Biochemical isolation and mass spectrometric analysis of the CpG island proteome

Deniz Kaya



Linacre College, University of Oxford

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

Hilary Term 2022

In loving memory of Erdal Kaya

Acknowledgements

I thank Prof Rob Klose and Prof Shabaz Mohammed for their guidance, support, and most of all, patience. Both of them gave me an abundance of freedom in my DPhil which I have highly appreciated. I would also like to thank the Medical Research Council and the Department of Biochemistry for funding my studentship. I am grateful to Sabrina and Svenja for their assistance with the mass spectrometry side of this project. I would like to offer my special thanks to Emilia, Nadya, and Paula for helping me with ChIP sequencing, and to Aleks for his help with microscopy. I would also like to express my gratitude to Tatyana for always taking the time to help me with innumerable things throughout my DPhil. I am incredibly grateful to Antonio who helped me with my protein purifications and whose invaluable experimental advice was fundamental to my progress in particular during my first year. Equally, I am deeply grateful to Krzysztof who also provided important technical tips, helped me with my data analysis, and sacrificed his time for numerous stimulating and refreshing chats. Moreover, I would like to thank my friends for their support and providing joyful distractions to rest my mind outside of my research. Finally, I am eternally grateful to my mum for her unwavering support and patience which made this all possible.

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.

Deniz Kaya
14th April 2022

Contents

Acknowledgements I

Declaration of authorship II

List of figures VI

List of tables VIII

List of abbreviations IX

Abstract XI

1 Introduction 1
1.1 Chromatin-based regulation of transcription 1
1.1.1 Eukaryotic DNA is packaged into nucleosomes 1
1.1.2 Histone modifications and DNA methylation are mediators of epigenetic transcription regulation 2
1.1.3 Chromatin remodellers regulate DNA accessibility 4
1.1.4 Chromatin features associated with gene promoters in vertebrates 6
1.1.5 The RNAPII transcription cycle 9
1.2 The role of CpG islands in gene regulation 14
1.2.1 Identification of CpG islands 14
1.2.2 Readers of non-methylated CpGs 16
1.2.3 Writers and erasers of CpG methylation 18
1.2.4 CpG islands are associated with gene regulatory elements 21
1.3 Antagonism of Polycomb and Trithorax groups in transcription regulation 25
1.3.1 PcG complex composition and target site identification 26
1.3.2 COMPASS complex composition and target site identification 29
1.3.3 Mechanisms of PcG- and TrxG-mediated chromatin regulation 32
1.4 Proteomic methods to study chromatin and gene regulation 36
1.4.1 Technical challenges in chromatin proteomics 36
1.4.1.1 Preserving chromatin interactions 36
1.4.1.2 Identifying low-abundance proteins 38
1.4.2 Whole-chromatin purification 41
1.4.3 Proximal proteome enrichment 43
1.5 Aim of this study 45

2 Materials and methods 46
2.1 Molecular cloning 46
2.1.1 Cloning of a bicistronic vector for co-expression of ZF-CXXC and BirA 46
2.1.2 Site-directed mutagenesis of the ZF-CXXC expression construct 47
2.2 Protein expression, purification, and biotinylation 49
2.2.1 Expression and purification of BirA 49
2.2.2 Expression, purification, and *in vitro* biotinylation of ZF-CXXC 49

2.2.3 Expression and purification of *in vivo* biotinylated ZF-CXXC 51
2.3 Mammalian cell culture 52
2.4 CpG island chromatin enrichment 52
2.4.1 ChIP-MS-based approach 52
2.4.2 ChIP-SICAP-based approach 53
2.4.3 ChEP-based approach 54
2.5 Quantitative real-time PCR 55
2.6 Microscopy 56
2.7 Chromatin immunoprecipitation 57
2.8 Mass spectrometry 58
2.8.1 Sample preparation 58
2.8.1.1 In-gel digestion of *in vitro* biotinylated ZF-CXXC sample 58
2.8.1.2 Filter-aided sample preparation 59
2.8.1.3 Desalting 59
2.8.2 Data acquisition 60
2.8.2.1 Intact protein mass spectrometry 60
2.8.2.2 Shotgun proteomics 60
2.8.3 Protein identification and analysis 60
2.8.4 Data visualisation 61
2.9 Next generation sequencing 61
2.9.1 Library preparation and sequencing 61
2.9.2 Read alignment and normalisation 62
2.9.3 Peak calling 62
2.9.4 Differential enrichment and binding analysis 62
2.9.5 Data visualisation 62
2.10 Published datasets used in this study 63

3 Producing a ZF-CXXC bait protein to enrich CpG island chromatin ... 64
3.1 Introduction 64
3.2 Results 64
3.2.1 Expression and purification of BirA enzyme for use in biotinylation reactions of ZF-CXXC 64
3.2.2 Expression and purification of the ZF-CXXC domain 66
3.2.3 *In vitro* biotinylation of ZF-CXXC by BirA is inefficient and causes protein precipitation 67
3.2.4 A new approach to isolate biotinylated ZF-CXXC protein 70
3.2.5 Expression and purification of a ZF-CXXC domain with mutations in its CpG recognition site 73
3.3 Summary 75

4 Development of a chromatin purification approach to enrich CGIs 76
4.1 Introduction 76
4.2 Results 77
4.2.1 CpG islands isolated with a ChIP-based approach are contaminated with non-chromatin proteins 77
4.2.2 An additional chromatin purification step using SICAP does not improve CGI enrichment and purity 85
4.2.3 A new approach to purify chromatin for CGI enrichment 91

4.2.4 The new chromatin purification approach improves CGI enrichment 99
4.3 Discussion 104

5 C-CAP identifies known and potentially novel CGI-associated proteins 107
5.1 Introduction 107
5.2 Results 107
5.2.1 The ZF-CXXC^{mt} protein is deficient in binding non-methylated CpGs 107
5.2.2 Next generation sequencing reveals enrichment of non-methylated islands and regions that contain transcription-associated histone modifications 110
5.2.3 Sequencing analysis reveals enrichment of rDNA genes in C-CAP 114
5.2.4 The ZF-CXXC^{mt} protein improves the specificity of the CGI proteome 115
5.2.5 The identification of new CGI-associated proteins 122
5.3 Discussion 129

6 C-CAP in PRC1 and PRC2 CKO cells provides new insight into chromatin composition and function of CGIs 132
6.1 Introduction 132
6.2 Results 133
6.2.1 C-CAP equally recovers CGIs in PRC1 and PRC2 CKO cells 133
6.2.2 Conditional knockout of PRC1 in mESCs causes extensive changes in the CGI proteome 137
6.2.3 Conditional knockout of PRC2 in mESCs causes less pronounced CGI proteome changes than loss of PRC1 142
6.2.4 ChIP-sequencing of differentially enriched proteins in PRC2 CKO cells corroborates MS results 146
6.3 Discussion 149

7 Discussion 155
7.1 The choice of the chromatin purification approach 156
7.2 The strength of MS-based techniques in identifying chromatin interactions *de novo* 159
7.3 Application of C-CAP to understand CGI function in development and disease 162

8 Summary and perspectives 165

Appendix: Supplementary data 166

References 175

List of figures

Figure 1: Schematic of PRC1 complexes in mammals 27
Figure 2: Schematic of PRC2 complexes in mammals 28
Figure 3: Schematic of COMPASS H3K4 methyltransferase complexes in mammals 30
Figure 4: Schematic workflow for producing *in vitro* biotinylated ZF-CXXC 65
Figure 5: SDS-PAGE analysis of BirA purification by affinity and size exclusion chromatography 66
Figure 6: SDS-PAGE analysis of ZF-CXXC affinity purification and TEV cleavage 67
Figure 7: SDS-PAGE analysis of protein precipitation during *in vitro* biotinylation of ZF-CXXC 68
Figure 8: SDS-PAGE analysis of *in vitro* biotinylated ZF-CXXC size exclusion chromatography 69
Figure 9: Intact protein mass spectrometry of *in vitro* biotinylated ZF-CXXC 70
Figure 10: Co-expression of BirA with ZF-CXXC for *in vivo* biotinylation 71
Figure 11: SDS-PAGE analysis of *in vivo* biotinylated ZF-CXXC purification by affinity and size exclusion chromatography 72
Figure 12: Intact protein mass spectrometry of *in vivo* biotinylated ZF-CXXC after improving the purification process 73
Figure 13: SDS-PAGE analysis of ZF-CXXC^{mt} purification by affinity and size exclusion chromatography 74
Figure 14: Intact protein mass spectrometry of *in vivo* biotinylated ZF-CXXC^{mt} 75
Figure 15: A CGI chromatin enrichment approach combining ChIP and BioCAP 78
Figure 16: Quantitative PCR analysis of C-CAP elution fractions 79
Figure 17: Protein count and log2 intensity of subcellular protein categories in C-CAP 80
Figure 18: Comparison of selected protein groups in C-CAP and the full proteome 82
Figure 19: Comparison of PcG and TrxG core subunit percentiles in C-CAP and the full proteome 83
Figure 20: Gene ontology, pathway, and protein domain enrichment analyses of C-CAP eluates 84
Figure 21: Schematic workflow of C-CAP in combination with SICAP 86
Figure 22: Quantitative PCR analysis of C-CAP elution fractions 87
Figure 23: Flowchart of sample processing steps after C-CAP 87
Figure 24: Protein identifications in C-CAP and SICAP eluates 88
Figure 25: Proteins categorised by subcellular localisation and group affiliation 89
Figure 26: Gene ontology and pathway enrichment analyses for C-CAP and SICAP 91
Figure 27: Workflow of the new chromatin purification and CGI enrichment approach 92
Figure 28: Validation of nuclei purification by fluorescence microscopy 93
Figure 29: Detailed steps of the new chromatin purification approach 94
Figure 30: Categorisation of protein identifications according to subcellular localisation 95
Figure 31: Cumulative distribution of abundance-ranked protein groups per sample 96
Figure 32: Enrichment and depletion of non-nuclear, nuclear, and chromatin-binding proteins in the purification process 98
Figure 33: Quantitative PCR analysis of C-CAP elution fractions 99
Figure 34: MA plots of CGI proteome enrichment in the C-CAP elution fractions 101
Figure 35: Volcano plots of CGI proteome enrichment in C-CAP elution fractions 103

Figure 36: Gene ontology, pathway, and protein domain enrichment analyses of proteins identified in C-CAP eluates 104
Figure 37: Schematic of CGI-chromatin affinity purification including the ZF-CXXC^{mt} ... 108
Figure 38: qPCR analysis for C-CAP elution fractions of ZF-CXXC^{WT} and ZF-CXXC^{mt} . 109
Figure 39: Genome-wide enrichment of NMIs in C-CAP eluates 110
Figure 40: C-CAP sequencing reads are enriched at the TSS of genes 111
Figure 41: Analysis of C-CAP reads enriched at selected histone modification peaks... 113
Figure 42: Specific regions of the rDNA gene locus are enriched by C-CAP 114
Figure 43: MA plots of WT/Input and WT/mt analyses 116
Figure 44: Volcano plots and Venn diagram of significantly enriched proteins in the mt eluate of WT/mt and in the WT eluate of the WT/Input analyses 118
Figure 45: Volcano plot and Venn diagram of significantly enriched proteins in the WT eluate of the WT/mt analysis 119
Figure 46: GO analysis of significantly enriched proteins in the WT/Input and WT/mt approach 121
Figure 47: Network analysis of significantly enriched proteins in C-CAP 123
Figure 48: ChIP-qPCR analysis for CDCA7L 126
Figure 49: ChIP-qPCR analysis for CDCA7 127
Figure 50: ChIP-qPCR analysis for KAT6B 128
Figure 51: BANP, SAMD1, and ZBTB11 localise at NMIs in mESCs 129
Figure 52: Conditional deletion systems for PRC1 and PRC2 134
Figure 53: CGI chromatin enrichment in PRC1 and PRC2 CKO cells is comparable between untreated and TAM-treated cells 136
Figure 54: Volcano plot of differentially enriched proteins in C-CAP of PRC1 CKO..... 138
Figure 55: Differentially enriched PcG proteins in the CGI proteome of PRC1 CKO 139
Figure 56: Differentially enriched protein groups in the CGI proteome of PRC1 CKO ... 141
Figure 57: Volcano plot of differentially enriched proteins in C-CAP of PRC2 CKO..... 143
Figure 58: Differentially enriched PcG proteins in the CGI proteome of PRC2 CKO 144
Figure 59: Differentially enriched protein groups in the CGI proteome of PRC2 CKO ... 145
Figure 60: Calibrated ChIP-seq analysis of KDM2B, PCGF1, and PCGF6 shows increased binding at NMIs in PRC2 CKO 147
Figure 61: Differential binding analysis of KDM2B, PCGF1, and PCGF6 in PRC2 CKO 148
Figure 62: Overlap of genomic sites that gained occupancy of KDM2B, PCGF1, and PCGF6 in PRC2 CKO 149

Figure S1: Validation of the SUZ12 CKO cell line (courtesy of Iona Manley)..... 166
Figure S2: Correlation between mRNA and protein log2 fold changes in PRC1 CKO ... 167
Figure S3: Correlation between mRNA and protein log2 fold changes in PRC2 CKO ... 168
Figure S4: DNA binding motifs enriched in ZF-CXXC^{WT} and ZF-CXXC^{mt} C-CAP eluates 169
Figure S5: DNA binding motifs enriched in C-CAP eluates of PRC1 and PRC2 CKO ... 170
Figure S6: DNA binding motifs enriched in KDM2B, PCGF1, and PCGF6 ChIP assays 170

List of tables

Table 1: PCR mix for BirA amplification 46
Table 2: PCR primers for BirA cloning 46
Table 3: Thermocycler programme for BirA amplification 46
Table 4: ZF-CXXC plasmid restriction digestion 47
Table 5: PCR mix for site-directed mutagenesis of ZF-CXXC 48
Table 6: PCR primers for site-directed mutagenesis of ZF-CXXC 48
Table 7: Thermocycler programme for site-directed mutagenesis of ZF-CXXC 48
Table 8: qPCR master mix components 56
Table 9: qPCR thermocycler programme 56
Table 10: Antibodies used in this study for ChIP 58
Table 11: References of published datasets used in this study 63
Table 12: Shortlist of potentially new CGI-binding proteins 124

Table S1: List of enriched proteins from Figure 35 171
Table S2: List of enriched proteins from Figure 44B 172
Table S3: List of enriched proteins from Figure 45A 173
Table S4: List of background proteins (enriched in Input and ZF-CXXC^{mt}) 174

List of abbreviations

CGI	CpG island
2YT	2X yeast extract and tryptone medium
5mC	5-methylcytosine
BP	Biological process
BRE	TFIIB recognition element
CC	Cellular component
C-CAP	CGI-chromatin affinity purification
ChEP	Chromatin enrichment for proteomics
ChIP	Chromatin immunoprecipitation
CKO	Conditional knockout
COMPASS	Complex of proteins associated with Set1
cPRC1	Canonical Polycomb repressive complex 1
CTD	RNAPII C-terminal domain
CV	Column volume
DNMT	DNA methyltransferase
DSG	Disuccinimidyl glutarate
EdU	5-ethynyl-2′-deoxyuridine
FASP	Filter-aided sample preparation
FC	Fold change
FDR	False discovery rate
GO	Gene ontology
GTF	General transcription factor
HAT	Histone acetyltransferase
HDM	Histone demethylase
hESCs	Human embryonic stem cells
IPTG	Isopropyl β-D-1-thiogalactopyranoside
KDM	Lysine demethylase
KMT	Lysine methyltransferase
kPSI	Kilo pound per square inch
LB	Luria-Bertani medium/agar
LOBSTR	Low background strain
mESCs	Mouse embryonic stem cells
MF	Molecular function
MS	Mass spectrometry
mt	Mutant
Ni-NTA	Nickel nitrilotriacetic acid beads
NMI	Non-methylated island
OD	Optical density
PBS	Phosphate-buffered saline
PcG	Polycomb group
PCR	Polymerase chain reaction
PHD	Plant homeodomain
PIC	Pre-initiation complex
PIC	Protease inhibitor cocktail (only in Chapter 2)
PRC	Polycomb repressive complex
PTM	Post-translational modification

qPCR	Quantitative polymerase chain reaction
rDNA	Ribosomal DNA
RNAP	RNA polymerase
RT	Room temperature
SDS-PAGE	Sodium dodecylsulfate polyacrylamid gel electrophoresis
SEC	Size exclusion chromatography
SICAP	Selective isolation of chromatin-associated proteins
TAM	4-hydroxytamoxifen
TDT	Terminal deoxynucleotidyl transferase
TEV	Tobacco etch virus protease
TF	Transcription factor
TrxG	Trithorax group
TSS	Transcription start site
UNT	Untreated
vPRC1	Variant Polycomb repressive complex 1
WCL	Whole-cell lysate
WH	Winged helix
WT	Wild type
ZF-CXXC	Zinc finger CXXC

Abstract

Biochemical isolation and mass spectrometric analysis of the CpG island proteome

Deniz Kaya, Linacre College, Hilary Term 2022

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

Cytosine methylation is a pervasive DNA modification in vertebrate genomes that occurs predominantly in the context of sparsely distributed CpG dinucleotides. In contrast to the bulk genome, short genomic elements referred to as CpG islands (CGIs) feature a high GC content and clusters of non-methylated CpGs. CGIs co-localise with the majority of gene promoters where their distinct DNA sequence features enables the recruitment and assembly of transcription regulating complexes. Thus, as sites of transcription initiation, CGIs exert an important function in controlling promoter activity and gene expression. Despite their importance, it remains unclear how this is mechanistically achieved and how CGI dysfunction contributes to diseases such as cancer. This is largely due to the fact that the majority of our mechanistic insights into CGI function come from candidate-based approaches such as chromatin immunoprecipitation coupled to sequencing, which are inherently biased and limit us to investigating one protein at a time. The capacity to attribute detailed mechanistic understanding to CGI function has been hampered by our lack of a comprehensive inventory of proteins that contribute to their maintenance and function.

To overcome this fundamental lack of understanding, I have developed a new biochemical approach to isolate CGI chromatin from the bulk genome in order to investigate the CGI proteome in an unbiased manner using mass spectrometry. Applying next generation sequencing, I validated on the genome level that this method is suitable for specific enrichment of CGIs from mouse embryonic stem cells. Mass spectrometry data revealed many known CGI interactors, and it provided new leads for further investigation of potentially novel CGI-associated proteins. By deleting selected components of the CGI proteome, I uncovered CGI protein composition changes that pointed to new interrelations between the components of the gene regulation network at CGI promoters. Finally, using calibrated ChIP-sequencing as an orthogonal approach, I verified for selected proteins that the CGI proteome alterations were caused by changes in genomic occupancy. In conclusion, this work presents the first method to investigate the CGI proteome and its alterations comprehensively and unbiasedly in different conditions. Ultimately, this method can advance our knowledge of how CGIs contribute to gene regulation in developmental processes and pathogenesis.

1 Introduction

1.1 Chromatin-based regulation of transcription

The accurate execution of gene expression in metazoan development and physiology is crucial for cell differentiation, maintenance of cell identity, and adaptation to environmental changes. A focal point for regulating gene expression is the transcription of the DNA-encoded information into RNA. This first step in the gene expression process is performed in metazoans by three types of DNA-dependent RNA polymerases (RNAPI-III) that transcribe different groups of genes (Roeder and Rutter, 1969, 1970; Zylber and Penman, 1971). Of these polymerases, only RNAPII transcribes protein-coding genes, whereas RNAPI and RNAPIII primarily transcribe ribosomal RNA (rRNA) and transfer RNA (tRNA) genes, respectively. This chapter focuses mainly on mechanisms and principles of RNAPII-mediated transcription in the context of vertebrate genomes. A key determinant in gene expression is the chromatin substrate on which transcription is performed. By default, chromatin is inhibitory to transcription as it limits the accessibility of DNA (Li et al., 2007; Zhou et al., 2019). For transcription to initiate, the chromatin structure needs to be opened up through epigenetic modifications and chromatin remodelling. Changes to the chromatin architecture need to be temporally and spatially regulated to ensure that only those genes are expressed that are required in a particular developmental state or physiological condition. In addition to chromatin accessibility, gene expression is regulated through DNA sequence features that recruit specific transcription factors and RNAP to the right genes (Reiter et al., 2017). Thus, dynamic regulation of the chromatin composition and architecture represent important elements in establishing and maintaining cell identity during metazoan development.

1.1.1 Eukaryotic DNA is packaged into nucleosomes

Chromatin is a heterogeneous assembly of DNA, histones, and non-histone proteins (Bernstein et al., 2010). The repeating building block of chromatin in eukaryotes is the nucleosome, which comprises a histone octamer around which ~146 bp of genomic DNA are wound (Luger et al., 1997). The histone octamer is formed by two copies of each of the four core histone proteins H2A, H2B, H3, and H4. The core histones assemble into four dimers defined by the H2A/H2B and H3/H4 histone pairs. Additionally, a fifth histone, termed linker histone H1, can bind outside of the nucleosome core on the short stretch of DNA that connects two nucleosomes (Allan et al., 1980). In mammals, all histones except H4 have functionally specialised variants that can replace their canonical paralogues in

certain genomic contexts (Talbert and Henikoff, 2010). For example, CENP-A is an H3 variant that marks chromosome centromeres, and H2A.Z and H3.3 variants are associated with active gene promoters (Earnshaw and Rothfield, 1985; Jin and Felsenfeld, 2007). The packing of DNA into arrays of nucleosomes leads to a considerable compaction of the linear DNA molecule and enables higher-order chromatin structures. This in turn determines the accessibility of DNA which is a key parameter in regulating all DNA-based processes such as replication, repair, and transcription (Zhou et al., 2019).

1.1.2 Histone modifications and DNA methylation are mediators of epigenetic transcription regulation

Histones are highly basic proteins due to the abundance of positively charged arginine and lysine residues. The electrostatic attraction between positively charged histones and the negatively charged DNA backbone promotes tight wrapping of the DNA molecule around nucleosomes and contributes to the compaction of chromatin (Luger et al., 1997). The degree of chromatin compaction inside the cell is nonuniform, containing patches of high and low compaction referred to as heterochromatin and euchromatin, respectively (Babu and Verma, 1987). An important factor influencing chromatin compaction are post-translational modifications (PTMs) of histones, which mainly occur on lysine, arginine, serine, and threonine (Huang et al., 2014). Histones can be modified on the structured domains that form the nucleosome core and the unstructured N-terminal tails that protrude from the core (Lawrence et al., 2016). Some examples of histone PTMs include acetylation, methylation, phosphorylation, and ubiquitylation. These modifications are dynamically placed and removed by chromatin modifiers, and they impact the chromatin architecture and how the underlying genetic information is utilised. The precise effects of these modifications depend on their position, as well as the electrostatic and steric changes they impart (Zhou et al., 2011). For example, lysine acetylation can neutralise the positive charge on this residue and weaken the interaction between histones and the DNA (Grunstein, 1997). Therefore, acetylation of H3 on lysine 9 and 27 (H3K9ac, H3K27ac) is often associated with active gene regulatory elements which require an open chromatin structure that facilitates DNA access for transcription regulating factors (Ernst et al., 2011). In contrast, trimethylation of the same residues (H3K9me3, H3K27me3) marks different types of heterochromatin (Saksouk et al., 2015). H3K9me3 is typically found in highly condensed and stably silenced regions referred to as constitutive heterochromatin, whereas facultative heterochromatin marked by H3K27me3 is less condensed and more readily convertible into an active chromatin state (Mikkelsen et al., 2007; Rea et al., 2000). Histone PTMs also

impact the stability of the nucleosome core and the contacts between neighbouring nucleosomes (Lawrence et al., 2016). Ultimately, the changes in the chromatin architecture affect the accessibility of DNA, thus, representing an important mechanism that contributes to gene regulation.

In addition to histones, DNA itself can be modified by methylation on the C5 position of cytosine (5-methylcytosine, 5mC) (Hotchkiss, 1948). DNA methylation is a reversible epigenetic modification that is placed by DNA methyltransferases and removed by the concerted action of dioxygenases and DNA glycosylases (Wu and Zhang, 2014). Cytosine methylation is abundant in vertebrate genomes where it occurs predominantly in the context of CpG dinucleotides (Bird, 2002; Lister et al., 2009; Meissner et al., 2008; Stadler et al., 2011). CpGs are symmetrically methylated on both strands, and during DNA replication this allows re-establishment of the methylation patterns on the daughter strands which makes cytosine methylation an inheritable epigenetic mark (Holliday and Pugh, 1975; Riggs, 1975; Song et al., 2011). DNA methylation is functionally linked to stable gene silencing with important roles in development and disease (Bird and Wolffe, 1999). For example, it contributes to permanent silencing of the inactive X chromosome in female mammals, and to mono-allelic silencing of imprinted genes (Birger et al., 1999; Mohandas et al., 1981; Shemer et al., 2000; Tremblay et al., 1997; Venolia et al., 1982). Additionally, promoters of germline-specific and pluripotency genes are often methylated in somatic cells to stably repress their expression, and failure to achieve this can lead to ectopic expression and promote carcinogenesis (Borgel et al., 2010; Farthing et al., 2008; Meissner et al., 2008; Mohn et al., 2008; Shen et al., 2007). However, transcriptional inactivation precedes DNA methylation, indicating that DNA methylation does not initiate silencing, but rather contributes to its stable maintenance (Okamoto and Heard, 2009; Payer and Lee, 2008). Importantly, there is communication and interdependence between DNA methylation and histone modifications. H3K4me3 is a modification associated with active gene promoters that inhibits *de novo* methylation by DNA methyltransferases in mouse embryonic stem cells (mESCs) (Farthing et al., 2008; Guo et al., 2015; Ooi et al., 2007; Otani et al., 2009). Other histone modifications can promote DNA methylation. Actively transcribed gene bodies in mESCs are preferentially target by *de novo* methyltransferases when they are marked by H3K36me3 (Baubec et al., 2015; Morselli et al., 2015). While histone modifications such as H3K4me3 and H3K36me3 seem to influence DNA methyltransferase localisation and activity, in other cases, DNA methylation seems to help guide histone modifications. In mESCs, DNA methylation at silent promoters opposes deposition of H3K27me3, and DNA methylation can also outcompete existing H3K27me3 (King et al., 2016). Taken together, these examples indicate a reciprocal effect between DNA

methylation and histone modifications that is context-dependent. In conclusion, both DNA methylation and histone modifications are mediators of epigenetic regulation that act through modulation of the chromatin architecture which in turn can influence the binding of transcription regulating proteins.

1.1.3 Chromatin remodellers regulate DNA accessibility

Chromatin remodelling describes the process of changing the chromatin architecture without chemically modifying DNA or histones (Clapier et al., 2017). This process is mediated by chromatin remodellers which are ATP-dependent enzymes that control the accessibility of genomic DNA by altering the position and composition of nucleosomes. They represent central regulators of almost all chromatin-related processes, and their deregulation is often associated with gene expression defects that can lead to a variety of diseases (Alfert et al., 2019; Goodwin and Picketts, 2018). Based on their domain composition and sequence homology, eukaryotic chromatin remodellers are categorised into four families (Reyes et al., 2021): switch/sucrose non-fermentable (SWI/SNF), imitation switch (ISWI), chromodomain helicase DNA-binding (CHD), and inositol requiring 80 (INO80). All four families are highly conserved from yeast to human, but in higher eukaryotes each remodeller family diverges into multiple subtypes that can carry out cell type-specific or developmental state-specific functions (Clapier et al., 2017). The unifying features of the four families is their ATPase domain that catalyses the translocation of double stranded DNA on the histone octamer. However, they differ in the way they utilise their ATPase domain to perform specialised functions.

ISWI and CHD family remodellers primarily catalyse nucleosome assembly and periodicity, which is a parameter describing the (regular) spacing between nucleosomes (Clapier et al., 2017). These activities are essential, for instance, during DNA replication when ISWI and CHD complexes work in conjunction with histone chaperones to assemble H3/H4 and H2A/H2B pairs into canonical histone octamers, and then generate arrays of equally spaced nucleosomes (Corona et al., 1999; Fyodorov and Kadonaga, 2002; Lusser et al., 2005; Torigoe et al., 2011; Varga-Weisz et al., 1997; Yang et al., 2006). The distance between nucleosomes is regulated through a process called nucleosome sliding and this has a direct effect on chromatin accessibility as it can lead to compaction or relaxation of the chromatin structure. Consequently, nucleosome sliding can regulate transcription by providing or limiting access to transcription factor binding sites. For example, the ISWI family nucleosome remodelling factor (NURF) binds to promoter-associated H3K4me3 via

its BPTF subunit (Li et al., 2006). In *Xenopus* embryos it was shown that this interaction couples chromatin remodelling with *Hox* gene expression (Wysocka et al., 2006). A distinctive feature of CHD remodellers is that they do not form their own multi-subunit complexes like the other three chromatin remodeller families (Reyes et al., 2021). Instead, CHD remodellers associate with other protein complexes to regulate gene expression. Studies in *S. cerevisiae* have shown that Chd1 cooperates with PAF1, DSIF, and FACT complexes to facilitate transcription elongation (Krogan et al., 2002; Simic et al., 2003). Moreover, Chd1 is part of the SAGA histone acetyltransferase complex, where Chd1 enhances stable promoter occupancy of this complex through the interaction of its double chromodomain with H3K4me3 (Pray-Grant et al., 2005). However, CHD family members can also have gene repressing functions, for example as subunits of the metazoan nucleosome remodelling deacetylase (NuRD) complex (Hoffmeister et al., 2017; Low et al., 2016). Only in metazoans, CHD proteins also perform nucleosome editing functions, which describes the exchange of canonical histones for their variant paralogues (Clapier et al., 2017). CHD1 removes canonical H3 from nucleosomes and replaces it with variant H3.3, which is often associated with active gene promoters (Chow et al., 2005; Konev et al., 2007).

Nucleosome editing and sliding can also be performed by members of the INO80 chromatin remodeller family, however, in different biological processes and on different histone substrates than ISWI and CHD remodellers (Poli et al., 2017). The founding member of this family, INO80, catalyses the exchange of the variant H2A.Z/H2B dimer for canonical H2A/H2B, whereas another member of this remodeller family, the SRCAP complex, catalyses the reciprocal reaction, i.e. ejection of canonical H2A/H2B and incorporation of variant H2A.Z/H2B (Mizuguchi et al., 2004; Papamichos-Chronakis et al., 2011; Ruhl et al., 2006). H2A.Z-containing nucleosomes often flank active core promotors and H2A.Z together with variant H3.3 were reported to destabilise nucleosomes (Jin and Felsenfeld, 2007; Jin et al., 2009). This property may be important for regulating transcription initiation. Additionally, INO80 remodellers have been implicated in creating the nucleosome-depleted region at the transcription start site (TSS) of active promoters, and in positioning the first nucleosome (+1 nucleosome) of the open reading frame (Krietenstein et al., 2016; Ranjan et al., 2013; Yen et al., 2013). Aside from its contribution to transcription regulation, INO80 also acts in DNA repair and replication pathways. In yeast, INO80 removes nucleosomes around DNA double-strand breaks, and it can evict RNAPII from chromatin at sites where the transcription machinery collided with a replication fork (van Attikum et al., 2007; Lafon et al., 2015; Poli et al., 2016).

The SWI/SNF family of chromatin remodellers (known as BAF in mammals) catalyse nucleosome sliding, histone dimer eviction, and complete histone octamer ejection (Varga et al., 2021). They have the ability to specifically eject H2A/H2B dimers, but unlike the SRCAP complex which performs a similar reaction, SWI/SNF complexes do not replace it with variant histone dimers (Lorch et al., 2006; Yang et al., 2007b). Replication-independent nucleosome sliding mediated by SWI/SNF complexes results in irregularly spaced nucleosomes (Clapier et al., 2017). This is in contrast to ISWI/CHD remodellers, which create equally spaced nucleosome arrays. A distinctive feature of the SWI/SNF remodellers is the diversity of subunits that can be differentially combined to form a large variety of complex subtypes (Jungblut et al., 2020). Importantly, many subunits are expressed only in specific tissues or developmental states which provides additional targeting specificity and functional specialisation (Ho and Crabtree, 2010). Targeting of SWI/SNF complexes is mediated through a multitude of chromatin-interaction domains on various subunits. A particular complex composition can simultaneously contain several histone modification readers such as plant homeodomains, chromo- and bromodomains, as well as DNA-binding domains (Ho and Crabtree, 2010). The combination of these reader domains directs SWI/SNF complexes to specific gene promoters and enhancers where they can expose transcription factor bindings sites for activators or repressors by sliding and ejecting nucleosomes (Clapier et al., 2017). In this context, it was shown in mammalian cells that different transcription factor binding motifs are enriched at target sites of distinct SWI/SNF complexes (Gatchalian et al., 2018; Michel et al., 2018). Through their ability to modulate chromatin access for transcription factors, and their large heterogeneity in subunit composition for targeting specific promoters and enhancers in different developmental states, SWI/SNF complexes play critical roles in cell differentiation and specification (Alfert et al., 2019).

1.1.4 Chromatin features associated with gene promoters in vertebrates

Transcription typically begins at the gene promoter, which comprises the core promoter and the proximal promoter (Haberle and Stark, 2018). The core promoter is located approximately ±50 bp around the transcription start site (TSS) and it serves as the binding platform for RNAP and the general transcription factors (GTFs). The proximal promoter is ~250 bp upstream of the core promoter and it contains binding sites for DNA sequence-specific transcription factors (TFs) that modulate transcription together with transcriptional cofactors. Similar to proximal promoters, more distal DNA elements called enhancers recruit TFs and cofactors. Enhancers and associated factors can make contact with the

core promoter over large genomic distances in order to increase transcription activity (Banerji et al., 1981, 1983; Shlyueva et al., 2014). Hence, the coordinated effort of promoters, enhancers, RNAP, and auxiliary factors ensures precise execution of transcriptional programmes.

The core promoter has a crucial role in guiding the sequential assembly of the pre-initiation complex (PIC), which consists of RNAPII and six GTFs (TFIID, TFIIA, TFIIB, TFIIF, TFIIE, TFIIH) (Roeder, 1996). Specific DNA sequence elements known as core promoter motifs signal the location of the core promoter to recruit the components of the pre-initiation complex (Danino et al., 2015). These core promoter motifs surround the TSS at defined distances and facilitate the precise positioning of the PIC for transcription initiation. The first core promoter motif that was identified in eukaryotes is the TATA-box, which is estimated to occur at 10-20% of promoters in humans (Goldberg, 1979; Osman and Cramer, 2020; Yang et al., 2007a). The TATA-box is recognised by the TBP subunit of TFIID, which is a large multimeric complex (>1 MDa) with an additional 13-14 subunits known as TAFs (TBP-associated factors) (Bieniossek et al., 2013; Kim et al., 1993; Patikoglou et al., 1999). The subunit composition of TFIID is variable which enables the recognition of other core promoter motifs. For example, TAF1/2 bind the initiator (Inr) motif (Chalkley and Verrijzer, 1999), TAF1 can also bind the downstream core element (DCE) (Lee et al., 2005), and TAF6/9 recognise the downstream promoter element (DPE) (Shao et al., 2005). Aside from TFIID, the only other GTF known to bind a core promoter motif is TFIIB, which binds the TFIIB recognition element (BRE) (Deng and Roberts, 2005a; Lagrange et al., 1998a). In addition to signalling the location of the core promoter, certain motifs or motif combinations can also contribute to the selection of a specific promoter or promoter classes. The TCT motif, for instance, uniquely marks core promoters of most ribosomal protein genes and a few other protein genes involved in translation (Parry et al., 2010). Moreover, in *Drosophila* it was shown that the TCT motif recruits a special TFIID complex containing TRF2 (homolog of TBPL1 in vertebrates) instead of TBP (Wang et al., 2014). Thus, the TCT motif represents a key element of a customised RNAPII transcription system that selectively supports the translational machinery. As the TCT example illustrates, core promoter motifs can mediate gene-specific recruitment and assembly of the PIC, however, many promoters do not have any known motifs. One reason might be yet undiscovered motifs, but more importantly, there are other chromatin features that characterise gene promoters.

Genomic elements referred to as CpG islands (CGIs) critically shape the chromatin architecture at many gene promoters (described in more detail in section 1.2). Unlike the surrounding genome, CGIs feature a high GC content, a disproportionately high CG

dinucleotide density, and non-methylated cytosines (Bird et al., 1985; Gardiner-Garden and Frommer, 1987; Tykocinski and Max, 1984). These attributes help CGIs to stand out from the bulk genome and mark sites from which transcription can be potentially initiated (Deaton and Bird, 2011). Hence, CGIs are associated with the majority of gene promoters in mammals (Saxonov et al., 2006). Although the contribution of CGIs to promoter function is not fully understood yet, it is evident that their distinct DNA sequence features enables the recruitment and assembly of transcription regulating complexes. Non-methylated CpGs can be selectively recognised by proteins containing Zinc finger CXXC (ZF-CXXC) or certain types of winged helix (WH) domains (Long et al., 2013a; Perino et al., 2018; Stielow et al., 2021; Voo et al., 2000a). These proteins influence transcription at CGI promoters either by catalysing chromatin modifications (if a catalytic function is present), or by serving as recruitment modules for larger chromatin-modifying complexes. For example, Polycomb group (PcG) and Trithorax group (TrxG) proteins assemble into large chromatin-modifying complexes with antagonistic functions in transcription regulation (Schuettengruber et al., 2017). Both complexes contain distinctive subunits that can direct them to CGI promoters where they place histone modifications that are either permissive (TrxG) or repressive (PcG) for transcription initiation. This highlights how in addition to DNA sequence motifs different epigenetic modifications work in conjunction to provide another layer of regulation in transcription initiation.

Core promoters are commonly marked by distinct combinations of histone modifications that can signal the activity state of the associated gene (Haberle and Stark, 2018). Such signals can then recruit transcriptional regulators that help to maintain and modulate the active or the repressed gene expression state. For example, the combination of H3K4me3, H3K9ac, and H3K27ac is frequently associated with active core promotors (Dunham et al., 2012; Ernst et al., 2011; Xiao et al., 2012). Acetylation works through weakening the interaction between histones and the DNA, thereby promoting an open chromatin structure. Additionally, acetylation can provide binding sites for transcription cofactors, such as BRD4, which in turn recruits the P-TEFb complex to the core promoter (Moon et al., 2005; Yang et al., 2005). The P-TEFb complex includes the CDK9 kinase subunit which supports the transition from the initiation into the elongation phase of transcription by phosphorylating the C-terminal domain of RNAPII. H3K4me3 is catalysed by TrxG complexes (see 1.3) and is known to recruit the mammalian TFIID complex via its TAF3 subunit (Vermeulen et al., 2007). Moreover, H3K9ac and H3K14ac potentiate the interaction between TFIID and H3K4me3 (Vermeulen et al., 2007). Similarly, structural studies have shown that the mammalian TAF1 subunit of TFIID can recognise and bind acetylation on various positions of the histone H4 tail (Jacobson et al., 2000). Even though H3K4me3 can recruit TFIID, a

key component of the PIC, H3K4me3 on its own does not seem to have an instructive role in transcription initiation. Several studies from yeast to mammalian cells indicate that perturbation of H3K4me3 levels at core promoters has minimal effects on nascent transcript levels for the majority of genes (Clouaire et al., 2012, 2014; Margaritis et al., 2012). Yet, H3K4me3 and other histone modifications are components of chromatin signals that can specifically mark core promoters and modulate transcription through recruitment of TFs and cofactors.

