

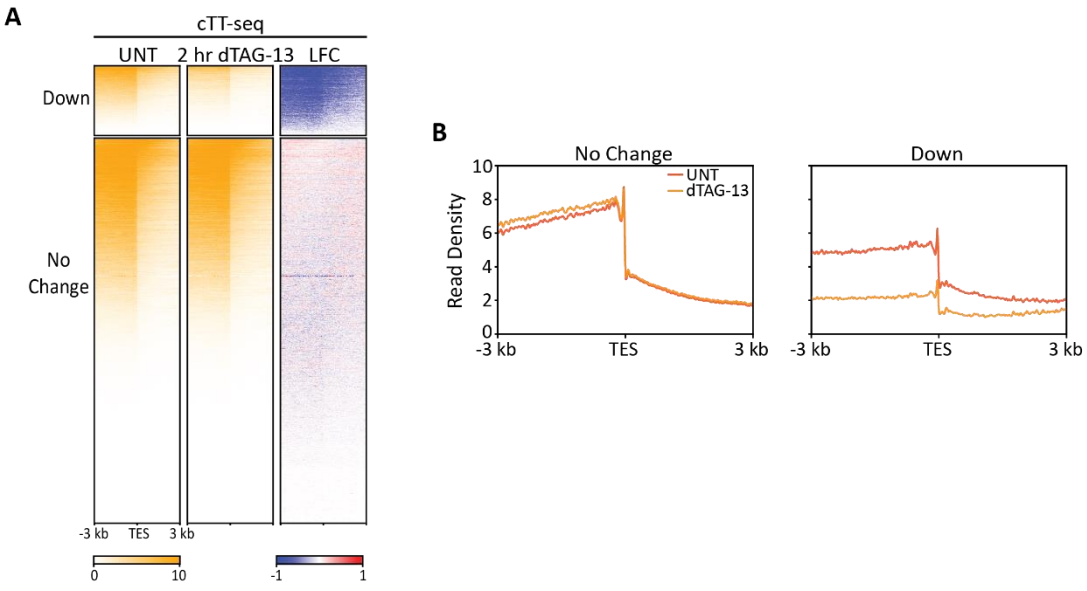


Figure A14 SET1A/B depletion does not result in transcription read-through at TESs
(A) Heatmaps of cTT-seq signal at TESs in untreated and 2 hr dTAG-13-treated dTAG-SET1A/B cells. Also shown is log2 fold change in cTT-seq signal (LFC). Genes are split into those that experienced a significant reduction (Down, n=3098), or no change (No Change, n=17460) in transcription based on cTT-seq analysis. Intervals are sorted by total RNAPII cChIP-seq signal in untreated cells.
(B) Metaplots of cTT-seq signal at TESs in untreated cells (UNT) and 2 hr dTAG-13-treated dTAG-SET1A/B cells. Genes are split into the same groups as in (A).

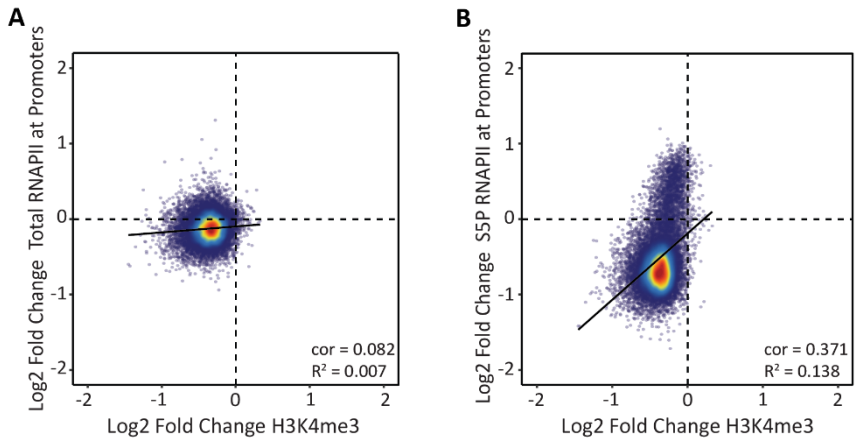


Figure A15 There is a moderate correlation between change in H3K4me3 and promoter-associated S5P RNAPII

Scatterplots comparing the log2 fold change in H3K4me3 cChIP-seq signal and promoter-associated total RNAPII (A) or promoter-associated S5P RNAPII (B) following 2 hr SET1A/B depletion. Only genes with a H3K4me3 peak in untreated cells are included (n=14546). The dashed lines indicate log2 fold change=0, while the solid line represents linear regression. Cor denotes the Pearson correlation coefficient and R^2 the coefficient of determination for linear regression.

Bibliography

Abdella, R. *et al.* (2021) 'Structure of the human Mediator-bound transcription preinitiation complex', *Science* 372: pp. 52-56. doi: 10.1126/science.abg3074.

Acquaviva, L. *et al.* (2013) 'Spp1 at the crossroads of H3K4me3 regulation and meiotic recombination', *Epigenetics* 8: pp. 355-360. doi: 10.4161/epi.24295.

Adelman, K. and Lis, J. T. (2012) 'Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans', *Nature Reviews Genetics* 13: pp. 720–731. doi: 10.1038/nrg3293.

Allen, B. L. and Taatjes, D. J. (2015) 'The Mediator complex: A central integrator of transcription', *Nature Reviews Molecular Cell Biology* 16: pp. 155–166. doi: 10.1038/nrm3951.

Allis, C. D. and Jenuwein, T. (2016) 'The molecular hallmarks of epigenetic control', *Nature Reviews Genetics* 17: pp. 487–500. doi: 10.1038/nrg.2016.59.

Alsulami, M. *et al.* (2019) 'SETD1A methyltransferase is physically and functionally linked to the DNA damage repair protein RAD18', *Molecular and Cellular Proteomics* 18: pp. 1428–1436. doi: 10.1074/mcp.RA119.001518.

Anderson, J. D., Lowary, P. T. and Widom, J. (2001) 'Effects of histone acetylation on the equilibrium accessibility of nucleosomal DNA target sites', *Journal of Molecular Biology* 307: pp. 977–985. doi: 10.1006/jmbi.2001.4528.

Antequera, F. (2003) 'Structure, function and evolution of CpG island promoters', *Cellular and Molecular Life Sciences* 60: pp. 1647–1658. doi: 10.1007/s00018-003-3088-6.

Antequera, F., Macleod, D. and Bird, A. P. (1989) 'Specific protection of methylated CpGs in mammalian nuclei', *Cell* 58: pp. 509–517. doi: 10.1016/0092-8674(89)90431-5.

Ardehali, M. B. *et al.* (2011) 'Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription', *The EMBO Journal* 30: pp. 2817–2828. doi: 10.1038/emboj.2011.194.

Ashokkumar, D. *et al.* (2020) 'MLL4 is required after implantation, whereas MLL3 becomes essential during late gestation', *Development* 147: dev186999. doi: 10.1242/dev.186999.

Audia, J. E. and Campbell, R. M. (2016) 'Histone modifications and cancer', *Cold Spring Harbor Perspectives in Biology* 8: a019521. doi: 10.1101/cshperspect.a019521.

Austenaa, L. M. I. *et al.* (2015) 'Transcription of Mammalian cis-Regulatory Elements Is Restrained by Actively Enforced Early Termination', *Molecular Cell* 60: pp. 460–474. doi: 10.1016/j.molcel.2015.09.018.

Azuara, V. *et al.* (2006) 'Chromatin signatures of pluripotent cell lines', *Nature Cell Biology* 8: pp. 532–538. doi: 10.1038/ncb1403.

Bae, H. J. *et al.* (2020) 'The Set1 N-terminal domain and Swd2 interact with RNA polymerase II CTD to recruit COMPASS', *Nature Communications* 11: pp. 1–10. doi: 10.1038/s41467-020-16082-2.

Bannister, A. J. *et al.* (2005) 'Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes', *Journal of Biological Chemistry* 280: pp. 17732–17736. doi: 10.1074/jbc.M500796200.

Barnes, C. E., English, D. M. and Cowley, S. M. (2019) 'Acetylation and Co: An expanding repertoire of histone acylations regulates chromatin and transcription', *Essays in Biochemistry* 63: pp. 97–107. doi: 10.1042/EBC20180061.

Barski, A. *et al.* (2007) 'High-Resolution Profiling of Histone Methylations in the Human Genome', *Cell* 129: pp. 823–837. doi: 10.1016/j.cell.2007.05.009.

Bartke, T. *et al.* (2010) 'Nucleosome-interacting proteins regulated by DNA and histone

methylation', *Cell* 143: pp. 470–484. doi: 10.1016/j.cell.2010.10.012.

Basu, S., Nandy, A. and Biswas, D. (2020) 'Keeping RNA polymerase II on the run: Functions of MLL fusion partners in transcriptional regulation', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1863: 194563. doi: 10.1016/j.bbagrm.2020.194563.

Battaglia, S. *et al.* (2017) 'RNA-dependent chromatin association of transcription elongation factors and pol II CTD kinases', *eLife* 6: e25637. doi: 10.7554/eLife.25637.

Becker, J. S., Nicetto, D. and Zaret, K. S. (2016) 'H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes', *Trends in Genetics* 32: pp. 29-41. doi: 10.1016/j.tig.2015.11.001.

Ben-Shem, A., Papai, G. and Schultz, P. (2020) 'Architecture of the multi-functional SAGA complex and the molecular mechanism of holding TBP', *The FEBS Journal* 288: pp. 3135-3147. doi: 10.1111/febs.15563.

Bernecky, C. *et al.* (2016) 'Structure of transcribing mammalian RNA polymerase II', *Nature* 529: pp. 551–554. doi: 10.1038/nature16482.

Bernstein, B. E. *et al.* (2005) 'Genomic maps and comparative analysis of histone modifications in human and mouse', *Cell* 120: pp. 169–181. doi: 10.1016/j.cell.2005.01.001.

Bernstein, B. E. *et al.* (2006) 'A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells', *Cell* 125: pp. 315–326. doi: 10.1016/j.cell.2006.02.041.

Bhuiyan, T. and Timmers, H. T. M. (2019) 'Promoter Recognition: Putting TFIID on the Spot', *Trends in Cell Biology* 29: pp. 752–763. doi: 10.1016/j.tcb.2019.06.004.

Bian, C. *et al.* (2011) 'Sgf29 binds histone H3K4me2/3 and is required for SAGA complex recruitment and histone H3 acetylation', *The EMBO Journal* 30: pp. 2829–2842. doi: 10.1038/emboj.2011.193.

Bilodeau, S. *et al.* (2009) 'SetDB1 contributes to repression of genes encoding developmental regulators and maintenance of ES cell state', *Genes & Development* 23: pp. 2484–2489. doi: 10.1101/gad.1837309.

Binda, O. *et al.* (2010) 'Trimethylation of histone H3 lysine 4 impairs methylation of histone H3 lysine 9: Regulation of lysine methyltransferases by physical interaction with their substrates', *Epigenetics* 5: pp. 767–775. doi: 10.4161/epi.5.8.13278.

Binder, H. *et al.* (2013) 'Transcriptional regulation by histone modifications: Towards a theory of chromatin re-organization during stem cell differentiation', *Physical Biology* 10: 026006. doi: 10.1088/1478-3975/10/2/026006.

Bird, A. *et al.* (1985) 'A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA', *Cell* 40: pp. 91–99. doi: 10.1016/0092-8674(85)90312-5.

Bird, A. (2002) 'DNA methylation patterns and epigenetic memory', *Genes & Development* 16: pp. 6–21. doi: 10.1101/gad.947102.

Bird, A. P. and Wolffe, A. P. (1999) 'Methylation-induced repression-belts, braces, and chromatin', *Cell* 99: pp. 451–454. doi: 10.1016/S0092-8674(00)81532-9.

Blackledge, N. P. *et al.* (2014) 'Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation', *Cell* 157: pp. 1445–1459. doi: 10.1016/j.cell.2014.05.004.

Blackledge, N. P. *et al.* (2020) 'PRC1 Catalytic Activity Is Central to Polycomb System Function', *Molecular Cell* 77: pp. 857-874.e9. doi: 10.1016/j.molcel.2019.12.001.

Blackledge, N. P. and Klose, R. J. (2011) 'CpG island chromatin: A platform for gene regulation', *Epigenetics* 6: pp. 147–152. doi: 10.4161/epi.6.2.13640.

Blackledge, N. P., Rose, N. R. and Klose, R. J. (2015) 'Targeting Polycomb systems to regulate gene

expression: Modifications to a complex story', *Nature Reviews Molecular Cell Biology* 16: pp. 643–649. doi: 10.1038/nrm4067.

Bledau, A. S. *et al.* (2014) 'The H3K4 methyltransferase Setd1a is first required at the epiblast stage, whereas Setd1b becomes essential after gastrulation', *Development* 141: pp. 1022-1035. doi: 10.1242/dev.098152.

Boa, S., Coert, C. and Patterton, H. G. (2003) 'Saccharomyces cerevisiae Set1p is a methyltransferase specific for lysine 4 of histone H3 and is required for efficient gene expression', *Yeast* 20: pp. 827–835. doi: 10.1002/yea.995.

Bochyńska, A., Lüscher-Firzlaff, J. and Lüscher, B. (2018) 'Modes of Interaction of KMT2 Histone H3 Lysine 4 Methyltransferase/COMPASS Complexes with Chromatin', *Cells* 7: 17. doi: 10.3390/cells7030017.

Boeing, S. *et al.* (2010) 'RNA polymerase II C-terminal heptarepeat domain Ser-7 phosphorylation is established in a mediator-dependent fashion', *Journal of Biological Chemistry* 285: pp. 188–196. doi: 10.1074/jbc.M109.046565.