In summary, transcription begins at the gene promoter which is characterised by distinct DNA sequence and chromatin features. These features can include core promoter motifs, CpG islands, histone modifications, nucleosome composition and occupancy. Together they form a complex regulation system that supports the assembly of the basal transcription machinery and ensures the precise initiation of transcriptional programmes.

1.1.5 The RNAPII transcription cycle

Transcription for all three metazoan RNAPs proceeds in three main stages. The process initiates with the recognition of the core promoter and assembly of the pre-initiation complex. This is followed by the elongation phase in which the DNA sequence is transcribed into a growing RNA transcript. Once the gene has been fully transcribed, the RNA transcript is released from RNAP, and transcription is terminated. During the entire process, RNAP is supported by a host of accessory factors that act at specific stages of the transcription process. In the following, key events of each stage will be described in more detail for RNAPII.

Transcription initiation is a stepwise process that begins with the recognition of the core promoter and sequential assembly of the pre-initiation complex (PIC) comprising RNAPII and six GTFs (TFIID, TFIIA, TFIIB, TFIIF, TFIIE, TFIIH) (He et al., 2013; Roeder, 1996). Since RNAPII cannot recognise the promoter by itself, TFIID plays a central role in identifying the core promoter and nucleating PIC assembly (Sainsbury et al., 2015). It is a ~1.3 MDa complex consisting of TATA box-binding protein (TBP) and 13-14 TBP-associated factors (TAFs) (Patel et al., 2020). Alternative TFIID complexes formed by variable subunit compositions can select specific subsets of promoters (Hochheimer and Tjian, 2003). Additionally, TFIID composition is also dependent on developmental stage and cell type. For example, human embryonic stem cells (hESCs) express only six TAFs and perturbation of this expression pattern causes misregulation of pluripotency genes and

induction of differentiation (Maston et al., 2012). Moreover, TFIID can be recruited to promoters through diverse chromatin features that include DNA sequence motifs, histone modifications, and other transcription activators (Chalkley and Verrijzer, 1999; Lee et al., 2005; Liu et al., 2009; Shao et al., 2005; Vermeulen et al., 2007). Although TFIID has been associated with the recruitment to TATA box-containing promoters through binding of its TBP subunit, TFIID components are found at virtually all promoters irrespective of the presence of a TATA box motif (Cianfrocco et al., 2013; Patikoglou et al., 1999; Pugh and Tjian, 1991). The omnipresence of TFIID at promoters, its adaptable subunit composition, and diverse mechanisms of recruitment make TFIID a major hub for regulating the first step of transcription initiation.

Once the central TBP subunit of TFIID is bound at the core promoter, the TBP-DNA complex is stabilised by TFIIA and TFIIB which flank the complex on either side (Geiger et al., 1996; Nikolov et al., 1995). TFIIB has multiple important functions in PIC formation (Sainsbury et al., 2013). It facilitates binding of TBP to DNA and interacts with regions flanking the TBP binding site that can contain BREs (Deng and Roberts, 2005b; Lagrange et al., 1998b). The grip of TFIIB surrounding TBP is thought to contribute to the oriented assembly of the PIC, and thus, to the directionality of transcription (Littlefield et al., 1999). Furthermore, TFIIB is required to recruit the TFIIF-RNAPII complex to the promoter (Ha et al., 1993; Mühlbacher et al., 2014; Rani et al., 2004). Opening of the promoter DNA is a key function of the PIC and this is mediated by TFIIE and TFIIH (Holstege et al., 1996). TFIIE binds the TFIIF-RNAPII complex on the promoter and then recruits TFIIH which is the last GTF to complete the PIC (Maxon et al., 1994). TFIIH is a 10-subunit complex with DNA helicase and protein kinase functions (Gibbons et al., 2012). The helicase function is required to melt the DNA double helix around the TSS which then forms an open transcription bubble (Guzmán and Lis, 1999). TFIIB helps to maintain the transcription bubble and to position the TSS inside the active centre of RNAPII (Kostrewa et al., 2009). Finally, the CDK7 kinase subunit of TFIIH phosphorylates serine 5 (Ser5) on the C-terminal domain (CTD) of RNAPII which triggers initial RNA synthesis and promoter escape (Feaver et al., 1991; Serizawa et al., 1995). The RNAP CTD has a crucial function in transcription regulation. In humans, it contains 52 repeats of the heptapeptide sequence $Y_1S_2P_3T_4S_5P_6S_7$ which is targeted by different kinases and phosphatases (Eick and Geyer, 2013). The phosphorylation pattern on the CTD changes during the transcription process and recruits various factors that support co-transcriptional events such as chromatin remodelling, histone modifications, and RNA processing (Cramer, 2019).

In addition to the six GTFs described above, the transcription initiation process is further regulated by distal TFs and cofactors. A key transcription regulator is the Mediator complex which comprises 30 subunits in humans (Soutourina, 2018). Mediator does not directly bind the promoter, but instead is recruited to distal enhancer elements which loop over long distances to make contacts with the transcription machinery at the promoter. An important function of Mediator is to transduce activating signals from enhancer-bound TFs to the PIC. Mediator activities support PIC recruitment and assembly, and Mediator also facilitates the transition into the elongation phase by stimulating TFIIH to phosphorylate Ser5 on the CTD of RNAPII (Esnault et al., 2008; Wong et al., 2014).

Promoter escape of RNAPII is accompanied by the release of the GTFs that formed the PIC and the concomitant recruitment of elongation factors. However, at many genes RNAPII pauses elongation ~50 bp downstream of the TSS, which is an event referred to as promoter-proximal pausing (Guenther et al., 2007; Rougvie and Lis, 1988). The exact trigger for promoter-proximal pausing is not known, but different catalysts have been proposed, including high GC content downstream of the TSS and certain promoter motifs such as the DPE or a Pause button (PB) motif (Core and Adelman, 2019). Although the cause of promoter-proximal pausing remains elusive, it is a highly regulated event that represents another mechanism to modulate the frequency of transcription initiation. Promoter-proximal pausing keeps RNAPII in a poised state that can be rapidly released into productive elongation (Gressel et al., 2017; Shao and Zeitlinger, 2017). This might be advantageous when fast and/or synchronous activation of gene expression is required. For example, promoter-proximal pausing in *Drosophila* embryos ensures synchronous activation of genes required for tissue morphogenesis, and in mESCs it coordinates genes involved in cell cycle progression (Lagha et al., 2013; Williams et al., 2015). Mechanistically, promoter-proximal pausing is mediated by DSIF and NELF that associate with paused RNAPII and the nascent transcript (Yamaguchi et al., 2013). Phosphorylation of DSIF, NELF, and Ser2 on RNAPII CTD by P-TEFb releases pausing and enables the transition into productive elongation (Marshall and Price, 1995).

In the elongation phase, RNAPII transcribes into the gene body and processively extends the precursor messenger RNA (pre-mRNA). However, nucleosomes pose obstacles to the passage of RNAPII through the gene body and they need to be bypassed to elongate the transcript (Bondarenko et al., 2006; Kireeva et al., 2005). At the same time, nucleosomes carry epigenetic information in the form of histone modifications that provide instructions for transcription regulation. Therefore, RNAPII needs to bypass nucleosomes while histone modifications need to be preserved or rewritten. This is mediated by chromatin remodellers,

histone chaperones and modifiers that travel along with RNAPII during transcription. Importantly, many of these factors are recruited by differential phosphorylation patterns of the RNAPII CTD. For example, the PAF1 complex plays a critical role in transcription elongation through its interactions with a wide range of cofactors (van Oss et al., 2017). The PAF1 complex binds RNAPII in a Ser5P-dependent manner and then recruits the E3 ubiquitin ligase complex RNF20/40 which monoubiquitylates H2B (H2Bub) in transcribed gene bodies (Ng et al., 2003; Pavri et al., 2006; Qiu et al., 2012). Both PAF1 and H2Bub are required to support the FACT complex, which is a histone chaperone that travels with RNAPII and partially disassembles and reassembles the nucleosomes by transiently ejecting the H2A/H2B dimer (Orphanides et al., 1998, 1999). This process considerably enhances transcription through chromatin (Fuchs et al., 2014). Furthermore, H2Bub is able to recruit the histone methyltransferase DOT1L which co-transcriptionally catalyses H3K79me2/3 in gene bodies (Ng et al., 2002). H3K79me2/3, H2Bub, and FACT activity influence the elongation rate which in turn affects other co-transcriptional processes such as alternative splicing (Ljungman et al., 2019). Moreover, elongating RNAPII exhibits Ser2P on its CTD and this modification is bound by the histone methyltransferase SETD2 that deposits H3K36me3 in gene bodies (Edmunds et al., 2008; Krogan et al., 2003). H3K36me3 is a docking site for the SIN3B histone deacetylase complex and the DNA methyltransferase DNMT3B (Baubec et al., 2015; Jelinic et al., 2011; Yang and Seto, 2008). Histone acetylation is associated with an open chromatin structure as it relaxes the wrapping of DNA around histones. This is beneficial in gene promoters where RNAPII needs to gain initial access to the DNA. In gene bodies, however, acetylation increases the risk for misguided transcription initiation, and therefore, the SIN3B complex removes acetylation in gene body regions (Jelinic et al., 2011). Similarly, nonmethylated DNA facilitates transcription initiation at gene promoters while methylated DNA regions are generally refractory to transcription initiation (Schübeler, 2015). Thus, hypoacetylation and DNA methylation mediated by the SIN3B complex and DNMT3B, respectively, suppress cryptic initiation events from within gene bodies. Interestingly, the MORF4L1 subunit of the SIN3B complex also plays an important role in splicing. It localises to gene bodies through specific recognition of both H3K36me3 and H2Bub and then recruits the PTB splicing factor that functions in alternative splice site selection (Luco et al., 2010; Wu et al., 2011b). Thus, MORF4L1 establishes a connection between histone modifications and alternative splicing. These examples illustrate how elongation is coupled to several co-transcriptional events that modulate the transcriptional output.

At the 3’ end of the gene, RNAPII passes over a polyadenylation site which triggers transcription termination (Beaudoing et al., 2000; Hu et al., 2005). This stage of the

transcription cycle is less well understood and there are two different, but not mutually exclusive models, to explain how transcription is terminated (Porrua and Libri, 2015). In either model, termination requires cleavage and polyadenylation of the mRNA, and again, the phosphorylation pattern on RNAPII CTD is critical in recruiting the factors that catalyse these processes (Eick and Geyer, 2013). Phosphorylated Ser2 and dephosphorylated Y1 are required to recruit and stimulate the termination complexes called cleavage and polyadenylation specificity factor (CPSF), cleavage stimulatory factor (CstF), cleavage factor I/II (CFI/II), and poly(A) polymerase (PAP) (Christofori and Keller, 1988; Gilmartin and Nevins, 1989; Takagaki et al., 1989). These complexes release the mRNA from RNAPII through cleavage, and then catalyse polyadenylation of the mRNA 3' end. Soon after passing the polyadenylation site, RNAPII dissociates from DNA either through a conformational change triggered by the mRNA 3' processing enzymes (allosteric model), and/or through displacement by the RNase XRN2 (torpedo model) (Porrua and Libri, 2015).

In summary, initiation, promoter-proximal pausing, elongation, and termination represent the key stages of the transcription cycle. All stages need to be precisely coordinated in order to achieve appropriate levels of gene expression. This involves several co-transcriptional events that reciprocally influence each other. The regulation in the elongation and termination stages is important to fine-tune gene expression, for example, to produce specific isoforms required at distinct developmental stages. However, the commitment for gene expression is made at the initiation stage and elucidating the mechanisms that lead to this commitment is central to understanding the basic principles of gene regulation.

1.2 The role of CpG islands in gene regulation

1.2.1 Identification of CpG islands

CpG dinucleotides in vertebrate genomes exhibit two striking features compared to other dinucleotide combinations: the bulk of CpGs are methylated on the cytosine residue (5mC), and they are statistically underrepresented. In the human genome, for instance, CpGs occur at a frequency of ~0.008 whereas one would expect a frequency of 0.04 by simply multiplying the fraction of cytosines and guanines (0.21 x 0.21) (Lander et al., 2001; Saxonov et al., 2006). The statistical deficit of CpGs is thought to be caused by spontaneous deamination of methylcytosine to thymine over the course of vertebrate evolution (Bird, 1980; Coulondre et al., 1978; Duncan and Miller, 1980). Leads about an implication of CpG methylation in genomic function accumulated when several studies revealed that CpG methylation was not uniformly distributed in vertebrate genomes. These studies suggested that methylated CpGs were enriched in heterochromatin, repeats, as well as unique sequences whereas non-methylated CpGs appeared to be localised predominantly in euchromatin and transcribed genes (Ehrlich et al., 1982; Naveh-Many and Cedar, 1981; Razin and Cedar, 1977; Solage and Cedar, 1978). A new light was shed on the distribution of CpG methylation with the use of methylation-sensitive isoschizomers of restriction enzymes (Bird and Southern, 1978). For example, both HpaII and MspI recognise the sequence 5'-CCGG, but unlike MspI, HpaII cannot efficiently cleave if the central cytosine is methylated. Using this approach initially in various chicken tissues revealed clusters of non-methylated CpGs that were localised at the 5' end of genes and amounted to ~1% of the genome (Cooper et al., 1983). These clusters enriched in non-methylated CpGs were initially referred to as HpaII tiny fragments (HTFs) and became later known as CpG islands (CGIs) (Bird, 1987).

Further studies revealed that CGIs seemed to be a common feature of vertebrate genomes including those of mouse and human. For example, a study of mouse tissues echoed previous findings in chicken by showing that CGI fragments had an average length of 1 kb, represented ~1% of the genome, and co-localised with promoters. Moreover, the results indicated the presence of ~30,000 CGIs, most of which seemed to have unique sequences (Bird et al., 1985). Furthermore, a broad investigation of mammalian, avian, and amphibian genomes indicated that CGIs were associated with the promoters of all investigated housekeeping genes and many tissue-specific genes (Gardiner-Garden and Frommer, 1987). The authors also showed that CGIs occupied the same position relative to orthologous genes from different species, which implied an evolutionary conserved function of CGIs. This study proposed for the first time a definition of CGIs based on DNA sequence

features. According to that, CGIs have a G+C content greater than 50%, an observed/expected CpG ratio of more than 0.6, and a minimum length of 200 bp. Although other definitions have been proposed since then (for example G+C ≥ 55%, Obs/Exp CpG > 0.65, >500 bp by Takai and Jones, 2002), the original definition put forward by Gardiner-Garden and Frommer is still commonly used today for bioinformatic prediction of CGIs. For example, this definition was used in the initial sequencing and analysis of the human and mouse genomes and resulted in the prediction of ~29,000 CGIs in human compared to ~15,500 CGIs in mouse (Lander et al., 2001; Waterston et al., 2002).

A major shortcoming of CGI prediction algorithms is that they cannot account for the absence of cytosine methylation, which is a defining feature of CGIs. Consequently, prediction algorithms cannot differentiate between true CGIs and GC-rich repetitive elements such as transposons, which make up a substantial proportion of vertebrate genomes (Lander et al., 2001). The prediction of ~29,000 CGIs in the human genome was derived after eliminating 21,000 repeat elements that also satisfied the CGI definition by Gardiner-Garden and Frommer. Moreover, this CGI definition is based on the human and mouse genome compositions and works rather poorly for predicting CGIs in other vertebrates. The reason is most likely the variation in G+C content and CpG density between different vertebrate genomes (Long et al., 2013b).

Aside from prediction algorithms that are based on somewhat arbitrary definitions, CGIs can be identified empirically using biochemical approaches. Two related approaches that were developed more recently (termed as CAP and BioCAP) provide an unbiased methodology for the identification of CGIs using Zinc finger CXXC (ZF-CXXC) protein domains which specifically recognise and bind non-methylated CpGs (Blackledge et al., 2012; Illingworth et al., 2008). In both methods, the ZF-CXXC domains are immobilised on an affinity matrix to capture non-methylated islands (NMIs) from purified genomic DNA. Enriched NMI fragments are then subjected to massively parallel sequencing and their locations are mapped to the genome. BioCAP was used to compare NMIs in the genomes of seven diverse vertebrate species ranging from fish to human (Long et al., 2013b). The comparison revealed that the NMI nucleotide composition varied substantially between species, which compromised bioinformatic CGI prediction in non-mammalian species. As a consequence, orthologous NMIs determined by BioCAP were far more conserved than algorithmic CGI prediction, indicating that NMIs may have conserved biological relevance. Thus, biochemical approaches provide a more direct and reliable way to identify CGIs genome-wide because they can account for the absence of CpG methylation and are agnostic to the overall nucleotide composition of the genome.

1.2.2 Readers of non-methylated CpGs

CpG methylation and the absence thereof may have an epigenetic function in regulating gene expression. This function is mediated through proteins with specialised domains that can read the methylation state of CpGs. The focus here is on domains that selectively recognise and bind non-methylated CpGs. In 2000, CFP1 (also known as CXXC1) was characterised as the first protein with the ability to selectively bind non-methylated CpGs in humans (Voo et al., 2000a). This ability relied on a zinc finger CXXC (ZF-CXXC) domain, which at that time was known to be present also in DNMT1, MLL1, and MBD1. The demonstration that the ZF-CXXC domain was responsible for non-methylated CpG binding prompted the search for further proteins with ZF-CXXC domains. This led to the identification of a total of 12 ZF-CXXC-containing proteins in mammals (Long et al., 2013a).

ZF-CXXC domains comprise two C4-type zinc fingers separated by a variable linker sequence (Xu et al., 2018). Based on the similarity of their linker regions, ZF-CXXC domains can be categorised into three groups (Long et al., 2013a). The eight conserved cysteine residues that make up the zinc fingers are shared between all three groups and mutational analysis in CFP1 has shown that the cysteines are required for DNA binding (Lee et al., 2001). The linker regions of group 1 proteins, which comprise KDM2A/B, FBXL19, CFP1, DNMT1, and MLL1/2, have the same length and contain two conserved motifs composed of KFGG and the dipeptide KQ, respectively. Proteins in group 3, which contains TET1/3, IDAX, and CXXC5, have shorter linker sequences without a complete KFGG motif. However, they do have a conserved HQ dipeptide motif, which resembles the KQ dipeptide motif in that a basic residue is followed by glutamine. MBD1 is the only member of group 2, but it has three variable ZF-CXXC domains. The third ZF-CXXC domain of MBD1 has almost identical linker region properties as group 1 proteins, i.e. a KFGG and a RQ motif. In contrast, the first and second ZF-CXXC domains do not have these motifs and structural studies indicate that both domains do not have the capacity to bind non-methylated CpGs (Xu et al., 2018). Indeed, several structures of DNA-bound ZF-CXXC domains have shown that the dipeptide motif (KQ, RQ, or HQ) is responsible for recognising non-methylated CpGs and point mutations in this motif abrogate DNA binding (Allen et al., 2006a; Cierpicki et al., 2010; Xu et al., 2011). Moreover, structures of group 1 proteins revealed that the ZF-CXXC domain clamps around the DNA and makes contact with both the minor and the major groove (Xu et al., 2018). *In vivo* studies of KDM2A suggested that group 1 ZF-CXXC domains preferentially bind the linker DNA between nucleosomes because the histone octamer would obstruct simultaneous access to minor and major groove (Zhou et al., 2012).

More recently, it was shown that certain types of winged-helix domains also have the ability to bind non-methylated CpGs. This ability was first demonstrated for the winged-helix domains of Polycomb-like proteins 1-3 (PCL1-3) from mouse and human (Choi et al., 2017; Li et al., 2017; Perino et al., 2018). In contrast to ZF-CXXC domains, the PCL winged-helix domains do not form a clamp around the DNA and therefore do not require simultaneous access to the minor and major groove. This could imply that they are able to bind nucleosomal CpGs. All three PCL proteins have a conserved IKK(R/K)K motif that mediates CpG binding, with the first two lysine residues being responsible for recognising the CpG dinucleotide (Li et al., 2017). Moreover, the structures of the PCL proteins have indicated that a non-methylated cytosine is essential but not sufficient for binding (Perino et al., 2018). It seems that the helical shape of CG-rich sequences plays an equally important role for DNA binding. SAMD1 is another protein with a winged-helix domain that recognises non-methylated CpGs, albeit with a different DNA binding mode from that shown for PCL proteins (Stielow et al., 2021). The SAMD1 winged-helix domain contacts both the minor and major groove and mutations in the residues that mediate these contacts reduce DNA binding affinity. The recognition of the non-methylated CpG is again carried out by a basic RK dipeptide motif. As new CGI-binding proteins are characterised, it is likely that other domains with the ability to bind non-methylated CpGs will emerge in the future.

Many proteins with the ability to selectively bind non-methylated CpGs either have an inherent chromatin-modifying property or they serve as targeting modules in chromatin-modifying and transcription-regulating complexes. H3K4 methyltransferase complexes have crucial functions as transcription activators that catalyse H3K4me3 at promoters, which is a hallmark of transcriptional activation (Shilatifard, 2012). For example, CFP1 does not have any known catalytic function of its own, but it is part of the H3K4 methyltransferase complex SETD1A/B and contributes via its ZF-CXXC domain to the recruitment of the complex to CGIs (Brown et al., 2017). MLL1/2 form the core subunits of another H3K4 methyltransferase complex but in this case the recruitment function of the ZF-CXXC domain and the catalytic activity are both united in MLL1 and MLL2 (Allen et al., 2006b; Bach et al., 2009; Guenther et al., 2005; Milne et al., 2005). Non-methylated CpG-binding proteins can be also constituents of transcription repressing complexes such as Polycomb repressive complex 1 and 2 (PRC1/2). The PCL1-3 proteins mentioned above have been shown to recruit PRC2 to CGIs where this complex places the repressive mark H3K27me3 (Brien et al., 2012; Choi et al., 2017; Li et al., 2017; Loh et al., 2021; Perino et al., 2018). Similarly, PRC1 promotes transcription repression through deposition of H2AK119ub at gene promoters (Boyer et al., 2006; de Napoles et al., 2004). A particular PRC1 variant is recruited to CGI promoters by the ZF-CXXC protein KDM2B (Farcas et al., 2012). Apart

from its targeting function as a component of PRC1, KDM2B is also able to demethylate H3K36me3 which is absent at active promoters but enriched in transcribed gene bodies (Bannister et al., 2005; He et al., 2008). In summary, non-methylated CpG-binding proteins play important roles in transcription regulation through their enzymatic activities and/or targeting function for chromatin-modifying complexes.

1.2.3 Writers and erasers of CpG methylation

As described in section 1.1.2, DNA methylation on the C5 position of cytosine is an epigenetic modification that is essential in mammalian development and often associated with stable repression of gene promoters. During mammalian embryogenesis, the genome is subject to global DNA demethylation and re-establishment of new methylation patterns (Hajkova et al., 2002; Mayer et al., 2000; Oswald et al., 2000). In pre-implantation embryos and primordial germ cells, global DNA demethylation contributes to setting up pluripotent states and regulating genomic imprinting. DNA methylation is re-established in post-implantation stages and somatic cells where it is required for stable gene silencing and maintaining genome stability. Once established, DNA methylation patterns are relatively static in somatic cells and only change locally to activate or repress specific genomic processes (Smith and Meissner, 2013). In contrast to the bulk genome, CGIs mostly remain methylation-free irrespective of the developmental stage and cell type (Borgel et al., 2010; Deaton and Bird, 2011; Illingworth et al., 2010; Long et al., 2013b). These differential methylation patterns are established by proteins that have the catalytic ability to write and erase 5-methylcytosine.

Methylation of DNA is carried out by a family of four DNA methyltransferases (DNMTs) in mammals: DNMT1, DNMT3A/B, and DNMT3L (Moore et al., 2013). Although these enzymes share sequence homology and structural similarities, they have distinct functions and expression patterns (Xie et al., 1999; Yen et al., 1992). DNMT1 is the only member of this protein family with a functional ZF-CXXC domain, however, DNMT1 is not localised at CGIs (Jin et al., 2012). Structural studies with truncated DNMT1 versions suggested that the ZF-CXXC domain has an autoinhibitory function preventing non-methylated DNA from being methylated and promoting methylation of hemimethylated DNA (Song et al., 2011). However, this could not be confirmed *in vitro* and *in vivo* with full-length DNMT1 (Bashtrykov et al., 2012; Frauer et al., 2011). Therefore, the contribution of the ZF-CXXC domain to the molecular function of DNMT1 remains elusive.

DNMT1 is constitutively expressed but the highest expression is observed in the S phase of dividing cells as it is essential for faithful propagation of the methylation pattern during DNA replication (Kishikawa et al., 2003; Petryk et al., 2021). DNMT1 is recruited to the replication fork through interaction with its obligate partner UHRF1, and through interaction with the DNA clamp protein PCNA (Bostick et al., 2007; Chuang et al., 1997; Sharif et al., 2007). During semiconservative DNA replication, the DNA double strand separates, and each parental strand serves as a template for the synthesis of a new daughter strand. Consequently, semiconservative DNA replication creates hemimethylated DNA with methylated cytosines being retained on the parental strands and unmodified cytosines on the newly synthesised daughter strands. The DNMT1-UHRF1 dimer specifically recognises and binds the parental strand of hemimethylated DNA and then re-establishes the original methylation pattern on the daughter strand (Arita et al., 2008; Avvakumov et al., 2008). Additionally, there is evidence that DNMT1-UHRF1-PCNA also work together to re-establish DNA methylation upon DNA repair (Mortusewicz et al., 2005). Since DNMT1 maintains existing methylation patterns, it is referred to as the maintenance DNMT.

In contrast to DNMT1, the DNMT3 enzymes are referred to as *de novo* methyltransferases because they can establish DNA methylation in completely unmodified DNA (Okano et al., 1999). What distinguishes the DNMT3 enzymes is their expression pattern and functional specialisation. DNMT3A is ubiquitously expressed, whereas DNMT3B and DNMT3L are primarily expressed in early development, with the exception of adult thymus and germ cells (Aapola et al., 2000; Okano et al., 1999; Xie et al., 1999). DNMT3A/B are structurally very similar, and both have the capacity to methylate DNA (Cheng and Blumenthal, 2008; Xu et al., 2020). A DNA methyltransferase domain is absent in DNMT3L and therefore it does not have any catalytic function of its own (Aapola et al., 2000; Jia et al., 2007). Instead, DNMT3L mainly supports the catalytic activity of DNMT3A/B (Hata et al., 2002; Suetake et al., 2004). All three DNMT3 enzymes do not have a strong sequence preference aside from CpG dinucleotides (Dodge et al., 2002). They are recruited to target sites primarily via histone modifications, transcription factors, and chromatin remodellers (Smith and Meissner, 2013). For example, DNMT3A/B and DNMT3L preferentially bind unmodified H3K4 and are largely excluded from genomic regions marked by H3K4me3 (Guo et al., 2015; Ooi et al., 2007; Otani et al., 2009). In the case of DNMT3A structural studies have shown that only unmodified H3K4 is able to efficiently relieve an autoinhibitory conformation (Guo et al., 2015). Consistent with the transcription inhibitory role of cytosine methylation at gene promoters, DNMT3A/B are also recruited by epigenetic repressors such as histone deacetylases and H3K9 methyltransferases (Epsztejn-Litman et al., 2008; Lehnertz et al., 2003; Myant and Stancheva, 2008; Myant et al., 2011; Zhu et al., 2006). This supports a

model in which gene repression and heterochromatin formation are initiated before DNA methylation ensures stable silencing (Smith and Meissner, 2013).

The prevailing model of DNA demethylation suggests that 5-methylcytosine can be actively removed through an enzymatic cascade that involves the concerted action of 5-methylcytosine dioxygenases and DNA glycosylases (He et al., 2011; Ito et al., 2011; Kriaucionis and Heintz, 2009; Maiti and Drohat, 2011; Tahiliani et al., 2009). Mammalian genomes encode for three 5-methylcytosine dioxygenases, TET1-3, all of which are catalytically active and perform the same enzymatic reaction on their 5-methylcytosine substrate (Ito et al., 2010). They mainly differ in their expression pattern and recruitment mechanisms. High expression levels of TET1 are specifically detected in murine ESCs, blastocyst stage, and primordial germ cells (PGCs). TET2 is expressed in both early development and somatic cells, and TET3 is the dominant TET protein in most adult mouse tissues (Ito et al., 2010; Szwagierczak et al., 2010; Yamaguchi et al., 2012). These expression patterns are likely regulated by developmental stage-specific transcription factors. For instance, the proximal promoter of TET1 contains many binding motifs for core pluripotency transcription factors, which could explain the high expression of TET1 observed in mESC but not in differentiated tissues (Ficz et al., 2011). Both TET1 and TET3 have a ZF-CXXC domain that can specifically bind non-methylated CpG dinucleotides (An et al., 2017; Long et al., 2013a). In mESCs, TET1 occupancy is detected genome-wide but with a strong preference for CGI promoters that are in an active or poised state (Williams et al., 2011; Wu et al., 2011a). Similarly, TET3 preferentially binds gene promoters in Xenopus embryos, however, TET3 seems to target unmethylated cytosines also in a non-CpG context (Xu et al., 2012). Although TET2 does not have a ZF-CXXC domain, structural studies indicate that other Zinc finger type domains in its N-terminus mediate binding to the DNA backbone and CpG dinucleotides (Hu et al., 2013c, 2015). The genomic occupancy of TET2 seems to differ from TET1 and TET3 as genome-wide profiling has located TET2 at CGI promoters as well as enhancers (Chen et al., 2013; Deplus et al., 2013; Peng et al., 2016; Rasmussen et al., 2019; Xiong et al., 2016).

In the first phase of DNA demethylation, TET enzymes iteratively oxidise 5-methylcytosine (5mC) first to 5-hydroxymethylcytosine (5hmC) and then further to 5-formylcytosine (5fC) and 5-carboxycytosine (5caC) (Ito et al., 2011). Subsequently, thymine DNA glycosylase (TDG) recognises 5fC:G and 5caC:G base pairs and specifically excises the 5fC and 5caC bases (He et al., 2011; Maiti and Drohat, 2011). *In vitro* studies have shown that TDG efficiently excises 5fC and 5caC but has poor activity towards removal of 5mC and 5hmC (He et al., 2011; Maiti and Drohat, 2011). This could explain the requirement for iterative

oxidation of 5mC. After excision of the oxidised 5mC bases, the gap is repaired via the base excision repair pathway which results in the restoration of the unmodified cytosine base (Wu and Zhang, 2014).

Both DNMTs and TETs have essential roles in gene expression through antagonistic regulation of DNA methylation profiles. However, their essential roles are not necessarily obvious under developmental steady-state conditions but become evident when gene expression needs to direct cell differentiation programmes. For example, mESCs with triple knockout of TETs (Dawlaty et al., 2013), or triple knockout of the three catalytically active DNMTs (Tsumura et al., 2006) are viable, display normal ES cell morphology, and maintain pluripotency. Although the cells seemingly retain their stem cell properties, differentiation is severely compromised, and embryos die before birth. This underlines that the balanced interplay between DNMTs and TETs is crucial for dynamically adapting DNA methylation in space and time in order to support transcriptional programmes (Horii and Hatada, 2016).

1.2.4 CpG islands are associated with gene regulatory elements

Early studies on CGIs showed that they co-localise with gene promoters and that DNA methylation in this region had a repressive effect on transcription (Bird et al., 1985; Busslinger et al., 1983; Cooper et al., 1983; Kruczek and Doerfler, 1982; Stein et al., 1983; Tykocinski and Max, 1984). Their co-occurrence with gene promoters let to the notion that CGIs might serve as gene markers (Bird, 1987). Indeed, a study from 1992 demonstrated that CGIs were associated with all interrogated human housekeeping genes and 40% of the genes with a tissue-specific or limited expression pattern (Larsen et al., 1992). Advancements in sequencing technologies allowed a more comprehensive investigation of the human genome and provided new insights into CGI distribution. Using these new technologies revealed that 72% of annotated promoters in the human genome were associated with CGIs, which included virtually all housekeeping genes and many tissue-specific genes (Saxonov et al., 2006). Furthermore, several studies have investigated the distribution of CGIs in relation to gene promoters and other genomic elements in different cell types and species. This revealed that CGIs are a conserved feature of vertebrate genomes, which supports their central role in gene regulation. More than 90% of human CGIs localise at orthologous gene promoters in diverse vertebrate species from mouse to fish (Long et al., 2013b). The annotated gene promoters that are associated with CGIs range from 55% in fish to 72% in human. Conversely, roughly half of the CGIs in mouse and human are located around TSSs of known genes and the other half is evenly split

between intra- and intergenic CGIs (Illingworth et al., 2010). Although intra- and intergenic CGIs are distal to annotated gene promoters, 40-60% show signs of promoter activity as evidenced by the presence of H3K4me3, RNAPII, nascent transcripts and gene prediction annotations (Illingworth et al., 2010). Indeed, several studies have shown that intra- and intergenic CGIs co-localise with alternative TSSs of protein isoforms and TSSs of non-coding RNAs. For example, transcription of PAX6 and PARP12 protein isoforms is initiated from intragenic CGIs (Kleinjan et al., 2004; Rauch et al., 2009). Similarly, many long non-coding RNAs (lncRNAs) such as *Xist*, *Tsix*, and *Hotair* are transcribed from intergenic CGIs (Herzing et al., 1997; Lee et al., 1999; Panning and Jaenisch, 1996; Rinn et al., 2007). Aside from serving as sites of transcription initiation, some promoter-distal CGIs might also function as enhancers (Bell and Vertino, 2017; Bogdanović et al., 2016; Pachano et al., 2021). For example, in mESCs 17% of intergenic CGIs were shown to carry the enhancer-associated histone modification H3K4me1 (Long et al., 2013b).

On average, intra- and intergenic CGIs are shorter and have a slightly lower CpG density than promoter CGIs (Illingworth et al., 2010; Jeziorska et al., 2017; Long et al., 2013b). Moreover, intra- and intergenic CGIs become more frequently methylated than promoter CGIs during mammalian development. In human and mouse somatic cells, methylation affects 13-26% of intra- and intergenic CGIs but less than 3% of promoter CGIs (Illingworth et al., 2010; Maunakea et al., 2010). Importantly, DNA methylation is accompanied by a loss of H3K4me3 and RNAPII occupancy and correlates with transcriptional silencing. For intragenic CGIs, which are twice as likely as intergenic CGIs to become methylated, several studies have demonstrated that differential methylation regulates the expression of cell type-specific transcripts (Illingworth et al., 2010; Long et al., 2013b; Maunakea et al., 2010). The observation that intragenic CGIs are predominantly found in genes associated with organ development and cell lineage control, but largely absent from housekeeping genes, supports the idea that differential methylation of intragenic CGIs mediates cell differentiation processes (Choi et al., 2020; Deaton et al., 2011; Jeziorska et al., 2017; Lee et al., 2017).

A long-standing question in CGI biology is how the majority of CGIs remain methylation-free in most cell types and developmental states while the rest of the genome is heavily methylated. Additionally, it is unclear how differential methylation of a subset of CGIs is regulated during development. A key piece in this puzzle seem to be histone modifications, in particular H3K4me3. In mouse and human ESCs, more than 90% of annotated promoter CGIs are enriched in H3K4me3 (Illingworth et al., 2010). For intra- and intergenic CGIs this number drops to 40%, which correlates with the more frequent DNA methylation observed in promoter-distal CGIs. Conversely, 75-85% of H3K4me3 loci in human and mouse map

to CGIs (Illingworth et al., 2010). CGIs and H3K4me3 not only correlate in terms of co-localisation but also with regard to their density and abundance. The higher the CpG density the more abundant is the modification of nucleosomes with H3K4me3 (Illingworth et al., 2010). A causal relationship between H3K4me3 and the absence of CpG methylation is potentially provided by structural and functional studies of *de novo* DNMTs. H3K4me3 can block DNA methylation because *de novo* DNMTs require an unmodified H3K4 residue in order to efficiently methylate CpGs (Guo et al., 2015; Ooi et al., 2007; Otani et al., 2009). However, non-methylated CpGs do represent *bona fide* substrates for *de novo* DNMTs and low-efficiency catalysis could gradually lead to aberrant DNA methylation at CGIs. TET enzymes, which are mostly localised at CGIs, could counteract this process and help to keep CGIs in a methylation-free state (Long et al., 2013a).

But how is H3K4me3 deposited at CGIs in the first place? ZF-CXXC domain-containing proteins could represent a potential link between non-methylated CpGs and H3K4me3. SETD1A/B and MLL1/2 H3K4 methyltransferase complexes both have subunits with ZF-CXXC domains that can target them to CGIs (Long et al., 2013a). The SETD1A/B complex, which catalyses the bulk of H3K4me3, is recruited to CGIs through its CFP1 subunit (Ardehali et al., 2011; Thomson et al., 2010; Wu et al., 2008). Accordingly, depletion of CFP1 was shown to reduce H3K4me3 levels at CGI promoters. Moreover, an artificial promoterless CpG-rich sequence was sufficient to recruit CFP1 and generate new H3K4me3 peaks *in vivo* (Thomson et al., 2010). However, other studies have shown that loss of CFP1 in mESCs causes a reduction of H3K4me3 only at a subset of promoters, notably those with the highest expression levels (Clouaire et al., 2012). Furthermore, depletion of H3K4me3 at active CGIs in CFP1 null mESCs did not lead to increased DNA methylation and it did not significantly affect gene expression from these promoters (Clouaire et al., 2014). Surprisingly, rescue of the CFP1 null mESCs with a DNA-binding deficient CFP1 version restored nearly normal H3K4me3 levels at promoters where it was previously lost. However, CFP1 DNA-binding activity was required to prevent aberrant accumulation of H3K4me3 outside of CGI promoters. Comprehensive investigation of the interrelation between CFP1, H3K4me3, and DNA methylation in these studies indicates that the ability of H3K4 methyltransferases to target non-methylated CpGs is not sufficient to ensure correct placement of H3K4me3. The transcription process itself seems to play a role in regulating H3K4me3 levels at promoters (Clouaire et al., 2014), but it remains unclear how this is mechanistically achieved, and which proteins are involved.