Bonhoure, N. *et al.* (2014) 'Quantifying ChIP-seq data: A spiking method providing an internal reference for sample-to-sample normalization', *Genome Research* 24: pp. 1157–1168. doi: 10.1101/gr.168260.113.

Bornelö, S. *et al.* (2018) 'The Nucleosome Remodeling and Deacetylation Complex Modulates Chromatin Structure at Sites of Active Transcription to Fine-Tune Gene Expression', *Molecular Cell* 71: pp. 56–72. doi: 10.1016/j.molcel.2018.06.003.

Breen, T. R. (1999) 'Mutant Alleles of the *Drosophila trithorax* Gene Produce Common and Unusual Homeotic and Other Developmental Phenotypes', *Genetics* 152: pp. 319-344. doi: 10.1093/genetics/152.1.319.

Breen, T. R. and Harte, P. J. (1991) 'Molecular characterization of the trithorax gene, a positive regulator of homeotic gene expression in Drosophila', *Mechanisms of Development* 35: pp. 113–127. doi: 10.1016/0925-4773(91)90062-B.

Briggs, S. D. *et al.* (2001) 'Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae', *Genes & Development* 15: pp. 3286–3295. doi: 10.1101/gad.940201.

Brown-Bryan, T. A. *et al.* (2008) 'Alternative splicing and caspase-mediated cleavage generate antagonistic variants of the stress oncoprotein LEDGF/p75', *Molecular Cancer Research* 6: pp. 1293–1307. doi: 10.1158/1541-7786.MCR-08-0125.

Brown, D. A. *et al.* (2017) 'The SET1 Complex Selects Actively Transcribed Target Genes via Multivalent Interaction with CpG Island Chromatin', *Cell Reports* 20: pp. 2313–2327. doi: 10.1016/j.celrep.2017.08.030.

Buratowski, S. *et al.* (1989) 'Five intermediate complexes in transcription initiation by RNA polymerase II', *Cell* 56: pp. 549–561. doi: 10.1016/0092-8674(89)90578-3.

Cano-Rodriguez, D. *et al.* (2016) 'Writing of H3K4Me3 overcomes epigenetic silencing in a sustained but context-dependent manner', *Nature Communications* 7: 12284. doi: 10.1038/ncomms12284.

Cao, K. *et al.* (2017) 'SET1A/COMPASS and shadow enhancers in the regulation of homeotic gene expression', *Genes & Development* 31: pp. 787–801. doi: 10.1101/gad.294744.116.

Cao, R. *et al.* (2002) 'Role of histone H3 lysine 27 methylation in polycomb-group silencing', *Science* 298: pp. 1039–1043. doi: 10.1126/science.1076997.

Carlone, D. L. *et al.* (2005) 'Reduced Genomic Cytosine Methylation and Defective Cellular

Differentiation in Embryonic Stem Cells Lacking CpG Binding Protein', *Molecular and Cellular Biology* 25: pp. 4881–4891. doi: 10.1128/mcb.25.12.4881-4891.2005.

Carlone, D. L. and Skalnik, D. G. (2001) 'CpG Binding Protein Is Crucial for Early Embryonic Development', *Molecular and Cellular Biology* 21: pp. 7601–7606. doi: 10.1128/mcb.21.22.7601-7606.2001.

Caslini, C. *et al.* (2007) 'Interaction of MLL amino terminal sequences with menin is required for transformation', *Cancer Research* 67: pp. 7275–7283. doi: 10.1158/0008-5472.CAN-06-2369.

Casper, J. *et al.* (2018) 'The UCSC Genome Browser database: 2018 update', *Nucleic Acids Research* 46: pp. D762–D769. doi: 10.1093/nar/gkx1020.

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

Champagne, K. and Kutateladze, T. (2009) 'Structural Insight Into Histone Recognition by the ING PHD Fingers', *Current Drug Targets* 10: pp. 432–441. doi: 10.2174/138945009788185040.

Champagne, K. S. *et al.* (2008) 'The crystal structure of the ING5 PHD finger in complex with an H3K4me3 histone peptide', *Proteins: Structure, Function, and Bioinformatics* 72: pp. 1371–1376. doi: 10.1002/prot.22140.

Chang, P. Y. *et al.* (2010) 'Binding of the MLL PHD3 Finger to Histone H3K4me3 Is Required for MLL-Dependent Gene Transcription', *Journal of Molecular Biology* 400: pp. 137–144. doi: 10.1016/j.jmb.2010.05.005.

Chao, S. H. and Price, D. H. (2001) 'Flavopiridol Inactivates P-TEFb and Blocks Most RNA Polymerase II Transcription in Vivo', *Journal of Biological Chemistry* 276: pp. 31793–31799. doi: 10.1074/jbc.M102306200.

Chédin, F., Lieber, M. R. and Hsieh, C. L. (2002) 'The DNA methyltransferase-like protein DNMT3L stimulates de novo methylation by Dnmt3a', *Proceedings of the National Academy of Sciences of the United States of America* 99: pp. 16916–16921. doi: 10.1073/pnas.262443999.

Chen, S. *et al.* (2018) 'Unique Structural Platforms of Suz12 Dictate Distinct Classes of PRC2 for Chromatin Binding', *Molecular Cell* 69: pp. 840-852.e5. doi: 10.1016/j.molcel.2018.01.039.

Chen, Y. *et al.* (2011) 'Crystal structure of the N-terminal region of human Ash2L shows a winged-helix motif involved in DNA binding', *EMBO Reports* 12: pp. 797–803. doi: 10.1038/embor.2011.101.

Chen, Y. X. *et al.* (2006) 'The tumor suppressor menin regulates hematopoiesis and myeloid transformation by influencing Hox gene expression', *Proceedings of the National Academy of Sciences of the United States of America* 103: pp. 1018–1023. doi: 10.1073/pnas.0510347103.

Cheng, H., He, X. and Moore, C. (2004) 'The Essential WD Repeat Protein Swd2 Has Dual Functions in RNA Polymerase II Transcription Termination and Lysine 4 Methylation of Histone H3', *Molecular and Cellular Biology* 24: pp. 2932–2943. doi: 10.1128/mcb.24.7.2932-2943.2004.

Chi, P., Allis, C. D. and Wang, G. G. (2010) 'Covalent histone modifications-miswritten, misinterpreted and mis-erased in human cancers', *Nature Reviews Cancer* 10: pp. 457–469. doi: 10.1038/nrc2876.

Chih, L. L. *et al.* (2005) 'Single-nucleosome mapping of histone modifications in S. cerevisiae', *PLoS Biology* 3: e328. doi: 10.1371/journal.pbio.0030328.

Cho, H. S. *et al.* (2012) 'Enhanced HSP70 lysine methylation promotes proliferation of cancer cells through activation of Aurora kinase B', *Nature Communications* 3: pp. 1–10. doi: 10.1038/ncomms2074.

Choi, J. K. (2010) 'Contrasting chromatin organization of CpG islands and exons in the human

genome', *Genome Biology* 11: R70. doi: 10.1186/gb-2010-11-7-r70.
Ciccone, D. N. *et al.* (2009) 'KDM1B is a histone H3K4 demethylase required to establish maternal genomic imprints', *Nature* 461: pp. 415–418. doi: 10.1038/nature08315.
Clapier, C. R. and Cairns, B. R. (2009) 'The biology of chromatin remodeling complexes', *Annual Review of Biochemistry* 78: pp. 273–304. doi: 10.1146/annurev.biochem.77.062706.153223.
Clouaire, T. *et al.* (2012) 'Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells.', *Genes & Development* 26: pp. 1714–1728. doi: 10.1101/gad.194209.112.
Clouaire, T., Webb, S. and Bird, A. (2014) 'Cfp1 is required for gene expression-dependent H3K4 trimethylation and H3K9 acetylation in embryonic stem cells', *Genome Biology* 15: p. 451. doi: 10.1186/s13059-014-0451-x.
Cockerill, P. N. (2011) 'Structure and function of active chromatin and DNaseI hypersensitive sites', *FEBS Journal* 278: pp. 2182–2210. doi: 10.1111/j.1742-4658.2011.08128.x.
Coker, H. *et al.* (2020) 'The role of the Xist 5' m6A region and RBM15 in X chromosome inactivation', *Wellcome Open Research* 5: p. 31. doi: 10.12688/wellcomeopenres.15711.1.
Colgan, D. F. and Manley, J. L. (1997) 'Mechanism and regulation of mRNA polyadenylation', *Genes & Development* 11: pp. 2755–2766. doi: 10.1101/gad.11.21.2755.
Corden, J. L. *et al.* (1985) 'A unique structure at the carboxyl terminus of the largest subunit of eukaryotic RNA polymerase II', *Proceedings of the National Academy of Sciences of the United States of America* 82: pp. 7934–7938. doi: 10.1073/pnas.82.23.7934.
Core, L. J., Waterfall, J. J. and Lis, J. T. (2008) 'Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters', *Science* 322: pp. 1845–1848. doi: 10.1126/science.1162228.
Cortazar, M. A. *et al.* (2019) 'Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a "Sitting Duck Torpedo" Mechanism', *Molecular Cell* 76: pp. 896-908.e4. doi: 10.1016/j.molcel.2019.09.031.
Coulondre, C. *et al.* (1978) 'Molecular basis of base substitution hotspots in Escherichia coli', *Nature* 274: pp. 775–780. doi: 10.1038/274775a0.
Crump, N. T. *et al.* (2011) 'Dynamic acetylation of all lysine-4 trimethylated histone H3 is evolutionarily conserved and mediated by p300/CBP', *Proceedings of the National Academy of Sciences of the United States of America* 108: pp. 7814–7819. doi: 10.1073/pnas.1100099108.
Crump, N. T. and Milne, T. A. (2019) 'Why are so many MLL lysine methyltransferases required for normal mammalian development?', *Cellular and Molecular Life Sciences* 76: pp. 2885–2898. doi: 10.1007/s00018-019-03143-z.
Deaton, A. M. and Bird, A. (2011) 'CpG islands and the regulation of transcription', *Genes & Development* 25: pp. 1010–1022. doi: 10.1101/gad.2037511.
Deng, C. *et al.* (2013) 'USF1 and hSET1A Mediated Epigenetic Modifications Regulate Lineage Differentiation and HoxB4 Transcription', *PLoS Genetics* 9: e1003524. doi: 10.1371/journal.pgen.1003524.
Denissov, S. *et al.* (2014) 'Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant', *Development* 141: pp. 526-537. doi: 10.1242/dev.102681.
De Dieuleveult, M. *et al.* (2016) 'Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells', *Nature* 530: pp. 113–116. doi: 10.1038/nature16505.
Dillon, S. C. *et al.* (2005) 'The SET-domain protein superfamily: Protein lysine methyltransferases',

Genome Biology 6: p. 227. doi: 10.1186/gb-2005-6-8-227.

Dimitrova, E. *et al.* (2018) 'FBXl19 recruits CDK-Mediator to CpG islands of developmental genes priming them for activation during lineage commitment', *eLife* 7: e37084. doi: 10.7554/eLife.37084.

Dobin, A. *et al.* (2013) 'STAR: Ultrafast universal RNA-seq aligner', *Bioinformatics* 29: pp. 15–21. doi: 10.1093/bioinformatics/bts635.

Dobrinić, P., Szczurek, A. T. and Klose, R. J. (2021) 'PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency', *Nature Structural and Molecular Biology* 28: pp. 811-824. doi: 10.1038/s41594-021-00661-y.

Dou, Y. *et al.* (2005) 'Physical Association and Coordinate Function of the H3 K4 Methyltransferase MLL1 and the H4 K16 Acetyltransferase MOF', *Cell* 121: pp. 873–885. doi: 10.1016/j.cell.2005.04.031.

Dou, Y. *et al.* (2006) 'Regulation of MLL1 H3K4 methyltransferase activity by its core components', *Nature Structural and Molecular Biology* 13: pp. 713–719. doi: 10.1038/nsmb1128.

Douillet, D. *et al.* (2020) 'Uncoupling histone H3K4 trimethylation from developmental gene expression via an equilibrium of COMPASS, Polycomb and DNA methylation', *Nature Genetics* 52: pp. 615–625. doi: 10.1038/s41588-020-0618-1.

Doyon, Y. *et al.* (2006) 'ING Tumor Suppressor Proteins Are Critical Regulators of Chromatin Acetylation Required for Genome Expression and Perpetuation', *Molecular Cell* 21: pp. 51–64. doi: 10.1016/J.MOLCEL.2005.12.007.

Duncan, B. K. and Miller, J. H. (1980) 'Mutagenic deamination of cytosine residues in DNA', *Nature* 287: pp. 560–561. doi: 10.1038/287560a0.

Dunham, I. *et al.* (2012) 'An integrated encyclopedia of DNA elements in the human genome', *Nature* 489: pp. 57–74. doi: 10.1038/nature11247.

Van Dyke, M. W., Roeder, R. G. and Sawadogo, M. (1988) 'Physical analysis of transcription preinitiation complex assembly on a class II gene promoter', *Science* 241: pp. 1335–1338. doi: 10.1126/science.3413495.

Eaton, J. D. *et al.* (2020) 'A unified allosteric/torpedo mechanism for transcriptional termination on human protein-coding genes', *Genes & Development* 34: pp. 132–145. doi: 10.1101/gad.332833.119.

Eberl, H. C. *et al.* (2013) 'A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-free Interaction Proteomics', *Molecular Cell* 49: pp. 368–378. doi: 10.1016/j.molcel.2012.10.026.

Ebmeier, C. C. *et al.* (2017) 'Human TFIIH Kinase CDK7 Regulates Transcription-Associated Chromatin Modifications', *Cell Reports* 20: pp. 1173–1186. doi: 10.1016/J.CELREP.2017.07.021.