Differential methylation during mammalian development affects predominantly intragenic CGIs because they are mostly associated with tissue-specific transcripts (Deaton et al.,

2011; Illingworth et al., 2008, 2010; Maunakea et al., 2010). Recent studies provided insights on how differential methylation at intragenic CGIs might be regulated. Transcription initiation from promoter CGIs seems to suppress transcription from downstream located intragenic CGIs (Lee et al., 2017). During transcription, SETD2 travels with elongating RNAPII and co-transcriptionally places H3K36me3 in gene bodies (Edmunds et al., 2008; Krogan et al., 2003). In the absence of H3K4me3, DNMT3B can bind H3K36me3 and methylate CpGs throughout the gene body, including intragenic CGIs located in this region (Baubec et al., 2015; Morselli et al., 2015; Neri et al., 2017). Thus, transcription traversing an intragenic CGI can lead to its silencing (Neri et al., 2017). However, this is not true for all intragenic CGIs. If the intragenic CGI itself is a source of high transcription initiation levels, it remains H3K4me3-positive which prevents its silencing through methylation (Jeziorska et al., 2017). Moreover, it was reported that intragenic CGIs are enriched in tissue-specific transcription factor binding sites that facilitate promoter activity from the intragenic CGI during cell differentiation (Lee et al., 2017). An intragenic CGIs that acts as typical promoter seems to suppress transcription from an upstream promoter CGI. Therefore, it was suggested that different CGIs within the same transcription unit compete with each other for initiating transcription (Jeziorska et al., 2017; Lee et al., 2017). Unlike intragenic CGIs, promoter CGIs are rarely silenced by methylation but through histone modifications such as H2AK119ub and H3K27me3 which are placed by Polycomb repressive complexes. However, the exact mechanisms of these processes are not well understood, in part because we do not have a method to comprehensively capture the proteins that bind to CGIs and how their localisation changes during development.

1.3 Antagonism of Polycomb and Trithorax groups in transcription regulation

Polycomb group (PcG) and Trithorax group (TrxG) proteins are large protein families of conserved epigenetic regulators that have antagonistic functions in gene expression (Schuettengruber et al., 2017). The balance between transcription-activating TrxG proteins and transcription-repressing PcG proteins is critical to normal developmental processes, and its disruption can cause a range of pathologies including cancer (Poynter and Kadoch, 2016). Both protein groups were initially discovered in *Drosophila* as regulators of *Hox* genes, which is a group of related genes that play an essential role in body plan specification along the anterior-posterior axis in Bilateria (Ingham, 1983; Ingham and Whittle, 1980; Lewis, 1978; Pearson et al., 2005). Accordingly, mutations of PcG and TrxG proteins cause prominent phenotypes of segmental transformations in *Drosophila*. Later it was found that the antagonistic relationship of PcG and TrxG proteins is not limited to *Hox* genes but represents a general mechanism of controlling gene expression throughout development (Holdeman et al., 1998; Lee and Skalnik, 2005; Robert et al., 1996). Subsequent studies have identified mammalian orthologs and have shown that PcG and TrxG proteins branched into additional subcomplexes with more specialised roles in chromatin-based regulation of transcription in vertebrate development.

In mammals, the Polycomb group comprises ~40 proteins that belong to one of two large multiprotein complexes with distinct catalytic activities related to the post-translational modification of histones (Piunti and Shilatifard, 2021). Polycomb repressive complex 1 (PRC1) monoubiquitylates H2A at lysine 119 (H2AK119ub), whereas PRC2 is able to catalyse all three methylation states of lysine 27 on H3 (H3K27me1/2/3) (Cao et al., 2002; Kuzmichev et al., 2002; Müller et al., 2002; de Napoles et al., 2004; Wang et al., 2004). Importantly, PRC1 and PRC2 cooperate in mediating gene repression by localising mostly to the same genomic sites where they influence each other's catalytic activities (Blackledge et al., 2014; Boyer et al., 2006; Bracken et al., 2006; Morey et al., 2012, 2013; Tavares et al., 2012). Sites at which PRC1, PRC2, and their catalysis products colocalise are referred to as Polycomb domains, and most of these domains are found at CGIs (Ku et al., 2008; Mikkelsen et al., 2007).

The founding member of the Trithorax group is the Drosophila *trithorax* (*trx*) gene which has a SET histone methyltransferase domain that was first characterised in *Drosophila* proteins **S**u(var)3-9, **E**nhancer-of-zeste, and **T**rithorax (Herz et al., 2013; Ingham and Whittle, 1980). The Trithorax group was later extended to include homologs of *trx*, subunits of TrxG complexes, and other proteins that counteract PcG-mediated transcription inhibition (Cenik and Shilatifard, 2021). As a result, the present-day Trithorax group comprises two

subgroups that exert completely different enzymatic functions to promote an active chromatin state. The COMPASS (complex of proteins associated with Set1) family of proteins are histone methyltransferases that catalyse mono-, di-, and trimethylation of H3K4, which are modifications associated with active enhancers and promoters (Cenik and Shilatifard, 2021). The second subgroup contains the SWI/SNF family of ATP-dependent chromatin remodellers that influence transcription through modulation of chromatin accessibility (Varga et al., 2021). Both COMPASS and SWI/SNF have in common that they counteract PcG activities, for example by targeting similar binding sites and replacing repressive histone modification or by evicting PcG complexes from chromatin. The focus here will be on mechanism of transcription activation employed by the COMPASS subgroup.

The following sections will briefly describe the subunit compositions of PcG and COMPASS complexes, how they identify their target sites, and the mechanism by which they regulate gene expression.

1.3.1 PcG complex composition and target site identification

PRC1 and PRC2 display a highly modular complex composition with 2-4 constitutive core subunits that are joined by a variable set of auxiliary subunits to form subcomplexes with distinct genomic localisations and specialised functions. The catalytic core of PRC1 is a heterodimer formed by the E3 ubiquitin ligase RING1A or its paralogue RING1B, and one of six PCGF proteins (PCGF1-6) (Figure 1) (Gao et al., 2012b). The particular PCGF core subunit determines which auxiliary components are incorporated into PRC1 subcomplexes. Depending on the subunit composition, PRC1 subcomplexes are grouped into canonical and variant PRC1. Canonical PRC1 (cPRC1) is composed of PCGF2 or 4, one of five CBX paralogues (CBX2/4/6/7/8), and one of three PHC paralogues (PHC1/2/3). Variant PRC1 (vPRC1) contains any of the six PCGFs and RYBP or its paralogue YAF2. Additionally, vPRC1 complexes contain several other auxiliary subunits as depicted in Figure 1. Based on the PCGF component, different vPRC1 complexes will be denoted here as vPRC1.1 for the PCGF1-containing complex, vPRC1.2 for the PCGF2-containing complex, and so on. The difference in subunit composition affects the E3 ubiquitin ligase activity of RING1A/B. For example, the RYBP component of vPRC1 stimulates ubiquitin ligase activity more than incorporation of CBX proteins in cPRC1 (Gao et al., 2012b; Rose et al., 2016; Taherbhoy et al., 2015).

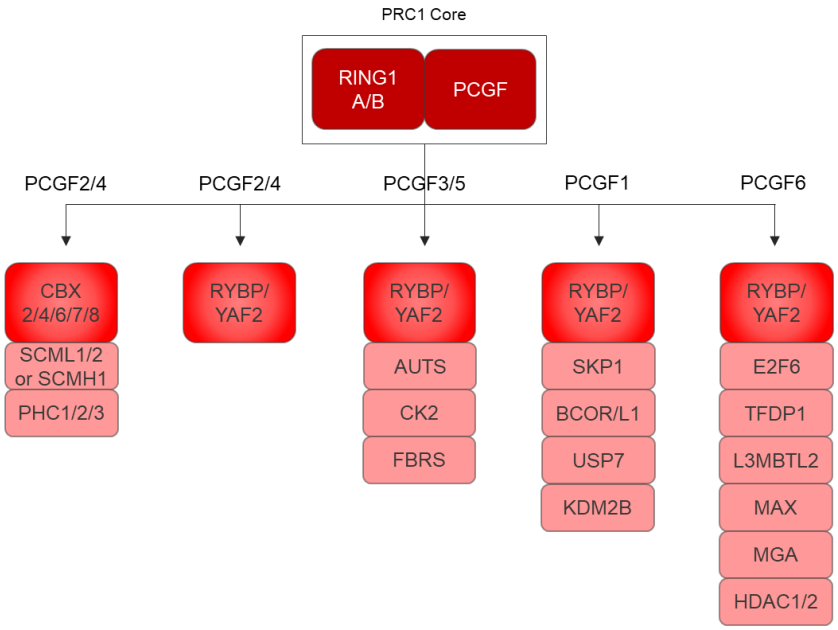


Figure 1: Schematic of PRC1 complexes in mammals
Schematic of the core and auxiliary subunits that form the various PRC1 complexes. The core comprises RING1A or RING1B and one of the six PCGF proteins. RING1A/B are the catalytic subunits that can monoubiquitylate H2AK119. Complexes that contain CBX subunits belong to canonical PRC1 whereas RYBP/YAF2-contaning complexes are part of variant PRC1. Canonical and variant PRC1 complexes have a variable set of auxiliary subunits coloured in light red.

The catalytic core of PRC2 comprises EED, EZH1 or its paralogue EZH2, SUZ12, and RBBP4 or RBBP7 (Figure 2) (van Mierlo et al., 2019). Methylation of H3K27 is performed by the SET domain of EZH1/2. EED, EZH1/2, and SUZ12 are essential for complex assembly and sufficient for catalytic activity, whereas RBBP4/7 are required for binding unmodified nucleosomes and for further stimulating the methyltransferase activity (Cao and Zhang, 2004; Murzina et al., 2008; Nekrasov et al., 2005; Pasini et al., 2004a; Schmitges et al., 2011). Additionally, PRC2 can bind the product of its own catalysis (H3K27me3) through the EED subunit, which leads to allosteric activation of the complex and enhancement of the catalytic activity (Jiao and Liu, 2015). Similar to PRC1, the PRC2 core associates with additional auxiliary subunits that give rise to two PRC2 subcomplexes with distinct recruitment and regulation mechanisms (van Mierlo et al., 2019). PRC2.1 includes one of three PCL proteins (PCL1/2/3 also known as PHF1, MTF2, PHF19) together with EPOP or PALI1/2 (Figure 2). In contrast, PRC2.2 is associated with AEBP2 and JARID2. Incorporation of the auxiliary subunits promotes catalytic activity and defines target site binding (Figure 2).

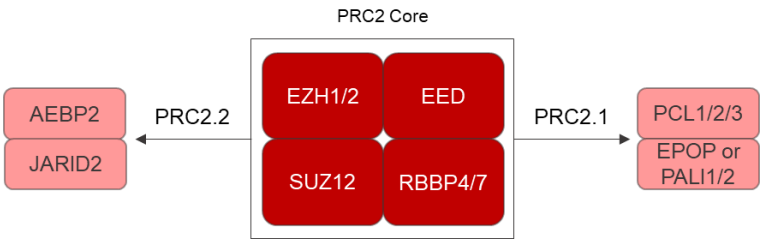


Figure 2: Schematic of PRC2 complexes in mammals
Schematic of the core and auxiliary subunits that form PRC2.1 and PRC2.2 complexes. PRC2 has a tetrameric core formed by EZH1 or EZH2, EED, SUZ12, and RBBP4 or RBBP7. EZH1/2 are the catalytic subunits that can mono-, di-, and trimethylate H3K27. In addition to the core, PRC2.1 contains one of three PCL proteins, and EPOP or one of two PALI proteins. AEBP and JARID2 join the core to form the PRC2.2 complex.

Auxiliary subunits mediate recruitment of Polycomb complexes to gene regulatory elements through mechanisms that include recognition of DNA sequence features and histone modifications. For example, vPRC1.6 contains DNA sequence-specific transcription factors that help the complex to identify a subset of its genomic targets, in particular germ cell-related genes. The capacity to target specific sequences relies on the MAX-MGA dimer recognising E-box and T-box motifs, and the E2F6-TFDP1 dimer that binds the E2F motif (Endoh et al., 2017; Stielow et al., 2018). Moreover, genome-wide studies of Polycomb complex occupancy revealed that they primarily localise at CGIs (Ku et al., 2008; Meissner et al., 2008). Components that are known to recruit Polycomb complexes to CGIs are the ZF-CXXC protein KDM2B incorporated in vPRC1.1 and the winged-helix proteins PCL1-3 associated with PRC2.1 (Choi et al., 2017; Farcas et al., 2012; Li et al., 2017; Perino et al., 2018).

Polycomb complexes converge at the same target sites not only through their ability to bind CGIs, but also through their ability to recognise each other's catalysis products. JARID2 and AEBP2 subunits of PRC2.2 can directly bind H2AK119ub (Cooper et al., 2014; Kasinath et al., 2021). In addition to recruiting PRC2.2, this interaction also stimulates the methyltransferase activity of PRC2 (Kalb et al., 2014). Accordingly, disruption of the E3 ubiquitin ligase function of vPRC1, which catalyses the bulk of H2AK119ub, was shown to cause a considerable decrease in PRC2 occupancy and H3K27me3 in mESCs (Blackledge et al., 2014; Fursova et al., 2019). Similarly, CBX subunits interact with H3K27me3 and recruit cPRC1 to Polycomb domains (Bernstein et al., 2006b; Morey et al., 2012). However, since CBX-containing cPRC1 has weak E3 ubiquitin ligase activity, recruitment of cPRC1 by H3K27me3 contributes little to H2AK119ub levels *in vivo* (Fursova et al., 2019; Rose et al., 2016; Tavares et al., 2012). Instead, there are several studies indicating that cPRC1 supports long-range interactions and liquid-liquid phase separation of Polycomb domains

giving rise to focal accumulations of Polycomb proteins referred to as Polycomb bodies (Denholtz et al., 2013; Eskeland et al., 2010; Plys et al., 2019a; Ren et al., 2008; Saurin et al., 1998; Schoenfelder et al., 2015; Seif et al., 2020; Tatavosian et al., 2019a). These condensates are thought to enhance cellular processes by bringing together components of the same catalytic and signalling pathways (Shin and Brangwynne, 2017).

Both PRC1 and PRC2 can also bind their own catalysis products. As mentioned above, EED interacts with H3K27me3 which allosterically activates the methyltransferase function of EZH1/2 (Jiao and Liu, 2015). In vPRC1, binding of RYBP/YAF2 to H2AK119ub promotes catalytic activity (Arrigoni et al., 2006; Gao et al., 2012b; Rose et al., 2016; Zhao et al., 2020). Interaction with their own catalysis products supports spreading of H2AK119ub and H3K27me3 onto neighbouring nucleosomes (Oksuz et al., 2018; Zhao et al., 2020). Thus, binding their own and each other's catalytic output reinforces recruitment and catalytic activity of PRC1 and PRC2. Through these positive feedback mechanisms, PRC1 and PRC2 synergise their repressive activities to create robust barriers against aberrant transcription initiation (Blackledge and Klose, 2021).

1.3.2 COMPASS complex composition and target site identification

The COMPASS family in mammals comprises the six H3K4 methyltransferases SETD1A, SETD1B, KMT2A, KMT2B, KMT2C, and KMT2D (they are also known as SET1A, SET1B, MLL1, MLL2, MLL3, MLL4) that form different complexes with distinct subunit compositions and functional specialisations. Common to all six COMPASS complexes are the four core subunits WDR5, RBBP5, ASH2L, and DPY30 (collectively also known as WRAD) (Figure 3), which are required for complex assembly but do not have histone methyltransferase activity of their own (Bochyńska et al., 2018; Ernst and Vakoc, 2012; Li et al., 2016). However, they are essential for the full methylation activity of the holo-complex *in vivo*. Thus, each COMPASS complex contains one catalytic subunit supported by four core subunits and a variable set of auxiliary components (Figure 3).

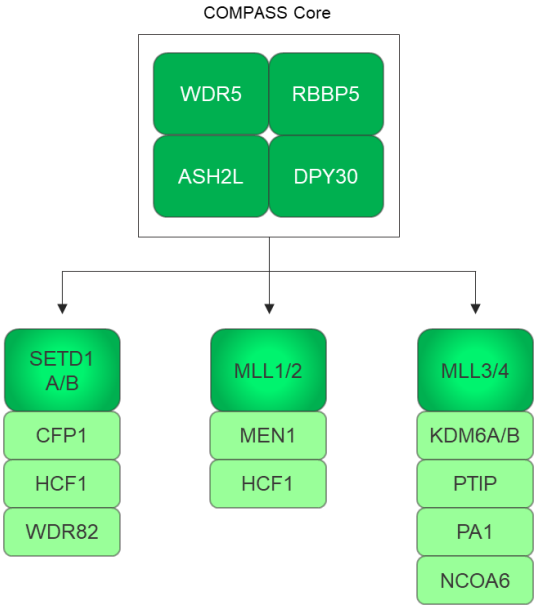


Figure 3: Schematic of COMPASS H3K4 methyltransferase complexes in mammals
Schematic of the core and auxiliary subunits that form the various COMPASS complexes. COMPASS has a tetrameric core formed by WDR5, RBBP5, ASH2L, and DPY30 (collectively known as WRAD). In addition to the core, each complex has one of the six H3K4 methyltransferases SETD1A, SETD1B, MLL1, MLL2, MLL3, or MLL4. The core and the H3K4 methyltransferase subunit are joined by a variable set auf auxiliary components coloured in light green.

All six mammalian COMPASS complexes can methylate H3K4, but they have specific roles in different developmental stages and in targeting distinct genomic loci, although some functional overlap seems to exist (Cenik and Shilatifard, 2021). The distinct role of each complex in development is underlined by the fact that knockout of any individual COMPASS member causes prenatal lethality with different phenotypes in mice (Bledau et al., 2014; Glaser et al., 2006; Lee et al., 2013; Yagi et al., 1998; Yu et al., 1995). For example, the SETD1A and SETD1B complexes are responsible for the bulk of H3K4me2/3 on promoters genome-wide and have distinct roles in embryogenesis (Lee and Skalnik, 2005; Wu et al., 2008). Mouse embryos with SETD1A knockout do not develop past the epiblast stage whereas SETD1B-depleted embryos survive until E11.5 (Bledau et al., 2014). Moreover, SETD1A was shown to be essential for mESC survival, whereas SETD1B was dispensable (Bledau et al., 2014). Likewise, knockout of SETD1A but not SETD1B led to a global reduction of H3K4me3 in mESCs. The loss of SETD1A could not be rescued by overexpressing SETD1B, suggesting that they regulate different target genes (Bledau et al., 2014).

The MLL1/2 complex has a very similar composition to the SETD1A/B complex. The main difference is that WDR82 is replaced by MEN1 in the MLL1/2 complex and the ZF-CXXC

targeting module is integrated in the MLL1/2 proteins. Both paralogs catalyse H3K4me2/3 mainly at promoters of developmental genes in a locus-specific manner (Cenik and Shilatifard, 2021). For instance, detailed analyses of *Hox* gene expression during mESC to embryoid body differentiation showed that loss of MLL1 impaired expression of *Hoxa7*, *Hoxa9*, and *Hoxc8*, whereas MLL2 knockout did not affect these genes (Ernst et al., 2004; Glaser et al., 2006; Hanson et al., 1999; Yu et al., 1998). Conversely, MLL2 but not MLL1 was required for the expression of *Hoxb2* and *Hoxb5*, suggesting that MLL1 and MLL2 regulate different processes in development. Further studies have shown that MLL2 was responsible for most H3K4me3 at bivalent genes, which is a distinctive feature of many developmental gene promoters in pluripotent cells (Azuara et al., 2006; Bernstein et al., 2006a; Denissov et al., 2014; Hu et al., 2013a). In contrast, only a relatively small subset of developmental genes relies on the catalysis of H3K4me3 by MLL1 in pluripotent cells (Denissov et al., 2014; Wang et al., 2009).

The MLL3/4 complex has four complex-specific subunits in addition to the WRAD assembly and the methyltransferase component (Figure 3). The KDM6A subunit is of special interest because it is a histone demethylase that can remove H3K27me3, thereby counteracting the gene repressive activity of PRC2. MLL3/4 are major H3K4 mono- and dimethyltransferases with partially overlapping functions (Hu et al., 2013b; Lee et al., 2013; Sze and Shilatifard, 2016). H3K4me1 in combination with H3K27ac is a hallmark of active enhancers (Creyghton et al., 2010). Indeed, in mESCs KDM6A was shown to facilitate recruitment of the H3K27 histone acetyltransferase EP300 and MLL3/4 to enhancers, thereby integrating both catalytic activities to stimulate enhancer function (Wang et al., 2017a). Accordingly, both paralogs are enriched at enhancers and are essential for enhancer function. However, evidence from mESCs suggests that their contribution to enhancer function does not entirely depend on their ability to catalyse H3K4me1 (Cao et al., 2018; Dorighi et al., 2017; Hu et al., 2013b; Lee et al., 2013; Wang et al., 2016). Rather it appears that enhancer-bound MLL3/4 make long-range promoter interactions to regulate release of RNAPII from promoter-proximal pausing in a catalysis-independent manner (Dorighi et al., 2017).

Recruitment of COMPASS proteins relies on some similar mechanisms as described for PcG proteins. The ZF-CXXC protein CFP1 is an essential component of the SETD1A/B complex that mediates recruitment to CGIs (Lee and Skalnik, 2005; Yu et al., 2017). Likewise, MLL1/2 contain a ZF-CXXC domain that is required for their association with CGIs (Ayton et al., 2004; Hu et al., 2017; Milne et al., 2010). Hence, ZF-CXXC domains (and WH domains in PCL proteins) are critical elements that guide both PcG and COMPASS proteins to CGI promoters where their balance determines whether genes are in a transcriptionally

repressive or permissive state. Moreover, COMPASS complexes are recruited to their target genes by TFs that have specificities for particular genes, developmental stages, and cell types. For instance, the hematopoietic activator NFE2 recruits MLL2 for H3K4 trimethylation of the beta-globin locus in erythroid cells (Demers et al., 2007). Similarly, H3K4 monomethylation of enhancers in breast cancer cells was shown to be dependent on MLL3 recruitment by the pioneer TF FOXA1 (Jozwik et al., 2016). In particular the auxiliary subunits of MLL3/4 complexes are known to partner with TFs for complex recruitment. For example, NCOA6 directly interacts with nuclear receptors (NRs) such as the oestrogen receptor through an LXXLL motif and recruits MLL3/4 to NR-targeted gene loci (Wang et al., 2021a). Furthermore, histone modifications can serve as docking sites for COMPASS recruitment. The CFP1 subunit of SETD1A/B complex and the ZF-CXXC/PHD domains of MLL1 have been shown to make multivalent contacts with both non-methylated CpGs and H3K4me3 *in vivo* (Brown et al., 2017; Milne et al., 2010). Thus, similar to PcG proteins, the SETD1A/B and MLL1 complexes can interact with their own catalysis products, presenting an opportunity for a positive feedback mechanism.

1.3.3 Mechanisms of PcG- and TrxG-mediated chromatin regulation

PcG and TrxG proteins extensively colocalise on the same genomic sites but not all genes bound by PcG proteins are in a repressed state and not all TrxG-occupied genes are actively transcribed (Brand et al., 2019; Schuettengruber et al., 2017). This suggests that other mechanism beyond recruitment and occupancy must influence how these complexes exert their repressive or activating functions. Therefore, a key questions of the PcG-TrxG antagonism is what tips the balance between gene repression and activation. Several factors have been proposed to participate in controlling the balance between PcG and TrxG activities.

The transcription process itself appears to counteract PcG activity as demonstrated in mESCs treated with different transcription inhibitors (Riising et al., 2014). Within 12 h of transcription inhibition, PRC2 was recruited to thousands of new target sites with a concomitant increase in H3K27me3 deposition. The vast majority of new PRC2 binding occurred at CGIs which were *bona fide* PcG target sites in differentiated cells, showing that the gain in binding was not at random genomic locations. Interestingly, a subset of CGIs did not show stable PRC2 binding after transcription inhibition, suggesting that other factors might be involved in PRC2 recruitment. Indeed, nucleosome mapping data showed that CGIs with high nucleosome density were insusceptible to PRC2 binding. Thus, PRC2

recruitment was dictated not only by transcription but also by DNA accessibility at CGIs. These results were supported by other studies in mouse fibroblast and ES cells which similarly found that PRC2 binding and H3K27me3 deposition at promoters and CpG-rich sequences are suppressed by transcription activity (Hosogane et al., 2013; Jermann et al., 2014; Yuan et al., 2012). Mechanistically, transcription could influence PcG activity through nascent RNA, which was shown to interfere with PRC2 binding and catalytic activity (Beltran et al., 2016, 2019; Kaneko et al., 2013; Wang et al., 2017b; Zhang et al., 2019a). However, there are also reports that nascent transcripts might support PRC2 binding at target sites, and therefore, the role of nascent RNA in this context requires more investigation (Long et al., 2020; Wei et al., 2016).

Histone modifications play a pivotal role regulating the balance between PcG and TrxG activities. Various studies have uncovered mechanism by which H3K4me3 deposited by SETD1A/B and MLL1/2 TrxG complexes can antagonise Polycomb-mediated transcription repression. For example, symmetrical modification of nucleosomes with H3K4me3 was shown to impede PRC2 binding and H3K27 trimethylation (Schmitges et al., 2011; Voigt et al., 2012). Accordingly, a decrease of H3K4me3 caused by a loss of MLL2 in mESCs led to reduced nascent transcription and chromatin accessibility while Polycomb occupancy and H3K27me3 levels increased at bivalent genes (Mas et al., 2018). Moreover, H3K4me3 can counteract PRC2 recruitment by preventing binding of the transcriptional repressor NuRD complex, which has histone deacetylase and chromatin remodelling function (Musselman et al., 2009; Zegerman et al., 2002). The NuRD complex catalyses deacetylation of the active promoter mark H3K27ac, thereby paving the way for PRC2 recruitment and subsequent H3K27 trimethylation (Reynolds et al., 2012). Existing H3K27me3 at PcG target sites can be removed by histone demethylases KDM6A/B, which were reported to associate not only with MLL3/4 but also with MLL2 (Agger et al., 2007; Hong et al., 2007; Issaeva et al., 2007; Min et al., 2007). Modification of H3K27 seems to play a critical role in determining the balance between activating and repressing signals as acetylation and methylation on this histone residue are mutually exclusive (Tie et al., 2009). Loss of SUZ12 in mESCs was shown to lead to a global increase in H3K27ac and correlated with the transcriptional activation of Polycomb target genes (Pasini et al., 2010a). Responsible for H3K27 acetylation were histone acetyltransferases CBP and EP300 that can be recruited by TrxG proteins (Pasini et al., 2010a; Tie et al., 2009). Additionally, other acetylation marks such as H4K16ac and H4K20ac were also implicated in abrogating PRC2 catalytic activity (Weaver et al., 2019). Considering that transcription activating H3K4me3 and repressing H3K27me3 are not mutually exclusive on the same nucleosome, and indeed

mark poised promoters *in vivo*, acetylation of H3K27 might have a decisive contribution to resolving a bivalent promoter state.

As described in previous sections, chromatin accessibility is a critical parameter in controlling gene expression. Active gene promoters are characterised by an open chromatin state which is modulated by ATP-dependent chromatin remodellers that can be recruited by activating histone modifications. For example, the NURF complex is able to bind H3K4me3 and acetylated histone tails via its BPTF subunit (Li et al., 2006; Ruthenburg et al., 2011; Wysocka et al., 2006). Nucleosome sliding mediated by NURF at promoters can facilitate the access to DNA for TFs and the transcription machinery. Similarly, the chromatin remodeller CHD1 can recognise H3K4me3 and promote a transcriptionally permissive chromatin state at promoters through nucleosome sliding (Flanagan et al., 2005; Gaspar-Maia et al., 2009; Lin et al., 2011; Sims et al., 2005). Aside from directly modulating DNA accessibility through nucleosome repositioning, chromatin remodellers can also promote transcriptional activation by counteracting PcG occupancy. For instance, the SWI/SNF family chromatin remodeller BRG1 was shown to directly evict PRC1 from transcriptionally inactive promoters (Kadoch et al., 2017; Stanton et al., 2017).

Chromatin compaction has been postulated as one of the main principles of PcG-mediated repression (Francis et al., 2004). In particular, the CBX and PHC subunits of cPRC1 have been associated with compacting chromatin (Grau et al., 2011; Lau et al., 2017; Wani et al., 2016). Studies in various *in vitro* and *in vivo* systems have indicated that cPRC1 mediates chromatin compaction on different structural levels. On the local level, cPRC1 was shown to promote compaction of nucleosomal arrays (Francis et al., 2004; Tatavosian et al., 2019b). On a chromosomal level, cPRC1 appears to support the formation of phase-separated Polycomb body condensates by facilitating long-range interactions between individual Polycomb domains (Kundu et al., 2017; Plys et al., 2019b; Tatavosian et al., 2019b). In agreement with the relatively weak H2A ubiquitylation activity of cPRC1, chromatin compaction seems to be largely independent of PRC1 catalytic activity (Eskeland et al., 2010). However, studies in mESCs have shown that disruption of cPRC1 components caused relatively few gene expression changes whereas defects in vPRC1 led to the derepression of several thousand genes (Fursova et al., 2019). Moreover, other studies have questioned whether PRC1 depletion or its accumulation affected chromatin accessibility at promoters (Hodges et al., 2018; King et al., 2018). Therefore, chromatin compaction that might be mediated by cPRC1 does not seem to be the primary mechanism by which Polycomb proteins inhibit gene expression in mESCs.

In contrast, H2AK119ub, which is mainly catalysed by vPRC1 in mESCs, appears to have a central role in gene repression (Blackledge et al., 2020; Fursova et al., 2019). The RSF1 subunit of the RSF chromatin remodelling complex could represent a potential link between H2AK119ub and gene silencing. It was shown in HeLa and mES cells that binding of RSF1 to H2AK119ub is required for stable maintenance of H2AK119ub nucleosome patterns at promoters (Zhang et al., 2017). Additional *in vitro* experiments revealed that RSF1 can repress transcription activation from H2AK119ub chromatin. Other *in vitro* studies have suggested that H2AK119ub could counteract gene expression by inhibiting H3K4 methyltransferase activity of MLL1, which would imply an interference with the initiation step of transcription (Nakagawa et al., 2008). This is supported by studies in which PRC1 was shown to allow binding of TFIID and RNAP, but blocked transcription initiation possibly by disrupting the interaction between Mediator and the PIC (Dellino et al., 2004; Lehmann et al., 2012). Further studies are required to elucidate the exact mechanisms by which H2AK119ub affects transcription.

In summary, there is extensive communication between PcG and TrxG proteins through post-translational modifications and alterations in chromatin accessibility. Bivalent chromatin occupied by both H3K4me3 and H3K27me3 marks sites where PcG and TrxG activities balance out. During development, the bivalent state needs to be resolved in order to guide transcriptional programmes towards cell differentiation and specification. CGIs as sites of transcription initiation and targets of PcG and TrxG regulation play a pivotal role in this process. Therefore, understanding how the bivalent state is resolved at CGIs can possibly elucidate the more general principles that govern the opposing forces of transcription repression and activation.

1.4 Proteomic methods to study chromatin and gene regulation

Knowing the components of a system is a prerequisite for understanding how it works. Therefore, an essential step in understanding the regulation of gene expression is the identification of proteins that interact with gene regulatory elements. Through alterations of the chromatin composition at these regulatory elements, transcriptional programmes modulate the gene activity in development and normal physiology. Mass spectrometry (MS)-based analysis of chromatin is a powerful tool to comprehensively identify proteins associated with gene regulation and to study them in their *in vivo* chromatin-bound context (van Mierlo and Vermeulen, 2021). The aim of MS-based chromatin studies is usually the investigation of the chromatin composition at defined genomic regions, which can be for example genetic elements with distinct DNA sequence features, or regions occupied by a protein of interest. Over the past ~15 years, several methods have been developed to probe the chromatin proteome by mass spectrometry. These methods can be categorised into three groups: (1) single-locus isolation, (2) whole-chromatin purification, and (3) proximal proteome enrichment. The following sections will focus on whole-chromatin purification and proximal proteome enrichment strategies, as single-locus isolation is not relevant in the context of this work. First, a few technical challenges will be considered that are common to most chromatin proteomics approaches. Then, selected methods representing whole-chromatin purification and proximal proteome enrichment strategies will be described and evaluated regarding their strengths and limitations.

1.4.1 Technical challenges in chromatin proteomics

1.4.1.1 Preserving chromatin interactions

Chromatin is a heterogenous assembly of DNA and proteins. While the DNA part of chromatin is relatively static, the composition of the protein fraction is subject to dynamic changes (Espejo et al., 2020). However, these changes do not affect all proteins equally. Structural components, such as histones, are stably bound and have high temporal residency on chromatin. In contrast, regulatory proteins usually interact more transiently with chromatin, which inevitably leads to their dissociation or to the induction of composition rearrangements in the process of chromatin isolation. However, to understand how the chromatin composition influences gene regulation, it is essential to capture the transient interactions of regulatory proteins as they occur under physiological conditions inside the cell. To achieve this, transient interactions can be preserved *in vivo* by crosslinking which is the process of connecting molecules by covalent bonds. Upon crosslinking, the sample can be subjected to non-physiological buffer conditions without disrupting the integrity of

chromatin interactions. Two commonly used methods to induce crosslinks are UV light and chemical reagents.

Irradiation of cells with UV light in the range of 250 – 280 nm generates highly reactive radical sites at the nucleobases that rapidly react with nearby amino acids to form covalent bonds between the nucleobase and the amino acid (Steen and Jensen, 2002). Thus, an advantage of UV crosslinking is the specificity because it creates covalent bonds only with those proteins that are in direct contact with the nucleobases. Proteins that contact only the DNA backbone or associate through other DNA-bound proteins are not crosslinked. However, UV crosslinking has also several drawbacks. The crosslinking efficiency is very low when native nucleobases are irradiated with conventional UV lamps (Zhang et al., 2004). In fact, the main outcomes will be pyrimidine dimers, interstrand DNA-DNA crosslinks, and breaks of the sugar-phosphate backbone (Zhang et al., 2004). Since UV crosslinks are irreversible, the damage caused to the DNA can hamper downstream applications such as polymerase chain reactions (PCR) used for DNA amplification. The crosslinking efficiency can be considerably increased, and DNA damage reduced by incorporation of photoactivatable base analogues such as bromouridine (BrdU) into the DNA, and by employing pulsed UV lasers (Meisenheimer and Koch, 1997; Steen and Jensen, 2002). However, this requires manipulation of the cells (BrdU) and specialised equipment (UV lasers) which is not a standard in most molecular biology laboratories.

Chemical crosslinking reagents have reactive ends that are capable of forming covalent bonds with specific functional groups present on proteins and/or nucleic acids such as primary amines ($-NH_2$), carboxyls (-COOH), sulfhydryls (-SH), and carbonyls (-CHO). A large inventory of crosslinking reagents is available that differ for example in their specificity for functional groups, and the molecular distance over which they can conjugate two molecules. One of the most popular crosslinkers used in chromatin-related studies is formaldehyde. Key advantages of formaldehyde include its cell membrane permeability and small size which allows it to diffuse rapidly into biological materials (Hoffman et al., 2015). Due to its small size, formaldehyde connects direct interactions over short distances of ~2 Å. Formaldehyde can generate both DNA-protein and protein-protein crosslinks involving the amino acids Lys, Cys, His, Trp and the nucleobases adenine, guanine, and cytosine (Lu et al., 2010). Hence, proteins do not have to be in direct contact with DNA to be crosslinked to chromatin. This can be beneficial for capturing proteins that interact mainly with histones but not with the DNA bases, and for preserving the stability of multimeric complexes in which only one or a few subunits bind directly to DNA. However, several studies have reported that formaldehyde cannot efficiently capture some transcription

factors and cofactors that interact indirectly or transiently with DNA (Festuccia et al., 2019; Kurdistani and Grunstein, 2003; Nowak et al., 2005; Zeng et al., 2006). In these case it was shown that a sequential double crosslinking approach using a long-range crosslinker such as disuccinimidyl glutarate (DSG) followed by formaldehyde could overcome this problem. Finally, a useful property of formaldehyde crosslinks is their reversibility through heating (Kennedy-Darling and Smith, 2014), which allows separate investigation of the DNA and protein components by sequencing and mass spectrometry, respectively.

The choice of the crosslinker is influenced by the research objective, for example, is the target a specific transcription factor, or the complete complement of proteins bound to a particular genomic region? Additionally, practical considerations such as the availability of specialised crosslinking equipment might guide the selection of the crosslinking approach. Furthermore, once a crosslinking approach has been chosen, it might be necessary to empirically determine the crosslinking conditions such as crosslinker concentration and incubation time in order to obtain the aspired result.

1.4.1.2 Identifying low-abundance proteins

The investigation of the protein composition at particular genomic regions requires a method that is able to selectively isolate the regions of interest and to enrich them above the abundance level of all non-target regions. For instance, CGIs constitute ~1% of the entire mouse genome. In order to obtain a chromatin sample in which the majority (>50%) of purified genomic regions contain CGIs, one would need an enrichment factor of more than 50-fold. A simple way to estimate the enrichment factor is to evaluate the target region abundance relative to that of a similarly sized non-target region by quantitative PCR. Although the genomic enrichment factor gives a good indication of the capture efficiency for the target region, it does not provide any information on the specificity of the proteins that were enriched. This is because proteins can bind non-specifically to the enriched target region when the cell is lysed and the contents are mixed during the chromatin isolation step. Therefore, it is essential to implement chromatin purification strategies in the workflow that are able to reduce background proteins. One strategy is subcellular fractionation which has the advantage that cytoplasmic and nuclear fractions can be analysed separately from the chromatin fraction (van Mierlo and Vermeulen, 2021). This can be useful to examine for instance the translocation of proteins between cellular compartments when different treatments are compared. If a specific group of proteins or peptides with shared biophysical properties are the objective of the enrichment, fractionation can be further continued on the

protein and peptide levels, respectively (Eberl et al., 2011). For example, special protocols are available that selectively enrich histones based on their basic nature and/or small size (Shechter et al., 2007). If the objective is the total protein composition of a genomic region, the specificity of the enriched proteome can be evaluated by identifying known interactors of the target region (Vermeulen and Déjardin, 2020).

The chromatin association of regulatory proteins such as TFs is usually transient and very dynamic. Therefore, the abundance of TFs on chromatin can be several fold lower than the abundance of structural proteins such as histones. Thus, a genomic region targeted for enrichment will not only have dozens of different proteins that define its function, but the local abundance range of bound proteins can also be large (Vermeulen and Déjardin, 2020). This is a major technical challenge in mass spectrometry because proteins present in minute concentrations are more likely to escape identification if their signal is masked by abundant structural components and non-specific background proteins. Biochemical procedures such as fractionation can improve the purity and enrichment of the target proteome by removing non-target background proteins. This reduces sample complexity and facilitates the measurement of low-abundance proteins by MS. Additional steps to improve the detection of low-abundance proteins can be taken on the level of the mass spectrometer.

A frequently used proteome analysis technique is bottom-up liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) (Eng et al., 1994; Link et al., 1999; Zhang et al., 2014). In bottom-up proteomics, protein mixtures are digested into peptides, typically by trypsin protease. The peptide mixture is then separated usually by reverse-phase liquid chromatography based on the hydrophobicity of the peptides. This involves loading the peptide mixture on a chromatography column consisting of C18-covered particles. The C18 material (a chain of 18 carbon atoms) retains the peptides, which are then eluted with a gradually increasing concentration of an organic solvent. In the case of electrospray ionisation (ESI), a high voltage is applied to the peptides eluting from the C18 column, thereby vaporising the liquid and ionising the peptides (Fenn et al., 1989). The peptide ions then enter the mass spectrometer where they are separated by their mass-to-charge (m/z) ratio, which is referred to as the MS1 stage. Peptide ions (also known as precursor ions) of a particular m/z ratio are selected and split into smaller fragments. The fragment ions are again separated by their m/z ratio and detected, which is known as the MS2 stage. Thus, the peptide sequence is determined by fragmentation. A major bottleneck in this process is the number of peptides that can be sequenced per unit of time, which is a parameter dictated by the type of instrument. If the sample complexity exceeds the analytical capacity

of the mass spectrometer, undersampling occurs (Wang et al., 2010). That means, only a subset of the peptides can be selected for sequencing and identified. One way to reduce undersampling is to decrease the number of co-eluting peptides per unit of time (Liu et al., 2004). This can be achieved by increasing the gradient time, or in other words, the concentration of organic solvent needs to increase more slowly to achieve better separation of the eluting peptides. However, longer gradient times require more sample material, which can be a limiting factor in some experiments. Moreover, adjustments to the gradient time need to be harmonised with the type of the chromatography column as there is an optimal time range within which the resolution increases. Outside of this optimal time range, longer gradients can lead to poorer results.

Undersampling becomes particularly an issue with data-dependent acquisition (DDA), which is the standard acquisition method used in most proteomics studies (Liu et al., 2004). In DDA, peptides are selected for sequencing within a narrow range of signal intensity. In practice that means, at the MS1 stage typically the 10 – 15 most abundant peptides (top N) within that range are selected for sequencing. MS2 data acquisition occurs sequentially for each peptide, i.e. one precursor at a time is selected for fragmentation. Thus, the resulting fragments can be easily assigned to one precursor ion. Subsequently, the resulting spectra are searched against existing databases to identify the peptides/proteins. Hence, with DDA the identification of precursors is relatively straightforward, but very-low abundant peptides are less likely to be sequenced, and therefore underrepresented. Furthermore, the mass spectrometer decides on the fly which are the top N precursor ions to be sequenced. As a result, the generated data can contain missing values where peptides have been sequenced in some samples only. However, missing values can be reduced through computational methods. For example, the MS data analysis software MaxQuant offers a match-between-runs option that can transfer peptide identifications from one run to another (Cox and Mann, 2008). To achieve this, the match-between-runs algorithm compares a sequenced peptide in an MS1 spectrum with an unsequenced peptide in another run. A peptide identification is transferred if both peptides have the same properties (e.g. m/z ratio, charge state) and are observed within a specified retention time window. This feature can considerably improve quantitation of peptides that do not have MS2 information (Lim et al., 2019).

An alternative to DDA is data-independent acquisition (DIA) in which all peptides within a specified m/z range are selected for fragmentation, regardless of their abundance within that range (Venable et al., 2004). Thus, DIA offers better reproducibility between replicates and higher quantitative accuracy (Gillet et al., 2012). Moreover, DIA reduces undersampling

which benefits low-abundance proteins and could in principle provide a more complete analysis of the sample. However, DIA has some limitations mainly with regard to data processing. The data analysis is more challenging and computationally demanding because fragment ion spectra are highly multiplexed (Liebler, 2004). Since all precursor ions in a specified m/z range are selected and fragmented in parallel, the direct relation between one precursor ion and its fragment ions is lost. This makes the identification of the precursor sequences more difficult. Deconvolution of the multiplexed fragment ion spectra requires spectral libraries generated by DDA. A spectral library contains parameters such as precursor and fragment m/z ratio, fragment ion type, charge state and retention time for each unique peptide in the analysed sample (Krasny and Huang, 2021). For many model organism, reference spectral libraries are available as a community resource. However, if the subject of the study is a sub-proteome, for example the chromatin proteome, a study-specific spectral library needs to be generated because such a sub-proteome would be likely underrepresented in a whole-cell proteome. Therefore, prior to DIA MS analysis, the sample needs to be subjected to DDA-based proteomics to generate a study-specific spectral library. Consequently, DIA MS requires larger amounts of sample, and also increases MS data acquisition time and costs. As bioinformatic tools evolve and computational resources increase, it is going to become easier in the future to deconvolute the complexity of DIA datasets. With more widespread use of DIA MS, sub-proteomes might also become available as a community resource. Moreover, there are already efforts to merge DDA and DIA into a hybrid method that potentially combines the strengths of both approaches (Guan et al., 2020).

1.4.2 Whole-chromatin purification

The aim of whole-chromatin purification approaches is to selectively isolate chromatin from cells, and to remove as much as possible any co-purifying proteins that interact non-specifically with chromatin. Whole-chromatin purification approaches are usually based on either ultracentrifugation, precipitation, or biotin-labelling. Crosslinking is a prerequisite for all methods presented in this section.

Crosslinked chromatin has a specific buoyant density of 1.39 g/cm^3 and this particular property can be exploited to separate it from other cell components in a caesium chloride (CsCl) gradient (Orlando et al., 1997). The procedure requires fragmentation of crosslinked chromatin, e.g. by sonication, and a CsCl solution with an initially homogenous density. Fragmented chromatin is loaded onto the CsCl solution in ultracentrifugation tubes and

subjected to high centrifugal forces for at least 48 h. Extensive ultracentrifugation creates a stable isopycnic density gradient in which free DNA molecules settle at the bottom, free proteins remain at the top, and crosslinked chromatin accumulates in the middle of the gradient. This method can very efficiently purify chromatin as it was recently demonstrated with a human cell line chromatin proteome (Ginno et al., 2018). However, this procedure requires more than two days of sample processing time, and the recovery of purified chromatin can be challenging because the isopycnic gradient is exceedingly sensitive to disruption by motion.

An alternative method to purify total chromatin relies on labelling nucleobase analogues such as EdU with biotin (Kliszczak et al., 2011; Sirbu et al., 2011). Cells that are grown in EdU-supplemented medium incorporate the base analogue into their DNA with each cell division. After crosslinking, the cells are incubated with a reactive biotin species that forms a covalent bond with the EdU-containing DNA. Biotin-labelled chromatin can be then immobilised on streptavidin beads and subjected to stringent wash conditions to remove non-specific background proteins. This method recently helped to uncover an unexpected chromatin-association and gene regulatory function of the metabolic enzyme adenosylhomocysteinase (Aranda et al., 2019). A caveat of this method is that it requires incorporation of nucleobase analogues, which can potentially disrupt normal cell physiology. Therefore, the authors of aforementioned study had to empirically determine and validate a suitable EdU concentration that efficiently labels DNA while showing no toxicity in their cell model system. Moreover, technical challenges can arise if purified chromatin needs to be eluted from streptavidin beads to further enrich a subset of genomic regions. Disruption of the biotin/streptavidin interaction can be achieved only with denaturing conditions, such as boiling, or with high concentrations of free biotin (Cheah and Yamada, 2017; Lin et al., 2019). Thus, eluted chromatin might not be amenable to downstream processes and/or a substantial amount of sample will not be recovered.

Biotin-labelling of chromatin is not an absolute requirement for using stringent wash conditions as it was demonstrated by the Chromatin Enrichment for Proteomics (ChEP) method (Kustatscher et al., 2014a). Under hypotonic buffer conditions, crosslinked chromatin precipitates as a gelatinous pellet. This property was used to repeatedly precipitate and resuspend chromatin with a highly denaturing wash buffer containing 8 M urea and 4% SDS. While crosslinked components are retained on chromatin, most non-crosslinked factors will be rinsed away under these stringent wash conditions. The major advantage of ChEP is that it is quick and easy to implement and does not require the manipulation of cells or any specialised equipment. However, achieving reproducible

chromatin precipitation might require empirical determination of the right conditions for different cell types. Furthermore, chromatin is not the only cell component that is insoluble in low ionic conditions, and therefore, impurities can arise from the co-precipitation of other organelles (Espejo et al., 2020). The authors of ChEP tried to mitigate this problem by performing a crude nuclei isolation before chromatin precipitation. However, their proteomics data shows that purified chromatin still contained a considerable amount of cytoplasmic contaminants (Kustatscher et al., 2014b).

1.4.3 Proximal proteome enrichment

The proximal proteome in the context of chromatin describes the complete complement of proteins that bind to defined genomic regions which can be distinguished by DNA sequence features or by the presence of distinctive proteins. Most proximal proteome enrichment approaches are candidate-based as they rely on a specific chromatin-bound protein. For example, in ChIP followed by MS (ChIP-MS), crosslinked chromatin is first fragmented, and then an antibody is used to enrich chromatin fragments occupied by the target protein. These fragments are then subjected to MS to identify other proteins that were bound in the vicinity of the immunoprecipitated target protein. ChIP-MS has been shown to be particularly useful for targeting histone modifications that are associated with particular genomic elements such as gene promoters and gene bodies in different activity states (Engelen et al., 2015; Ji et al., 2015; Soldi and Bonaldi, 2013). However, ChIP-MS cannot unbiasedly probe the proteome of a genomic locus because it is based on the occupancy of a target protein, which can change in different cell types and treatment conditions. Moreover, the requirement for a specific antibody comes with several drawbacks. First, a specific antibody which recognises the protein of interest in the context of crosslinked chromatin might not always be available. In such a case, endogenous tagging of the protein of interest with an affinity handle such as a FLAG-tag can potentially circumvent the requirement for a protein-specific antibody (van Mierlo and Vermeulen, 2021). Second, only one protein at a time can be investigated, resulting in a low throughput. Third, as relatively large amounts of antibody are used in ChIP-MS assays, the antibody is often the most abundant protein species in the eluate (van Mierlo and Vermeulen, 2021). Thus, antibody-derived peptides can mask the signal of less abundant peptides from chromatin proteins. A modified ChIP-MS approach that aims to remove the antibody and other contaminating proteins from the eluate is referred to as Selective Isolation of Chromatin-Associated Proteins (SICAP) (Rafiee et al., 2016). In SICAP, immunoprecipitated DNA fragments are end-labelled with biotin, immobilised on streptavidin beads, and then subjected to stringent

wash conditions to remove the antibody and non-crosslinked contaminants. Thus, SICAP can potentially improve the measurement of low-abundance peptides in ChIP-MS.

Another approach to enrich the proximal proteome of chromatin-bound factors is proximity biotinylation. This method requires a genetic fusion of the protein of interest with a promiscuous biotin ligase or engineered peroxidase (Kim and Roux, 2016). The genetic fusion construct is stably expressed in cells where it biotinylates proteins in its vicinity upon initiation of the labelling reaction, thereby marking proteins that share a specific localisation. To purify biotinylated proteins, the cell lysate can be directly incubated with streptavidin beads and then washed under stringent conditions without the need for time-consuming fractionation procedures. In contrast to the other methods presented here, proximity biotinylation does not require crosslinking and therefore avoids potential crosslinking artefacts. Although the biotinylation and purification procedures are relatively quick and easy, this method requires a considerable amount of preparatory work (Gingras et al., 2019). For example, one needs to make treatment and control cells lines with similar expression levels of the biotinylation enzyme in both. Further, validation is necessary to ensure that the fusion of the biotinylation enzyme does not impair the function, localisation, and interactions of the natural protein. Moreover, the fusion of the biotinylation enzyme to the N- or C-terminus can yield different proteomes (Varnaitė and MacNeill, 2016). A critical point to consider is that the biotinylation enzyme does not directly label nearby proteins. Rather, it releases an activated biotin species that diffuses away from the biotinylation enzyme and reacts with certain amino acids on nearby proteins. That means, the biotinylation signal is a gradient that is strongest where the fusion construct resides the longest. Thus, it is critical to have appropriate controls and a good analysis pipeline to differentiate between target and off-target biotinylation.

1.5 Aim of this study

Cytosine methylation is a pervasive DNA modification in vertebrate genomes that occurs predominantly in the context of sparsely distributed CpG dinucleotides. In contrast to the bulk genome, short genomic elements referred to as CpG islands (CGIs) feature a high GC content and clusters of non-methylated CpGs. CGIs co-localise with the majority of gene promoters where their distinct DNA sequence features enables the recruitment and assembly of transcription regulating complexes. Thus, CGIs exert an important function in controlling promoter activity and gene expression. Despite their importance, it remains unclear how this is mechanistically achieved and how CGI dysfunction contributes to diseases such as cancer. This is largely due to the fact that the majority of our mechanistic insights into CGI function come from candidate-based approaches such as chromatin immunoprecipitation coupled to sequencing (ChIP-seq), which are inherently biased and limit us to investigating one protein at a time. The capacity to attribute detailed mechanistic understanding to CGI function has been hampered by our lack of a comprehensive inventory of proteins that contribute to their maintenance and function. To overcome this fundamental lack of understanding, the central aim of this study is to develop a new biochemical approach to isolate CGI chromatin from the bulk genome in order to investigate the CGI proteome in an unbiased manner using mass spectrometry.

To achieve this, I developed an approach combining proteomics and functional genomics. First, I produced an affinity reagent that selectively binds non-methylated CpGs (Chapter 3). I then designed and optimised a chromatin purification workflow to remove most contaminant proteins from isolated chromatin (Chapter 4). Combining my new chromatin purification approach with the affinity reagent, I enriched CGI chromatin from the bulk genome of mouse embryonic stem cells (mESCs) (Chapter 4). Using next-generation sequencing, I validated on the genome level that this method is suitable to selectively enrich CGI-containing regions (Chapter 5). Mass spectrometry data revealed many known CGI interactors, and it provided new leads for further investigation of potentially novel CGI-associated proteins (Chapter 5). In Chapter 6, I perturbed the CGI proteome in mESCs by deleting selected protein components and demonstrated that my new method can unbiasedly and comprehensively identify CGI proteome changes caused by the perturbation. Finally, using calibrated ChIP-seq as an orthogonal approach, I verified for selected proteins that the CGI proteome alterations were caused by changes in genomic occupancy (Chapter 6). In conclusion, this work presents the first approach to comprehensively and unbiasedly investigate the CGI proteome and its alterations in different conditions. Ultimately, this approach can advance our knowledge of how CGIs contribute to gene regulation in developmental processes and pathogenesis.

2 Materials and methods

2.1 Molecular cloning

2.1.1 Cloning of a bicistronic vector for co-expression of ZF-CXXC and BirA

To create a vector that allows co-expression of ZF-CXXC and BirA for *in vivo* biotinylation, first the BirA coding sequence was amplified from pET21a-BirA plasmid (Addgene #20857) by polymerase chain reaction (PCR) using parameters in Table 1, 2, 3.

Table 1: PCR mix for BirA amplification

Component	Volume (µL)
5X Phusion buffer (NEB #B0518S)	10
dNTPs [10 mM]	1
Primer forward [10 µM]	2.5
Primer reverse [10 µM]	2.5
pET21a-BirA template [24 ng/µL]	1
Phusion polymerase (NEB #M0530S)	0.5
ddH2O	32.5
Total	**50**

Table 2: PCR primers for BirA cloning

Primer	Sequence
Forward	5'-ACAGTAAAGGTGGATACGCCTCTAGAAATAATTTTGTTTAAC
Reverse	5'-CGACGGAGCTCGAATTCGTTATTTTTCTGCACTACGC

Table 3: Thermocycler programme for BirA amplification

Step	Temperature (°C)	Time (mm:ss)	Number of cycles
Initiation	98	01:00	1
Denaturation	98	00:10	30
Annealing	56	00:30	
Extension	72	00:30	
Final extension	72	05:00	1
Hold	4	∞	

After verifying a PCR amplicon of the expected size by agarose gel electrophoresis, the remainder of the reaction was purified using a PCR purification kit (Qiagen #28104) according to the manual.

The target plasmid pNIC28-hKdm2b-CxxC-PHD (Addgene #49216) was cut to allow insertion of the BirA coding sequence. For this, a restriction digestion reaction was prepared (Table 4) and incubated overnight at 37°C. The next day, 5 U alkaline phosphatase (NEB #M0290S) was added to the reaction and incubated at 37°C for 1 h. Plasmid digestion was verified by agarose gel electrophoresis and bands were cut out for gel extraction. The plasmid DNA was purified using a gel extraction kit (Qiagen #28706X4) according to the manual.

Table 4: ZF-CXXC plasmid restriction digestion

Component	Volume (µL)
CutSmart buffer (NEB #B7204)	5
Plasmid (289 ng/µL)	4
BamHI (NEB #3136S)	1
ddH2O	40
Total	**50**

The BirA coding sequence was cloned into the cut plasmid by Gibson assembly using NEBuilder HiFi DNA Assembly Master Mix (NEB #E2621L) according to the manual. The assembly reaction was used directly for heat-shock transformation of *E. coli* (Agilent #200314). Transformed bacteria were incubated on kanamycin selection plates overnight at 37°C. The next day, colonies were picked and expanded in 10 mL LB culture medium overnight at 30°C. DNA was isolated from bacterial cultures using the QIAprep Spin Miniprep Kit (Qiagen #27106X4) according to manual. The sequence of the new co-expression vector was confirmed by sequencing at Source BioScience.

2.1.2 Site-directed mutagenesis of the ZF-CXXC expression construct

A ZF-CXXC domain that does not have the capacity to bind non-methylated CpGs was created by site-directed mutagenesis based on the protocol published by Edelheit et al. (2009). First, separate PCRs were set up for the forward and reverse primer to introduce the desired mutations in the new expression construct (Table 5, 6, 7).

Table 5: PCR mix for site-directed mutagenesis of ZF-CXXC

Component	Volume (µL)
2X Q5 HF Master Mix (NEB #M0492S)	12.5
Primer (forward or reverse) [10 µM]	1.25
ZF-CXXC/BirA template [126 ng/µL]	1
ddH2O	10.25
Total	**25**

Table 6: PCR primers for site-directed mutagenesis of ZF-CXXC

Primer	Sequence
Forward	5'-CGGGCGCATGGCGGCGAGCTGCATCATG
Reverse	5'- CATGATGCAGCTCGCCGCCATGCGCCCG

Table 7: Thermocycler programme for site-directed mutagenesis of ZF-CXXC

Step	Temperature (°C)	Time (mm:ss)	Number of cycles
Initiation	98	00:30	1
Denaturation	98	00:10	35
Annealing/Extension	72	06:00	
Final extension	72	10:00	1
Hold	4	∞	

After PCR, forward and reverse primer reactions were combined. The separate strands were annealed by a gradual temperature decrease on a thermocycler. For this, the combined reaction was first incubated at 95°C for 5 min. Then, the temperature was reduced at a pace of 0.1°C/s until 40°C was reached. To digest the template plasmid, 20 U of DpnI (NEB #R0176L) were added and incubated overnight at 37°C. The following day, *E. coli* (Agilent #200314) were transformed by heat-shock using the DpnI-digested sample and incubated on kanamycin selection plates overnight at 30°C. Colonies from the selection plate were picked and expanded in 10 mL LB medium overnight at 30°C. Plasmid DNA was purified from bacterial cultures using a QIAprep Spin Miniprep Kit (Qiagen #27106X4) according to manual. The mutation was confirmed by sequencing at Source BioScience.

2.2 Protein expression, purification, and biotinylation

2.2.1 Expression and purification of BirA

To express His-tagged BirA, *E. coli* (LOBSTR, Andersen et al., 2013) were transformed by heat shock with pET21a-BirA plasmid (Addgene #20857) and incubated on an ampicillin selection plate overnight at 30°C. A bacterial colony from the selection plate was used to inoculate 30 mL LB medium supplemented with 100 µg/mL ampicillin and grown overnight at 30°C in a shaking incubator. The overnight culture was diluted 1:70 in fresh LB medium and incubated at 37°C until OD_{600} ~0.5 was reached. Protein expression was induced by addition of 0.5 mM IPTG, and the expression culture was incubated at 30°C for 4 h. Bacteria were harvested by centrifugation at 6,000 x g for 15 min. The supernatant was discarded, and the cell pellet was stored at -80°C.

To purify BirA, the pellet was resuspended in 10 mL lysis buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 1X PIC). All protein purification steps were performed on ice or at 4°C. Bacteria were lysed by sonication using a microtip with 6 x 10 s pulses at 60% amplitude with 1 min pause between pulses. The lysate was cleared by centrifugation at 30,000 x g for 30 min. For affinity purification of His-tagged BirA, the cleared lysate was mixed with 2 mL Ni-NTA slurry (Thermo Scientific #88221) and incubated overnight on a rotator. The Ni-NTA beads were transferred into gravity flow chromatography columns (Bio-Rad #7321010) and washed with 30 column volumes (CV) of wash buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 25 mM imidazole, 1X PIC) to remove contaminating proteins. BirA was eluted with 8 CV elution buffer (50 mM Tris-HCl pH 7.5, 300 mM NaCl, 250 mM imidazole) and then further purified by size exclusion chromatography on an ÄKTA system (GE) using a HiLoad 16/600 S200 chromatography column (GE). After eluting the sample with PBS, the purification was analysed by SDS-PAGE, and BirA-containing fractions were pooled and concentrated in centrifugal filters (Sartorius #VS2022). The protein concentration was measured on a NanoDrop at A280 and purified BirA was stored at -80°C.

2.2.2 Expression, purification, and *in vitro* biotinylation of ZF-CXXC

To express His-tagged ZF-CXXC, *E. coli* (LOBSTR, Andersen et al., 2013) were transformed by heat shock with pNIC28-hKdm2b-CxxC-PHD (Addgene #49216) and incubated on a kanamycin selection plate overnight at 30°C. A bacterial colony from the selection plate was used to inoculate 100 mL 2YT medium supplemented with 50 µg/mL kanamycin and grown overnight at 30°C in a shaking incubator. The overnight culture was diluted 1:70 in fresh 2YT medium and incubated at 37°C until OD_{600} ~0.5 was reached.

Protein expression was induced by addition of 0.5 mM IPTG, and the expression culture was incubated at 30°C for 5 h. Bacteria were harvested by centrifugation at 6,000 x g for 15 min. The supernatant was discarded, and the cell pellet was stored at -80°C.

To purify ZF-CXXC, the cell pellet was resuspended in 20 mL lysis buffer (50 mM bicine pH 8, 300 mM NaCl, 10 mM imidazole, 2 mM 2-mercaptoethanol, 1X PIC). All protein purification steps were performed on ice or at 4°C. Bacteria were lysed in a French press at 30 kPSI after which the lysate was cleared by centrifugation at 30,000 x g for 30 min. For affinity purification of the His-tagged ZF-CXXC protein, the cleared lysate was mixed with 2 mL Ni-NTA slurry (Thermo Scientific #88221) and incubated overnight. The Ni-NTA beads were transferred into gravity flow chromatography columns (Bio-Rad #7321010) and washed with 30 column volumes (CV) of wash buffer (50 mM bicine pH 8, 300 mM NaCl, 15 mM imidazole, 5% glycerol, 0.1% IGEPAL CA-630, 2 mM 2-mercaptoethanol, 1X PIC) to remove contaminating proteins. The ZF-CXXC protein was eluted with 10 x 1 CV elution buffer (50 mM bicine pH 8, 300 mM KCl, 250 mM imidazole, 1X PIC) and analysed by SDS-PAGE to evaluate the preliminary purity of the sample. The protein concentration was measured on a NanoDrop at A280. To cleave off the His-tag of ZF-CXXC, 1:50 (w/w) TEV protease (produced in-house) was added and incubated overnight on a rotator. The sample was loaded on a HiPrep 26/10 desalting column (GE) and eluted into wash buffer. To capture the cleaved His-tag and His-TEV protease, 1 mL Ni-NTA slurry was added to the sample and incubated rotating overnight. The sample was transferred into gravity flow chromatography columns. The flow-through containing ZF-CXXC without His-tag was collected and loaded again on a HiPrep 26/10 desalting column for a buffer exchange into biotinylation buffer (20 mM Tris-HCl pH 8, 250 mM potassium glutamate). As this biotinylation buffer caused protein precipitation, the composition was later modified as follows: 50 mM bicine pH 8, 0.05% IGEPAL CA-630, 5% glycerol. To biotinylate the Avi-tag on the ZF-CXXC protein, 1:75 (w/w) His-BirA and biomix (5 mM $Mg(CH3COOH)_2$, 5 mM ATP, 2 mM biotin) were added and incubated overnight. To capture His-BirA, 1 mL Ni-NTA slurry was added to the reaction and incubated for 3 h. The sample was transferred into gravity flow chromatography columns. The flow-through containing biotinylated ZF-CXXC protein was collected and concentrated in centrifugal filters (Sartorius #VS2001). To increase the purity of ZF-CXXC protein, size exclusion chromatography was performed using a HiLoad 16/600 S75 column (GE). The elution fractions were first analysed by SDS-PAGE, and then pooled and concentrated using centrifugal filters (Sartorius #VS2001). The protein concentration was measured on a NanoDrop at A280 and purified ZF-CXXC was stored at -80°C.

2.2.3 Expression and purification of *in vivo* biotinylated ZF-CXXC

To co-express ZF-CXXC and BirA for *in vivo* biotinylation, *E. coli* (LOBSTR, Andersen et al., 2013) were transformed by heat shock with the co-expression construct (see 2.1.1) and incubated on a kanamycin selection plate overnight at 30°C. A bacterial colony from the selection plate was used to inoculate 100 mL 2YT medium supplemented with 50 µg/mL kanamycin and grown overnight at 30°C in a shaking incubator. The overnight culture was diluted 1:100 in fresh 2YT medium and incubated at 37°C until OD_{600} ~0.5 was reached. Protein expression was induced by addition of 0.5 mM IPTG. Additionally, 0.25 mM $ZnCl_2$ and 0.25 mM biotin were added to support expression and biotinylation of the ZF-CXXC protein. After incubation for 4 h at 30°C, the bacteria were harvest by centrifugation at 6,000 x g for 15 min. The dry cell pellet was stored at -80°C.

The following protein purification steps were performed on ice or at 4°C. The cell pellet was resuspended in 30 mL lysis buffer (50 mM Tris-HCl pH 7.5, 500 mM NaCl, 10 mM imidazole, 5 mM 2-mercaptoethanol, 5% glycerol, 1X PIC) and cells were lysed in a French press at 30 kPSI. Cell debris was collected by centrifugation at 30,000 x g for 30 min. To purify the protein by affinity chromatography, the cleared lysate was loaded on a 5 mL HisTrap chromatography column (GE) connected to an ÄKTA system and washed with 10 CV of wash buffer (50 mM Tris-HCl pH 7.5, 700 mM NaCl, 20 mM imidazole, 5 mM 2-mercaptoethanol, 5% glycerol, 0.1% IGEPAL CA-630, 1X PIC) to remove contaminating proteins. The ZF-CXXC protein was eluted with 15 CV elution buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 500 mM imidazole) and analysed by SDS-PAGE to evaluate the preliminary purity of the sample. A buffer exchange into wash buffer was performed using centrifugal filters (Sartorius #VS2001). The protein concentration was determined by Qubit Protein Assay (Invitrogen #Q33211) according to the manual. To cleave off the His-tag of ZF-CXXC, 1:50 (w/w) TEV protease (produced in-house) was added and incubated overnight on a rotator. The free His-tag molecules and His-TEV protease were captured on a HisTrap column. The flow-through containing the untagged ZF-CXXC protein was collected and further purified by size exclusion chromatography on a HiLoad 16/600 S75 column (GE) connected to an ÄKTA system. The protein was eluted into BC150 (20 mM HEPES pH 7.9, 150 mM KCl, 0.5 mM DTT, 10% glycerol) and the elution fractions were analysed by SDS-PAGE. Selected elution fractions were then pooled and concentrated in centrifugal filters (Sartorius #VS2001). Finally, the protein concentration was determined by Qubit Protein Assay, after which the sample was stored at -80°C.

2.3 Mammalian cell culture

Mouse embryonic stem cell (mESC) line E14TG2a was grown in gelatine-coated Petri dishes at 37°C and 5% CO_2 in Dulbecco's Modified Eagle Medium (Gibco #41965062) supplemented with 2 mM L-glutamine (Gibco #25030081), 1X non-essential amino acids (Gibco #11140035), 0.5 mM 2-mercaptoethanol (Gibco #31350010), 1X penicillin-streptomycin (Gibco #15140122), 15% foetal bovine serum (Labtech #FB-1001), and 10 ng/mL leukaemia inhibitory factor (produced in-house). Human HEK293T cells were grown in gelatine-coated Petri dishes at 37°C and 5% CO_2 in Dulbecco's Modified Eagle Medium supplemented with 2 mM L-glutamine, 0.5 mM 2-mercaptoethanol, 1X penicillin-streptomycin, and 10% foetal bovine serum. For passaging, cells were detached from dishes with trypsin (Gibco #12605036) and counted on a Countess II automated cell counter (Invitrogen #AMQAX1000) to plate a specific number of cells. All cells were tested regularly for mycoplasma infection.

Mouse ESC lines with *Ring1a*$^{-/-}$*/Ring1b*$^{fl/fl}$ or *Suz12*$^{fl/fl}$ conditional knockout genotypes were previously developed and validated by members of the Klose laboratory (Fursova et al., 2019 and Appendix). To induce conditional knockout, mESCs were treated with 800 nM 4-hydroxytamoxifen for 72 h (*Ring1a*$^{-/-}$*/Ring1b*$^{fl/fl}$) or 96 h (*Suz12*$^{fl/fl}$).

2.4 CpG island chromatin enrichment

2.4.1 ChIP-MS-based approach

In vivo chromatin interactions were preserved by crosslinking 100 x 10^6 mESCs with 1.5% formaldehyde (Thermo Scientific #28906) at 25°C for 10 min. The crosslinking reaction was quenched with 125 mM glycine for 5 min at 25°C. The following chromatin purification steps were performed on ice or at 4°C unless stated otherwise. Cells were centrifuged at 500 x g for 5 min, the supernatant was discarded, and the cells were resuspended in 10 mL PBS. The PBS wash was repeated once more after which the cell pellet was resuspended in 20 mL LB1 (50 mM HEPES pH 7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% IGEPAL CA-630, 0.25% Triton X-100, 1X PIC). To release nuclei, the cell suspension was incubated for 10 min in LB1 and then centrifuged at 1,000 x g for 5 min. The supernatant was discarded, and the nuclei pellet was resuspended in 20 mL LB2 (10 mM Tris-HCl pH 8, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 1X PIC). After 10 min incubation in LB2, the suspension was centrifuged at 1,000 x g for 5 min, the supernatant was discarded, and the nuclei pellet was resuspended in 10 mL LB3 (10 mM Tris-HCl pH 8, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine). To fragment

chromatin, 1.5 mL aliquots of nuclei suspension were transferred into sonication tubes containing ~200 µL sonication beads (Diagenode #C01020031). Sonication was performed in a Bioruptor Pico (Diagenode) using 30 s sonication pulses followed by 30 s rest for 30 cycles. Sheared chromatin was transferred into Eppendorf tubes, supplemented with 1% Triton X-100 (Sigma-Aldrich # X100-100ML) and cleared by centrifugation at 20,000 x g for 10 min.

CGI enrichment was performed as described previously with some alterations (Blackledge et al., 2012). To remove proteins that interact non-specifically with the affinity resin, sonicated chromatin was mixed 1:15 (v/v) with empty NeutrAvidin beads (Thermo Scientific #29200) and incubated for 2 h. The empty NeutrAvidin beads were discarded, and pre-cleared chromatin was transferred into new Eppendorf tubes. To prepare the affinity matrix, fresh NeutrAvidin beads were charged with 2 µg/µL ZF-CXXC protein. An input sample was set aside and pre-cleared chromatin was loaded on the affinity matrix and incubated for 1 h. The ZF-CXXC affinity capture reaction was centrifuged at 500 x g for 2 min and the flow-through was transferred to a new Eppendorf tube. The affinity matrix was incubated for 5 min in 1 mL CAP100 buffer (20 mM HEPES pH 7.9, 10% glycerol, 0.1% Triton X-100, 100 mM NaCl), then centrifuged at 500 x g for 2 min, and the supernatant discarded. The CAP100 wash was repeated once more. Chromatin was then eluted with CAP buffers (20 mM HEPES pH 7.9, 10% glycerol, 0.1% Triton X-100) that differ only in their NaCl concentration, which is indicated in the buffer name, for example, CAP300 contains 300 mM NaCl. To elute CGI chromatin, 200 µL CAP300 were applied on the affinity matrix and incubated for 5 min. The sample was centrifuged at 500 x g for 2 min and the eluate was transferred to a new Eppendorf tube. This elution step with CAP300 was repeated once more and both eluates were combined. Subsequent elution fractions with CAP500, CAP700, and CAP1000 were collected in the same way as CAP300. Of each elution fraction, 75 µL was set aside for quantitative PCR and the remainder was prepared for MS analysis.

2.4.2 ChIP-SICAP-based approach

Total chromatin was prepared and CGIs were enriched in the same way as described in section 2.4.1. After the CGI chromatin pulldown, the eluate fractions were split: one half was directly prepared for MS analysis, the other half was subjected to SICAP (Rafiee et al., 2016). As the CAP buffer composition is incompatible with the catalysis requirements of terminal deoxynucleotidyl transferase (TDT), the elution fractions had to be dialysed against

TDT reaction buffer (50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 0.1 mM $CoCl_2$, 0.1 mM $ZnCl_2$, adjusted to pH 7.6) before SICAP could be performed. For this, the eluates were transferred into dialysis cassettes (Thermo Scientific #66205), which were then placed into 0.4 L TDT reaction buffer and dialysed for 2 h at 4°C. The dialyses cassettes were transferred into new TDT reaction buffer and dialysed for another 2 h. The TDT reaction buffer was renewed once more and dialyses was continued overnight. The next day, eluates were recovered from the dialysis cassettes and DNA end-labelling and wash steps of the SICAP procedure were performed. To end-label the DNA with biotinylated nucleotides, a TDT reaction containing 25 µM biotin-11-ddUTP (Jena Bioscience #NU-1619-BIOX-S) and 40 U TDT enzyme (Thermo Scientific #EP0162) was set up and incubated at 37°C for 45 min. To capture biotinylated chromatin fragments, 60 µL NeutrAvidin slurry were added and incubated for 2 h at 4°C. NeutrAvidin beads with captured chromatin fragments were then wash with 3 x SDS buffer (10 mM Tris pH 8, 200 mM NaCl, 1 mM EDTA, 1% SDS), 2 x BW2x buffer (10 mM Tris pH 8, 2 M NaCl, 1 mM EDTA, 0.1% Triton X-100), 2 x 20% isopropanol, 4 x 40% acetonitrile. For each wash, 1 mL of the respective wash buffer was added, the tubes were inverted several times, beads were collected by centrifugation at 500 x g for 3 min at RT, and the supernatant was discarded. The NeutrAvidin beads were resuspended in 150 µL CAP300 and incubated overnight at 65°C to reverse the formaldehyde crosslinks. The next day, the supernatant containing free protein was transferred into new Eppendorf tubes. To elute any remaining proteins, the beads were incubated in 50 µL elution buffer (60 mM Tris-HCl pH 6.8, 4% SDS, 1 mM DTT) for 45 min at 90°C. The supernatant was transferred into new Eppendorf tubes for MS analysis.

2.4.3 ChEP-based approach

To prepare chromatin for CGI enrichment, an approach combining sucrose density barrier ultracentrifugation (Graham, 2001) and ChEP (Kustatscher et al., 2014a) was pursued. First, 100 x 10^6 mESCs were crosslinked with 1% formaldehyde at 37°C for 10 min. The reaction was quenched with 125 mM glycine for 5 min. The following steps were performed at 4°C unless stated otherwise. Cells were pelleted at 500 x g for 5 min and washed with 10 mL PBS. The cells were spun again, resuspended in 1 mL PBS, and transferred to new Falcon tubes. PBS was removed and cells were resuspended in 5 mL hypotonic lysis buffer (10 mM Tris-HCl pH 7.5, 10 mM KCl, 1 mM $MgCl_2$, 5 mM DTT, 0.2% Triton X-100, 1X PIC). After incubating the cells for 10 min, the suspension was transferred into a Dounce homogeniser and lysis was performed with 30 pestle strokes. An aliquot of the lysate was

set aside as the whole-cell lysate sample for MS analysis. The remaining lysate was centrifuged at 1,300 x g for 10 min and the nuclei pellet was washed in 2 mL sucrose buffer (20 mM Tris-HCl pH 7.5, 50 mM KCl, 10 mM $MgCl_2$, 5 mM DTT, 0.1% Triton X-100, 1X PIC, 250 mM sucrose). Following another spin, the nuclei pellet was resuspended in 1 ml sucrose buffer and the suspension was layered on top of a 4 mL sucrose density barrier (20 mM Tris-HCl pH 7.5, 50 mM KCl, 10 mM $MgCl_2$, 5 mM DTT, 1X PIC, 2 M sucrose) in ultracentrifugation tubes. To purify nuclei, the suspension was centrifuged at 20,000 rpm in a SW55Ti rotor (Beckman Coulter) for 1/2 h. The liquid was removed from the ultracentrifugation tubes and the nuclei pellet was resuspended in 0.5 mL sucrose buffer. A small aliquot of the purified nuclei was saved in a new Eppendorf tube for MS analysis. The remaining nuclei were spun at 2,300 x g for 5 min and then resuspended in 0.5 mL sarkosyl buffer (20 mM Tris-HCl pH 7.5, 10 mM EDTA, 5 mM EGTA, 4% sarkosyl) to lyse nuclei. 1.5 mL urea buffer (20 mM Tris-HCl pH 7.5, 1 mM EDTA, 1 mM EGTA, 8 M urea) was added to the chromatin solution and the tubes were centrifuged at 20,000 x g for 30 min. The supernatant was removed and the wash with sarkosyl and urea buffer was repeated twice. After the last wash, the pellet was resuspended in 2 mL sarkosyl buffer and spun again at 20,000 x g for 30 min. The chromatin pellet was dissolved in 2 mL storage buffer (10 mM Tris-HCl pH 7.5, 50 mM NaCl, 0.5 mM EDTA, 0.1% Triton X-100), centrifuged 20,000 x g for 30 min, and then resuspended again in 0.7 mL storage buffer. The chromatin solution was transferred into sonication tubes filled with ~200 µL sonication beads (Diagenode #C01020031). Sonication was performed in a Bioruptor Pico (Diagenode) using 30 s sonication pulses followed by 30 s rest for 12 cycles. Any unfragmented chromatin was collected by centrifugation at 20,000 x g for 5 min and the soluble chromatin fraction was transferred into new Eppendorf tubes. A small aliquot was set aside for MS analysis and the remainder was used for CGI enrichment as described above (see 2.4.1).