Eick, D. and Geyer, M. (2013) 'The RNA polymerase II carboxy-terminal domain (CTD) code', *Chemical Reviews* 113: pp. 8456–8490. doi: 10.1021/cr400071f.

Eidahl, J. O. *et al.* (2013) 'Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes', *Nucleic Acids Research* 41: pp. 3924–3936. doi: 10.1093/nar/gkt074.

Ernst, P. *et al.* (2001) 'MLL and CREB Bind Cooperatively to the Nuclear Coactivator CREB-Binding Protein', *Molecular and Cellular Biology* 21: pp. 2249–2258. doi: 10.1128/mcb.21.7.2249-2258.2001.

Faedo, A. *et al.* (2017) 'Differentiation of human telencephalic progenitor cells into MSNs by inducible expression of Gsx2 and Ebf1', *Proceedings of the National Academy of Sciences of the United States of America* 114: pp. E1234–E1242. doi: 10.1073/pnas.1611473114.

Fang, L. *et al.* (2016) 'H3K4 Methyltransferase Set1a Is A Key Oct4 Coactivator Essential for Generation of Oct4 Positive Inner Cell Mass', *Stem Cells* 34: pp. 565–580. doi: 10.1002/stem.2250.

Fang, L. *et al.* (2018) 'SET1A-Mediated Mono-Methylation at K342 Regulates YAP Activation by Blocking Its Nuclear Export and Promotes Tumorigenesis', *Cancer Cell* 34: pp. 103-118.e9. doi: 10.1016/j.ccell.2018.06.002.

Feng, W. *et al.* (2010) 'PHF8 activates transcription of rRNA genes through H3K4me3 binding and H3K9me1/2 demethylation', *Nature Structural and Molecular Biology* 17: pp. 445–450. doi: 10.1038/nsmb.1778.

Fenouil, R. *et al.* (2012) 'CpG islands and GC content dictate nucleosome depletion in a transcription-independent manner at mammalian promoters', *Genome Research* 22: pp. 2399–2408. doi: 10.1101/gr.138776.112.

Filippakopoulos, P. *et al.* (2012) 'Histone recognition and large-scale structural analysis of the human bromodomain family', *Cell* 149: pp. 214–231. doi: 10.1016/j.cell.2012.02.013.

Flanagan, J. F. *et al.* (2005) 'Double chromodomains cooperate to recognize the methylated histone H3 tail', *Nature* 438: pp. 1181–1185. doi: 10.1038/nature04290.

Fong, N. *et al.* (2015) 'Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition', *Molecular Cell* 60: pp. 256–267. doi: 10.1016/j.molcel.2015.09.026.

Fortschegger, K. *et al.* (2010) 'PHF8 Targets Histone Methylation and RNA Polymerase II To Activate Transcription', *Molecular and Cellular Biology* 30: pp. 3286–3298. doi: 10.1128/mcb.01520-09.

Fournier, M. *et al.* (2016) 'FOXA and master transcription factors recruit Mediator and Cohesin to the core transcriptional regulatory circuitry of cancer cells', *Scientific Reports* 6: pp. 1–11. doi: 10.1038/srep34962.

Franks, T. M. *et al.* (2017) 'Nup98 recruits the Wdr82-Set1A/ COMPASS complex to promoters to regulate H3K4 trimethylation in hematopoietic progenitor cells', *Genes & Development* 31: pp. 2222-2234. doi: 10.1101/gad.306753.117.

Frantz, G. D. *et al.* (2001) 'DNA methylation in genomic imprinting, development, and disease', *Journal of Pathology* 195: pp. 97–110. doi: 10.1002/path.890.

Fraser, N. W., Sehgal, P. B. and Darnell, J. E. (1978) 'DRB-induced premature termination of late adenovirus transcription', *Nature* 272: pp. 590–593. doi: 10.1038/272590a0.

Frietze, S. and Farnham, P. J. (2011) 'Transcription factor effector domains', *Sub-Cellular Biochemistry* 52: pp. 261–277. doi: 10.1007/978-90-481-9069-0_12.

Fujinaga, K. *et al.* (2004) 'Dynamics of Human Immunodeficiency Virus Transcription: P-TEFb Phosphorylates RD and Dissociates Negative Effectors from the Transactivation Response Element', *Molecular and Cellular Biology* 24: pp. 787–795. doi: 10.1128/mcb.24.2.787-795.2004.

Fursova, N. A. *et al.* (2019) 'Synergy between Variant PRC1 Complexes Defines Polycomb-Mediated Gene Repression', *Molecular Cell* 74: pp. 1020-1036.e8. doi: 10.1016/j.molcel.2019.03.024.

Fursova, N. A. *et al.* (2020) 'BAP1 constrains pervasive H2AK119ub1 to control the transcriptional potential of the genome', *Genes & Development* 35: pp. 749-770. doi: 10.1101/gad.347005.120.

Gahurova, L. *et al.* (2017) 'Transcription and chromatin determinants of de novo DNA methylation timing in oocytes', *Epigenetics & Chromatin* 10: p. 25. doi: 10.1186/s13072-017-0133-5.

Gardiner-Garden, M. and Frommer, M. (1987) 'CpG Islands in vertebrate genomes', *Journal of Molecular Biology* 196: pp. 261–282. doi: 10.1016/0022-2836(87)90689-9.

Gariglio, P., Bellard, M. and Chambon, P. (1981) 'Clustering of RNA polymerase B molecules in the 5' moiety of the adult β-globin gene of hen erythrocytes', *Nucleic Acids Research* 9: pp. 2589–2598. doi: 10.1093/nar/9.11.2589.

Garneau, J. E. *et al.* (2010) 'The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA', *Nature* 468: pp. 67–71. doi: 10.1038/nature09523.

Gasiunas, G. *et al.* (2012) 'Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria', *Proceedings of the National Academy of Sciences of the United States of America* 109: pp. E2579–E2586. doi: 10.1073/pnas.1208507109.

Gaspar-Maia, A. *et al.* (2009) 'Chd1 regulates open chromatin and pluripotency of embryonic stem cells.', *Nature* 460: pp. 863–868. doi: 10.1038/nature08212.

Giardina, C., Perez-Riba, M. and Lis, J. T. (1992) 'Promoter melting and TFIID complexes on Drosophila genes in vivo', *Genes & Development* 6: pp. 2190–2200. doi: 10.1101/gad.6.11.2190.

Gilmour, D. S. and Lis, J. T. (1986) 'RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells.', *Molecular and Cellular Biology* 6: pp. 3984–3989. doi: 10.1128/mcb.6.11.3984.

Glaser, S. *et al.* (2006) 'Multiple epigenetic maintenance factors implicated by the loss of MII2 in mouse development', *Development* 133: pp. 1423–1432. doi: 10.1242/dev.02302.

Glaser, S. *et al.* (2009) 'The histone 3 lysine 4 methyltransferase, Mll2, is only required briefly in development and spermatogenesis', *Epigenetics & Chromatin* 2: p. 5. doi: 10.1186/1756-8935-2-5.

Glover-Cutter, K. *et al.* (2009) 'TFIIH-Associated Cdk7 Kinase Functions in Phosphorylation of C-Terminal Domain Ser7 Residues, Promoter-Proximal Pausing, and Termination by RNA Polymerase II', *Molecular and Cellular Biology* 29: pp. 5455–5464. doi: 10.1128/mcb.00637-09.

Görisch, S. M. *et al.* (2005) 'Histone acetylation increases chromatin accessibility', *Journal of Cell Science* 118: pp. 5825–5834. doi: 10.1242/jcs.02689.

Greenfield, R. *et al.* (2018) 'Role of transcription complexes in the formation of the basal methylation pattern in early development', *Proceedings of the National Academy of Sciences of the United States of America* 115: pp. 10387–10391. doi: 10.1073/pnas.1804755115.

Greer, E. L. and Shi, Y. (2012) 'Histone methylation: A dynamic mark in health, disease and inheritance', *Nature Reviews Genetics* 13: pp. 343–357. doi: 10.1038/nrg3173.

Gregersen, L. H., Mitter, R. and Svejstrup, J. Q. (2020) 'Using TTchem-seq for profiling nascent transcription and measuring transcript elongation', *Nature Protocols* 15: pp. 604–627. doi: 10.1038/s41596-019-0262-3.

Gressel, S. *et al.* (2017) 'CDK9-dependent RNA polymerase II pausing controls transcription initiation', *eLife* 6: e29736. doi: 10.7554/eLife.29736.

Groth, A. *et al.* (2007) 'Chromatin Challenges during DNA Replication and Repair', *Cell* 128: pp. 721–733. doi: 10.1016/j.cell.2007.01.030.

Gu, B. and Lee, M. G. (2013) 'Histone H3 lysine 4 methyltransferases and demethylases in self-renewal and differentiation of stem cells', *Cell and Bioscience* 3: p. 39. doi: 10.1186/2045-3701-3-39.

Guenther, M. G. *et al.* (2005) 'Global and Hox-specific roles for the MLL1 methyltransferase', *Proceedings of the National Academy of Sciences of the United States of America* 102: pp. 8603–8608. doi: 10.1073/pnas.0503072102.

Guenther, M. G. *et al.* (2007) 'A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells', *Cell* 130: pp. 77–88. doi: 10.1016/j.cell.2007.05.042.

Guillemette, B. *et al.* (2011) 'H3 lysine 4 is acetylated at active gene promoters and is regulated by H3 lysine 4 methylation', *PLoS Genetics* 7: e1001354. doi: 10.1371/journal.pgen.1001354.

Guo, X. *et al.* (2015) 'Structural insight into autoinhibition and histone H3-induced activation of DNMT3A', *Nature* 517: pp. 640–644. doi: 10.1038/nature13899.

Haberle, V. and Stark, A. (2018) 'Eukaryotic core promoters and the functional basis of transcription initiation', *Nature Reviews Molecular Cell Biology* 19: pp. 621–637. doi: 10.1038/s41580-018-0028-8.

Hallson, G. *et al.* (2012) 'dSet1 is the main H3K4 di- and tri-methyltransferase throughout Drosophila development', *Genetics* 190: pp. 91–100. doi: 10.1534/genetics.111.135863.

Hamamoto, R. *et al.* (2004) 'SMYD3 encodes a histone methyltransferase involved in the proliferation of cancer cells', *Nature Cell Biology* 6: pp. 731–740. doi: 10.1038/ncb1151.

Harlen, K. M. *et al.* (2016) 'Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue', *Cell Reports* 15: pp. 2147–2158. doi: 10.1016/j.celrep.2016.05.010.

Harlen, K. M. and Churchman, L. S. (2017) 'The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain', *Nature Reviews Molecular Cell Biology* 18: pp. 263–273. doi: 10.1038/nrm.2017.10.

Hazzalin, C. A. and Mahadevan, L. C. (2005) 'Dynamic Acetylation of All Lysine 4–Methylated Histone H3 in the Mouse Nucleus: Analysis at c-fos and c-jun', *PLoS Biology* 3: p. e393. doi: 10.1371/journal.pbio.0030393.

He, R. and Kidder, B. L. (2017) 'H3K4 demethylase KDM5B regulates global dynamics of transcription elongation and alternative splicing in embryonic stem cells', *Nucleic Acids Research* 45: pp. 6427–6441. doi: 10.1093/nar/gkx251.

He, Y. *et al.* (2013) 'Structural visualization of key steps in human transcription initiation', *Nature* 495: pp. 481–486. doi: 10.1038/nature11991.

He, Y. *et al.* (2016) 'Near-atomic resolution visualization of human transcription promoter opening', *Nature* 533: pp. 359–365. doi: 10.1038/nature17970.

Hebbes, T. R., Thorne, A. W. and Crane-Robinson, C. (1988) 'A direct link between core histone acetylation and transcriptionally active chromatin.', *The EMBO Journal* 7: pp. 1395–1402. doi: 10.1002/j.1460-2075.1988.tb02956.x.

Heintzman, N. D. *et al.* (2007) 'Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome', *Nature Genetics* 39: pp. 311–318. doi: 10.1038/ng1966.

Heintzman, N. D. *et al.* (2009) 'Histone modifications at human enhancers reflect global cell-type-specific gene expression', *Nature* 459: pp. 108–112. doi: 10.1038/nature07829.

Herz, H. M. *et al.* (2012) 'Enhancer-associated H3K4 monomethylation by trithorax-related, the drosophila homolog of mammalian MLL3/MLL4', *Genes & Development* 26: pp. 2604–2620. doi: 10.1101/gad.201327.112.

Herzog, V. A. *et al.* (2017) 'Thiol-linked alkylation of RNA to assess expression dynamics', *Nature Methods* 14: pp. 1198–1204. doi: 10.1038/nmeth.4435.

Hieb, A. R. *et al.* (2007) 'TFIIA Changes the Conformation of the DNA in TBP/TATA Complexes and Increases their Kinetic Stability', *Journal of Molecular Biology* 372: pp. 619–632. doi: 10.1016/j.jmb.2007.06.061.

Higgs, M. R. *et al.* (2018) 'Histone Methylation by SETD1A Protects Nascent DNA through the Nucleosome Chaperone Activity of FANCD2', *Molecular Cell* 71: pp. 25-41.e6. doi:

10.1016/j.molcel.2018.05.018.

Ho, C. K. and Shuman, S. (1999) 'Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme', *Molecular Cell* 3: pp. 405–411. doi: 10.1016/S1097-2765(00)80468-2.

Hödl, M. and Basler, K. (2009) 'Transcription in the Absence of Histone H3.3', *Current Biology* 19: pp. 1221–1226. doi: 10.1016/j.cub.2009.05.048.

Hödl, M. and Basler, K. (2012) 'Transcription in the absence of histone H3.2 and H3K4 methylation', *Current Biology* 22: pp. 2253–2257. doi: 10.1016/j.cub.2012.10.008.

Holliday, R. and Pugh, J. E. (1975) 'DNA modification mechanisms and gene activity during development', *Science* 187: pp. 226–232. doi: 10.1126/science.187.4173.226.

Horikoshi, M. *et al.* (1988) 'Transcription factor ATF interacts with the TATA factor to facilitate establishment of a preinitiation complex', *Cell* 54: pp. 1033–1042. doi: 10.1016/0092-8674(88)90118-3.

Horikoshi, M. *et al.* (1992) 'Transcription factor TFIID induces DNA bending upon binding to the TATA element', *Proceedings of the National Academy of Sciences of the United States of America* 89: pp. 1060–1064. doi: 10.1073/pnas.89.3.1060.

Horton, J. R. *et al.* (2010) 'Enzymatic and structural insights for substrate specificity of a family of jumonji histone lysine demethylases', *Nature Structural and Molecular Biology* 17: pp. 38–44. doi: 10.1038/nsmb.1753.

Hoshii, T. *et al.* (2018) 'A Non-catalytic Function of SETD1A Regulates Cyclin K and the DNA Damage Response', *Cell* 172: pp. 1007-1021.e17. doi: 10.1016/j.cell.2018.01.032.

Hota, S. K. and Bruneau, B. G. (2016) 'ATP-dependent chromatin remodeling during mammalian development', *Development* 143: pp. 2882–2897. doi: 10.1242/dev.128892.

Hou, L. *et al.* (2019) 'Paf1C regulates RNA polymerase II progression by modulating elongation rate', *Proceedings of the National Academy of Sciences of the United States of America* 116: pp. 14583–14592. doi: 10.1073/pnas.1904324116.

Howe, F. S. *et al.* (2017) 'Is H3K4me3 instructive for transcription activation?', *BioEssays* 39: p. e201600095. doi: 10.1002/bies.201600095.

Hsin, J. P. and Manley, J. L. (2012) 'The RNA polymerase II CTD coordinates transcription and RNA processing', *Genes & Development* 26: pp. 2119–2137. doi: 10.1101/gad.200303.112.

Hu, B. *et al.* (2015) 'Biological chromodynamics: A general method for measuring protein occupancy across the genome by calibrating ChIP-seq', *Nucleic Acids Research* 43: e132. doi: 10.1093/nar/gkv670.

Hu, D., Garruss, A. S., *et al.* (2013) 'The Mll2 branch of the COMPASS family regulates bivalent promoters in mouse embryonic stem cells', *Nature Structural and Molecular Biology* 20: pp. 1093–1097. doi: 10.1038/nsmb.2653.

Hu, D., Gao, X., *et al.* (2013) 'The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers', *Molecular and Cellular Biology* 33: pp. 4745–4754. doi: 10.1128/MCB.01181-13.

Hu, D. *et al.* (2017) 'Not All H3K4 Methylations Are Created Equal: Mll2/COMPASS Dependency in Primordial Germ Cell Specification', *Molecular Cell* 65: pp. 460-475.e6. doi: 10.1016/j.molcel.2017.01.013.

Hu, J. L. *et al.* (2009) 'The N-terminus of histone H3 is required for de novo DNA methylation in chromatin', *Proceedings of the National Academy of Sciences of the United States of America* 106: pp. 22187–22192. doi: 10.1073/pnas.0905767106.

Huang, J. *et al.* (2012) 'The same pocket in menin binds both MLL and JUND but has opposite effects on transcription', *Nature* 482: pp. 542–546. doi: 10.1038/nature10806.

Huang, Y. *et al.* (2006) 'Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A', *Science* 312: pp. 748–751. doi: 10.1126/science.1125162.

Hughes, A. L., Kelley, J. R. and Klose, R. J. (2020) 'Understanding the interplay between CpG island-associated gene promoters and H3K4 methylation', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1863: p. 194567. doi: 10.1016/j.bbagrm.2020.194567.

Hughes, C. M. *et al.* (2004) 'Menin associates with a trithorax family histone methyltransferase complex and with the Hoxc8 locus', *Molecular Cell* 13: pp. 587–597. doi: 10.1016/S1097-2765(04)00081-4.

Hung, T. *et al.* (2009) 'ING4 Mediates Crosstalk between Histone H3 K4 Trimethylation and H3 Acetylation to Attenuate Cellular Transformation', *Molecular Cell* 33: pp. 248–256. doi: 10.1016/j.molcel.2008.12.016.

Hyun, K. *et al.* (2017) 'Writing, erasing and reading histone lysine methylations', *Experimental & Molecular Medicine* 49: p. e324. doi: 10.1038/emm.2017.11.

Illingworth, R. *et al.* (2008) 'A novel CpG island set identifies tissue-specific methylation at developmental gene loci', *PLoS Biology* 6: e22. doi: 10.1371/journal.pbio.0060022.

Illingworth, R. S. *et al.* (2010) 'Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome', *PLoS Genetics* 6: e1001134. doi: 10.1371/journal.pgen.1001134.

Imbalzano, A. N., Zaret, K. S. and Kingston, R. E. (1994) 'Transcription factor (TF) IIB and TFIIA can independently increase the affinity of the TATA-binding protein for DNA', *Journal of Biological Chemistry* 269: pp. 8280–8286. doi: 10.1016/s0021-9258(17)37190-9.

van Ingen, H. *et al.* (2008) 'Structural Insight into the Recognition of the H3K4me3 Mark by the TFIID Subunit TAF3', *Structure* 16: pp. 1245–1256. doi: 10.1016/J.STR.2008.04.015.

Ingham, P. and Whittle, R. (1980) 'Trithorax: A new homoeotic mutation of Drosophila melanogaster causing transformations of abdominal and thoracic imaginal segments - I. Putative role during embryogenesis', *MGG Molecular & General Genetics* 179: pp. 607–614. doi: 10.1007/BF00271751.

Jambhekar, A., Dhall, A. and Shi, Y. (2019) 'Roles and regulation of histone methylation in animal development', *Nature Reviews Molecular Cell Biology* 20: pp. 625–641. doi: 10.1038/s41580-019-0151-1.

Jiang, H. *et al.* (2011) 'Role for Dpy-30 in ES cell-fate specification by regulation of H3K4 methylation within bivalent domains.', *Cell* 144: pp. 513–525. doi: 10.1016/j.cell.2011.01.020.

Jiang, H. *et al.* (2013) 'Regulation of transcription by the MLL2 complex and MLL complex-associated AKAP95', *Nature Structural and Molecular Biology* 20: pp. 1156–1165. doi: 10.1038/nsmb.2656.

Jinek, M. *et al.* (2012) 'A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity', *Science* 337: pp. 816–821. doi: 10.1126/science.1225829.

Jones, P. A. (2012) 'Functions of DNA methylation: Islands, start sites, gene bodies and beyond', *Nature Reviews Genetics* 13: pp. 484–492. doi: 10.1038/nrg3230.

Jonkers, I., Kwak, H. and Lis, J. T. (2014) 'Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons', *eLife* 29: e02407. doi: 10.7554/eLife.02407.

Joo, Y. J. *et al.* (2019) 'In vitro analysis of rna polymerase II elongation complex dynamics', *Genes & Development* 33: pp. 578–589. doi: 10.1101/gad.324202.119.

Josse, J., Kaiser, A. D. and Kornberg, A. (1961) 'Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid.', *The Journal of Biological Chemistry* 236: pp. 864–875. doi: 10.1016/S0021-9258(18)64321-2.

Juven-Gershon, T. *et al.* (2008) 'The RNA polymerase II core promoter - the gateway to transcription', *Current Opinion in Cell Biology* 20: pp. 253–259. doi: 10.1016/j.ceb.2008.03.003.

Kadonaga, J. T. (2012) 'Perspectives on the RNA polymerase II core promoter', *Wiley Interdisciplinary Reviews: Developmental Biology* 1: pp. 40–51. doi: 10.1002/wdev.21.

Kaikkonen, M. U. *et al.* (2013) 'Remodeling of the enhancer landscape during macrophage activation is coupled to enhancer transcription', *Molecular Cell* 51: pp. 310–325. doi: 10.1016/j.molcel.2013.07.010.

Kamieniarz-Gdula, K. and Proudfoot, N. J. (2019) 'Transcriptional Control by Premature Termination: A Forgotten Mechanism', *Trends in Genetics* 35: pp. 553–564. doi: 10.1016/j.tig.2019.05.005.

Kanda, H. *et al.* (2013) ' The Drosophila Ortholog of MLL3 and MLL4 , trithorax related , Functions as a Negative Regulator of Tissue Growth', *Molecular and Cellular Biology* 33: pp. 1702–1710. doi: 10.1128/mcb.01585-12.

Karimi, M. M. *et al.* (2011) 'DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mescs', *Cell Stem Cell* 8: pp. 676–687. doi: 10.1016/j.stem.2011.04.004.

Karmodiya, K. *et al.* (2012) 'H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells', *BMC Genomics* 13: p. 424. doi: 10.1186/1471-2164-13-424.

Kassis, J. A., Kennison, J. A. and Tamkun, J. W. (2017) 'Polycomb and trithorax group genes in drosophila', *Genetics* 206: pp. 1699–1725. doi: 10.1534/genetics.115.185116.

Kent, W. J. *et al.* (2002) 'The Human Genome Browser at UCSC', *Genome Research* 12: pp. 996–1006. doi: 10.1101/gr.229102.

Kidder, B. L. *et al.* (2013) 'Extended Self-Renewal and Accelerated Reprogramming in the Absence of Kdm5b', *Molecular and Cellular Biology* 33: pp. 4793–4810. doi: 10.1128/mcb.00692-13.

Kidder, B. L., Hu, G. and Zhao, K. (2014) 'KDM5B focuses H3K4 methylation near promoters and enhancers during embryonic stem cell self-renewal and differentiation', *Genome Biology* 15: R32. doi: 10.1186/gb-2014-15-2-r32.

Kim, D. H. *et al.* (2013) 'Histone H3K27 Trimethylation Inhibits H3 Binding and Function of SET1-Like H3K4 Methyltransferase Complexes', *Molecular and Cellular Biology* 33: pp. 4936–4946. doi: 10.1128/mcb.00601-13.

Kim, J. *et al.* (2006) 'Tudor, MBT and chromo domains gauge the degree of lysine methylation', *EMBO Reports* 7: pp. 397–403. doi: 10.1038/sj.embor.7400625.

Kim, J. *et al.* (2013) 'The n-SET Domain of Set1 Regulates H2B Ubiquitylation-Dependent H3K4 Methylation', *Molecular Cell* 49: pp. 1121–1133. doi: 10.1016/j.molcel.2013.01.034.

Kim, S. *et al.* (2016) 'Mechanism of Histone H3K4me3 Recognition by the Plant Homeodomain of Inhibitor of Growth 3.', *The Journal of Biological Chemistry* 291: pp. 18326–18341. doi: 10.1074/jbc.M115.690651.

Kim, T. K., Ebright, R. H. and Reinberg, D. (2000) 'Mechanism of ATP-dependent promoter melting by transcription factor IIH', *Science* 288: pp. 1418–1421. doi: 10.1126/science.288.5470.1418.

Klose, R. J. *et al.* (2013) 'Chromatin Sampling-An Emerging Perspective on Targeting Polycomb Repressor Proteins', *PLoS Genetics* 9: e1003717. doi: 10.1371/journal.pgen.1003717.

Klug, M. and Rehli, M. (2006) 'Functional analysis of promoter CpG methylation using a CpG-free luciferase reporter vector', *Epigenetics* 1: pp. 127–130. doi: 10.4161/epi.1.3.3327.

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

Kornberg, R. D. and Lorch, Y. (1999) 'Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome', *Cell* 98: pp. 285–294. doi: 10.1016/S0092-8674(00)81958-3.

Kostrewa, D. *et al.* (2009) 'RNA polymerase II-TFIIB structure and mechanism of transcription initiation', *Nature* 462: pp. 323–330. doi: 10.1038/nature08548.

Kouzarides, T. (2007) 'Chromatin Modifications and Their Function', *Cell* 128: pp. 693–705. doi: 10.1016/j.cell.2007.02.005.

Kranz, A. and Anastassiadis, K. (2020) 'The role of SETD1A and SETD1B in development and disease', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1863: p. 194578. doi: 10.1016/j.bbagrm.2020.194578.

Krogan, N. J. *et al.* (2002) 'COMPASS, a histone H3 (lysine 4) methyltransferase required for telomeric silencing of gene expression', *Journal of Biological Chemistry* 277: pp. 10753–10755. doi: 10.1074/jbc.C200023200.

Krogan, N. J. *et al.* (2003) 'Methylation of Histone H3 by Set2 in Saccharomyces cerevisiae Is Linked to Transcriptional Elongation by RNA Polymerase II', *Molecular and Cellular Biology* 23: pp. 4207–4218. doi: 10.1128/mcb.23.12.4207-4218.2003.

Kubota, T. *et al.* (2013) 'The Elg1 Replication Factor C-like Complex Functions in PCNA Unloading during DNA Replication', *Molecular Cell* 50: pp. 273–280. doi: 10.1016/j.molcel.2013.02.012.

Kurup, J. T., Campeanu, I. J. and Kidder, B. L. (2019) 'Contribution of H3K4 demethylase KDM5B to nucleosome organization in embryonic stem cells revealed by micrococcal nuclease sequencing', *Epigenetics & Chromatin* 12: p. 20. doi: 10.1186/s13072-019-0266-9.

Kwon, M. *et al.* (2021) 'H2B ubiquitylation enhances H3K4 methylation activities of human KMT2 family complexes', *Nucleic Acids Research* 48: pp. 5442–5456. doi: 10.1093/nar/gkaa317.