2.5 Quantitative real-time PCR

From the CGI pulldown (see 2.4.3), 75 µL of eluate and 10 µL of input chromatin were transferred into new Eppendorf tubes. Input was brought up to a volume of 75 µL with CAP100. Crosslinks were reversed overnight at 65°C. The next day, 1 µL RNase (Thermo Scientific #EN0551) was added, and samples were incubated for 2 h at 37°C. Proteins were digested with 2 µL Proteinase K (Sigma-Aldrich #P4850) for 2 h at 37°C. DNA was purified using a GeneJET Purification kit (Thermo Scientific #K0702) according to the manual. The DNA concentration was measured by Qubit dsDNA assay (Thermo Scientific #Q32854). For qPCR, input samples were diluted 1:10 and eluate samples 1:5 in 10 mM Tris-HCl pH

8.5. The qPCR reaction was set up as in Table 8 and performed with the settings in Table 9 using a Rotor-Gene Q Pure Detection thermocycler (Qiagen). Each sample was measured in duplicate.

Table 8: qPCR master mix components

Component	Volume per reaction (µL)
2X SensiFAST SYBR no-ROX (Meridian Bioscience #BIO-98005)	7.5
Primer forward [10 µM]	0.6
Primer reverse [10 µM]	0.6
Template DNA	1
ddH2O	5.3
Total	**50**

Table 9: qPCR thermocycler programme

Step	Temperature (°C)	Time (mm:ss)	Number of cycles
Initiation	95	10:00	1
Denaturation	95	00:10	40
Annealing	60	00:15	
Extension	72	00:15	

2.6 Microscopy

The quality of nuclei purified by sucrose density barrier ultracentrifugation (see 2.4.3) was assessed by fluorescence microscopy. Purified nuclei and control samples (whole cells) were washed with 500 µL Hank's balanced salt solution (HBSS) (1 mM $CaCl_2$, 0.5 mM $MgCl_2$, 0.4 mM $MgSO_4$, 5 mM KCl, 0.4 mM KH_2PO_4, 140 mM NaCl, 0.3 mM Na_2HPO_4). The samples were centrifuged at 2,300 x g for 5 min at 4°C. The supernatant was removed, and samples were incubated in staining solution (2 µL wheat germ agglutinin Alexa Fluor 488 conjugate (Invitrogen #W11261), 0.8 µL DAPI, 398 µL HBSS) at RT for 20 min in the dark. After staining, the samples were washed twice with 500 µL HBSS as above and then resuspended in 50 µL HBSS. 5 µL of the sample was mixed with 5 µL Vectashield (Vector laboratories #H-1000-10) on a microscope slide. A cover glass was placed on the sample and sealed with clear nail polish. Whole cells and nuclei were imaged on an inverted Axio

Observer Z.1 fluorescence microscope using PlanApo x63/1.4 NA oil-immersion objective and AxioVision software.

2.7 Chromatin immunoprecipitation

For calibrated ChIP performed in 6.2.4, 50 x 10^6 mESC of the untreated and tamoxifen-treated conditions were mixed with 2 x 10^6 human HEK293T cells. To carry out double crosslinking, cells were first incubated for 45 min at 25°C in 10 mL PBS containing 2 mM disuccinimidyl glutarate (Thermo Scientific #20593), followed by incubation for another 15 min with 1% formaldehyde (Thermo Scientific #28906). For non-calibrated ChIP performed in 5.2.5, cells were crosslinked using three different conditions. The first crosslinking condition was identical to the double crosslinking approach described above for calibrated ChIP but without human spike-in cells. In the second condition, 50 x 10^6 mESC were crosslinked with 1% formaldehyde for 15 min at 25°C. The third batch of cells was crosslinked in the same way as described in 2.4.3 for C-CAP. In all cases, the crosslinking reaction was quenched with 125 mM glycine for 5 min. All subsequent steps were performed at 4°C unless indicated otherwise. The crosslinking buffer was removed after pelleting the cells at 500 x g for 5 min. The cell pellet was washed with 10 mL PBS, after which the cells were incubated in 10 mL LB1 (50 mM HEPES pH 7.9, 140 mM NaCl, 1 mM EDTA, 10% glycerol, 0.5% IGEPAL CA-630, 0.25% Triton X-100, 1X PIC) for 10 min. Released nuclei were pelleted at 500 x g for 5 min and then washed in 10 mL LB2 (10 mM Tris-HCl pH 8, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 1X PIC) for 5 min. For sonication, nuclei were resuspended in 1 mL LB3 (10 mM Tris-HCl pH 8, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% sodium deoxycholate, 0.5% N-lauroylsarcosine) and transferred into sonication tubes filled with ~200 µL sonication beads (Diagenode #C01020031). Sonication was performed in a Bioruptor Pico (Diagenode) using 30 s sonication pulses followed by 30 s rest for 30-35 cycles. Sheared chromatin was transferred into Eppendorf tubes, supplemented with 1% Triton X-100 and cleared by centrifugation at 20,000 x g for 10 min.

Sonicated and cleared chromatin was diluted 1:10 in ChIP dilution buffer (20 mM Tris-HCl pH 8, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100). For pre-clearing, Protein A agarose beads (Repligen #10-2003-02) or Protein A Dynabeads (Thermo Scientific #10001D) were blocked with 1 mg/mL BSA and 1 mg/mL yeast tRNA for 1 h and then added into the diluted chromatin solution. Pre-clearing was performed for 1 h, after which 1 mL of pre-cleared chromatin was incubated overnight with the appropriate amount of antibody (Table 10). To immunoprecipitate chromatin, blocked Protein A agarose beads or Dynabeads were added

and incubated for 2 h. Beads were collected either by centrifugation (agarose beads) or on a magnetic rack (Dynabeads) and then subjected to ChIP washes. For each wash, 1 mL of the respective wash buffer was applied and incubated for 5 min with rotation. The following wash buffers were used in this order: 1 x 1 mL low salt buffer (20 mM Tris-HCl pH 8, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS, 1X PIC), 1 x 1 mL high salt buffer (20 mM Tris-HCl pH 8, 500 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS, 1X PIC), 1 x 1 mL LiCl buffer (10 mM Tris-HCl, pH 8, 250 mM LiCl, 1 mM EDTA, 1% IGEPAL CA-630, 1% sodium deoxycholate, 1X PIC), 2 x 1 mL TE (10 mM Tris-HCl pH 8, 1 mM EDTA). Chromatin was eluted with 200 µL elution buffer (100 mM $NaHCO_3$, 1% SDS) for 30 min at 65°C on a Thermomixer set to 1000 rpm. The eluate was transferred into a new Eppendorf tube and supplemented with 200 mM NaCl. Crosslinks were reversed overnight at 65°C. An input sample corresponding to 10% of the original ChIP reaction was processed in the same way. The next day, chromatin was treated with 1 µL RNase (Thermo Scientific #EN0551) for 2 h at 37°C, and 2 µL Proteinase K (Sigma-Aldrich #P4850) for another 2 h at 37°C. DNA was purified using a GeneJET Purification kit (Thermo Scientific #K0702) according to the manual. ChIP qPCR was performed and sequencing libraries were prepared as described in 2.5 and 2.9.1, respectively.

Table 10: Antibodies used in this study for ChIP

Antibody	Source	µL per ChIP
Rabbit polyclonal anti-PCGF6	LifeSpan BioSciences LS-C482495	3
Rabbit polyclonal anti-PCGF1	In-house (Blackledge et al., 2014)	5
Rabbit polyclonal anti-KDM2B	In-house (Farcas et al., 2012)	1
Rabbit polyclonal anti-CDCA7	Proteintech 15249-1-AP	8
Rabbit polyclonal anti-CDCA7L	Bethyl A300-846A	25
Rabbit polyclonal anti-KAT6B	GeneTex GTX56004	5

2.8 Mass spectrometry

2.8.1 Sample preparation

2.8.1.1 In-gel digestion of *in vitro* biotinylated ZF-CXXC sample

Protein gel bands were excised from an SDS-PAGE and cut into small pieces. The gel pieces were covered in de-staining buffer (50% ammonium bicarbonate, 50% acetonitrile) and incubated at 37°C for 15 min. The de-staining buffer was removed and the gel pieces were dehydrated in 100% acetonitrile. Acetonitrile was discarded and the gel pieces were incubated in reduction/alkylation buffer (100 mM ammonium bicarbonate, 10 mM TCEP, 50

mM 2-chloroacetamide) for 30 min at room temperature (RT). After reduction/alkylation, the gel pieces were again dried in 100% acetonitrile and then rehydrated in digestion buffer (25 mM ammonium bicarbonate, 100 ng trypsin). Proteins were digested at 37°C for 5 h. The supernatant was transferred into a new Eppendorf tube and the gel pieces were incubated in 10% formic acid for 15 min. The supernatant was extracted and combined with the first supernatant. The peptide solution was stored 4°C prior to mass spectrometry analysis.

2.8.1.2 Filter-aided sample preparation

To reduce sample loss, Protein LoBind tubes (Eppendorf) were used for all MS sample preparation steps. From the CGI pulldown (see 2.4.3), 325 µL of the eluates and 50 µL of input chromatin were used to reverse crosslinks overnight at 65°C. The next day, 1 µL Benzonase (Sigma-Aldrich #E1014-25KU) was added, and samples were incubated for 2 h at 37°C to digest nucleic acids. Protein concentration was quantified by BCA assay, and ~10 µg of protein was prepared for MS by FASP (Wiśniewski et al., 2009). All centrifugation steps with centrifugal filters were performed at 13,000 x g. Centrifugal filters (Sartorius # VN01H22) were first wetted with 300 µL UB8 (8 M urea, 100 mM ammonium bicarbonate) and spun until all liquid had passed through the filter. Samples were loaded on the centrifugal filters, centrifuged for 20 – 30 min, and the flow-through was discarded. Proteins were reduced in 100 µL reduction buffer (5 mM TCEP in UB8) for 30 min at RT. Reduced proteins were then alkylated in 100 µL alkylation buffer (80 mM 2-chloroacetamide in UB8) for 30 min at RT. Reduction and alkylation buffers were removed by centrifugation and the samples were washed six times with 400 µL UB8, followed by two washes with 400 µL UB1 (1 M urea, 25 mM ammonium bicarbonate). The filter units were transferred into new collection tubes and 100 µL digestion buffer (200 ng trypsin (Promega #V511B) in UB1) were loaded on the filters. Protein digestion was performed overnight at 30°C. To collect the peptides, the centrifugal filters were spun for 5 min, and then rinsed with 150 µL 0.1% trifluoroacetic acid (TFA) and 150 µL 50% acetonitrile in 0.1% TFA. The filter units were discarded and the flow-through containing the peptides was dried in a SpeedVac (Thermo Scientific).

2.8.1.3 Desalting

Self-made C18 Stop And Go Extraction (STAGE) tips (Rappsilber et al., 2003) were used to desalt peptide samples prior to MS data acquisition. STAGE tips were activated with 60 µL acetonitrile, and then rinsed twice with 60 µL 0.1% TFA. Dried peptides were dissolved

in MS loading solution (5% dimethylsulfoxide, 5% formic acid) and loaded on STAGE tips. STAGE tips were centrifuged at 700 x g and the flow-through was discarded. Peptides were washed twice with 60 µL 0.1% TFA and then eluted in three steps with 60 µL of 20%, 40%, and 60% acetonitrile in 0.1% TFA. The eluate was dried in a SpeedVac and peptides were stored at -20°C.

2.8.2 Data acquisition

2.8.2.1 Intact protein mass spectrometry

Purified ZF-CXXC proteins (see 2.2.2 and 2.2.3) were submitted for intact protein mass spectrometry to the Biophysics Facility in the Department of Biochemistry. The MS data was acquired on a 6530 electrospray ionisation quadrupole time-of-flight mass spectrometer (Agilent) coupled to a 1290 uHPLC system (Agilent). Deconvoluted mass spectra of the samples were provided by David Staunton of the Biophysics Facility.

2.8.2.2 Shotgun proteomics

Peptides were resuspended in 5% dimethylsulfoxide and 5% formic acid and analysed by LC-MS with an Ultimate 3000 UHPLC system (Thermo Scientific) coupled to a QExactive mass spectrometer (Thermo Scientific) through an EASY-Spray nano-electrospray ion source (Thermo Scientific). The peptides were trapped on a C18 PepMap 100 pre-column (300 mm i.d. x 5 mm, 100 Å, Thermo Scientific) using solvent A (0.1% formic acid in water). The peptides were separated on an in-house packed analytical column (75 mm i.d. x 50 cm packed with ReproSil-Pur 120 C18-AQ, 1.9 mm, 120 Å, Dr. Maisch GmbH) using a linear gradient (length: 60 min, 15% to 50% solvent B (acetonitrile with 0.1% formic acid)). Acquisition was performed in data-dependent mode (DDA). Full scan MS spectra were acquired in the Orbitrap (scan range 350–1,500 m/z, resolution 70,000, AGC target 3×10^6, maximum injection time 50 ms) followed by 10 MS/MS events at 30% NCE (resolution 17,500, AGC target 5×10^4, maximum injection time 120 ms, isolation window 1.5 m/z) with first fixed mass at 180 m/z. Charge exclusion was selected for unassigned 1+ and 2+ ions.

2.8.3 Protein identification and analysis

Thermo RAW files were searched in MaxQuant (v1.6.3.4) (Cox and Mann, 2008) against the curated UniProt (Bateman et al., 2021) mouse proteome (downloaded 07.02.2020) with

the match between runs and LFQ features enabled. The proteingroups.txt file generated by MaxQuant was loaded into Perseus (Tyanova et al., 2016). Proteins marked as reverse and contaminant were filtered out. Only proteins identified in at least two out of three replicates were used for further analysis. An appropriate normalisation method for the specific dataset was determined in R using the Normalyzer package (Willforss et al., 2019). The imputation function of Perseus was used with default parameters (width = 0.3, downshift = 1.8) to replace missing values from a normal distribution. To identify differentially enriched proteins between two samples, Welch's T-test was performed with S0 = 0.1, minimum two valid values in at least one group, permutation-based FDR, 250 randomisations, and FDR = 0.05.

2.8.4 Data visualisation

Pre-processed data from Perseus was imported into Microsoft Excel together with annotation data that was downloaded from UniProt. This annotation data was used to categorise proteins based on their subcellular localisation. Excel was used to calculate protein ranks and percentiles and to create scatter, bar, and pie charts. R packages ggplot2 and ggpubr were used to create box plots, MA plots, volcano plots, and density plots (Wickham, 2009). For statistical analysis of box plot data, rstatix function pairwise_t_test was used to perform Welch's t-test with FDR ≤ 0.05 considered significant. Venn diagrams were generated in R using the VennDiagram package (Chen and Boutros, 2011). REACTOME was used for pathway analysis and DAVID for gene ontology analysis (Huang et al., 2009; Jassal et al., 2020). Both analyses were visualised in Excel as bar charts. Network analysis and visualisation were performed in Cytoscape (v3.8.0) using stringApp (Doncheva et al., 2019; Shannon et al., 2003).

2.9 Next generation sequencing

2.9.1 Library preparation and sequencing

All next generation sequencing (NGS) experiments were performed in biological triplicate. Sequencing libraries were prepared using NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB #E7645L) according to the manual. Samples were indexed using NEBNext Multiplex Oligos. Sequencing libraries were profiled using Bioanalyzer High Sensitivity DNA Analysis Kit (Agilent #5067-4626) followed by qPCR quantification using KAPA Library Standards (Roche #07960387001). Libraries were sequenced on a NextSeq 550 (Illumina) using 40 bp paired-end reads, with the help of Amanda Williams at the sequencing facility of the Department of Zoology.

2.9.2 Read alignment and normalisation

Sequencing reads were aligned to the mm10 mouse genome (C-CAP-seq) or to the concatenated mouse mm10 and human hg19 genomes (calibrated ChIP-seq) using bowtie2 (Langmead and Salzberg, 2012). Unmapped reads, duplicate reads, multi-mappers, and reads that overlapped with a blacklist (Amemiya et al., 2019), were filtered out using sambamba (Tarasov et al., 2015) and bedtools (Quinlan and Hall, 2010). For the analysis of the rDNA gene locus in 5.2.3, sequencing reads were aligned to a single copy of the mm10 rDNA genomic sequence (GenBank: BK000964.3). All other steps were identical with C-CAP-seq, except that blacklist-filtering was omitted. C-CAP-seq samples were normalised by library size with SAMtools (Li et al., 2009) whereas calibrated ChIP-seq samples were normalised using the reads from the spike-in genome as detailed by Fursova et al. (2019). Normalised read alignments of replicates were merged with sambamba.

2.9.3 Peak calling

Peaks were called with MACS2 subcommand callpeak (Zhang et al., 2008). Depending on the dataset, different optional parameters were specified. The analysis in 6.2.1 was performed with default settings and genomic input as control. For calling peaks of histone modifications in 5.2.2, options --broad, --broad-cutoff 0.05, --fe-cutoff 2, and -q 0.05 were used. Peaks were called with options --fe-cutoff 2 and q 0.01 for the analysis in 5.2.5.

2.9.4 Differential enrichment and binding analysis

Differential enrichment of CGI chromatin in 6.2.1 was analysed in R using DiffBind with the edgeR method selected and FDR ≤ 0.05 set as the significance threshold. For calibrated ChIP-seq in 6.2.4, differential binding was analysed in R using DESeq2 with significance cut-offs set at p-adj ≤ 0.05 and fold change ±1.5 as described by Fursova et al. (2019).

2.9.5 Data visualisation

The read density at regions of interest was calculated in deeptools2 with the function computeMatrix (Ramírez et al., 2016). Heatmaps and metaplots of read densities were plotted using deeptools2 functions plotHeatmap and plotProfile, respectively. Genome tracks were visualised with Integrative Genome Viewer (Robinson et al., 2011;

Thorvaldsdóttir et al., 2013). Overlap of genomic binding sites was determined in R with ChIPpeakAnno (Zhu et al., 2010) and visualised as Venn diagrams using VennDiagram (Chen and Boutros, 2011). GimmeMotifs was used to calculate position weight matrices (PWMs) and to visualise the PWMs as DNA sequence logos (van Heeringen and Veenstra, 2011).

2.10 Published datasets used in this study

NGS datasets used in this study are listed in Table 11.

Table 11: References of published datasets used in this study

Dataset	Accession code	Reference
BioCAP-seq	GSE43512	(Long et al., 2013b)
H3K4me1	GSE36114	(Xiao et al., 2012)
H3K4me3	GSE94300	(Perino et al., 2018)
H3K9me3	GSE90893	(Chronis et al., 2017)
H3K9ac	GSE90893	(Chronis et al., 2017)
H3K27me3	GSE94300	(Perino et al., 2018)
H3K27ac	GSE36114	(Xiao et al., 2012)
H3K36me3	GSE36114	(Xiao et al., 2012)
PRC1 CKO RNA-seq	GSE119620	(Fursova et al., 2019)
Mouse genes	NA	(Rose et al., 2016)

3 Producing a ZF-CXXC bait protein to enrich CpG island chromatin

3.1 Introduction

CpG islands (CGIs) are genomic regions in vertebrates that contain clusters of non-methylated CG dinucleotides which are specifically recognised by the zinc finger (ZF)-CXXC domain (Illingworth & Bird, 2009; Xu et al., 2018). The affinity of the ZF-CXXC domain for non-methylated CpG has been exploited to develop approaches that allow non-methylated CGI DNA to be specifically isolated from the genome (Blackledge et al., 2012; Illingworth et al., 2008). The most advanced of these approaches, called BioCAP, relies on a biotin-tagged ZF-CXXC domain being immobilised on a resin to create an affinity matrix. When incubated with genomic DNA, this matrix retains non-methylated DNA and separates it from methylated and hydroxymethylated DNA (Blackledge et al., 2012). Isolated CGI DNA can then be used to determine the location and methylation state of CGIs in the genome using next generation sequencing.

The goal of my DPhil project is to develop a new BioCAP-based approach to specifically isolate CGI chromatin and unbiasedly characterise the CGI proteome. In order to achieve this, I first needed to generate a CGI affinity matrix. This required that I produce large amounts of highly pure recombinant biotin-tagged ZF-CXXC domain. In this chapter, I will detail an optimised approach I have developed to efficiently produce the biotinylated ZF-CXXC domain which I will use in subsequent chapters to develop my CGI chromatin enrichment strategy.

3.2 Results

3.2.1 Expression and purification of BirA enzyme for use in biotinylation reactions of ZF-CXXC

To produce biotinylated ZF-CXXC, I initially followed the protocol published by Blackledge et al. (2012). They describe a 3-step process in which they first purify His-tagged ZF-CXXC, then remove its His-tag by tobacco etch virus (TEV) protease cleavage, and finally biotinylate the Avi-tag of ZF-CXXC using BirA biotin ligase (Figure 4).

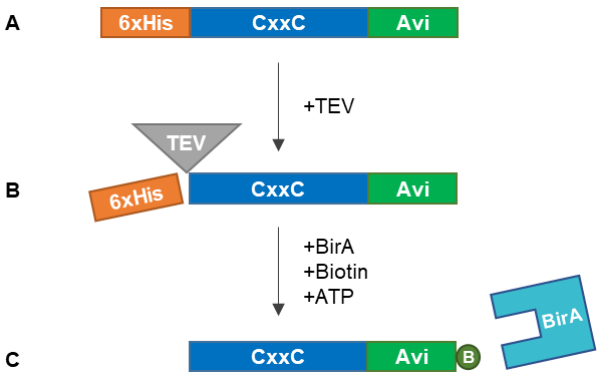


Figure 4: Schematic workflow for producing *in vitro* biotinylated ZF-CXXC
(A) 6xHis-ZF-CXXC is expressed and purified from *E. coli*.
(B) 6xHis-ZF-CXXC is treated with TEV protease to remove the N-terminal 6xHis-tag.
(C) The Avi-tag on the C-terminus is biotinylated by BirA biotin ligase. The B in the green circle denotes a biotin molecule.

To produce biotinylated ZF-CXXC domain, I first needed to obtain large amounts of active recombinant TEV protease and BirA biotin ligase. Recombinant TEV protease was already available in the Klose laboratory, so I initially focussed on producing the BirA protein for *in vitro* biotinylation reactions. I expressed His-tagged BirA in *E. coli* and affinity-purified using nickel nitriloacetic (Ni-NTA) agarose resin. SDS-PAGE analysis showed that the eluates from the Ni-NTA purification contained reasonably pure BirA, but that there were also some minor impurities (Figure 5A). To remove these impurities, the elution fractions were pooled and further purified by size exclusion chromatography (SEC). The peak fractions from SEC contained highly pure BirA protein which was sufficient in quantity for future *in vitro* biotinylation reactions (Figure 5B).

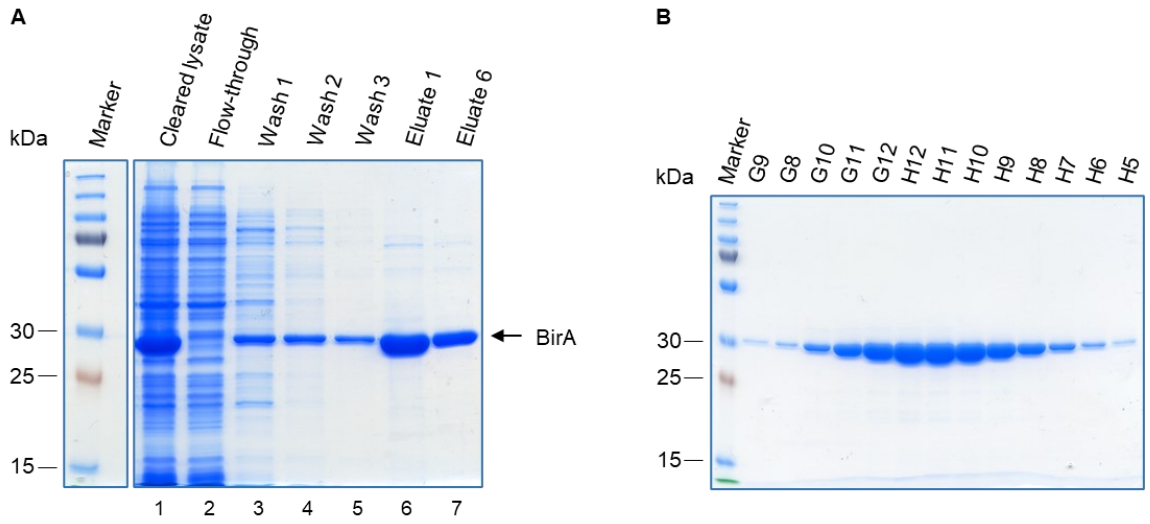


Figure 5: SDS-PAGE analysis of BirA purification by affinity and size exclusion chromatography

(A) Coomassie-stained SDS-PAGE showing the affinity purification of His-tagged BirA. The cleared lysate reveals a prominent band at approximately 30 kDa that corresponds to His-BirA (lane 1). The cleared lysate was loaded on Ni-NTA resin. BirA was efficiently retained on the affinity matrix as evidenced by its depletion in the flow-through (lane 2). Some BirA was lost in the wash fractions (lanes 3 – 5), but a sufficient amount is recovered in the eluates (lanes 6 and 7).
(B) Coomassie-stained SDS-PAGE showing size exclusion chromatography of BirA. Affinity-purified His-BirA from (A) was loaded on a size exclusion chromatography column and BirA protein was collected.

3.2.2 Expression and purification of the ZF-CXXC domain

Next, I needed to express and purify the ZF-CXXC domain. This domain corresponds to the ZF-CXXC and PHD domains of human KDM2B. The ZF-CXXC domain recognises non-methylated CG dinucleotides and the PHD domain is included as it improves the stability of the purified protein (Xu et al., 2018 and personal communication Rob Klose). This protein fragment will be referred to as the ZF-CXXC from this point forward. To enable purification, the ZF-CXXC fragment includes a His-tag at the N-terminus and a TEV protease cleavage site to facilitate His-tag removal after purification. Removal of the His-tag was previously shown to improve binding of the purified ZF-CXXC to non-methylated DNA (Blackledge et al. 2012). Furthermore, the protein has a C-terminal Avi-tag which is biotinylated by the BirA enzyme to enable conjugation to NeutrAvidin resin.

The His-TEV-ZF-CXXC-Avi-tag fragment was first expressed in *E. coli* and then purified via Ni-NTA affinity chromatography. This yielded a prominent band on SDS-PAGE corresponding to the His-tagged protein (Figure 6A). To remove the His-tag, the eluates were pooled and incubated with His-tagged TEV protease. After TEV cleavage, the reaction

was reapplied to Ni-NTA resin to capture both the cleaved His-tag and His-tagged TEV, while allowing the ZF-CXXC-Avi-tag fragment to continue into the flow-through (Figure 6B).

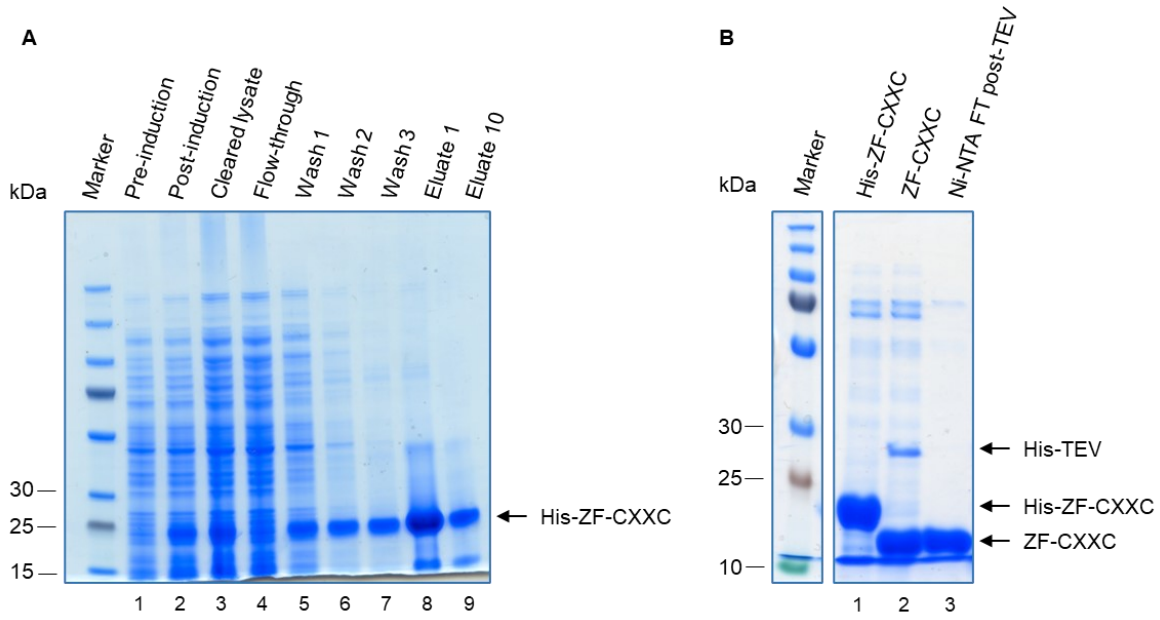


Figure 6: SDS-PAGE analysis of ZF-CXXC affinity purification and TEV cleavage
(A) Coomassie-stained SDS-PAGE showing the affinity purification of His-tagged ZF-CXXC. After induction of expression (compare lane 1 and 2), a new band of approximately 21 kDa is evident that corresponds to His-ZF-CXXC and this is retained in the cleared lysate (lane 3). The cleared lysate was loaded on Ni-NTA resin. His-ZF-CXXC was depleted from the flow-through (lane 4). Some His-ZF-CXXC domain was lost in the wash steps (lanes 5 – 7), but sufficient amounts of protein were present in the elution fractions (lanes 8 and 9).
(B) Coomassie-stained SDS-PAGE showing the cleavage of His-ZF-CXXC by TEV protease. To cleave the His-tag from the purified ZF-CXXC, His-TEV protease was added, and efficient cleavage was evident from the reduction in molecular weight of the His-ZF-CXXC fragment (compare lanes 1 & 2). The TEV cleavage reaction was applied on Ni-NTA resin to remove the His-TEV protease and cleaved His-tag and the flow-through containing ZF-CXXC was collected. This yielded a reasonably pure ZF-CXXC domain lacking the His-tag (lane 3).

3.2.3 *In vitro* biotinylation of ZF-CXXC by BirA is inefficient and causes protein precipitation

In order to immobilise the ZF-CXXC domain on NeutrAvidin beads and create the CGI affinity matrix, its C-terminal Avi-tag must be biotinylated. To achieve this, I incubated the purified ZF-CXXC-Avi-tag protein with purified BirA biotin ligase and a biomix that contains the necessary cofactors (biotin, ATP, Mg^{2+}) for *in vitro* biotinylation. Unexpectedly, a large amount of precipitate formed during the biotinylation reaction. To determine the nature of the precipitate, I collected it by centrifugation and compared it to the soluble material by SDS-PAGE. This revealed that there was very little protein in the soluble fraction and that most of the protein had precipitated (Figure 7A).

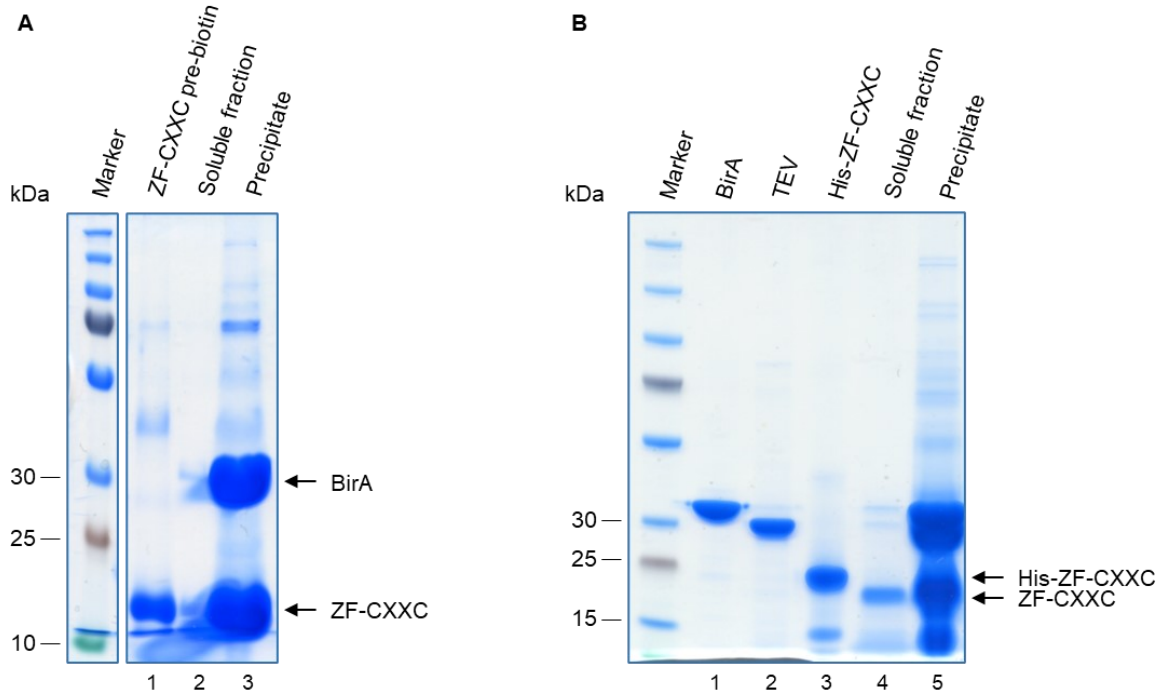


Figure 7: SDS-PAGE analysis of protein precipitation during *in vitro* biotinylation of ZF-CXXC
(A) Coomassie-stained SDS-PAGE of the biotinylation reaction of ZF-CXXC. The purified ZF-CXXC-Avi-tag protein (lane 1) was incubated with BirA in order to biotinylate its Avi-tag. During the *in vitro* biotinylation reaction, both ZF-CXXC and BirA precipitated (lane 3) and only a small amount of protein remained in the soluble fraction of the reaction (lane 2).
(B) Coomassie-stained SDS-PAGE of the biotinylation reaction of ZF-CXXC. The *in vitro* biotinylation of ZF-CXXC-Avi-tag protein was repeated after adjusting the reaction buffer to pH 8. The correct pH improved the solubility of the proteins only slightly and most protein formed again a precipitate (compare lanes 4 and 5).

In an attempt to overcome this precipitation, I examined what aspect of the reaction conditions could result in this outcome. One possible explanation was the pH of the reaction. I noted that while the reaction buffer is designed to buffer at pH 8, the addition of biomix caused the pH to drop to 5.5. I repeated the biotinylation reaction after I confirmed that the reaction buffer including biomix was set to pH 8. Despite the correct pH, a protein precipitate formed again, albeit to a lower degree than before (Figure 7B). The repeated protein precipitation prompted me to systematically test different buffer compositions. I found that using bicine instead of Tris as the buffering substance, and adding a small amount of non-ionic detergent and glycerol largely overcame the precipitation issue.

After optimising this step, I continued with the purification of ZF-CXXC protein by size exclusion chromatography (SEC). SDS-PAGE analysis of the SEC eluates revealed that two prominant contaminants co-purified with the ZF-CXXC protein (Figure 8). The faint

band at ca. 30 kDa is likely BirA that was carried over from the biotinylation reaction. The second contaminant at approximately 12 kDa was cut out from the gel and analysed by MS, which indicated that this contaminant is likely a 30S ribosomal protein from *E. coli*.

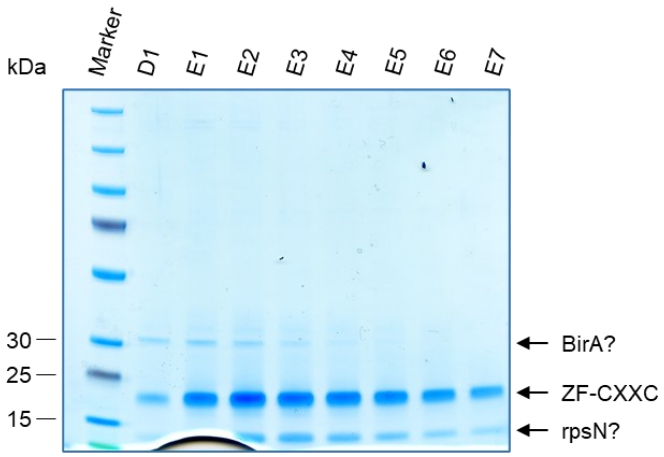


Figure 8: SDS-PAGE analysis of *in vitro* biotinylated ZF-CXXC size exclusion chromatography
Coomassie-stained SDS-PAGE showing size exclusion chromatography of ZF-CXXC. After *in vitro* biotinylation, the ZF-CXXC protein was further purified by size exclusion chromatography. The faint band at ca. 30 kDa is likely BirA that was carried over from the biotinylation reaction. In-gel digestion followed by MS analysis suggested that the band at ca. 12 kDa could be 30S ribosomal protein rpsN from *E. coli*.

The pooled SEC eluate was subjected to intact protein analysis by mass spectrometry in order to assess the biotinylation efficiency. The expected mass of the ZF-CXXC without 6xHis-tag and biotinylation is 18938.94 Da. One biotin molecule adds 226.30 Da, which brings the theoretical mass of biotinylated ZF-CXXC to 19165.24 Da. Unexpectedly, mass spectrometry detected a protein of 18562.98 Da and nothing at the expected mass of 19165.24 Da (Figure 9). This suggested that the sample did not contain a full-length biotinylated ZF-CXXC protein. Therefore, despite my optimisation, I was unable to produce protein of sufficient quality using the published approach to warrant developing my CGI chromatin enrichment method using this reagent.