Lambert, S. A. *et al.* (2018) 'The Human Transcription Factors', *Cell* 172: pp. 650–665. doi: 10.1016/j.cell.2018.01.029.

Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods* 9: pp. 357–359. doi: 10.1038/nmeth.1923.

Latham, J. A. *et al.* (2011) 'Chromatin signaling to kinetochores: Transregulation of Dam1 methylation by histone H2B ubiquitination', *Cell* 146: pp. 709–719. doi: 10.1016/j.cell.2011.07.025.

Lauberth, S. M. *et al.* (2013) 'H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation.', *Cell* 152: pp. 1021–1036. doi: 10.1016/j.cell.2013.01.052.

Lechner, C. C., Agashe, N. D. and Fierz, B. (2016) 'Traceless Synthesis of Asymmetrically Modified Bivalent Nucleosomes', *Angewandte Chemie - International Edition* 55: pp. 2903–2906. doi: 10.1002/anie.201510996.

Lee, J. E. *et al.* (2013) 'H3K4 mono- And di-methyltransferase MLL4 is required for enhancer activation during cell differentiation', *eLife* 2: e01503. doi: 10.7554/eLife.01503.

Lee, J. H. *et al.* (2007) 'Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex.', *The Journal of Biological Chemistry* 282: pp. 13419–13428. doi: 10.1074/jbc.M609809200.

Lee, J. H. *et al.* (2010) 'Identification and characterization of a novel human PP1 phosphatase complex', *Journal of Biological Chemistry* 285: pp. 24466–24476. doi: 10.1074/jbc.M110.109801.

Lee, J. H. and Skalnik, D. G. (2005) 'CpG-binding Protein (CXXC Finger Protein 1) Is a Component of the Mammalian Set1 Histone H3-Lys4 Methyltransferase Complex, the Analogue of the Yeast Set1/COMPASS Complex', *Journal of Biological Chemistry* 280: pp. 41725–41731. doi: 10.1074/jbc.M508312200.

Lee, J. H. and 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', *Molecular and Cellular Biology* 28: pp. 609–618. doi: 10.1128/mcb.01356-07.

Lee, J. H. and Skalnik, D. G. (2012) 'Rbm15-Mkl1 interacts with the Setd1b histone H3-Lys4 methyltransferase via a SPOC domain that is required for cytokine-independent proliferation', *PLoS One* 7: e42965. doi: 10.1371/journal.pone.0042965.

Lee, J. H., Voo, K. S. and Skalnik, D. G. (2001) 'Identification and Characterization of the DNA Binding Domain of CpG-binding Protein', *Journal of Biological Chemistry* 276: pp. 44669–44676. doi: 10.1074/jbc.M107179200.

Lee, K. Y. *et al.* (2018) 'H3K4 methylation dependent and independent chromatin regulation by JHD2 and SET1 in budding yeast', *G3: Genes, Genomes, Genetics* 8: pp. 1829–1839. doi: 10.1534/g3.118.200151.

Lenstra, T. L. *et al.* (2011) 'The Specificity and Topology of Chromatin Interaction Pathways in Yeast', *Molecular Cell* 42: pp. 536–549. doi: 10.1016/j.molcel.2011.03.026.

Li, B., Carey, M. and Workman, J. L. (2007) 'The Role of Chromatin during Transcription', *Cell* 128: pp. 707–719. doi: 10.1016/j.cell.2007.01.015.

Li, B. Z. *et al.* (2011) 'Histone tails regulate DNA methylation by allosterically activating de novo methyltransferase', *Cell Research* 21: pp. 1172–1181. doi: 10.1038/cr.2011.92.

Li, H. *et al.* (2006) 'Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF', *Nature* 442: pp. 91–95. doi: 10.1038/nature04802.

Li, Ying *et al.* (2016) 'Setd1a and NURF mediate chromatin dynamics and gene regulation during erythroid lineage commitment and differentiation.', *Nucleic Acids Research* 44: pp. 7173–7188. doi: 10.1093/nar/gkw327.

Li, Yanjing *et al.* (2016) 'Structural basis for activity regulation of MLL family methyltransferases', *Nature* 530: pp. 447–452. doi: 10.1038/nature16952.

Li, Y. *et al.* (2020) 'Crystal Structure of MLL2 Complex Guides the Identification of a Methylation Site on P53 Catalyzed by KMT2 Family Methyltransferases', *Structure* 28: pp. 1141-1148.e4. doi: 10.1016/j.str.2020.07.002.

Li, Z. *et al.* (2011) 'The nucleosome map of the mammalian liver', *Nature Structural and Molecular Biology* 18: pp. 742–746. doi: 10.1038/nsmb.2060.

Liang, G. *et al.* (2004) 'Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome', *Proceedings of the National Academy of Sciences of the United States of America* 101: pp. 7357–7362. doi: 10.1073/pnas.0401866101.

Lim, B. (2018) 'Imaging transcriptional dynamics', *Current Opinion in Biotechnology* 52: pp. 49–55. doi: 10.1016/j.copbio.2018.02.008.

Lin, J. J. *et al.* (2011) 'Mediator coordinates PIC assembly with recruitment of CHD1', *Genes & Development* 25: pp. 2198–2209. doi: 10.1101/gad.17554711.

Lister, R. *et al.* (2009) 'Human DNA methylomes at base resolution show widespread epigenomic differences.', *Nature* 462: pp. 315–322. doi: 10.1038/nature08514.

Liu, P. *et al.* (2010) 'Genetic organization, length conservation, and evolution of RNA polymerase II carboxyl-terminal domain', *Molecular Biology and Evolution* 27: pp. 2628–2641. doi: 10.1093/molbev/msq151.

Liu, X. *et al.* (2010) 'Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism', *Science* 327: pp. 206–209. doi: 10.1126/science.1182015.

Logan, J. *et al.* (1987) 'A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene.', *Proceedings of the National Academy of Sciences of the United States of America* 84: pp. 8306–8310. doi: 10.1073/pnas.84.23.8306.

Lohse, B. *et al.* (2013) 'Posttranslational Modifications of the Histone 3 Tail and Their Impact on the Activity of Histone Lysine Demethylases In Vitro', *PLoS One* 8: e67653. doi: 10.1371/journal.pone.0067653.

Long, H. K. *et al.* (2013) 'Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates', *eLife* 2: e00348. doi: 10.7554/eLife.00348.

Long, H. K., Blackledge, N. P. and Klose, R. J. (2013) 'ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection', *Biochemical Society Transactions* 41: pp. 727–740. doi: 10.1042/BST20130028.

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

Louder, R. K. *et al.* (2016) 'Structure of promoter-bound TFIID and model of human pre-initiation complex assembly', *Nature* 531: pp. 604–609. doi: 10.1038/nature17394.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology* 15: p. 550. doi: 10.1186/s13059-014-0550-8.

Lu, H. *et al.* (1991) 'The nonphosphorylated form of RNA polymerase II preferentially associates with the preinitiation complex', *Proceedings of the National Academy of Sciences of the United States of America* 88: pp. 10004–10008. doi: 10.1073/pnas.88.22.10004.

Luciano, P. *et al.* (2017) 'Binding to rna regulates set1 function', *Cell Discovery* 3: pp. 1–19. doi: 10.1038/celldisc.2017.40.

Luger, K. *et al.* (1997) 'Crystal structure of the nucleosome core particle at 2.8 Å resolution', *Nature* 389: pp. 251–260. doi: 10.1038/38444.

Madeira, F. *et al.* (2019) 'The EMBL-EBI search and sequence analysis tools APIs in 2019', *Nucleic Acids Research* 47: pp. W636–W641. doi: 10.1093/nar/gkz268.

Mahadevan, J. and Skalnik, D. G. (2016) 'Efficient differentiation of murine embryonic stem cells requires the binding of CXXC finger protein 1 to DNA or methylated histone H3-Lys4', *Gene* 594: pp. 1–9. doi: 10.1016/j.gene.2016.08.048.

Mahat, D. B. *et al.* (2016) 'Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq)', *Nature Protocols* 11: pp. 1455–1476. doi: 10.1038/nprot.2016.086.

Margaritis, T. *et al.* (2012) 'Two Distinct Repressive Mechanisms for Histone 3 Lysine 4 Methylation through Promoting 3'-End Antisense Transcription', *PLoS Genetics* 8: e1002952. doi: 10.1371/journal.pgen.1002952.

Marshall, N. F. *et al.* (1996) 'Control of RNA polymerase II elongation potential by a novel carboxyl-

terminal domain kinase', *Journal of Biological Chemistry* 271: pp. 27176–27183. doi: 10.1074/jbc.271.43.27176.

Mas, G. *et al.* (2018) 'Promoter bivalency favors an open chromatin architecture in embryonic stem cells', *Nature Genetics* 50: pp. 1452–1462. doi: 10.1038/s41588-018-0218-5.

Maunakea, A. K. *et al.* (2010) 'Conserved role of intragenic DNA methylation in regulating alternative promoters', *Nature* 466: pp. 253–257. doi: 10.1038/nature09165.

McCracken, S. *et al.* (1997) '5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II', *Genes & Development* 11: pp. 3306–3318. doi: 10.1101/gad.11.24.3306.

Meeks, J. J. and Shilatifard, A. (2017) 'Multiple roles for the mll/compass family in the epigenetic regulation of gene expression and in cancer', *Annual Review of Cancer Biology* 1: pp. 425–446. doi: 10.1146/annurev-cancerbio-050216-034333.

Meissner, A. *et al.* (2008) 'Genome-scale DNA methylation maps of pluripotent and differentiated cells', *Nature* 454: pp. 766–770. doi: 10.1038/nature07107.

Mendenhall, E. M. *et al.* (2010) 'GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells', *PLoS Genetics* 6: e1001244. doi: 10.1371/journal.pgen.1001244.

Mikkelsen, T. S. *et al.* (2007) 'Genome-wide maps of chromatin state in pluripotent and lineage-committed cells', *Nature* 448: pp. 553–560. doi: 10.1038/nature06008.

Miller, T. *et al.* (2001) 'COMPASS: A complex of proteins associated with a trithorax-related SET domain protein', *Proceedings of the National Academy of Sciences of the United States of America* 98: pp. 12902–12907. doi: 10.1073/pnas.231473398.

Milne, T. A. *et al.* (2002) 'MLL targets SET domain methyltransferase activity to Hox gene promoters', *Molecular Cell* 10: pp. 1107–1117. doi: 10.1016/s1097-2765(02)00741-4.

Milne, T. A., Hughes, C. M., *et al.* (2005) 'Menin and MLL cooperatively regulate expression of cyclin-dependent kinase inhibitors', *Proceedings of the National Academy of Sciences of the United States of America* 102: pp. 749–754. doi: 10.1073/pnas.0408836102.

Milne, T. A., Dou, Y., *et al.* (2005) 'MLL associates specifically with a subset of transcriptionally active target genes', *Proceedings of the National Academy of Sciences of the United States of America* 102: pp. 14765–14770. doi: 10.1073/pnas.0503630102.

Milne, T. A. *et al.* (2010) 'Multiple Interactions Recruit MLL1 and MLL1 Fusion Proteins to the HOXA9 Locus in Leukemogenesis', *Molecular Cell* 38: pp. 853–863. doi: 10.1016/j.molcel.2010.05.011.

Mishra, B. P. *et al.* (2014) 'The histone methyltransferase activity of MLL1 is dispensable for hematopoiesis and leukemogenesis', *Cell Reports* 7: pp. 1239–1247. doi: 10.1016/j.celrep.2014.04.015.

Mittal, A. *et al.* (2018) 'The structure of the RbBP5 β-propeller domain reveals a surface with potential nucleic acid binding sites', *Nucleic Acids Research* 46: pp. 3802–3812. doi: 10.1093/nar/gky199.

Mohan, M. *et al.* (2011) 'The COMPASS Family of H3K4 Methylases in Drosophila', *Molecular and Cellular Biology* 31: pp. 4310–4318. doi: 10.1128/mcb.06092-11.

Mohandas, T., Sparkes, R. S. and Shapiro, L. J. (1981) 'Reactivation of an inactive human X chromosome: Evidence for X inactivation by DNA methylation', *Science* 211: pp. 393–396. doi: 10.1126/science.6164095.

Morris, S. A. *et al.* (2014) 'Overlapping chromatin-remodeling systems collaborate genome wide at dynamic chromatin transitions', *Nature Structural and Molecular Biology* 21: pp. 73–81. doi:

10.1038/nsmb.2718.
Morselli, M. *et al.* (2015) 'In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse', *eLife* 4: e06205. doi: 10.7554/eLife.06205.
Muntean, A. G. *et al.* (2010) 'The PAF complex synergizes with MLL fusion proteins at HOX loci to promote leukemogenesis.', *Cancer Cell* 17: pp. 609–621. doi: 10.1016/j.ccr.2010.04.012.
Murai, M. J. *et al.* (2011) 'Crystal structure of menin reveals binding site for Mixed Lineage Leukemia (MLL) protein', *Journal of Biological Chemistry* 286: pp. 31742–31748. doi: 10.1074/jbc.M111.258186.
Murray, S. C. *et al.* (2019) 'H3K4me3 is neither instructive for, nor informed by, transcription', *bioRxiv*: 709014. doi: 10.1101/709014.
Muse, G. W. *et al.* (2007) 'RNA polymerase is poised for activation across the genome', *Nature Genetics* 39: pp. 1507–1511. doi: 10.1038/ng.2007.21.
Nabet, B. *et al.* (2018) 'The dTAG system for immediate and target-specific protein degradation', *Nature Chemical Biology* 14: pp. 431–441. doi: 10.1038/s41589-018-0021-8.
Nagy, P. L. *et al.* (2002) 'A trithorax-group complex purified from Saccharomyces cerevisiae is required for methylation of histone H3', *Proceedings of the National Academy of Sciences of the United States of America* 99: pp. 90–94. doi: 10.1073/pnas.221596698.
Nakagawa, T. *et al.* (2008) 'Deubiquitylation of histone H2A activates transcriptional initiation via trans-histone cross-talk with H3K4 di- and trimethylation', *Genes & Development* 22: pp. 37–49. doi: 10.1101/gad.1609708.
Nakamura, T. *et al.* (2002) 'ALL-1 is a histone methyltransferase that assembles a supercomplex of proteins involved in transcriptional regulation', *Molecular Cell* 10: pp. 1119–1128. doi: 10.1016/S1097-2765(02)00740-2.
Nakayama, J. *et al.* (2001) 'Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly', *Science* 292: pp. 110–113. doi: 10.1126/science.1060118.
Natsume, T. *et al.* (2016) 'Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors', *Cell Reports* 15: pp. 210–218. doi: 10.1016/j.celrep.2016.03.001.
Nikolov, M. *et al.* (2011) 'Chromatin Affinity Purification and Quantitative Mass Spectrometry Defining the Interactome of Histone Modification Patterns', *Molecular & Cellular Proteomics* 10: M110.005371. doi: 10.1074/mcp.M110.005371.
Ninova, M., Tóth, K. F. and Aravin, A. A. (2019) 'The control of gene expression and cell identity by H3K9 trimethylation', *Development* 146: dev181180. doi: 10.1242/dev.181180.
Nishimura, K. *et al.* (2009) 'An auxin-based degron system for the rapid depletion of proteins in nonplant cells', *Nature Methods* 6: pp. 917–922. doi: 10.1038/nmeth.1401.
Nishioka, K. *et al.* (2002) 'Set9, a novel histone H3 methyltransferase that facilitates transcription by precluding histone tail modifications required for heterochromatin formation', *Genes & Development* 16: pp. 479–489. doi: 10.1101/gad.967202.
Nislow, C., Ray, E. and Pillus, L. (1997) 'SET1, a yeast member of the Trithorax family, functions in transcriptional silencing and diverse cellular processes', *Molecular Biology of the Cell* 8: pp. 2421–2436. doi: 10.1091/mbc.8.12.2421.
Noh, K. M. *et al.* (2015) 'Engineering of a Histone-Recognition Domain in Dnmt3a Alters the Epigenetic Landscape and Phenotypic Features of Mouse ESCs', *Molecular Cell* 59: pp. 89–103. doi: 10.1016/j.molcel.2015.05.017.
Noma, K. ichi and Grewal, S. I. S. (2002) 'Histone H3 lysine 4 methylation is mediated by Set1 and

promotes maintenance of active chromatin states in fission yeast', *Proceedings of the National Academy of Sciences of the United States of America* 99: pp. 16438–16445. doi: 10.1073/pnas.182436399.

van Nuland, R., Schram, A. W., *et al.* (2013) 'Multivalent engagement of TFIID to nucleosomes', *PLoS One* 8: e73495. doi: 10.1371/journal.pone.0073495.

van Nuland, R., Smits, A. H., *et al.* (2013) 'Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes', *Molecular and Cellular Biology* 33: pp. 2067–2077. doi: 10.1128/MCB.01742-12.

Okano, M. *et al.* (1999) 'DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development', *Cell* 99: pp. 247–257. doi: 10.1016/S0092-8674(00)81656-6.

Okitsu, C. Y. and Hsieh, C.-L. (2007) 'DNA Methylation Dictates Histone H3K4 Methylation', *Molecular and Cellular Biology* 27: pp. 2746–2757. doi: 10.1128/mcb.02291-06.

Okuda, H. *et al.* (2014) 'MLL fusion proteins link transcriptional coactivators to previously active CpG-rich promoters', *Nucleic Acids Research* 42: pp. 4241–4256. doi: 10.1093/nar/gkt1394.

Ooi, S. K. T. *et al.* (2007) 'DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA', *Nature* 448: pp. 714–717. doi: 10.1038/nature05987.

Orlando, D. A. *et al.* (2014) 'Quantitative ChIP-Seq normalization reveals global modulation of the epigenome', *Cell Reports* 9: pp. 1163–1170. doi: 10.1016/j.celrep.2014.10.018.

Ormaza, G. *et al.* (2019) 'The Tumor Suppressor ING5 Is a Dimeric, Bivalent Recognition Molecule of the Histone H3K4me3 Mark', *Journal of Molecular Biology* 431: pp. 2298–2319. doi: 10.1016/j.jmb.2019.04.018.

Otani, J. *et al.* (2009) 'Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX–DNMT3–DNMT3L domain', *EMBO Reports* 10: pp. 1235–1241. doi: 10.1038/embor.2009.218.

Pack, L. R., Yamamoto, K. R. and Fujimori, D. G. (2016) 'Opposing chromatin signals direct and regulate the activity of lysine demethylase 4C (KDM4C)', *Journal of Biological Chemistry* 291: pp. 6060–6070. doi: 10.1074/jbc.M115.696864.

Paggetti, J. *et al.* (2010) 'Crosstalk between leukemia-associated proteins MOZ and MLL regulates HOX gene expression in human cord blood CD34 cells', *Oncogene* 29: pp. 5019–5031. doi: 10.1038/onc.2010.254.

Palacios, A. *et al.* (2008) 'Molecular Basis of Histone H3K4me3 Recognition by ING4', *Journal of Biological Chemistry* 283: pp. 15956–15964. doi: 10.1074/jbc.M710020200.

Park, K., Kim, J. A. and Kim, J. (2020) 'Transcriptional regulation by the KMT2 histone H3K4 methyltransferases', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1863: p. 194545. doi: 10.1016/j.bbagrm.2020.194545.

Park, S. *et al.* (2010) 'The PHD3 domain of MLL Acts as a CYP33-regulated switch between MLL-mediated activation and repression', *Biochemistry* 49: pp. 6576–6586. doi: 10.1021/bi1009387.

Patel, A. *et al.* (2008) 'A conserved arginine-containing motif crucial for the assembly and enzymatic activity of the mixed lineage leukemia protein-1 core complex', *Journal of Biological Chemistry* 283: pp. 32162–32175. doi: 10.1074/jbc.M806317200.

Patel, A. *et al.* (2009) 'On the mechanism of multiple lysine methylation by the human mixed lineage leukemia protein-1 (MLL1) core complex', *Journal of Biological Chemistry* 284: pp. 24242–24256. doi: 10.1074/jbc.M109.014498.

Patel, A. B. *et al.* (2018) 'Structure of human TFIID and mechanism of TBP loading onto promoter

DNA', *Science* 362: eaau8872. doi: 10.1126/science.aau8872.

Patel, A. B., Greber, B. J. and Nogales, E. (2020) 'Recent insights into the structure of TFIID, its assembly, and its binding to core promoter', *Current Opinion in Structural Biology* 61: pp. 17–24. doi: 10.1016/j.sbi.2019.10.001.

Pavri, R. *et al.* (2006) 'Histone H2B Monoubiquitination Functions Cooperatively with FACT to Regulate Elongation by RNA Polymerase II', *Cell* 125: pp. 703–717. doi: 10.1016/J.CELL.2006.04.029.

Payne, J. M., Laybourn, P. J. and Dahmus, M. E. (1989) 'The transition of RNA polymerase II from initiation to elongation is associated with phosphorylation of the carboxyl-terminal domain of subunit IIa', *Journal of Biological Chemistry* 264: pp. 19621–19629. doi: 10.1016/s0021-9258(19)47159-7.

Pedersen, M. T. *et al.* (2014) 'The Demethylase JMJD2C Localizes to H3K4me3-Positive Transcription Start Sites and Is Dispensable for Embryonic Development', *Molecular and Cellular Biology* 34: pp. 1031–1045. doi: 10.1128/mcb.00864-13.

Peña, P. V. *et al.* (2006) 'Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2', *Nature* 442: pp. 100–103. doi: 10.1038/nature04814.

Peña, P. V. *et al.* (2008) 'Histone H3K4me3 Binding Is Required for the DNA Repair and Apoptotic Activities of ING1 Tumor Suppressor', *Journal of Molecular Biology* 380: pp. 303–312. doi: 10.1016/j.jmb.2008.04.061.

Pokholok, D. K. *et al.* (2005) 'Genome-wide map of nucleosome acetylation and methylation in yeast', *Cell* 122: pp. 517–527. doi: 10.1016/j.cell.2005.06.026.

Pradeepa, M. M. *et al.* (2012) 'Psip1/Ledgf p52 Binds Methylated Histone H3K36 and Splicing Factors and Contributes to the Regulation of Alternative Splicing', *PLoS Genetics* 8: e1002717. doi: 10.1371/journal.pgen.1002717.

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

Quevedo, M. *et al.* (2019) 'Mediator complex interaction partners organize the transcriptional network that defines neural stem cells', *Nature Communications* 10: pp. 1–15. doi: 10.1038/s41467-019-10502-8.

Quinlan, A. R. (2014) 'BEDTools: The Swiss-Army tool for genome feature analysis', *Current Protocols in Bioinformatics* 47: pp. 11.12.1-34. doi: 10.1002/0471250953.bi1112s47.

Rabani, M. *et al.* (2011) 'Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells', *Nature Biotechnology* 29: pp. 436–442. doi: 10.1038/nbt.1861.

Rahl, P. B. *et al.* (2010) 'C-Myc regulates transcriptional pause release', *Cell* 141: pp. 432–445. doi: 10.1016/j.cell.2010.03.030.

Ramakrishnan, S. *et al.* (2016) 'Counteracting H3K4 methylation modulators Set1 and Jhd2 co-regulate chromatin dynamics and gene transcription', *Nature Communications* 7: p. 11949. doi: 10.1038/ncomms11949.

Ramírez, F. *et al.* (2014) 'DeepTools: A flexible platform for exploring deep-sequencing data', *Nucleic Acids Research* 42: pp. W187-191. doi: 10.1093/nar/gku365.

Ran, F. A. *et al.* (2013) 'Genome engineering using the CRISPR-Cas9 system', *Nature Protocols* 8: pp. 2281–2308. doi: 10.1038/nprot.2013.143.

Rasmussen, E. B. and Lis, J. T. (1993) 'In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes', *Proceedings of the National Academy of Sciences of the United*

States of America 90: pp. 7923–7927. doi: 10.1073/pnas.90.17.7923.

Rea, S. *et al.* (2000) 'Regulation of chromatin structure by site-specific histone H3 methyltransferases', *Nature* 406: pp. 593–599. doi: 10.1038/35020506.

Reinig, J. *et al.* (2020) 'A theoretical model of Polycomb/Trithorax action unites stable epigenetic memory and dynamic regulation', *Nature Communications* 11: 4782. doi: 10.1038/s41467-020-18507-4.

Rhodes, J. D. P. *et al.* (2020) 'Cohesin Disrupts Polycomb-Dependent Chromosome Interactions in Embryonic Stem Cells', *Cell Reports* 30: pp. 820-835.e10. doi: 10.1016/j.celrep.2019.12.057.

Rickels, R. *et al.* (2017) 'Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability', *Nature Genetics* 49: pp. 1647–1653. doi: 10.1038/ng.3965.

Riggs, A. D. (1975) 'X inactivation, differentiation, and DNA methylation', *Cytogenetic and Genome Research* 14: pp. 9–25. doi: 10.1159/000130315.

Rodríguez, J. *et al.* (2015) 'The AT-Hook motif as a versatile minor groove anchor for promoting DNA binding of transcription factor fragments', *Chemical Science* 6: pp. 4767–4771. doi: 10.1039/c5sc01415h.

Roguev, A. *et al.* (2001) 'The Saccharomyces cerevisiae Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4', *EMBO Journal* 20: pp. 7137–7148. doi: 10.1093/emboj/20.24.7137.

Roguev, A. *et al.* (2003) 'High conservation of the Set1/Rad6 axis of histone 3 lysine 4 methylation in budding and fission yeasts', *Journal of Biological Chemistry* 278: pp. 8487–8493. doi: 10.1074/jbc.M209562200.

Rose, N. R. *et al.* (2016) 'RYBP stimulates PRC1 to shape chromatin-based communication between polycomb repressive complexes', *eLife* 5: e18591. doi: 10.7554/eLife.18591.

Rothbart, S. B. and Strahl, B. D. (2014) 'Interpreting the language of histone and DNA modifications', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1839: pp. 627–643. doi: 10.1016/J.BBAGRM.2014.03.001.

Rougvie, A. E. and Lis, J. T. (1988) 'The RNA polymerase II molecule at the 5' end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged', *Cell* 54: pp. 795–804. doi: 10.1016/S0092-8674(88)91087-2.

Ruthenburg, A. J. *et al.* (2011) 'Recognition of a Mononucleosomal Histone Modification Pattern by BPTF via Multivalent Interactions', *Cell* 145: pp. 692–706. doi: 10.1016/j.cell.2011.03.053.

Sabo, P. J. *et al.* (2004) 'Genome-wide identification of DNaseI hypersensitive sites using active chromatin sequence libraries', *Proceedings of the National Academy of Sciences of the United States of America* 101: pp. 4537–4542. doi: 10.1073/pnas.0400678101.

Saksouk, N. *et al.* (2009) 'HBO1 HAT Complexes Target Chromatin throughout Gene Coding Regions via Multiple PHD Finger Interactions with Histone H3 Tail', *Molecular Cell* 33: pp. 257–265. doi: 10.1016/j.molcel.2009.01.007.

Santos-Rosa, H. *et al.* (2002) 'Active genes are tri-methylated at K4 of histone H3', *Nature* 419: pp. 407–411. doi: 10.1038/nature01080.

Sarvan, S. *et al.* (2011) 'Crystal structure of the trithorax group protein ASH2L reveals a forkhead-like DNA binding domain', *Nature Structural and Molecular Biology* 18: pp. 857–859. doi: 10.1038/nsmb.2093.