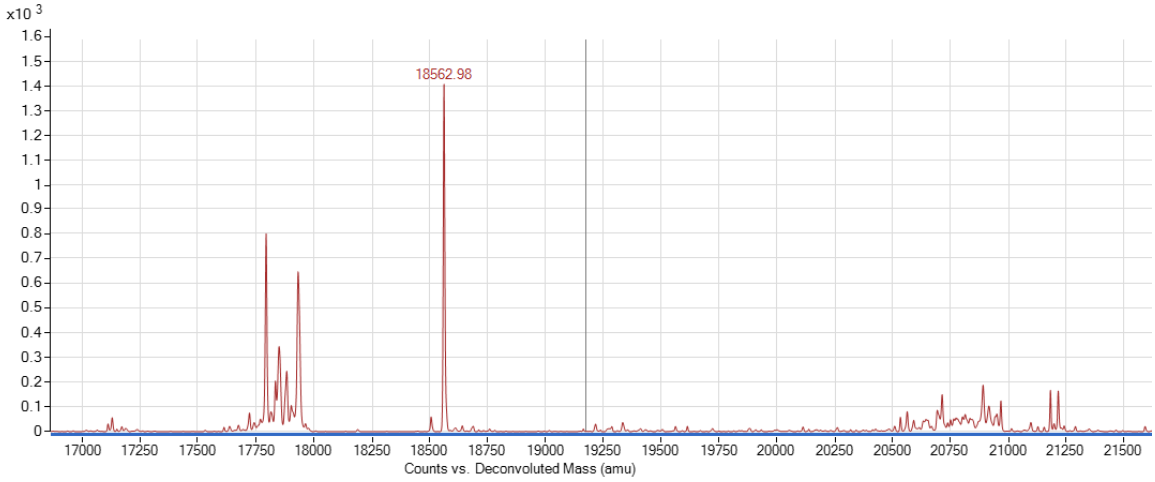


Figure 9: Intact protein mass spectrometry of *in vitro* biotinylated ZF-CXXC
Deconvoluted mass spectrum of purified and biotinylated ZF-CXXC. The main peak shows a protein at 18562.98 Da, which does not meet the expected mass of full-length biotinylated ZF-CXXC.

3.2.4 A new approach to isolate biotinylated ZF-CXXC protein

The inability to produce high yields of intact and biotinylated His-ZF-CXXC-Avi-tag protein prompted me to rethink my strategy. A key bottleneck for producing this reagent proved to be the *in vitro* biotinylation step which was subject to precipitation. The BirA enzyme is naturally expressed in *E. coli*, but the endogenous expression levels are not sufficient to achieve quantitative biotinylation of overexpressed proteins (Sibler et al., 1999). Therefore, I reasoned that quantitative *in vivo* biotinylation could possibly be achieved by overexpression of BirA together with the ZF-CXXC protein as has been demonstrated previously for other proteins (Smith et al., 1998). This could in theory circumvent the issues I had encountered during *in vitro* biotinylation and allow me to simply purify biotinylated ZF-CXXC protein directly from cells. Thus, I cloned wild-type BirA downstream of the His-ZF-CXXC-Avi-tag coding sequence, including a ribosomal binding site for each open reading frame (Figure 10A).

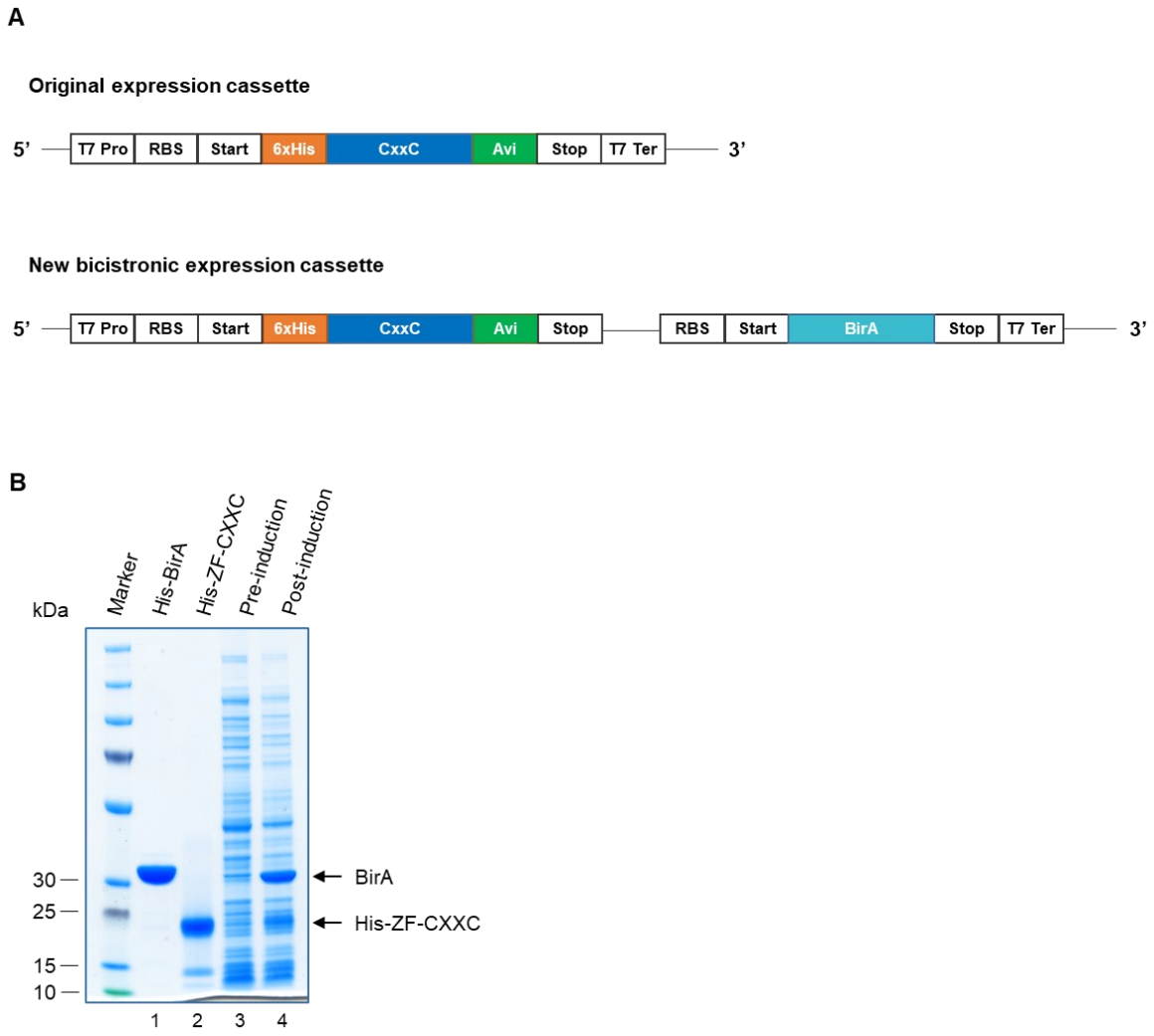


Figure 10: Co-expression of BirA with ZF-CXXC for *in vivo* biotinylation
(A) Schematic depicting the original and new expression cassette of ZF-CXXC. T7 Pro (T7 RNA polymerase promoter), T7 Ter (T7 RNA polymerase terminator), RBS (ribosomal binding site).
(B) Coomassie-stained SDS-PAGE analysis of the test expression with the new bicistronic vector. After induction of protein expression, two new bands appear that correspond to BirA and His-ZF-CXXC (compare lanes 3 and 4). The molecular weight of purified His-BirA (lane 1) is slightly higher due to the His-tag, which is absent on BirA in the bicistronic vector (lane 4).

First, I performed a small-scale test expression with the new bicistronic vector. Expression of both proteins was induced successfully, with BirA being expressed at higher levels than His-ZF-CXXC-Avi-tag protein (Figure 10B). This is not surprising considering that BirA is a bacterial enzyme while ZF-CXXC evolved in a eukaryotic system and might not be optimally transcribed/translated in *E. coli*. Having shown that both proteins were expressed after induction, I performed a large-scale expression and supplemented the culture media with biotin to support the function of BirA.

The previous purification of *in vitro* biotinylated ZF-CXXC had shown considerable contamination with co-purifying proteins (Figure 6). Therefore, in order to purify the His-ZF-CXXC-Avi-tag protein from my large-scale expression, I introduced two changes to the protocol aimed at increasing the purity of the protein. I performed the affinity purification on a HisTrap column connected to a chromatography system and I increased the stringency of the buffers to reduce non-specific protein binding. This raised the total yield of purified protein and considerably improved the purity as shown by SDS-PAGE (Figure 11).

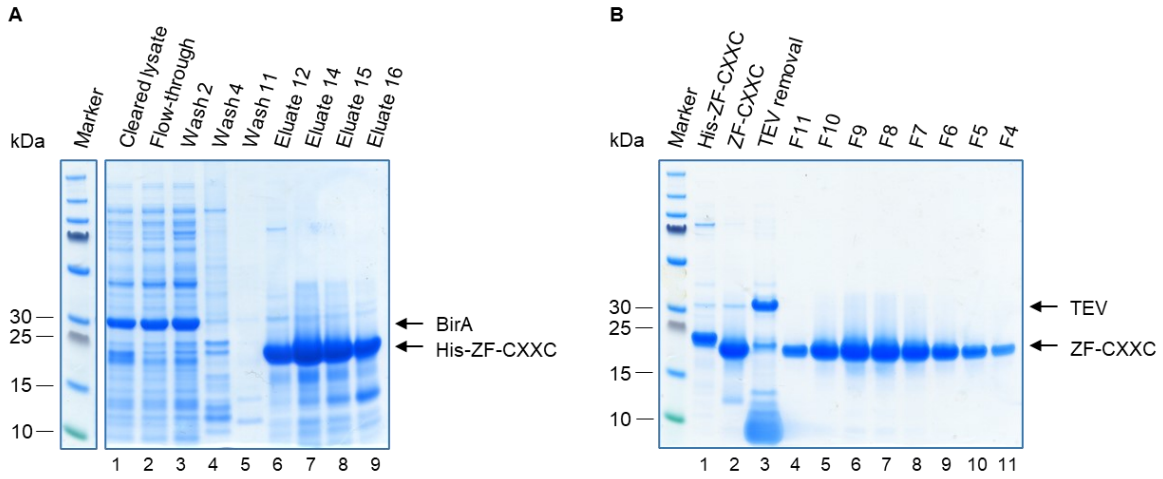


Figure 11: SDS-PAGE analysis of *in vivo* biotinylated ZF-CXXC purification by affinity and size exclusion chromatography
(A) Coomassie-stained SDS-PAGE showing the affinity purification of His-tagged ZF-CXXC. The cleared lysate was loaded on a HisTrap column for affinity purification. The band at ca. 21 kDa corresponding to His-ZF-CXXC is largely absent in the flow-through which indicates that His-ZF-CXXC was captured by the affinity matrix (compare lanes 1 and 2). BirA (ca. 30 kDa) does not have a His-tag and was therefore removed in the flow-through and wash fractions (compare lanes 1 – 5). His-ZF-CXXC protein was eluted from the column (lanes 6 – 9) and eluates were pooled.
(B) Coomassie-stained SDS-PAGE showing TEV protease cleavage and size exclusion chromatography of His-tagged ZF-CXXC. The pooled eluate from (A) was treated with His-TEV protease to remove the His-tag from His-ZF-CXXC. Efficient cleavage is reflected in the downshift in molecular weight of the His-ZF-CXXC protein (compare lanes 1 and 2). The TEV cleavage reaction was loaded on the HisTrap column, which retained the His-TEV protease and cleaved His-tag (lane 3). The flow-through containing ZF-CXXC was further purified by size exclusion chromatography. This yielded highly pure ZF-CXXC protein (lanes 4 – 11).

Intact protein mass spectrometry was performed to confirm that the purified protein was indeed biotinylated during expression. The result showed a major peak at 19166.17 Da, which fits the theoretical mass of 19165.24 Da of biotinylated ZF-CXXC when considering the mass error tolerance of the mass spectrometer (Figure 12). Furthermore, mass spectrometry did not detect any other abundant masses that would indicate non-biotinylated or degraded forms of ZF-CXXC. In summary, the new *in vivo* biotinylation approach yielded

highly pure and efficiently biotinylated ZF-CXXC protein for developing my CGI chromatin enrichment approach.

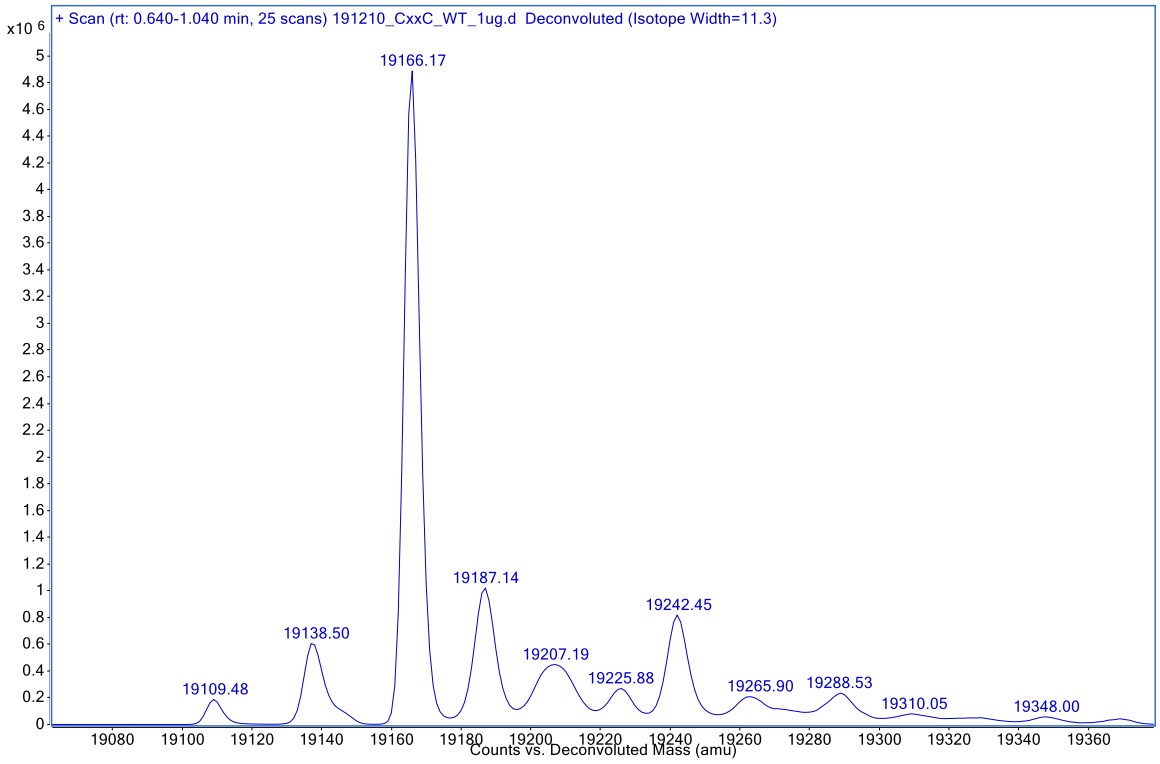


Figure 12: Intact protein mass spectrometry of *in vivo* biotinylated ZF-CXXC after improving the purification process
Deconvoluted mass spectrum of purified and biotinylated ZF-CXXC. The main peak at 19166.17 Da shows biotinylated full-length ZF-CXXC. Some less abundant species are also present, e.g. a potential sodium adduct (+21 Da) at 19187.14 Da.

3.2.5 Expression and purification of a ZF-CXXC domain with mutations in its CpG recognition site

The ultimate goal of using a ZF-CXXC domain for enrichment of non-methylated DNA is to investigate the proteome of CpG islands by mass spectrometry. However, the problem inherent to virtually all pulldown experiments is to differentiate target enrichment from non-specific background. The most common reasons for contaminating proteins in these approaches are high cellular abundance and non-specific interaction with the bait protein or the support matrix (Mellacheruvu et al., 2013). Therefore, it is essential to include controls in the experimental design to exclude the contaminating proteins in downstream analysis.

I reasoned that a bait protein which is similar to the ZF-CXXC domain, but deficient in recognising unmethylated CpG could help me to discern the CGI proteome from background proteins. The crystal structure of the human KDM2B ZF-CXXC domain

indicates that the residues responsible for CpG recognition are lysine 643 (K643) and glutamine 644 (Q644) (Xu et al., 2018). Both residues are conserved in proteins with a type-1 ZF-CXXC domain (Long et al., 2013a). Mutations of these residues have been shown to abrogate binding to non-methylated DNA without affecting the global fold of the ZF-CXXC domain (Allen et al., 2006a; Cierpicki et al., 2010; Xu et al., 2011). Therefore, I mutated both residues to alanines using site-directed mutagenesis. To produce mutant ZF-CXXC (ZF-CXXC^{mt}) protein, I used the same *in vivo* biotinylation and purification approach as described above (3.2.4).

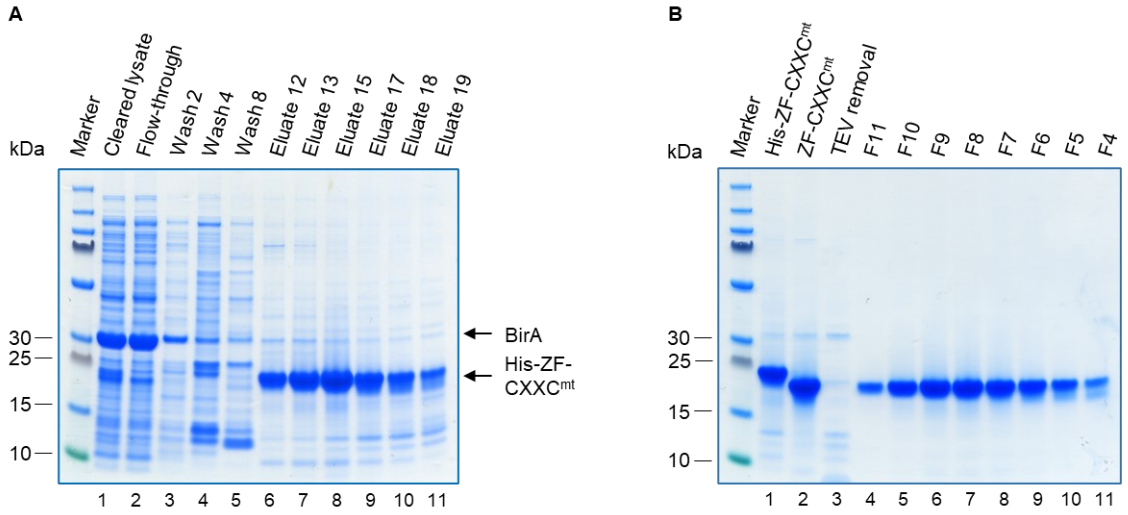


Figure 13: SDS-PAGE analysis of ZF-CXXC^{mt} purification by affinity and size exclusion chromatography

(A) Coomassie-stained SDS-PAGE showing the affinity purification of His-tagged ZF-CXXC^{mt}. The cleared lysate was loaded on a HisTrap column for affinity purification. The His-ZF-CXXC^{mt} protein at ca. 21 kDa was retained by the affinity matrix as evidenced by its depletion in the flow-through (compare lanes 1 and 2). Most of the untagged BirA protein (ca. 30 kDa) was efficiently removed in the flow-through and wash fractions (lanes 1 – 5). Eluate fractions containing His-tagged ZF-CXXC^{mt} were collected (lanes 6 – 11) and pooled for further processing.

(B) Coomassie-stained SDS-PAGE showing TEV protease cleavage and size exclusion chromatography of His-tagged ZF-CXXC^{mt}. His-TEV protease was added to the eluate from (A) to remove the His-tag of ZF-CXXC^{mt}. The His-tag was quantitatively cleaved as indicated by the reduction in molecular weight of the ZF-CXXC^{mt} protein (compare lanes 1 and 2). His-TEV protease and the cleaved His-tag were removed from the reaction by passing it through a HisTrap column (lane 3). The flow-through of the HisTrap contained ZF-CXXC^{mt} without His-tag, which was then further purified by size exclusion chromatography. Highly pure ZF-CXXC^{mt} was obtained (lanes 4 – 11) and frozen for storage.

SDS-PAGE analyses showed efficient affinity purification of the ZF-CXXC^{mt} protein and subsequent SEC further improved purity (Figure 13). Finally, I confirmed biotinylation of the purified protein by intact mass spectrometry. The theoretical mass of biotinylated ZF-CXXC^{mt} with two alanine substitutions is 19051.09 Da, which corresponds to the observed

mass of 19051.18 Da (Figure 14) that is well within the measurement error of the instrument. Moreover, no other protein species of appreciable amount were detected.

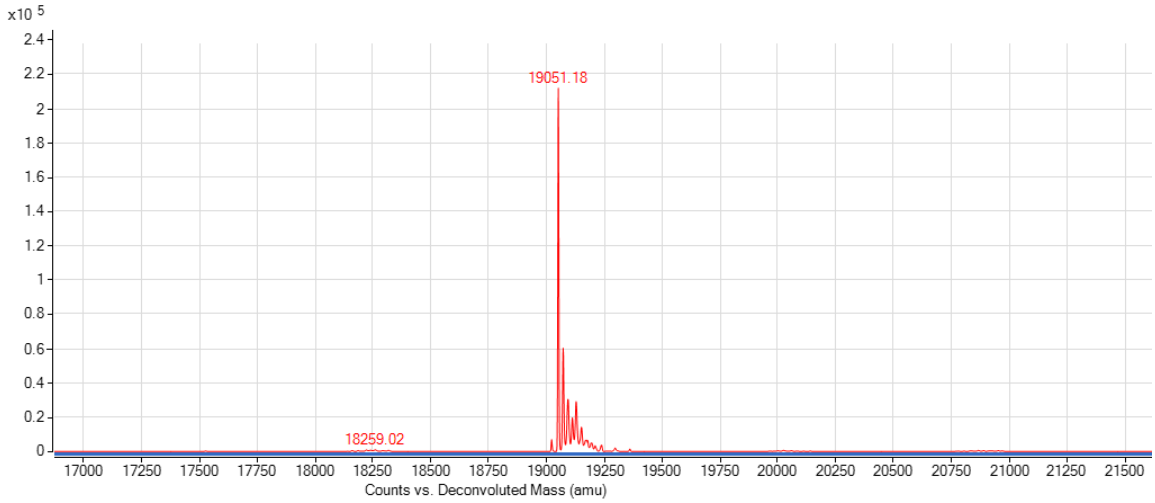


Figure 14: Intact protein mass spectrometry of *in vivo* biotinylated ZF-CXXC^{mt}
Deconvoluted mass spectrum of purified and biotinylated ZF-CXXC^{mt}. The peak at 19051.18 Da shows biotinylated ZF-CXXC with K643A and Q644A point mutations.

3.3 Summary

In order to develop a technique to purify CGI chromatin, I first needed to produce the reagents what would enable this. My initial attempts to produce the ZF-CXXC protein followed the published approach (Blackledge et al., 2012). However, I encountered multiple issues primarily related to the *in vitro* biotinylation step. These issues prompted me to revise my strategy to produce this essential reagent. To overcome the bottleneck of *in vitro* biotinylation, I decided to, instead, biotinylate ZF-CXXC *in vivo* by co-expressing it with BirA enzyme in *E. coli*. Subsequently, I developed a purification approach that enabled large-scale isolation of highly pure and quantitatively biotinylated ZF-CXXC protein. I then produced a ZF-CXXC domain with mutations in the two residues that are responsible for recognition of non-methylated CpG. The purpose of the mutant ZF-CXXC is to serve as a control in the CGI enrichment assay, as it could facilitate the differentiation between specific interactions and non-specific binding. The mutant ZF-CXXC was quantitatively biotinylated and obtained at high purity following the same protocol used for the wild-type ZF-CXXC. In the next chapter, I describe how I developed an approach for chromatin purification and how I utilised the ZF-CXXC domain to enrich CGI chromatin.

4 Development of a chromatin purification approach to enrich CGIs

4.1 Introduction

The identification of the proteins that interact with CGIs is essential for understanding their function. Many of the known factors that bind specifically to CGIs, were first shown to associate with non-methylated DNA or chromatin based on *in vitro* biochemical assays, and subsequently chromatin immunoprecipitation sequencing (ChIP-seq) demonstrated that they localise to CGIs *in vivo* (Blackledge et al., 2010; Farcas et al., 2012; Grand et al., 2021; Stielow et al., 2021; Thomson et al., 2010; Voo et al., 2000b). While these discoveries have been important for understanding CGI function, they required *a priori* knowledge that a factor might have specificity for CGIs, and they were limited by ChIP-seq analysis which can only investigate one protein at a time.

In contrast, mass spectrometry-based approaches can identify in an unbiased and comprehensive manner proteins that associate with chromatin. However, on its own, MS lacks the capacity to interrogate specific regions of the genome, such as CGIs. MS has been combined with ChIP to explore the proteome around the genomic binding sites of the immunoprecipitated protein (Engelen et al., 2015; Ji et al., 2015). However, in most cases CGI-binding factors are enriched at subsets of CGIs, limiting the utility of ChIP-MS for producing an unbiased view of proteins that associate with CGIs. There are also other methods to isolate defined regions of chromatin, but none of these are tailored to CGIs (reviewed in Gauchier et al., 2020)

A particular challenge for mass spectrometric investigation of specific genomic regions, like CGIs, is that they often account for only a small percentage (~1% for CGIs) of the genome (Bernstein et al., 2007; Lander et al., 2001). In order to identify the CGI proteome, other genomic regions must be efficiently removed, and non-specific proteins need to be limited such that they do not mask the CGI-associated protein signal. These challenges must be considered in developing any methodology for identifying the CGI proteome by MS.

In this chapter, I describe an approach I have developed to purify CGIs from cells and analyse their protein composition by mass spectrometry.

4.2 Results

4.2.1 CpG islands isolated with a ChIP-based approach are contaminated with non-chromatin proteins

To enrich CGI chromatin from cells, my initial approach was loosely based on a ChIP-MS protocol. In ChIP-MS, a conventional ChIP is performed to enrich a chromatin-bound protein with an antibody. Subsequently, the immunoprecipitated material is subjected to mass spectrometry analysis to identify proteins that are co-bound to the enriched chromatin fragments. Several studies have shown that ChIP-MS can identify proteins bound to chromatin containing specific histone modifications (Engelen et al., 2015; Ji et al., 2015; Soldi and Bonaldi, 2013). I reasoned that I might be able to use a similar approach to isolate CGI chromatin by replacing the antibody used for ChIP with the recombinant ZF-CXXC domain that can bind specifically to non-methylated DNA. This would enable me to probe all non-methylated CGIs and not just a subset as it would be the case if I targeted a CGI-associated protein with an antibody.

For developing my new CGI enrichment method, I chose mouse embryonic stem cells (mESCs) as my model system for mainly two reasons. First, embryonic stem cells have an open and accessible chromatin architecture with more CGIs being in an unmethylated state compared to differentiated cells (Efroni et al., 2008; Meshorer et al., 2006; Stadler et al., 2011; Xie et al., 2013). The better accessibility of chromatin and CGIs could potentially facilitate binding of the ZF-CXXC domain to non-methylated islands, thereby improving their enrichment. Second, many promoter- and CGI-binding proteins, such as Polycomb group (PcG) and Trithorax group (TrxG) complexes, have been well characterised in embryonic stem cells (Aloia et al., 2013; Engelen et al., 2015; Ji et al., 2015; Villaseñor et al., 2020). The existing knowledge about the promoter- and CGI-associated factors will serve as a resource to validate that my new approach is enriching CGI chromatin.

To test this approach, first I crosslinked cells with formaldehyde to fix proteins on their chromatin binding sites *in vivo* (Figure 15a). This was followed by lysing the cells and isolating the nuclei, which were then sonicated to fragment chromatin (Figure 15b-c). Subsequently, fragmented chromatin was used as the input for BioCAP to capture CGIs. The enrichment of CGIs from chromatin will be referred to as CGI-Chromatin Affinity Purification (C-CAP) from this point forward in order to distinguish it from BioCAP, which enriches CGIs from naked DNA. The CGI-containing chromatin fragments were eluted off the ZF-CXXC affinity matrix with a stepwise increasing salt gradient (Figure 15d). The higher the density of non-methylated CpG on the fragments, the stronger they bind to the affinity matrix and the later they elute (Blackledge et al., 2012). To determine whether this

approach was effective at enriching CGI chromatin, I performed qPCR analysis on the DNA from the elution fractions. For this, I compared a genomic region that does have a CGI (*Ncoa2* promoter) with genomic regions that lack CGIs (gene bodies and the non-CGI *Fgf7* promoter). This showed that CGI chromatin is enriched in the 500, 700, and 1000 mM NaCl elution fractions (Figure 16, *Ncoa2* promoter). In contrast, non-CGI chromatin makes up only a minor fraction of the eluates (< 1%) as evidenced by lack of qPCR signal for gene bodies and the *Fgf7* promoter.

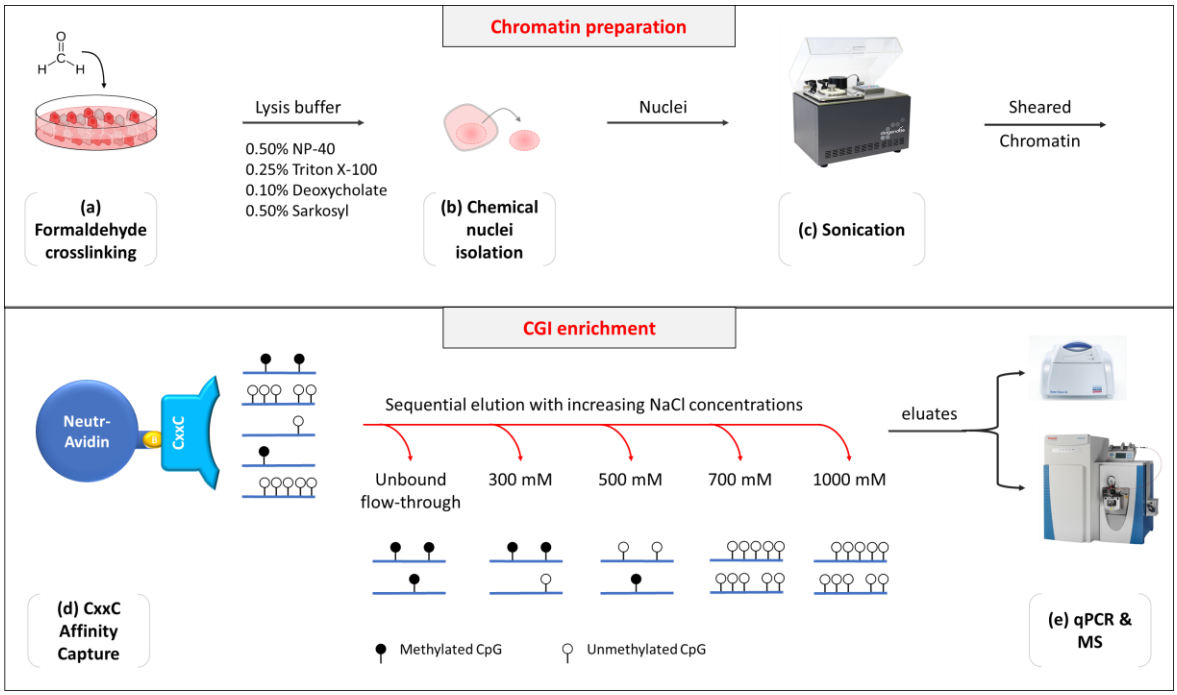


Figure 15: A CGI chromatin enrichment approach combining ChIP and BioCAP
In established ChIP procedures, cells are crosslinked with formaldehyde (a) and then lysed with a combination of various detergents (a – b). Intact nuclei are collected from the lysate by centrifugation and sonicated to fragment chromatin (b – c). To enrich CGIs, this chromatin is then loaded on the ZF-CXXC affinity matrix (d). Chromatin fragments with methylated CpG are removed in the flow-through and washes, while fragments with unmethylated CpG are retained by the ZF-CXXC domain. These are then eluted with a stepwise salt gradient. Finally, eluates are prepared for qPCR and MS (e).

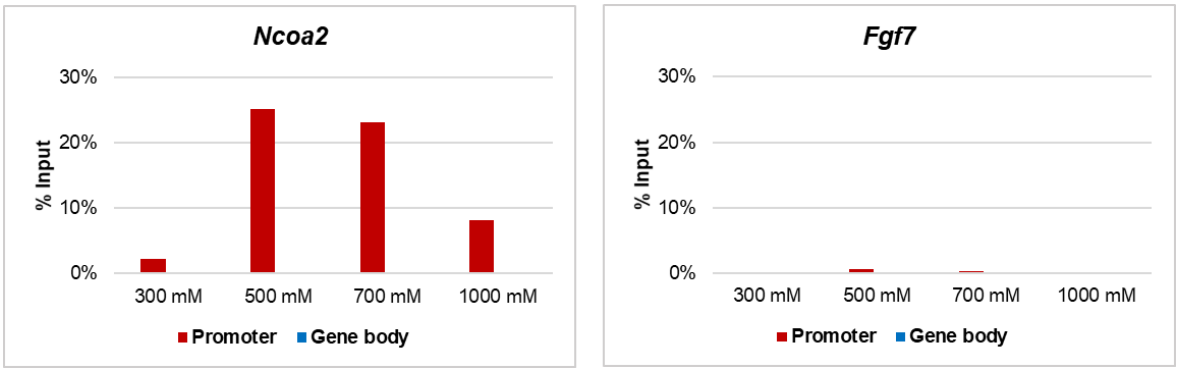


Figure 16: Quantitative PCR analysis of C-CAP elution fractions
Bar charts showing qPCR analysis of C-CAP eluates. DNA was purified from C-CAP eluates and CGI enrichment was evaluated by comparing amplification at promoter and gene body regions of *Ncoa2* and *Fgf7*. Enrichment is expressed as % of Input DNA. Gene body bars are not visible because the enrichment is below 1%.

After confirming enrichment of CGIs in the elution fractions, I prepared the samples for mass spectrometry using the filter-aided sample preparation (FASP) method (Wiśniewski et al., 2009). However, it proved to be difficult to remove the detergents introduced to the samples in the cell lysis step (Figure 15a-b). Despite extensive urea washes, some residual detergent remained in the input sample and blocked the analytical column on the liquid chromatography (LC) system attached to the mass spectrometer. Therefore, the mass spectrometry analysis of this experiment includes only the elution fractions. From the eluates, the 500-1000 mM fractions were combined in the MaxQuant search as they showed the highest enrichment of CGIs based on qPCR analysis (Figure 16).

The goal of the C-CAP approach outlined in Figure 15 is to remove contaminating proteins and to specifically isolate CGI-associated factors. MS analysis identified in total 840 proteins in the 500 – 1000 mM elution fractions of the ZF-CXXC affinity purification. To evaluate how efficiently contaminant proteins were removed in these eluates, I categorised proteins according to their subcellular localisation based on information from the UniProt database. Then, I calculated the number of proteins that localise to the nucleus, cytoplasm, cell membrane, and extracellular space. Of the 840 proteins identified in the eluates, 63% have a nuclear localisation and 19% are non-nuclear (sum of cytoplasm, cell membrane, extracellular space) (Figure 17A). Next, I checked these populations with respect to protein abundance using the crude proxy of reported intensities. This showed that the normalised log2 intensities of nuclear and chromatin-associated proteins are not significantly higher than those of the non-nuclear proteins (Figure 17B).

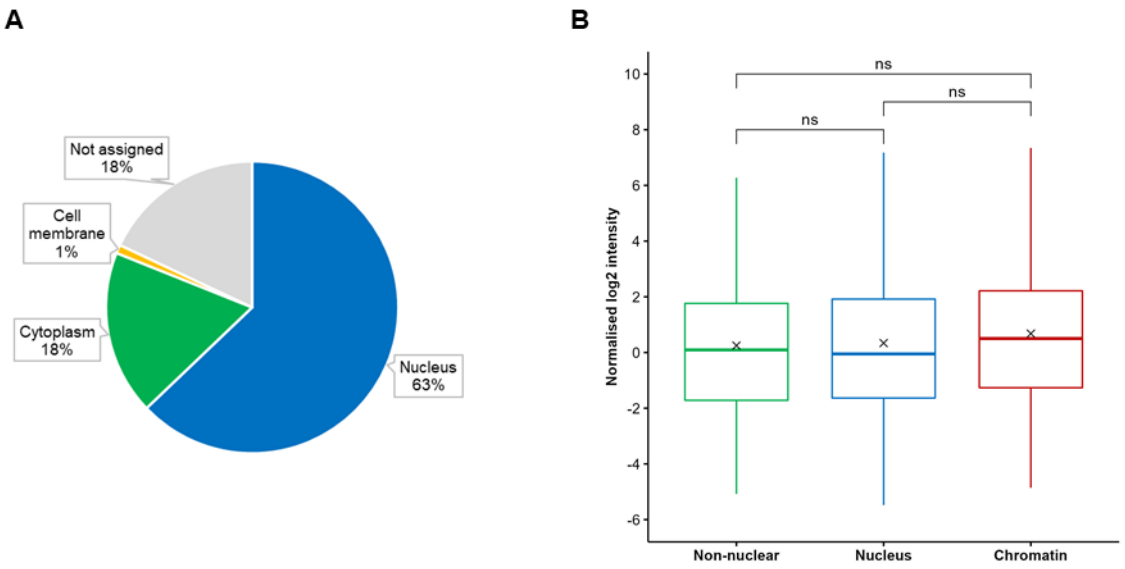


Figure 17: Protein count and log2 intensity of subcellular protein categories in C-CAP
(A) Pie chart showing the subcellular localisation categories of proteins identified in the 500 – 1000 mM elution fractions of the ZF-CXXC affinity purification (Figure 15d). The categorisation of proteins is based on information from the UniProt database and category percentages represent the share of all identified proteins.
(B) Boxplot showing the distribution of normalised log2 intensities of proteins that were identified in the 500 – 1000 mM elution fractions of the ZF-CXXC affinity purification. The proteins were categorised according to their subcellular localisation (non-nuclear, nucleus, chromatin). The non-nuclear category summarises cytoplasm, cell membrane, and extracellular space annotations. Welch's t-test was performed for statistical analysis. There is no statistically significant enrichment between the categories (ns: FDR > 0.05).

Comparison of the localisation categories within the C-CAP eluate showed no significant difference of protein abundance between the subcellular localisation categories (Figure 17B). Therefore, I wondered if there was any enrichment when compared against a full mESC proteome of 4536 proteins derived from a whole-cell lysate. To check this, I ranked the proteins in C-CAP and the full proteome by their normalised log2 intensities, and then calculated the cumulative count of the protein localisation categories among those ranks. For example, if chromatin-bound proteins were enriched in C-CAP, then one should observe a higher accumulation of those proteins among the top ranks compared to the full proteome. Indeed, the cumulative distributions of abundance-ranked proteins showed that nuclear and chromatin-associated proteins are moderately enriched among the top ranks of C-CAP, while non-nuclear proteins are slightly depleted (Figure 18A). Importantly, known CGI-binding proteins such as constituents of the Polycomb Group (PcG) and Trithorax Group (TrxG) complexes reveal higher accumulation among the most abundant proteins compared to the full proteome. However, more PcG and TrxG proteins were identified in the full proteome, which was unexpected considering that C-CAP aims to enrich the CGI proteome. To follow up on this observation, the number of proteins was counted that are

part of PcG, TrxG, and other promoter-binding protein groups. I then divided the count by the theoretically observable number of proteins in each group in order to define an identification rate which I could compare between C-CAP and the full proteome. For example, out of all 33 RNA polymerase subunits, 26 (79%) were identified in the full proteome, but only 12 (36%) in C-CAP (Figure 18B). Similarly, the full proteome shows a higher identification rate for DNMT, PRC1, and TrxG proteins. This indicates a lack of proteome depth in the C-CAP sample, which could have been caused by the MS sample preparation issues mentioned before.