Sato, M. H. *et al.* (1999) 'Atomic force microscopy sees nucleosome positioning and histone H1-induced compaction in reconstituted chromatin', *FEBS Letters* 452: pp. 267–271. doi:

10.1016/S0014-5793(99)00644-4.

Saxonov, S., Berg, P. and Brutlag, D. L. (2006) 'A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.', *Proceedings of the National Academy of Sciences of the United States of America* 103: pp. 1412–1417. doi: 10.1073/pnas.0510310103.

Sayou, C. *et al.* (2017) 'RNA Binding by Histone Methyltransferases Set1 and Set2', *Molecular and Cellular Biology* 37: e00165-17. doi: 10.1128/mcb.00165-17.

Schier, A. C. and Taatjes, D. J. (2020) 'Structure and mechanism of the RNA polymerase II transcription machinery', *Genes & Development* 34: pp. 465–488. doi: 10.1101/gad.335679.119.

Schmidt, K. *et al.* (2018) 'The H3K4 methyltransferase setd1b is essential for hematopoietic stem and progenitor cell homeostasis in mice', *eLife* 7: e27157. doi: 10.7554/eLife.27157.

Schmitges, F. W. *et al.* (2011) 'Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks', *Molecular Cell* 42: pp. 330–341. doi: 10.1016/j.molcel.2011.03.025.

Schmitz, S. U. *et al.* (2011) 'Jarid1b targets genes regulating development and is involved in neural differentiation', *EMBO Journal* 30: pp. 4586–4600. doi: 10.1038/emboj.2011.383.

Schneider, R. *et al.* (2004) 'Histone H3 lysine 4 methylation patterns in higher eukaryotic genes', *Nature Cell Biology* 6: pp. 73–77. doi: 10.1038/ncb1076.

Schwalb, B. *et al.* (2016) 'TT-seq maps the human transient transcriptome', *Science* 352: pp. 1225–1228. doi: 10.1126/science.aad9841.

Sedkov, Y. *et al.* (2003) 'Methylation at lysine 4 of histone H3 in ecdysone-dependent development of Drosophila', *Nature* 426: pp. 78–83. doi: 10.1038/nature02080.

Seibler, J. *et al.* (2003) 'Rapid generation of inducible mouse mutants', *Nucleic Acids Research* 31: e12. doi: 10.1093/nar/gng012.

Seila, A. C. *et al.* (2008) 'Divergent transcription from active promoters', *Science* 322: pp. 1849–1851. doi: 10.1126/science.1162253.

Shah, R. N. *et al.* (2018) 'Examining the Roles of H3K4 Methylation States with Systematically Characterized Antibodies', *Molecular Cell* 72: pp. 162-177.e7. doi: 10.1016/j.molcel.2018.08.015.

Shao, W. and Zeitlinger, J. (2017) 'Paused RNA polymerase II inhibits new transcriptional initiation', *Nature Genetics* 49: pp. 1045–1051. doi: 10.1038/ng.3867.

Shi, X. *et al.* (2006) 'ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression', *Nature* 442: pp. 96–99. doi: 10.1038/nature04835.

Shi, Y. *et al.* (2009) 'Molecular Architecture of the Human Pre-mRNA 3' Processing Complex', *Molecular Cell* 33: pp. 365–376. doi: 10.1016/j.molcel.2008.12.028.

Shilatifard, A. (2012) 'The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis', *Annual Review of Biochemistry* 81: pp. 65–95. doi: 10.1146/annurev-biochem-051710-134100.

Shinsky, S. A. *et al.* (2015) 'Biochemical reconstitution and phylogenetic comparison of human SET1 family core complexes involved in histone methylation', *Journal of Biological Chemistry* 290: pp. 6361–6375. doi: 10.1074/jbc.M114.627646.

Shlyueva, D., Stampfel, G. and Stark, A. (2014) 'Transcriptional enhancers: From properties to genome-wide predictions', *Nature Reviews Genetics* 15: pp. 272–286. doi: 10.1038/nrg3682.

Shogren-Knaak, M. *et al.* (2006) 'Histone H4-K16 acetylation controls chromatin structure and protein interactions', *Science* 311: pp. 844–847. doi: 10.1126/science.1124000.

Sims, R. J. *et al.* (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: pp. 41789–41792. doi: 10.1074/jbc.C500395200.

Singh, P. *et al.* (2013) 'De Novo DNA Methylation in the Male Germ Line Occurs by Default but Is Excluded at Sites of H3K4 Methylation', *Cell Reports* 4: pp. 205–219. doi: 10.1016/j.celrep.2013.06.004.

Slany, R. K. (2020) 'MLL fusion proteins and transcriptional control', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1863: p. 194503. doi: 10.1016/j.bbagrm.2020.194503.

Smallwood, S. A. *et al.* (2011) 'Dynamic CpG island methylation landscape in oocytes and preimplantation embryos', *Nature Genetics* 43: pp. 811–814. doi: 10.1038/ng.864.

Smith, S. T. *et al.* (2004) 'Modulation of heat shock gene expression by the TAC1 chromatin-modifying complex', *Nature Cell Biology* 6: pp. 162–167. doi: 10.1038/ncb1088.

Sneppen, K. and Ringrose, L. (2019) 'Theoretical analysis of Polycomb-Trithorax systems predicts that poised chromatin is bistable and not bivalent', *Nature Communications* 10: 2133. doi: 10.1038/s41467-019-10130-2.

Soares, L. M. *et al.* (2017) 'Determinants of Histone H3K4 Methylation Patterns', *Molecular Cell* 68: pp. 773-785.e6. doi: 10.1016/j.molcel.2017.10.013.

Soares, L. M. and Buratowski, S. (2012) 'Yeast Swd2 is essential because of antagonism between Set1 histone methyltransferase complex and APT (associated with Pta1) termination factor', *Journal of Biological Chemistry* 287: pp. 15219–15231. doi: 10.1074/jbc.M112.341412.

Soares, L. M. and Buratowski, S. (2013) 'Histone Crosstalk: H2Bub and H3K4 Methylation', *Molecular Cell* 49: pp. 1019–1020. doi: 10.1016/j.molcel.2013.03.012.

Søgaard, T. M. M. and Svejstrup, J. Q. (2007) 'Hyperphosphorylation of the C-terminal repeat domain of RNA polymerase II facilitates dissociation of its complex with mediator', *Journal of Biological Chemistry* 282: pp. 14113–14120. doi: 10.1074/jbc.M701345200.

Soutourina, J. (2018) 'Transcription regulation by the Mediator complex', *Nature Reviews Molecular Cell Biology* 19: pp. 262–274. doi: 10.1038/nrm.2017.115.

Spitz, F. and Furlong, E. E. M. (2012) 'Transcription factors: From enhancer binding to developmental control', *Nature Reviews Genetics* 13: pp. 613–626. doi: 10.1038/nrg3207.

Stalder, J. *et al.* (1980) 'Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase I', *Cell* 20: pp. 451–460. doi: 10.1016/0092-8674(80)90631-5.

Steffen, P. A. and Ringrose, L. (2014) 'What are memories made of? How Polycomb and Trithorax proteins mediate epigenetic memory', *Nature Reviews Molecular Cell Biology* 15: pp. 340–356. doi: 10.1038/nrm3789.

Steger, D. J. *et al.* (2008) 'DOT1L/KMT4 Recruitment and H3K79 Methylation Are Ubiquitously Coupled with Gene Transcription in Mammalian Cells', *Molecular and Cellular Biology* 28: pp. 2825–2839. doi: 10.1128/mcb.02076-07.

Stewart, K. R. *et al.* (2015) 'Dynamic changes in histone modifications precede de novo DNA methylation in oocytes', *Genes & Development* 29: pp. 2449–2462. doi: 10.1101/gad.271353.115.

Struhl, K. and Segal, E. (2013) 'Determinants of nucleosome positioning', *Nature Structural and Molecular Biology* 20: pp. 267–273. doi: 10.1038/nsmb.2506.

Suetake, I. *et al.* (2004) 'DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction', *Journal of Biological Chemistry* 279: pp. 27816–27823. doi: 10.1074/jbc.M400181200.

Sved, J. and Bird, A. (1990) 'The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model', *Proceedings of the National Academy of Sciences of the United States of America* 87: pp. 4692–4696. doi: 10.1073/pnas.87.12.4692.

Swartz, M. N., Trautner, T. A. and Kornberg, A. (1962) 'Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids', *The Journal of Biological Chemistry* 237: pp. 1961–1967. doi: 10.1016/S0021-9258(19)73967-2.

Sze, C. C. *et al.* (2017) 'Histone H3K4 methylation-dependent and -independent functions of Set1A/COMPASS in embryonic stem cell self-renewal and differentiation', *Genes & Development* 31: pp. 1732-1737. doi: 10.1101/gad.303768.117.

Sze, C. C. *et al.* (2020) 'Coordinated regulation of cellular identity–associated H3K4me3 breadth by the COMPASS family', *Science Advances* 6: eaaz4764. doi: 10.1126/sciadv.aaz4764.

Takahashi, S. and Yokoyama, A. (2020) 'The molecular functions of common and atypical MLL fusion protein complexes', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* 1863: p. 194548. doi: 10.1016/j.bbagrm.2020.194548.

Tang, Z. *et al.* (2013) 'SET1 and p300 Act Synergistically, through Coupled Histone Modifications, in Transcriptional Activation by p53', *Cell* 154: pp. 297–310. doi: 10.1016/J.CELL.2013.06.027.

Tarasov, A. *et al.* (2015) 'Sambamba: Fast processing of NGS alignment formats', *Bioinformatics* 31: pp. 2032–2034. doi: 10.1093/bioinformatics/btv098.

Tate, C. M., Lee, J. H. and Skalnik, D. G. (2010) 'CXXC finger protein 1 restricts the Setd1A histone H3K4 methyltransferase complex to euchromatin', *FEBS Journal* 277: pp. 210–223. doi: 10.1111/j.1742-4658.2009.07475.x.

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

Teif, V. B. *et al.* (2012) 'Genome-wide nucleosome positioning during embryonic stem cell development', *Nature Structural and Molecular Biology* 19: pp. 1185–1192. doi: 10.1038/nsmb.2419.

Terranova, R. *et al.* (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: pp. 6629–6634. doi: 10.1073/pnas.0507425103.

Thomson, J. P. *et al.* (2010) 'CpG islands influence chromatin structure via the CpG-binding protein Cfp1', *Nature* 464: pp. 1082–1086. doi: 10.1038/nature08924.

Tirode, F. *et al.* (1999) 'Reconstitution of the transcription factor TFIIH: Assignment of functions for the three enzymatic subunits, XPB, XPD, and cdk7', *Molecular Cell* 3: pp. 87–95. doi: 10.1016/S1097-2765(00)80177-X.

Trésaugues, L. *et al.* (2006) 'Structural Characterization of Set1 RNA Recognition Motifs and their Role in Histone H3 Lysine 4 Methylation', *Journal of Molecular Biology* 359: pp. 1170–1181. doi: 10.1016/j.jmb.2006.04.050.

Trinkle-Mulcahy, L. *et al.* (2006) 'Repo-Man recruits PP1γ to chromatin and is essential for cell viability', *Journal of Cell Biology* 172: pp. 679–692. doi: 10.1083/jcb.200508154.

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

Trouche, D., Cook, A. and Kouzarides, T. (1996) 'The CBP co-activator stimulates E2F1/DP1 activity', *Nucleic Acids Research* 24: pp. 4139–4145. doi: 10.1093/nar/24.21.4139.

Tschiersch, B. *et al.* (1994) 'The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene

complexes', *The EMBO Journal* 13: pp. 3822-3831. doi: 10.1002/j.1460-2075.1994.tb06693.x.

Turberfield, A. H. *et al.* (2019) 'KDM2 proteins constrain transcription from CpG island gene promoters independently of their histone demethylase activity', *Nucleic Acids Research* 47: pp. 9005–9023. doi: 10.1093/nar/gkz607.

Turner, B. M. (2000) 'Histone acetylation and an epigenetic code', *BioEssays* 22: pp. 836–845. doi: 10.1002/1521-1878(200009)22:9<836::AID-BIES9>3.0.CO;2-X.

Tusi, B. K. *et al.* (2015) '*Setd1a* regulates progenitor B-cell-to-precursor B-cell development through histone H3 lysine 4 trimethylation and *Ig heavy-chain* rearrangement', *The FASEB Journal* 29: pp. 1505–1515. doi: 10.1096/fj.14-263061.

Tyagi, S. *et al.* (2007) 'E2F Activation of S Phase Promoters via Association with HCF-1 and the MLL Family of Histone H3K4 Methyltransferases', *Molecular Cell* 27: pp. 107–119. doi: 10.1016/j.molcel.2007.05.030.

Tykocinski, M. L. and Max, E. E. (1984) 'CG dinucleotide clusters in MHC genes and in 5' demethylated genes', *Nucleic Acids Research* 12: pp. 4385–4396. doi: 10.1093/nar/12.10.4385.

Valerio, D. G. *et al.* (2017) 'Histone acetyltransferase activity of MOF is required for MLL-AF9 leukemogenesis', *Cancer Research* 77: pp. 1753–1762. doi: 10.1158/0008-5472.CAN-16-2374.

Vallianatos, C. N. and Iwase, S. (2015) 'Disrupted intricacy of histone H3K4 methylation in neurodevelopmental disorders', *Epigenomics* 7: pp. 503–518. doi: 10.2217/epi.15.1.

Venkatasubrahmanyam, S. *et al.* (2007) 'Genome-wide, as opposed to local, antisilencing is mediated redundantly by the euchromatic factors Set1 and H2A.Z', *Proceedings of the National Academy of Sciences of the United States of America* 104: pp. 16609–16614. doi: 10.1073/pnas.0700914104.