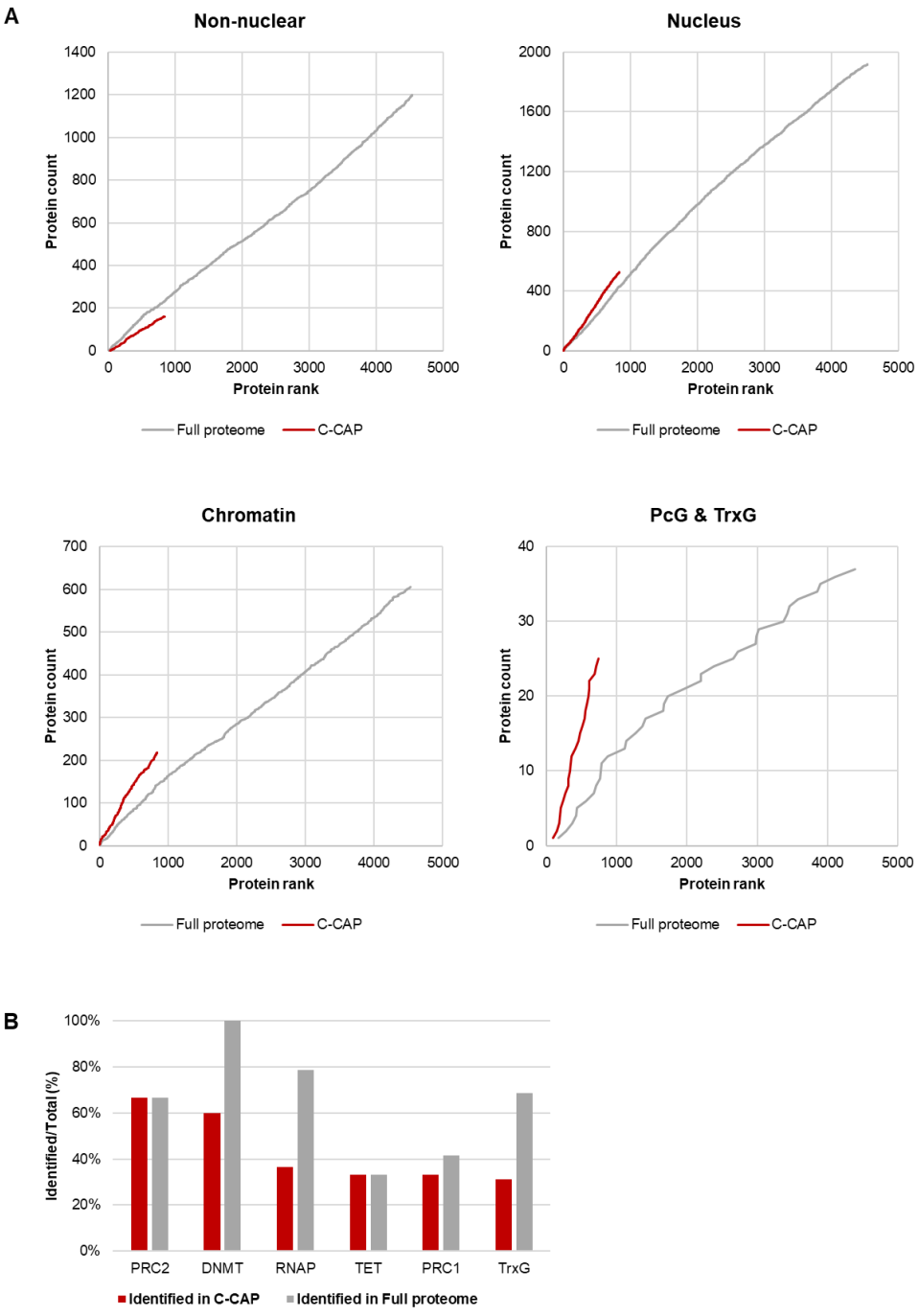


Figure 18: Comparison of selected protein groups in C-CAP and the full proteome

(A) Scatter plots showing the cumulative distribution of abundance-ranked proteins. Proteins in the C-CAP eluates and the full proteome were ranked by abundance (normalised log2 intensity), and then the cumulative count from first to last rank was calculated for proteins belonging to one of the protein categories. Proteins in the categories Nucleus, Chromatin, PcG & TrxG show higher accumulation among the top ranks in C-CAP compared to the full proteome. In contrast, Non-nuclear proteins accumulate less among the top ranks.

(B) Bar chart showing chromatin- and promoter-associated protein groups identified in C-CAP and the full proteome. The percentage identification was calculated by taking the number of identified proteins in the respective sample and dividing it by the total number of proteins found for that group in the UniProt database. For example, UniProt lists 5 DNMTs of which all 5 were detected in the full proteome.

Although overall fewer PcG and TrxG components were detected in C-CAP than in the full proteome, the core subunits RNF2, SUZ12, WDR5, ASH2L, RBBP5, and DPY30 were detected in both samples. Therefore, I narrowed down my analysis on the core subunits as a proxy for CGI-binding factors to gain a more detailed view of enrichment. I plotted the normalised log2 intensities of the core subunits against their percentiles to compare them between the samples (Figure 19). RNF2 and SUZ12, core subunits of PRC1 and PRC2 respectively, are approximately in the same percentiles in both proteomes. In contrast, the core TrxG components WDR5, ASH2L, RBBP5, and DPY30 are below the median (50th percentile) in C-CAP while they are above the median in the full proteome. Although C-CAP depleted approximately 3700 proteins compared to the full proteome, this depletion was not sufficiently specific to markedly improve enrichment of known CGI-binding proteins.

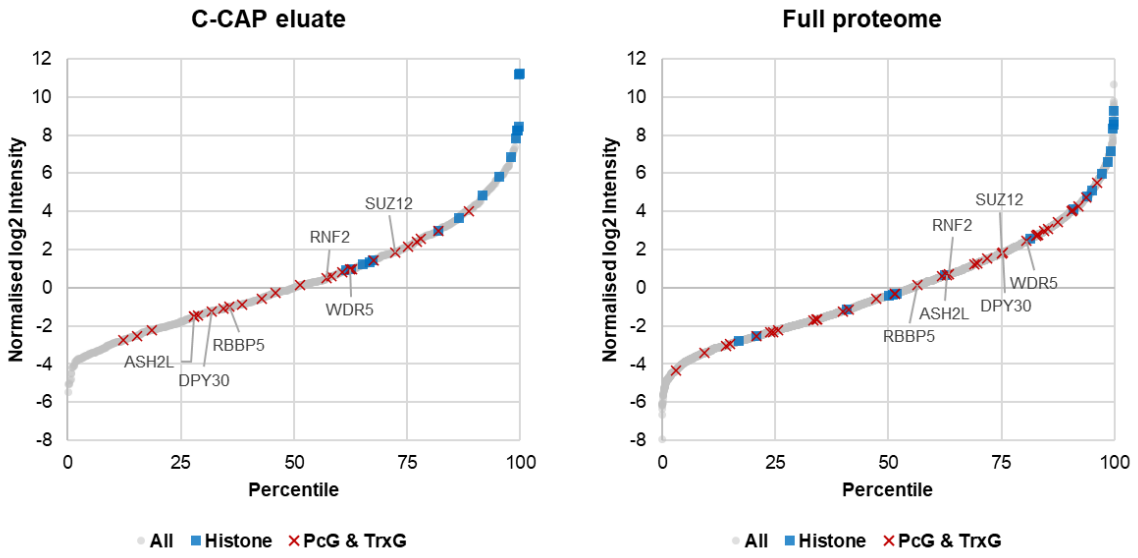


Figure 19: Comparison of PcG and TrxG core subunit percentiles in C-CAP and the full proteome
The scatter plots show proteins plotted by abundance (normalised log2 intensity) against their abundance ranking percentiles for the C-CAP eluate and the full proteome. Core subunits of CGI-binding complexes PRC1 (RNF2), PRC2 (SUZ12), and TrxG (WDR5, RBBP5, ASH2L, DPY30) are labelled. PRC1 and PRC2 core subunits are approximately in the same percentiles in both samples. Core subunits of TrxG complexes are below the median (50th percentile) in the C-CAP sample.

Since my results indicated that chromatin-associated proteins are not sufficiently enriched, I wondered whether there was any specificity for other factors and processes in the data obtained from C-CAP. To answer this question, I performed Gene Ontology (GO), protein domain, and pathway enrichment analyses (Figure 20). GO analysis revealed Nucleus, Nucleolus, and Nucleoplasm among the most significantly enriched Cellular Components (CC) (Figure 20A). However, other CC that are not related to chromatin, such as Cytoplasm

and Ribosome, equally feature among the top-ranked terms. Inspecting the enrichment of Biological Processes (BP) and pathways (Figure 20B-C), it becomes evident that the C-CAP proteome predominantly reflects abundant processes that are related to translation and RNA processing. This is further supported by the strong enrichment of proteins with RNA binding domains (Figure 20D). These proteins often co-purify non-specifically in affinity-purifications as it is evidenced by their prominence in the CRAPome database (Mellacheruvu et al., 2013).

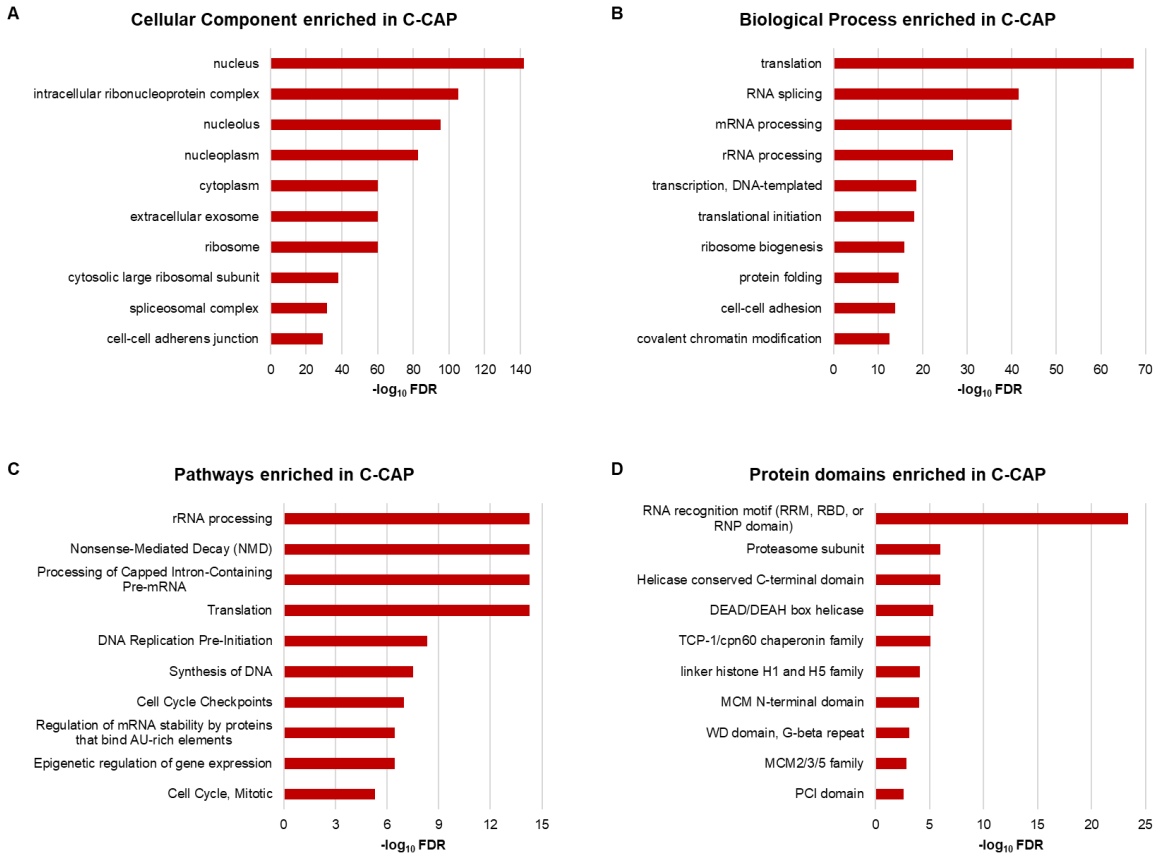


Figure 20: Gene ontology, pathway, and protein domain enrichment analyses of C-CAP eluates
Bar charts showing significantly enriched cellular compartments (A), biological processes (B), pathways (C), and protein domains (D) for proteins identified in the C-CAP eluates. The top 10 most significantly enriched terms are shown for each analysis.

In summary, my initial C-CAP experiment provided several important insights on which I could build to improve my approach. First, qPCR analysis of DNA from my elution fractions showed that my purified ZF-CXXC protein is capable of enriching CGI from crosslinked chromatin. Second, the proteomics results revealed that I needed to improve the purity of my chromatin sample as too many non-specific proteins co-purified in the eluates. Finally,

the mass spectrometry sample preparation required optimisation as indicated by the low proteome depth. Additionally, the relatively high concentration of detergents in the samples caused considerable difficulties when preparing them for mass spectrometry.

4.2.2 An additional chromatin purification step using SICAP does not improve CGI enrichment and purity

To address the issue with abundant contaminants encountered in my initial C-CAP approach, I reasoned that Selective Isolation of Chromatin-Associated Proteins (SICAP) (Rafiee et al., 2016) may help to improve the purity of my chromatin fractions. SICAP was developed to remove contaminants in ChIP-MS-type assays by applying stringent wash conditions. To achieve this, immunoprecipitated chromatin is end-labelled with biotinylated nucleotides and captured on streptavidin-coated beads (Figure 21e-f). This enables extremely stringent washing steps that remove non-crosslinked contaminants from enriched chromatin regions. Proteins are then eluted by heating which reverses crosslinks. To compare the performance of C-CAP against SICAP, the samples were split. One half was prepared for qPCR and MS after performing C-CAP, and the other half was further processed with the SICAP protocol. By using SICAP, I aimed to increase the purity of my chromatin sample in order to improve the detection of low-abundant CGI-binding proteins.

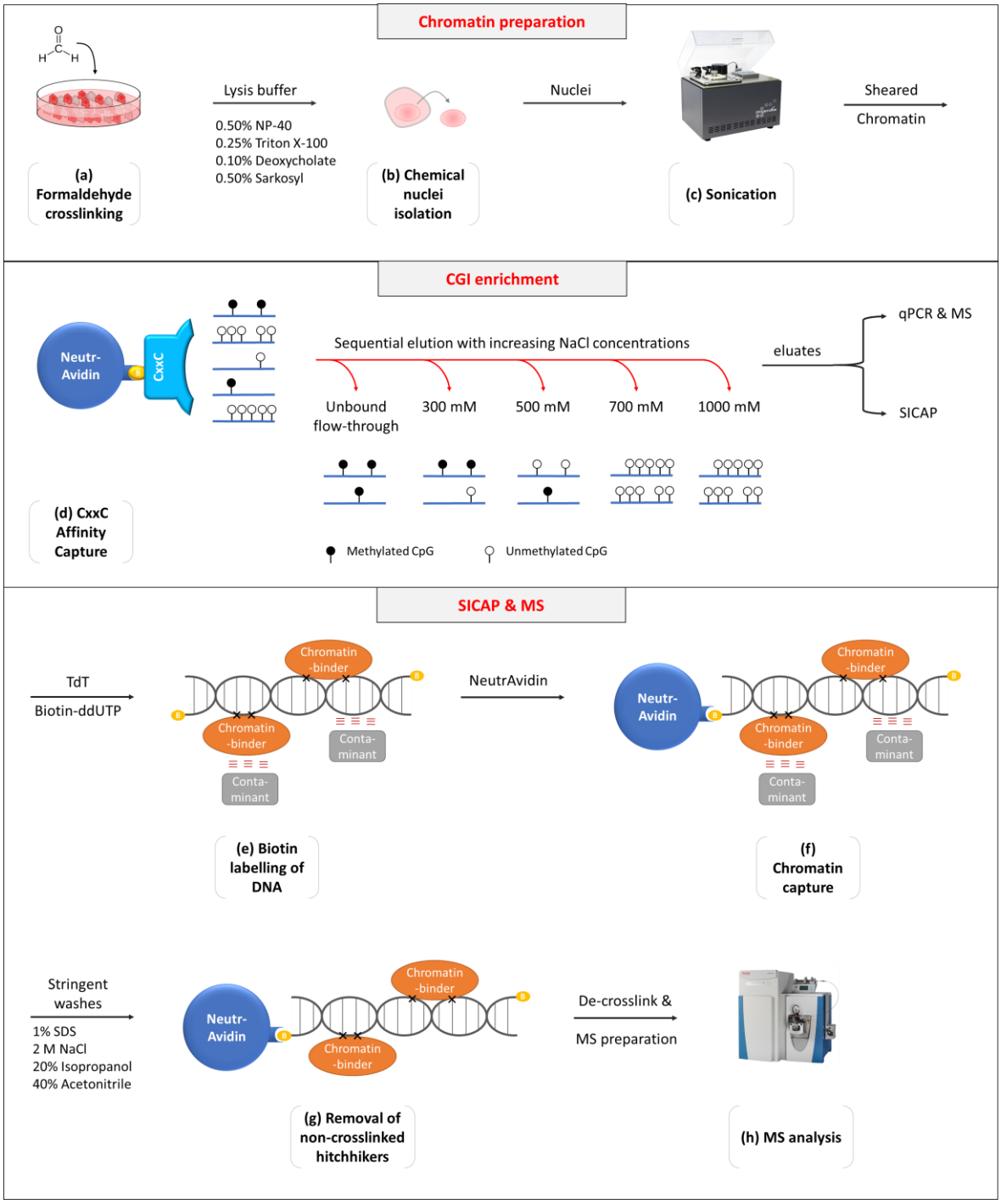


Figure 21: Schematic workflow of C-CAP in combination with SICAP

Chromatin was prepared and CGI were enriched as described in Figure 15. Afterwards, C-CAP eluates were dialysed into a terminal deoxynucleotidyl transferase (TdT)-compatible buffer (not shown) and then split into two aliquots: The C-CAP aliquot was prepared for qPCR and MS (d), and the other aliquot was subjected to the SICAP procedure (e - h). For SICAP, chromatin was end-labelled with biotin-ddUTP (e) and then captured on NeutrAvidin beads (f). Stringent washing was applied to remove non-crosslinked contaminants (f - g). After washing, formaldehyde crosslinks were reversed by heating the sample and proteins were prepared for MS analysis (g - h).

I carried out the first steps of this new approach as before by isolating CGI using C-CAP (Figure 21a-d). I validated the CGI enrichment by qPCR analysis comparing a CGI-associated gene promoter (*Ncoa2*), a non-CGI gene promoter (*Fgf7*), and their corresponding gene bodies (Figure 22).

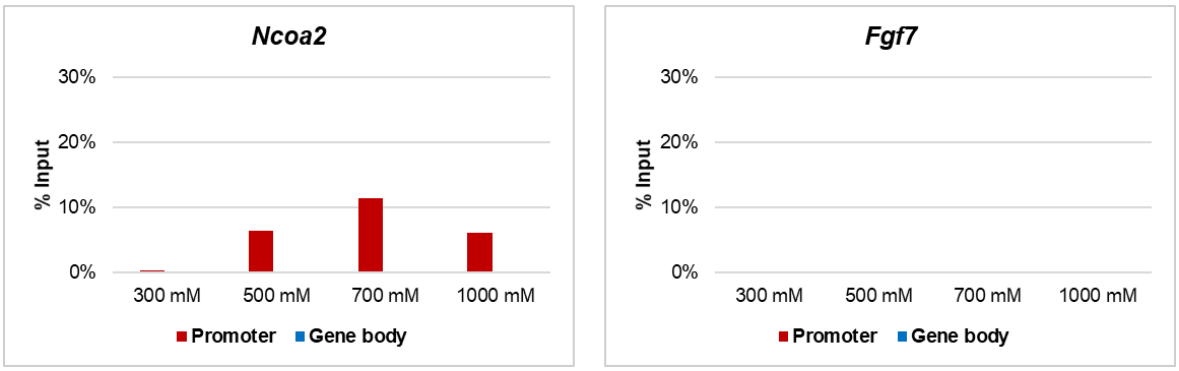


Figure 22: Quantitative PCR analysis of C-CAP elution fractions
Enrichment of CGI chromatin was validated by qPCR analysis. The qPCR profiles show strong amplification at the CGI-promoter of *Ncoa2* compared to gene bodies and *Fgf7* promoter, which are non-CGI regions. Enrichment is expressed as % of Input DNA. Some bars are not visible because the enrichment is below 1%.

After validating the enrichment of CGIs in the C-CAP eluates, I took half of the sample volume and processed it for MS (Figure 23). The remaining half was end-labelled with biotin and subjected to stringent washing which are inherent steps to SICAP (Figure 21e-g). To verify that all crosslinked proteins were recovered in the SICAP eluate, the NeutrAvidin beads were boiled, and the resulting supernatant was subjected to MS analysis as well (Figure 23).

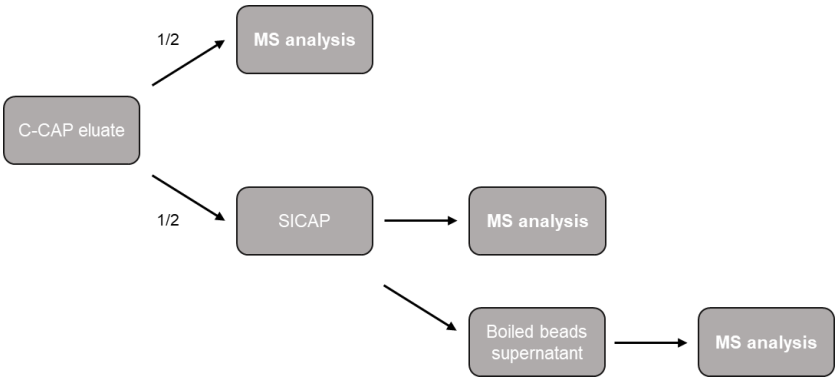


Figure 23: Flowchart of sample processing steps after C-CAP
The sample was split after C-CAP. The C-CAP-only half was prepared for MS analysis, and the other half was processed with SICAP followed by MS. The NeutrAvidin beads from the SICAP procedure were boiled and the supernatant was analysed by MS as well.

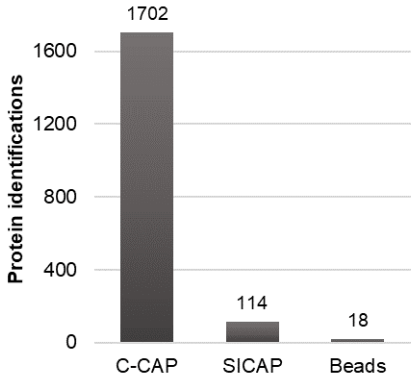


Figure 24: Protein identifications in C-CAP and SICAP eluates
Bar chart showing protein identifications for each experimental step. MS analysis identified 1702 proteins in the C-CAP eluates and 114 proteins in the SICAP eluates. To verify that all proteins had eluted off the affinity matrix, the beads were boiled, and the resulting supernatant analysed by MS. This revealed 18 protein identifications which indicated complete elution of proteins.

MS analysis showed a ~2-fold increase of identifications with 1702 proteins detected in this experiment (Figure 24) compared to 840 proteins in the previous one (see 4.2.1). This suggested that improved MS sample preparation and acquisition had considerably increased proteome depth. More importantly, SICAP identified only 114 proteins, which is ~15 times less than the 1702 proteins identified in C-CAP (Figure 24). To exclude the possibility that proteins were not completely eluted after SICAP, the NeutrAvidin beads were boiled, and the supernatant was also subjected to MS analysis. This showed only 18 protein identifications, indicating that proteins had efficiently eluted from the affinity matrix. To examine the possibility that a marked reduction in protein identifications in SICAP indicated an improvement in chromatin purity and specificity of CGI isolation, I compared the subcellular localisation of annotated proteins between C-CAP and SICAP. Surprisingly, proportionally fewer nuclear proteins were identified in SICAP (Figure 25A). Furthermore, there was no significant difference in the relative abundance of non-nuclear, nuclear, and chromatin-associated proteins between the two approaches (Figure 25B). At this level of analysis, SICAP did not appear to majorly alter the enrichment of nuclear and chromatin proteins.

Although nuclear and chromatin-bound proteins did not seem to be enriched as a broad group, I wondered if specific protein groups known to interact with chromatin, especially with gene promoters, were enriched in SICAP. Therefore, I attempted to determine the identification rate of promoter-binding protein groups as performed above (Figure 18B). Unexpectedly, not a single member of the analysed protein groups (PcG, TrxG, DNMT, TET, RNAP) was identified in the SICAP sample (Figure 25C).

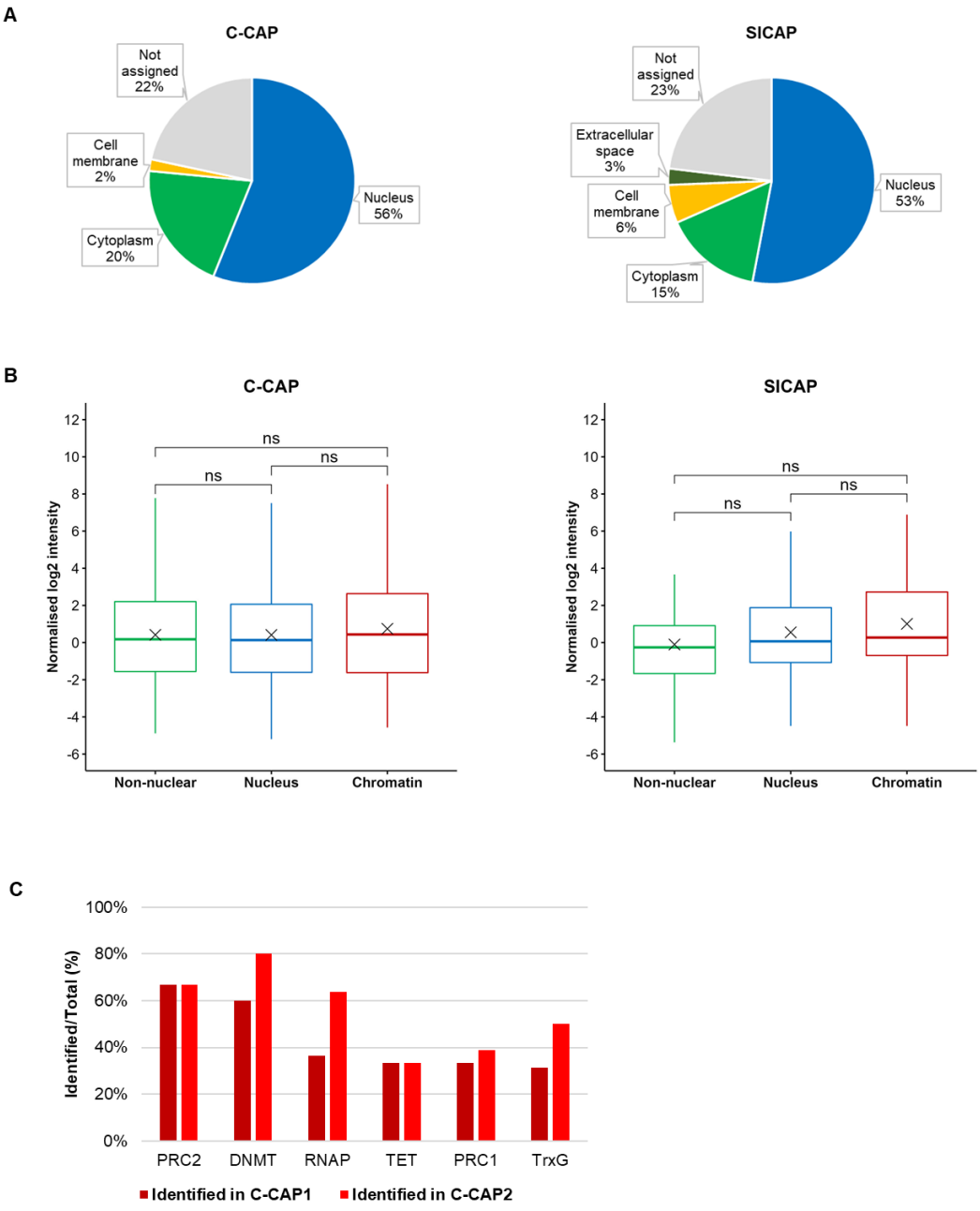


Figure 25: Proteins categorised by subcellular localisation and group affiliation

(A) Pie charts showing the subcellular localisation categories of proteins identified in the C-CAP and SICAP eluates. The categorisation of proteins is based on information from the UniProt database and category percentages represent the share of all identified proteins in the C-CAP or SICAP eluates, respectively.

(B) Boxplots showing the distribution of normalised log2 intensities of proteins that were identified in the C-CAP and SICAP eluates. The proteins were categorised according to their subcellular localisation (non-nuclear, nucleus, chromatin). The non-nuclear category summarises cytoplasm, cell membrane, and extracellular space annotations. Welch's t-test was performed for statistical analysis. There is no statistically significant enrichment between the categories in C-CAP or SICAP (ns: FDR > 0.05).

(C) Bar chart of chromatin- and promoter-associated protein groups identified in this and the previous C-CAP experiment. The identification rate remained constant for PRC2 and TET proteins, and improved for the remaining groups in the second C-CAP experiment. SICAP did not detect any proteins from these groups and therefore is not shown in the bar chart.

Since SICAP did not reveal any apparent enrichment of chromatin- and promoter-binding proteins, I examined if SICAP showed any specificity for other cellular components, processes, or pathways. To qualitatively assess this, I carried out gene ontology and pathway enrichment analyses (Figure 26). While Nucleus and Nucleoplasm are the top two ranked Cellular Components in C-CAP (Figure 26A), these positions change to Extracellular exosome and Extracellular matrix in SICAP (Figure 26B). Similarly, the Biological Processes in SICAP are largely dominated by terms that are unrelated to chromatin-based cellular activities (Figure 26D). Even the processes DNA-templated transcription and Covalent chromatin modification, which are among the top ten in C-CAP, fall out of this list in SICAP.

In conclusion, SICAP did not enhance chromatin purity as anticipated. Rather, it reduced protein abundance globally and non-specifically without helping to enrich promoter- and CGI-associated factors. The deficit to improve chromatin purity likely lies in SICAP itself, because the upstream input to SICAP showed CGI enrichment as determined by qPCR analysis, and the downstream MS sample preparation was the same as for C-CAP.

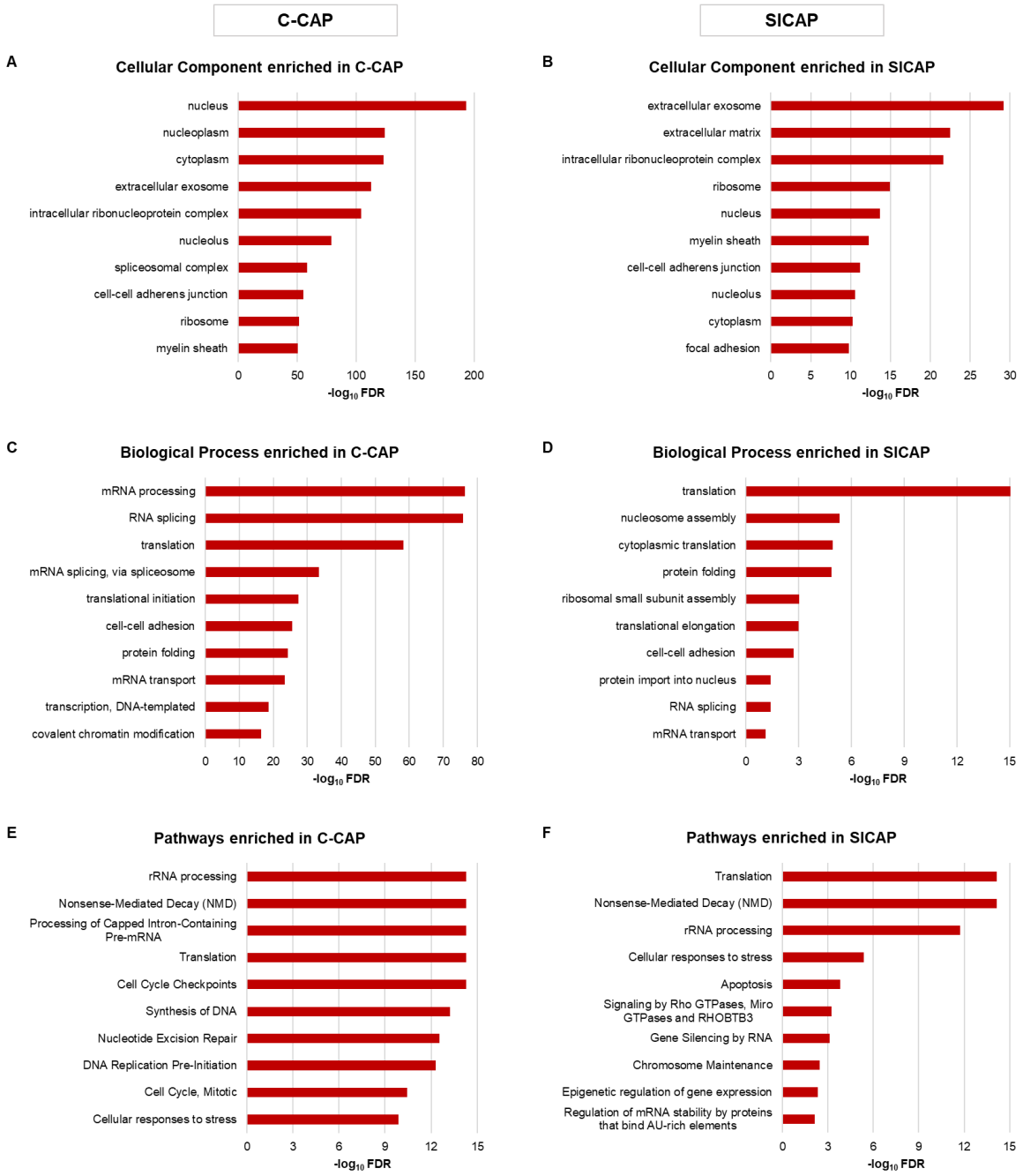


Figure 26: Gene ontology and pathway enrichment analyses for C-CAP and SICAP
Bar charts showing significantly enriched cellular compartments (A - B), biological processes (C - D), and pathways (E - F) for proteins identified in the C-CAP and SICAP elution fractions. The top 10 most significantly enriched terms are shown for each analysis.

4.2.3 A new approach to purify chromatin for CGI enrichment

The previous experiments suggested that C-CAP on its own does not provide sufficient CGI enrichment possibly due to the presence of contaminating proteins in the chromatin preparation. My attempt to remove abundant contaminant proteins by combining C-CAP

with SICAP did not improve chromatin purity. The possibility that contaminating proteins in my chromatin preparation might be limiting my analysis, led me to revise my strategy. Therefore, I decided to take a step back and to systematically improve and validate the chromatin preparation before moving on to the CGI enrichment. Hence, I designed a new workflow aimed at increasing the chromatin purity stepwise and validating the benefit of each step by MS (Figure 27). I will briefly outline the new approach here, and then describe in detail each step and its associated findings in later paragraphs. As before, I started by crosslinking the cells to maintain *in vivo* chromatin interactions during processing. After cell lysis, I purified nuclei by centrifuging them through a sucrose density barrier with the aim of reducing cytoplasmic components prior to chromatin preparation. From the purified nuclei, I extracted and washed chromatin using ChEP (Chromatin Enrichment for Proteomics, Kustatscher et al., 2014) to reduce non-crosslinked contaminants. Finally, I applied BioCAP to isolate CGI regions from total chromatin. In order to evaluate the benefit of each purification step, I compared the whole cell, nuclear, chromatin, and CGI proteomes with regard to the reduction of non-nuclear proteins and enrichment of nuclear and chromatin-associated proteins.

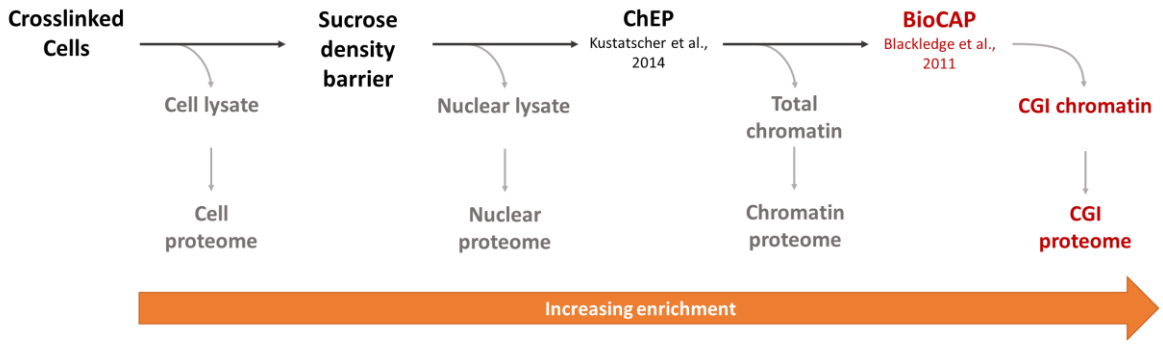


Figure 27: Workflow of the new chromatin purification and CGI enrichment approach
Aliquots are taken at each step in the workflow to verify the improvement in sample purity by MS.

In developing this new approach, my first consideration was how I could remove cytoplasmic contaminants before isolating chromatin as this could possibly improve the purity of the CGI fractions and their suitability for MS. I reasoned that one way to do this would be to separate the nuclei away from the rest of the cell prior to isolating and fragmenting chromatin. To achieve this, I resuspended crosslinked cells in a hypotonic lysis buffer which causes swelling of the cells. I then used a Dounce homogeniser to break the cell membrane and to release intact nuclei (Figure 29b). This crude nuclei suspension was applied to the top of a highly viscous sucrose density barrier in ultracentrifuge tubes (Figure 29c). During ultracentrifugation, the nuclei pass through the sucrose density barrier and

sediment at the bottom of the tubes (Graham, 2001). Cytoplasmic material is less dense than nuclei and therefore remains on top of the sucrose density barrier. To assess the quality of the nuclei purification, they were stained with DAPI to visualise DNA and with dye-conjugated agglutinin to visualise the cell membrane. Fluorescence microscopy revealed that purified nuclei had their cell membrane efficiently removed compared to non-processed control cells, and that the nuclei remained structurally intact (Figure 28).