Venolia, L. *et al.* (1982) 'Transformation with DNA from 5-azacytidine-reactivated X chromosomes', *Proceedings of the National Academy of Sciences of the United States of America* 79: pp. 2352–2354. doi: 10.1073/pnas.79.7.2352.

Vermeulen, M. *et al.* (2007) 'Selective Anchoring of TFIID to Nucleosomes by Trimethylation of Histone H3 Lysine 4', *Cell* 131: pp. 58–69. doi: 10.1016/J.CELL.2007.08.016.

Vermeulen, M. *et al.* (2010) 'Quantitative Interaction Proteomics and Genome-wide Profiling of Epigenetic Histone Marks and Their Readers', *Cell* 142: pp. 967–980. doi: 10.1016/j.cell.2010.08.020.

Vermeulen, M. and Timmers, H. T. M. (2010) 'Grasping trimethylation of histone H3 at lysine 4', *Epigenomics* 2: pp. 395–406. doi: 10.2217/epi.10.11.

Vertino, P. M. *et al.* (2002) 'DNMT1 is a component of a multiprotein DNA replication complex', *Cell Cycle* 1: pp. 416–423. doi: 10.4161/cc.1.6.270.

Vethantham, V. *et al.* (2012) 'Dynamic Loss of H2B Ubiquitylation without Corresponding Changes in H3K4 Trimethylation during Myogenic Differentiation', *Molecular and Cellular Biology* 32: pp. 1044–1055. doi: 10.1128/mcb.06026-11.

Voigt, P. *et al.* (2012) 'Asymmetrically Modified Nucleosomes', *Cell* 151: pp. 181–193. doi: 10.1016/j.cell.2012.09.002.

Voigt, P., Tee, W. W. and Reinberg, D. (2013) 'A double take on bivalent promoters.', *Genes & Development* 27: pp. 1318–1338. doi: 10.1101/gad.219626.113.

Voo, K. S. *et al.* (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: pp. 2108–2121. doi: 10.1128/MCB.20.6.2108-2121.2000.

Vos, S. M. *et al.* (2018) 'Structure of paused transcription complex Pol II–DSIF–NELF', *Nature* 560: pp. 601–606. doi: 10.1038/s41586-018-0442-2.

Wachter, E. *et al.* (2014) 'Synthetic CpG islands reveal DNA sequence determinants of chromatin structure', *eLife* 3: e03397. doi: 10.7554/eLife.03397.

Wada, T. *et al.* (1998) 'DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs', *Genes & Development* 12: pp. 343–356. doi: 10.1101/gad.12.3.343.

Wade, P. A., Pruss, D. and Wolffe, A. P. (1997) 'Histone acetylation: Chromatin in action', *Trends in Biochemical Sciences* 22: pp. 128–132. doi: 10.1016/S0968-0004(97)01016-5.

Wan, M. *et al.* (2013) 'The trithorax group protein Ash2l is essential for pluripotency and maintaining open chromatin in embryonic stem cells', *Journal of Biological Chemistry* 288: pp. 5039–5048. doi: 10.1074/jbc.M112.424515.

Wang, H. *et al.* (2001) 'Purification and functional characterization of a histone H3-lysine 4-specific methyltransferase', *Molecular Cell* 8: pp. 1207–1217. doi: 10.1016/S1097-2765(01)00405-1.

Wang, H. *et al.* (2004) 'Role of histone H2A ubiquitination in Polycomb silencing', *Nature* 431: pp. 873–878. doi: 10.1038/nature02985.

Wang, L. *et al.* (2017) 'A cytoplasmic COMPASS is necessary for cell survival and triple-negative breast cancer pathogenesis by regulating metabolism.', *Genes & Development* 31: pp. 2056–2066. doi: 10.1101/gad.306092.117.

Wang, P. *et al.* (2009) 'Global Analysis of H3K4 Methylation Defines MLL Family Member Targets and Points to a Role for MLL1-Mediated H3K4 Methylation in the Regulation of Transcriptional Initiation by RNA Polymerase II', *Molecular and Cellular Biology* 29: pp. 6074–6085. doi: 10.1128/mcb.00924-09.

Wang, X. *et al.* (2019) 'Regulation of histone methylation by automethylation of PRC2', *Genes & Development* 33: pp. 1416–1427. doi: 10.1101/gad.328849.119.

Wang, Z. *et al.* (2009) 'Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes', *Cell* 138: pp. 1019–1031. doi: 10.1016/j.cell.2009.06.049.

Wang, Z. *et al.* (2010) 'Pro isomerization in MLL1 PHD3-bromo cassette connects H3K4me readout to CyP33 and HDAC-mediated repression.', *Cell* 141: pp. 1183–1194. doi: 10.1016/j.cell.2010.05.016.

Weber, M. *et al.* (2007) 'Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome', *Nature Genetics* 39: pp. 457–466. doi: 10.1038/ng1990.

Weirich, S. *et al.* (2015) 'Somatic cancer mutations in the MLL3-SET domain alter the catalytic properties of the enzyme', *Clinical Epigenetics* 7: p. 36. doi: 10.1186/s13148-015-0075-3.

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

Williamson, K. *et al.* (2013) 'Catalytic and Functional Roles of Conserved Amino Acids in the SET Domain of the S. cerevisiae Lysine Methyltransferase Set1', *PLoS One* 8: p. 57974. doi: 10.1371/journal.pone.0057974.

Windhager, L. *et al.* (2012) 'Ultrashort and progressive 4sU-tagging reveals key characteristics of RNA processing at nucleotide resolution', *Genome Research* 22: pp. 2031–2042. doi: 10.1101/gr.131847.111.

Wissink, E. M. *et al.* (2019) 'Nascent RNA analyses: tracking transcription and its regulation', *Nature Reviews Genetics* 20: pp. 705–723. doi: 10.1038/s41576-019-0159-6.

Woodcock, C. L. and Ghosh, R. P. (2010) 'Chromatin higher-order structure and dynamics', *Cold Spring Harbor Perspectives in Biology* 2: a000596. doi: 10.1101/cshperspect.a000596.

Workman, J. L. and Kingston, R. E. (1998) 'Alteration of nucleosome structure as a mechanism of transcriptional regulation', *Annual Review of Biochemistry* 67: pp. 545-579. doi: 10.1146/annurev.biochem.67.1.545.

Wu, H. *et al.* (2010) 'Structural Biology of Human H3K9 Methyltransferases', *PLoS One* 5: e8570. doi: 10.1371/journal.pone.0008570.

Wu, M. *et al.* (2008) 'Molecular Regulation of H3K4 Trimethylation by Wdr82, a Component of Human Set1/COMPASS', *Molecular and Cellular Biology* 28: pp. 7337–7344. doi: 10.1128/MCB.00976-08.

Wysocka, J. *et al.* (2003) 'Human Sin3 deacetylase and trithorax-related Set1/Ash2 histone H3-K4 methyltransferase are tethered together selectively by the cell-proliferation factor HCF-1', *Genes & Development* 17: pp. 896–911. doi: 10.1101/gad.252103.

Wysocka, J. *et al.* (2006) 'A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling', *Nature* 442: pp. 86–90. doi: 10.1038/nature04815.

Xie, L. *et al.* (2011) 'KDM5B regulates embryonic stem cell self-renewal and represses cryptic intragenic transcription', *EMBO Journal* 30: pp. 1473–1484. doi: 10.1038/emboj.2011.91.

Xu, C. *et al.* (2011) 'The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain', *Nature Communications* 2: p. 227. doi: 10.1038/ncomms1237.

Yagi, H. *et al.* (1998) 'Growth disturbance in fetal liver hematopoiesis of Mll-mutant mice', *Blood* 92: pp. 108–117. doi: 10.1182/blood.v92.1.108.413k11_108_117.

Yamada, T. *et al.* (2006) 'P-TEFb-mediated phosphorylation of hSpt5 C-terminal repeats is critical for processive transcription elongation', *Molecular Cell* 21: pp. 227–237. doi: 10.1016/j.molcel.2005.11.024.

Yamaguchi, Y. *et al.* (1999) 'NELF, a multisubunit complex containing RD, cooperates with DSIF to repress RNA polymerase II elongation', *Cell* 97: pp. 41–51. doi: 10.1016/S0092-8674(00)80713-8.

Yang, Y. *et al.* (2020) 'A non-canonical monovalent zinc finger stabilizes the integration of Cfp1 into the H3K4 methyltransferase complex COMPASS', *Nucleic Acids Research* 48: pp. 421–431. doi: 10.1093/nar/gkz1037.

Ye, J. *et al.* (2012) 'Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction', *BMC Bioinformatics* 13: 134. doi: 10.1186/1471-2105-13-134.

Yokoyama, A. *et al.* (2004) 'Leukemia Proto-Oncoprotein MLL Forms a SET1-Like Histone Methyltransferase Complex with Menin To Regulate Hox Gene Expression', *Molecular and Cellular Biology* 24: pp. 5639–5649. doi: 10.1128/mcb.24.13.5639-5649.2004.

Yokoyama, A. *et al.* (2005) 'The menin tumor suppressor protein is an essential oncogenic cofactor for MLL-associated leukemogenesis', *Cell* 123: pp. 207–218. doi: 10.1016/j.cell.2005.09.025.

Yokoyama, A. and Cleary, M. L. (2008) 'Menin Critically Links MLL Proteins with LEDGF on Cancer-Associated Target Genes', *Cancer Cell* 14: pp. 36–46. doi: 10.1016/j.ccr.2008.05.003.

Young, S. R. L. and Skalnik, D. G. (2007) 'CXXC finger protein 1 is required for normal proliferation and differentiation of the PLB-985 myeloid cell line', *DNA and Cell Biology* 26: pp. 80–90. doi: 10.1089/dna.2006.0535.

Yu, B. D. *et al.* (1995) 'Altered Hox expression and segmental identity in Mll-mutant mice', *Nature* 378: pp. 505–508. doi: 10.1038/378505a0.

Zeitlinger, J. *et al.* (2007) 'RNA polymerase stalling at developmental control genes in the

Drosophila melanogaster embryo', *Nature Genetics* 39: pp. 1512–1516. doi: 10.1038/ng.2007.26.

Zeleznik-Le, N. J., Harden, A. M. and Rowley, J. D. (1994) '11q23 translocations split the "AT-hook" cruciform DNA-binding region and the transcriptional repression domain from the activation domain of the mixed-lineage leukemia (MLL) gene', *Proceedings of the National Academy of Sciences of the United States of America* 91: pp. 10610–10614. doi: 10.1073/pnas.91.22.10610.

Zhang, H. *et al.* (2019) 'MLL1 Inhibition and Vitamin D Signaling Cooperate to Facilitate the Expanded Pluripotency State', *Cell Reports* 29: pp. 2659-2671.e6. doi: 10.1016/j.celrep.2019.10.074.

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

Zhang, K. *et al.* (2005) 'The Set1 methyltransferase opposes Ipl1 Aurora kinase functions in chromosome segregation', *Cell* 122: pp. 723–734. doi: 10.1016/j.cell.2005.06.021.

Zhang, P. *et al.* (2012) 'The plasticity of WDR5 peptide-binding cleft enables the binding of the SET1 family of histone methyltransferases', *Nucleic Acids Research* 40: pp. 4237–4246. doi: 10.1093/nar/gkr1235.

Zhang, T., Cooper, S. and Brockdorff, N. (2015) 'The interplay of histone modifications - writers that read.', *EMBO Reports* 16: pp. 1467–1481. doi: 10.15252/embr.201540945.

Zhang, X. *et al.* (2002) 'Structure of the Neurospora SET domain protein DIM-5, a histone H3 lysine methyltransferase', *Cell* 111: pp. 117–127. doi: 10.1016/S0092-8674(02)00999-6.

Zhang, Y. *et al.* (2008) 'Model-based analysis of ChIP-Seq (MACS)', *Genome Biology* 9: R137. doi: 10.1186/gb-2008-9-9-r137.

Zhang, Y. *et al.* (2009) 'DNA Methylation Analysis of Chromosome 21 Gene Promoters at Single Base Pair and Single Allele Resolution', *PLoS Genetics* 5: e1000438. doi: 10.1371/journal.pgen.1000438.

Zhang, Y. *et al.* (2010) 'Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail', *Nucleic Acids Research* 38: pp. 4246–4253. doi: 10.1093/nar/gkq147.

Zhang, Y. and Reinberg, D. (2001) 'Transcription regulation by histone methylation: Interplay between different covalent modifications of the core histone tails', *Genes & Development* 15: pp. 2343–2360. doi: 10.1101/gad.927301.

Zhao, Y. and Garcia, B. A. (2015) 'Comprehensive catalog of currently documented histone modifications', *Cold Spring Harbor Perspectives in Biology* 7: a025064. doi: 10.1101/cshperspect.a025064.

Zheng, Y. *et al.* (2014) 'Site-specific human histone H3 methylation stability: fast K4me3 turnover.', *Proteomics* 14: pp. 2190–2199. doi: 10.1002/pmic.201400060.

Zhou, X. *et al.* (2006) 'Optimization of the Tet-On system for regulated gene expression through viral evolution', *Gene Therapy* 13: pp. 1382–1390. doi: 10.1038/sj.gt.3302780.

Zhu, B. *et al.* (2005) 'Monoubiquitination of human histone H2B: The factors involved and their roles in HOX gene regulation', *Molecular Cell* 20: pp. 601–611. doi: 10.1016/j.molcel.2005.09.025.

Zhu, L. *et al.* (2016) 'ASH1L links histone H3 lysine 36 dimethylation to MLL Leukemia', *Cancer Discovery* 6: pp. 770–783. doi: 10.1158/2159-8290.CD-16-0058.