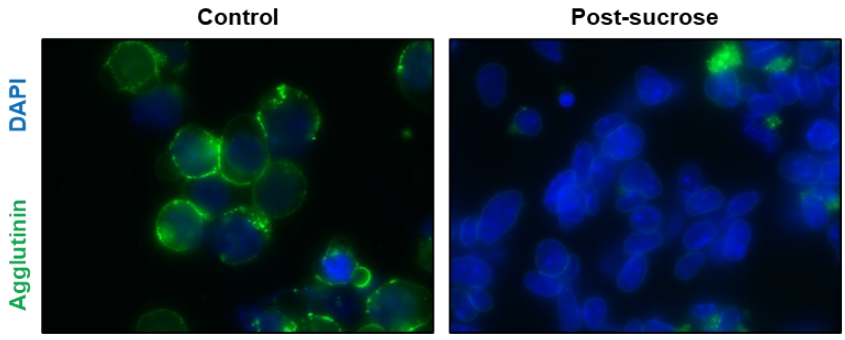


Figure 28: Validation of nuclei purification by fluorescence microscopy
Green fluorescent Alexa 488 dye is conjugated to wheat germ agglutinin, which binds sialic acid and N-acetylglucosaminyl residues primarily found on the cell membrane. DNA is visualised with DAPI. The control shows cells resuspended in hypotonic lysis buffer. The Post-sucrose sample shows the nuclei after ultracentrifugation through the sucrose density barrier.

Having isolated and purified nuclei from crosslinked cells, I then sought to isolate chromatin and remove proteins that were not chemically crosslinked (Figure 29d). This should also help to avoid contaminating nuclear proteins that can stick to chromatin during its isolation and impair the sensitivity of my CGI enrichment. Previous work has shown that this can be achieved by ChEP (Kustatscher et al., 2014a). In contrast to SICAP, ChEP does not rely on an enzymatic reaction to biotinylate DNA. Instead, nuclei are lysed, and chromatin is repeatedly precipitated and washed with 4% SDS and 8 M urea. In implementing ChEP, I found initially that the chromatin precipitation was unreliable, and additionally, the samples warmed up considerably during the extended centrifugation steps required to pellet the chromatin. I was worried that this could cause degradation of the samples and possibly even reversal of formaldehyde crosslinks. However, I found that I was able to improve the reliability of chromatin precipitation by adding EGTA to the buffer and by replacing SDS with sarkosyl which enabled pelleting of chromatin in a cooled centrifuge.

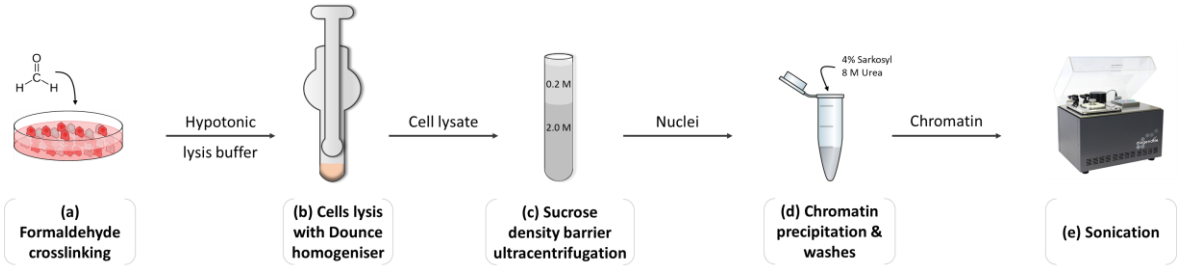


Figure 29: Detailed steps of the new chromatin purification approach
After crosslinking with formaldehyde (a), cells are resuspended in a hypotonic lysis buffer and gently lysed with a Dounce homogeniser (b). The lysate is briefly centrifuged to separate the cytoplasm from the nuclei. The crude nuclei suspension is loaded on a sucrose density barrier and nuclei are pelleted by ultracentrifugation (c). Purified nuclei are lysed, and released chromatin is repeatedly precipitated and washed with sarkosyl and urea (d). Finally, chromatin is fragmented by sonication (e) to solubilise it for ZF-CXXC affinity purification.

After the ChEP step, the purified chromatin was fragmented by sonication to make it amenable to the isolation of CGI by C-CAP. Before proceeding to the C-CAP component of the approach, I sought to ensure that the addition of these upstream purification steps improved the purity of the chromatin preparation. To achieve this, I compared the proteomes of the whole-cell lysate (WCL), cytoplasm, nucleus, and chromatin fractions. In the WCL, 4427 proteins were identified, which falls down to 2162 identifications in the Chromatin fraction (Figure 30A). These numbers are largely in line with the expected proteome depth for these fractions considering the available mass spectrometers and applied gradient times when operating with sufficient material (Christoforou et al., 2016; Ginno et al., 2018; Kustatscher et al., 2014b; Rafiee et al., 2016). Categorisation of the identified proteins according to their annotated cellular localisation revealed a marked increase of nuclear proteins in the nucleus and chromatin fractions (Figure 30B). A similar increase is observed when focussing on chromatin-associated proteins (Figure 30C). Strikingly, it seemed that the cytoplasm fraction was composed of more nuclear than cytoplasmic proteins (Figure 30B), and that it contained almost 10% chromatin-associated proteins (Figure 30C). This is likely due to the methodology chosen to categorise proteins, which demands that an annotated protein cannot be part of more than one subcellular compartment. To ensure unique localisation, proteins that localise to both cytoplasm and nucleus according to the UniProt database were designated as nuclear proteins. Equally, chromatin-associated proteins must be annotated as nuclear proteins. However, this methodology does not affect the observation that nuclear and chromatin-bound proteins make up a proportionally higher share of total proteins in the nuclei and chromatin samples compared to the WCL.

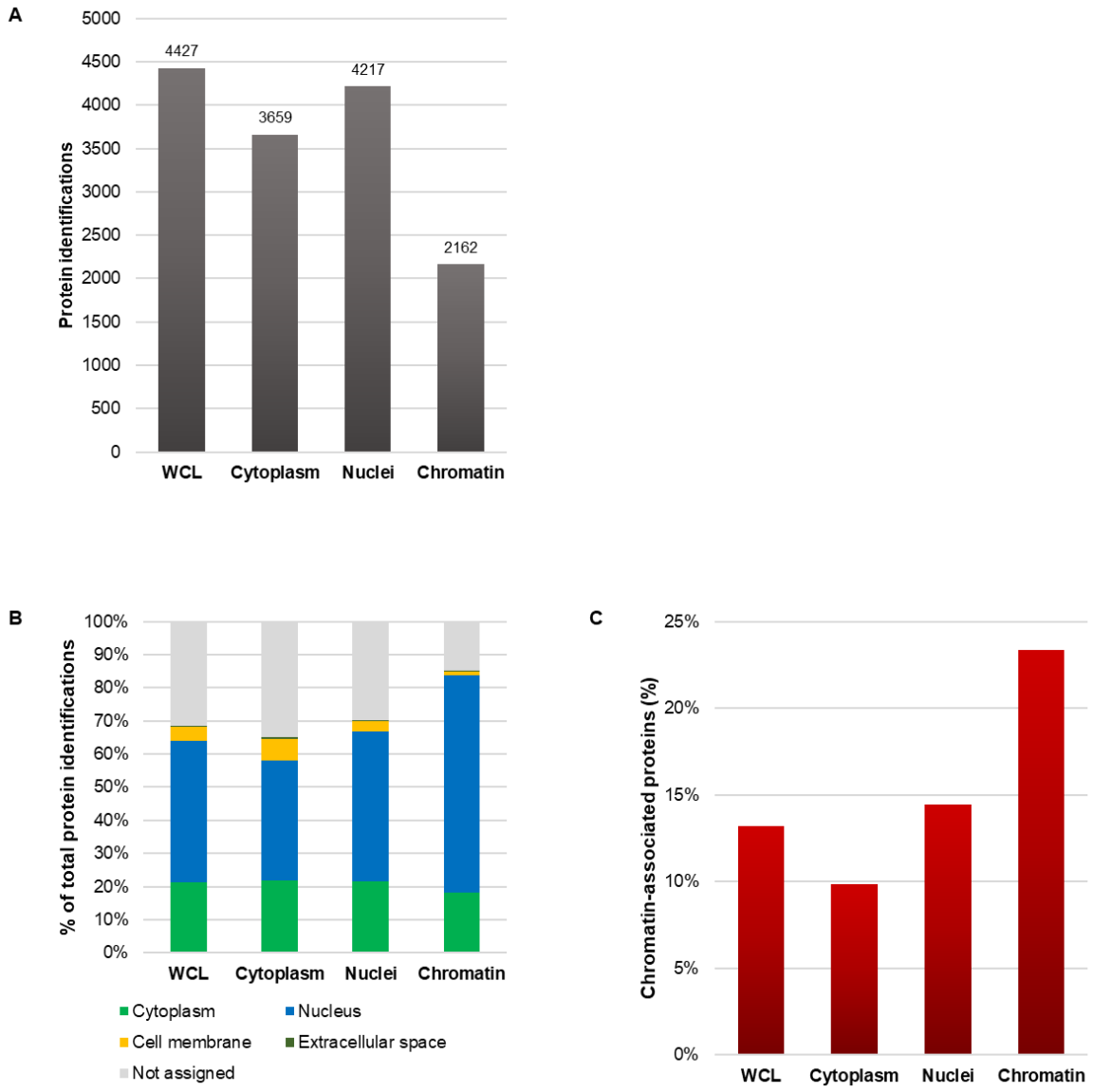


Figure 30: Categorisation of protein identifications according to subcellular localisation
(A) Bar chart showing the number of proteins identified by MS in each of the samples collected in the chromatin purification process outlined in Figure 27. Whole-cell lysate (WCL).
(B) Stacked bar chart showing the share of each subcellular localisation category in the four fractions of the chromatin purification process. Proteins were assigned to the categories based on information from the UniProt database.
(C) Bar chart comparing the share of chromatin-associated proteins in the samples.

If the new chromatin purification helps to limit non-nuclear proteins and retain nuclear proteins as well as chromatin-associated proteins, one would expect to see a shift in their relative abundance (normalised log2 intensity) between the fractions. To examine this possibility, proteins were ranked by abundance and plotted against their cumulative distribution as performed above (see 4.2.1). Indeed, non-nuclear proteins are considerably reduced in the chromatin fraction (Figure 31A), while nuclear and chromatin-binding proteins noticeably accumulate in the top abundance ranks (Figure 31B-C).

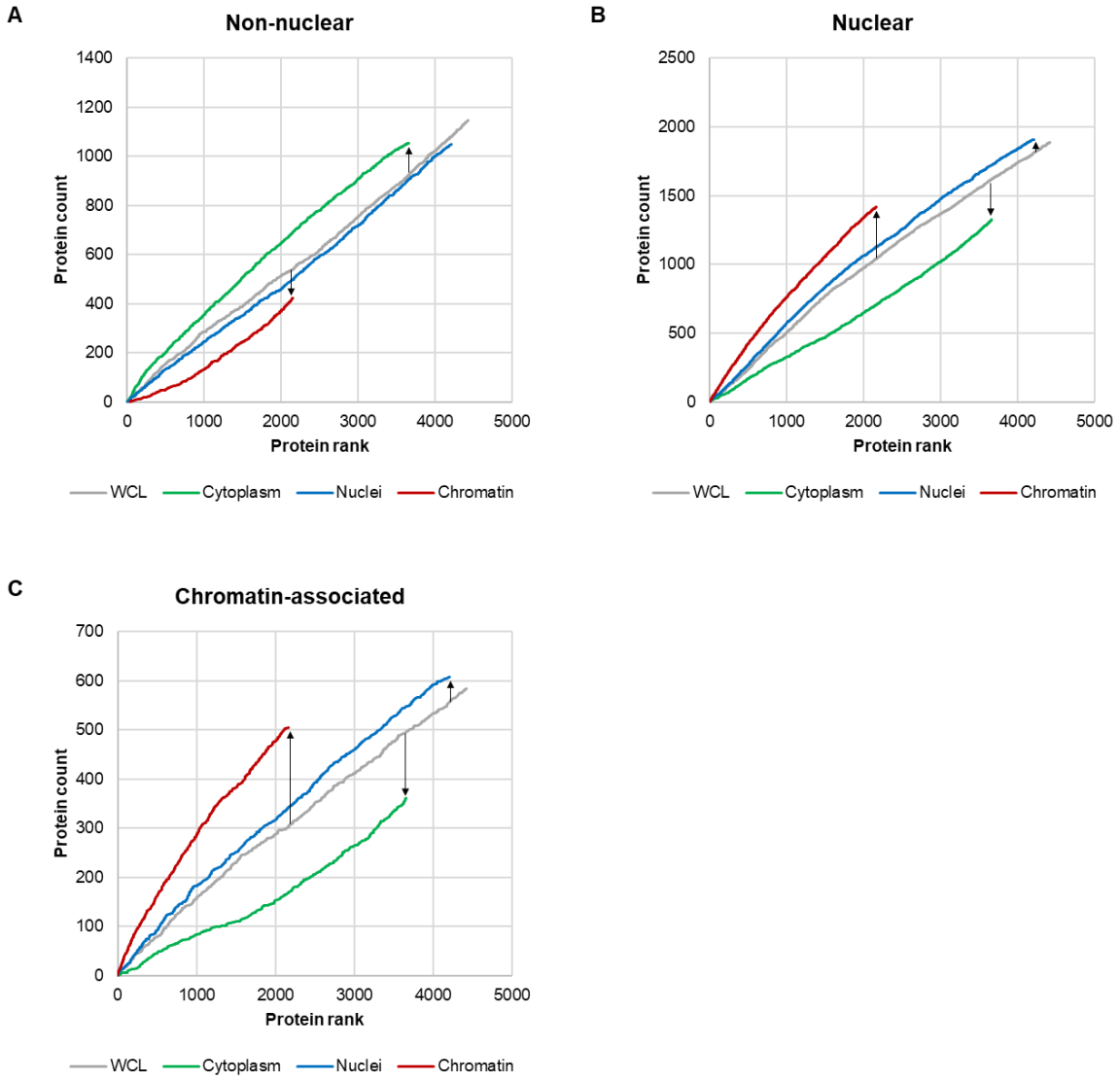


Figure 31: Cumulative distribution of abundance-ranked protein groups per sample
Scatter plots comparing enrichment of the protein localisation categories Non-nuclear (A), Nuclear (B), and Chromatin-associated (C) in each of the four samples collected in the chromatin purification process. Enrichment is assessed based on the cumulative distribution of abundance-ranked proteins that belong to one of the localisation categories.

To examine in more detail how the sample purity improves at each step of the purification, I performed a side-by-side comparison of the input and output samples of each processing step. For example, the purification starts with the whole-cell lysate (input) which is processed to obtain cytoplasm and nuclei fractions (outputs). Therefore, I quantified the statistical significance of the abundance fold change of non-nuclear and nuclear proteins in the sample pairs Cytoplasm/WCL and Nuclei/WCL. In the subsequent purification step, the nuclei fraction (now input) is further processed to extract the chromatin fraction (output). In

the Chromatin/Nuclei sample pair, I quantified the abundance fold change of non-nuclear and chromatin-associated proteins, because the last purification step aims to isolate only those proteins bound to chromatin. Figure 32A shows that non-nuclear proteins are significantly enriched in the cytoplasm fraction and nuclear proteins are largely depleted from that fraction compared to the WCL. The purification of the crude nuclei suspension on the sucrose density barrier improves the purity of the nuclei fraction only modestly (Figure 32B). However, the contribution of this step to reduce non-nuclear proteins was still significant, and therefore, I decided to keep it in the new protocol. By far the largest contribution to remove contaminants comes from the ChEP procedure (Figure 32C). It efficiently reduces the abundance of non-nuclear proteins while retaining chromatin-binding factors.

Taken together, the MS results indicate that the new chromatin purification approach has significantly reduced potential contaminants. I anticipate that this will help to detect less abundant CGI-binding proteins by mass spectrometry.

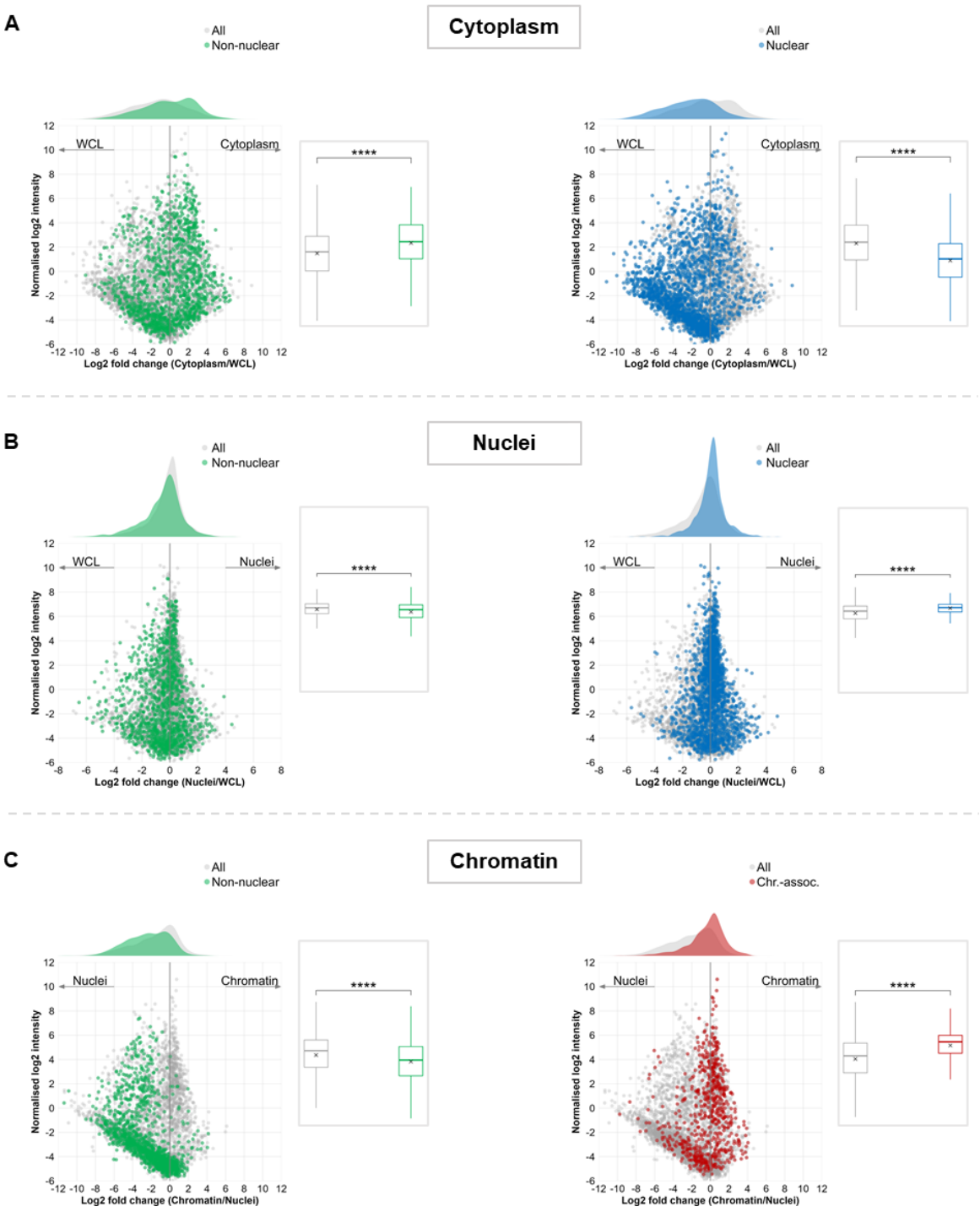


Figure 32: Enrichment and depletion of non-nuclear, nuclear, and chromatin-binding proteins in the purification process

(A – C) MA scatter plots show the log2 fold change (Log2 FC) of protein abundance in pairwise comparisons of the purification fractions. Protein groups of interest are highlighted in colour: green (non-nuclear), blue (nuclear), red (chromatin-associated). Density plots show the distribution of Log2 FC in the MA plots for the entire population and for protein groups of interest. Boxplots show the statistical significance of the Log2 FC between the entire population and the protein group of interest. Welch's t-test was performed to determine the statistical significance of the Log2 FC. ****: $p \leq 0.0001$. n = 3.

4.2.4 The new chromatin purification approach improves CGI enrichment

After having optimised the chromatin preparation, I sought to determine whether this would increase the purity of my C-CAP eluates. To investigate this, I applied chromatin that was prepared by the new ChEP-based approach on the ZF-CXXC affinity matrix and isolated CGIs from total chromatin as described in previous experiments (see 4.2.1 and 4.2.2). Then, I analysed the DNA in the elution fractions by qPCR to determine whether the ZF-CXXC domain could still isolate CGIs from chromatin prepared in the new way (Figure 33). This demonstrated that CGI-promoter DNA (*Ncoa2*, *Bmp7*, *Kmt5b*) was enriched in the elution fractions, whereas there was little or no detectable DNA from a non-CGI promoter (*Fgf7*) or gene bodies.

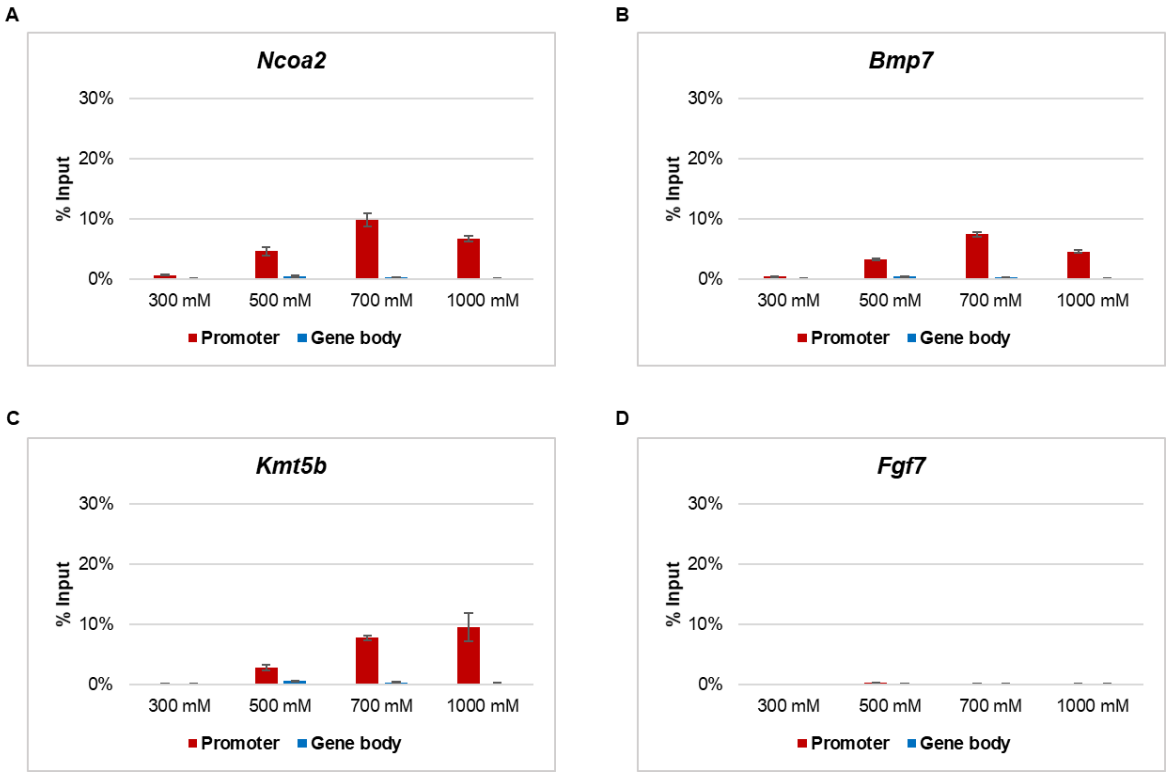


Figure 33: Quantitative PCR analysis of C-CAP elution fractions
Bar charts showing the enrichment of CGI DNA in the C-CAP eluates as determined by qPCR analysis for CGI-promoter (A - C) and non-CGI promoter (D) genes. The qPCR profiles show strong amplification at the CGI-promoters of *Ncoa2, Bmp7, and Kmt5b* compared to gene bodies and the non-CGI promoter of *Fgf7*. Enrichment is expressed as % of Input DNA. Some bars are not visible because the enrichment is below 1%. Error bars indicate the standard error, n = 3.

I then subjected the samples to MS analysis to identify proteins that were enriched in the C-CAP eluate fractions relative to the initial chromatin preparation (input). To evaluate the degree of enrichment of the CGI proteome in each elution fraction, I calculated the abundance log2 fold changes (log2 FC) separately for the four Eluate/Input pairs. A log2

FC above zero would indicate that a protein was present at higher abundance in the eluate compared to the input, thus, indicating a potential CGI-binding protein. To assess if my new approach enriched the CGI proteome, I initially focused my analysis on known CGI-binding factors, specifically, on PcG and TrxG complex components. The overall distribution of the proteins reveals in all fractions that the majority of proteins are depleted (log2 FC < 0) in the eluates compared to the input (Figure 34). This is a first indication of selectivity in the ZF-CXXC affinity purification. More importantly, all four eluates show an enrichment (log2 FC > 0) of PcG and TrxG proteins. In line with qPCR profiles, this enrichment seems to increase (higher log2 FC) from E300 to E1000, with a peak reached in the last two elution fractions (Figure 34C-D).

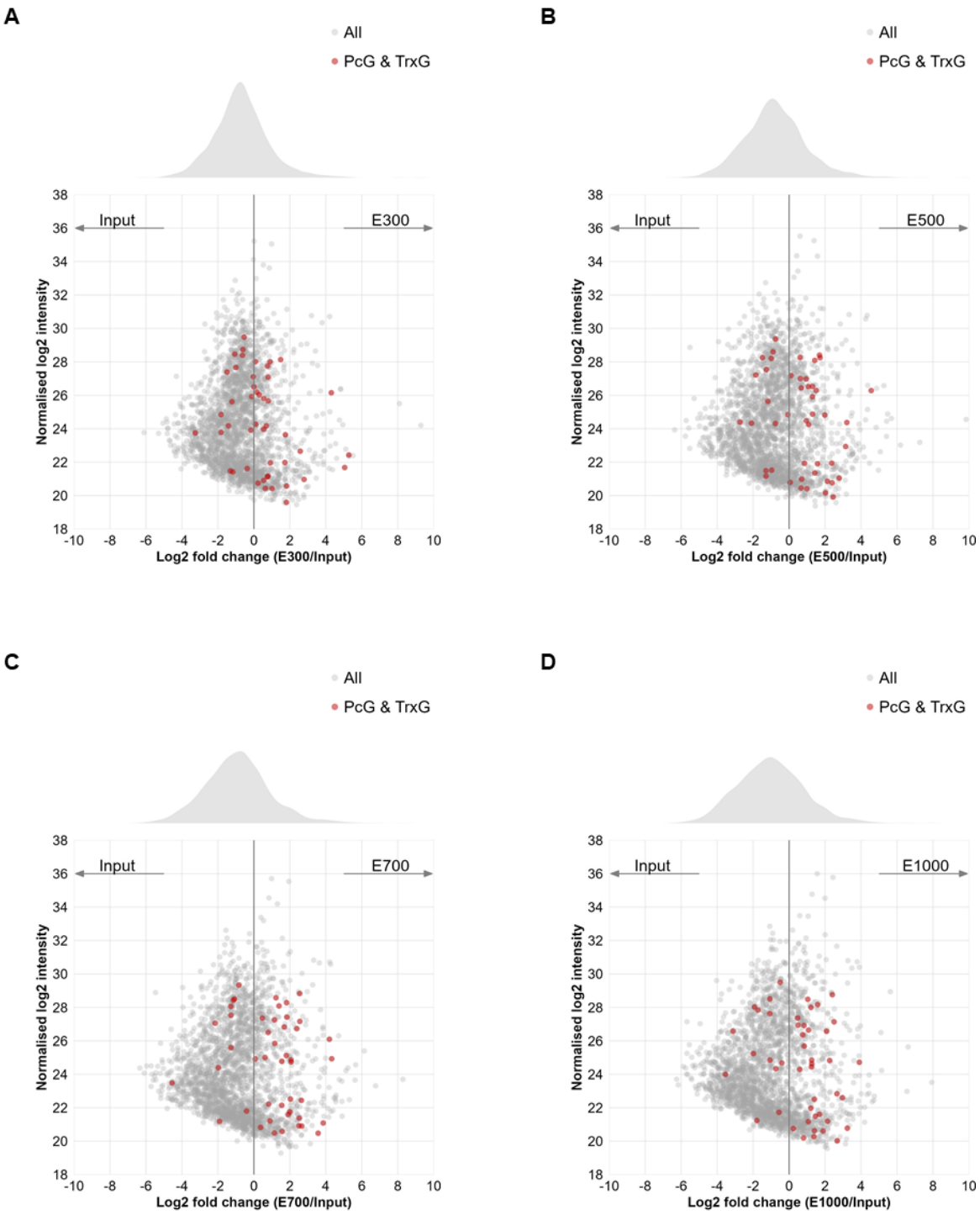


Figure 34: MA plots of CGI proteome enrichment in the C-CAP elution fractions
MA scatter plots comparing the enrichment of proteins in the respective eluate against the input sample. The density on top of the MA plots shows the distribution of the log2 fold changes for all proteins. Known CGI-binding proteins of PcG and TrxG complexes are highlighted in red to visually assess the enrichment of the CGI proteome. Based on this, the enrichment increases from the first to the last elution fraction, with a peak reached in the E700 and E1000 fractions. n = 3.

The MA plot analysis identified in total 921 unique proteins with a log2 FC above zero. However, not all of them will be constituents of the CGI proteome. Therefore, I performed statistical tests to narrow down the list of proteins to those that show statistical significance in their enrichment. This yielded 232 unique proteins that are significantly enriched in the combined eluate fractions (Figure 35, Table S1). To investigate the specificity of my CGI proteome, I used information from UniProt and PubMed databases to group the proteins into three categories, “known”, “requires validation”, and “contaminants”, based on evidence for their interaction with CGIs and gene promoters. The “known” category comprises proteins that bind CGIs or gene promoters either directly or through association with a complex. Those proteins for which I could not find evidence for their interaction with CGIs or gene promoters require further validation. The third category contains proteins which are likely to be contaminants. I found 114 proteins known to bind CGIs/gene promoters, 84 require validation, and 34 seem to be contaminants. Among the contaminant category, I noticed that 16 proteins are involved in ribosome biogenesis. The specificity of the enriched ribosome biogenesis factors will be explored later.

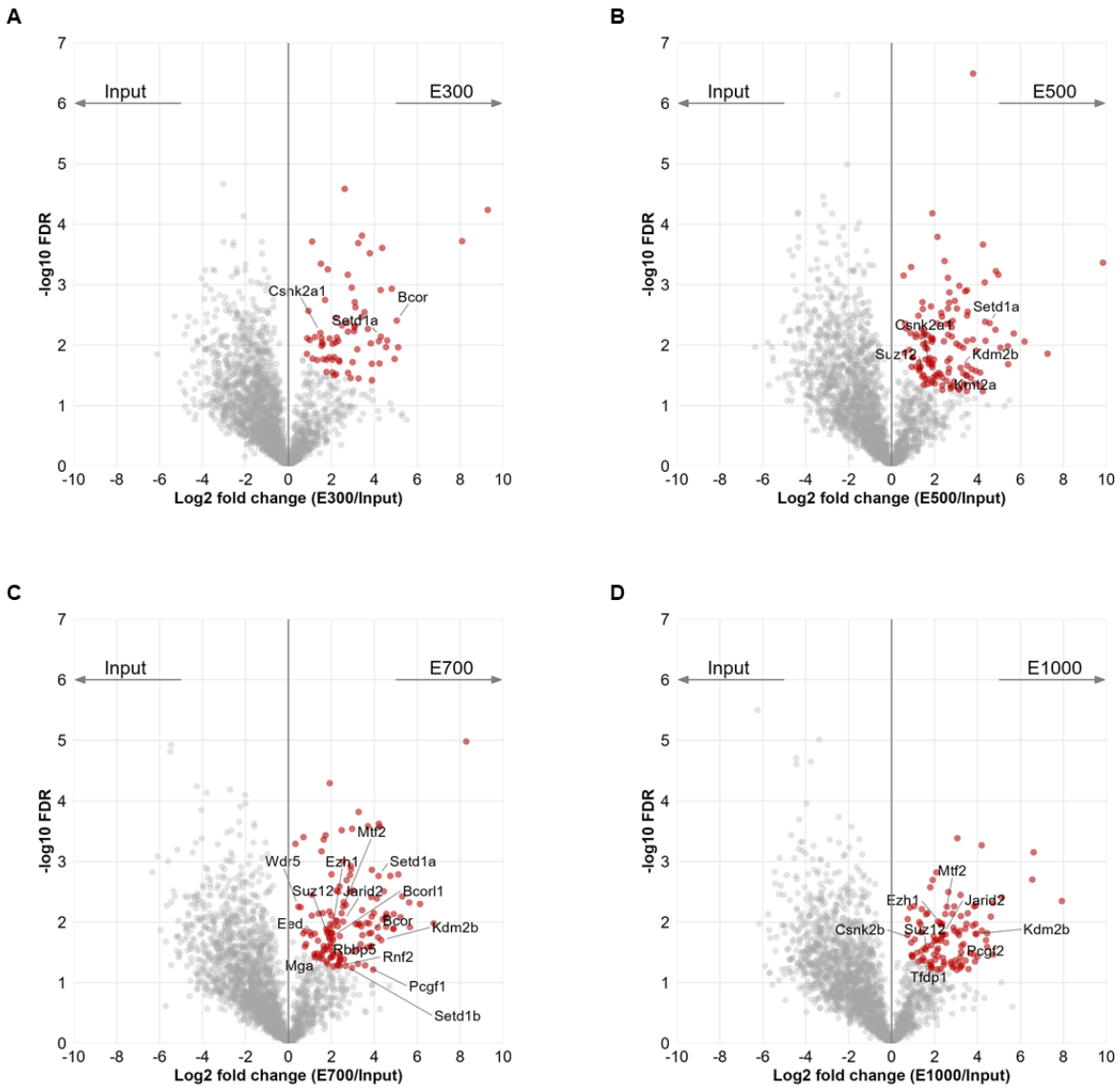


Figure 35: Volcano plots of CGI proteome enrichment in C-CAP elution fractions
Volcano plots of significantly enriched proteins in C-CAP eluates. Proteins that are significantly enriched in the respective eluate fraction over the input sample are highlighted in red and labelled in the case of PcG and TrxG complex subunits. Welch's t-test was performed to determine statistical significance with an FDR ≤ 0.05 considered significant. n = 3.

My database search revealed that among the enriched proteins, about half are known to associate with gene promoters, and the vast majority have chromatin-related functions. This prompted me to further investigate if the enriched proteins are involved in transcription-related processes and chromatin organisation. To verify this, I performed gene ontology (GO), pathway, and protein domain enrichment analyses (Figure 36). This showed enrichment in relevant categories, such as transcription (Figure 36B), chromatin modification and remodelling (Figure 36B), epigenetic regulation of gene expression (Figure

36C), and chromatin-binding domains (Figure 36D). The analyses support the previous evidence that selective enrichment of the CGI proteome was achieved.

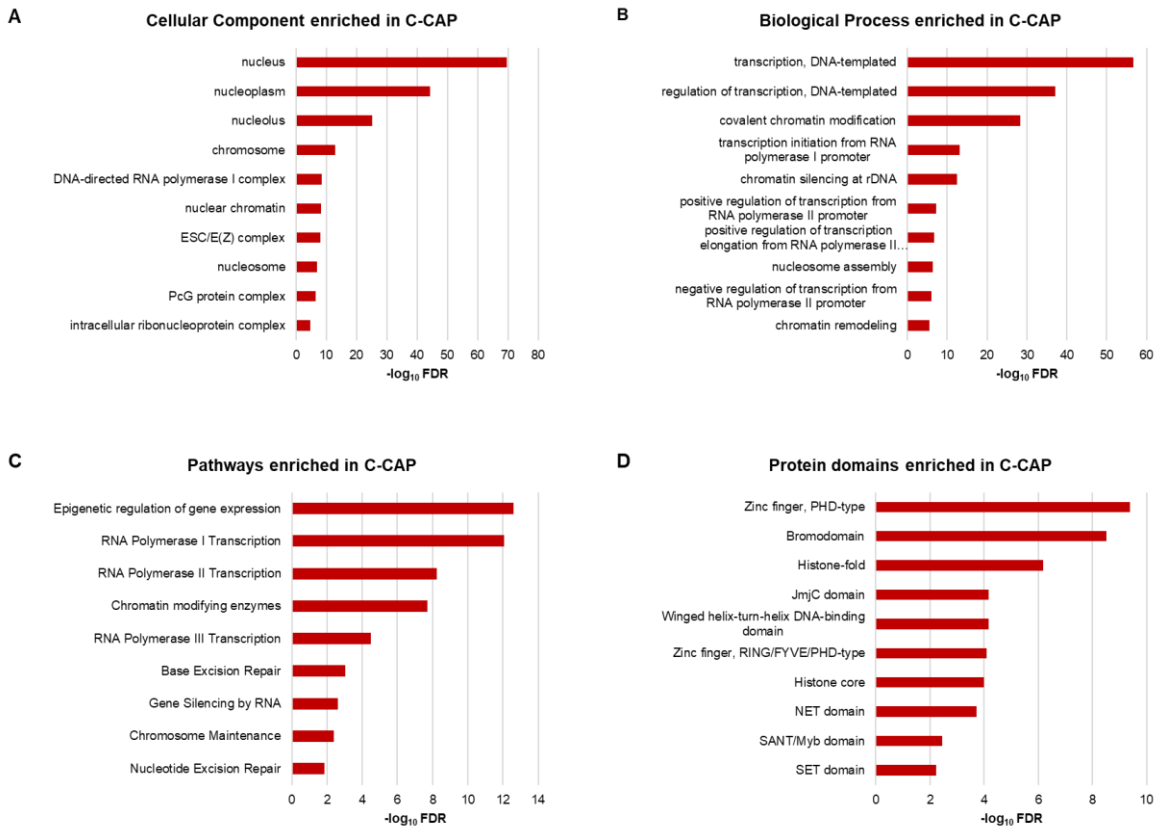


Figure 36: Gene ontology, pathway, and protein domain enrichment analyses of proteins identified in C-CAP eluates
Bar charts showing significantly enriched cellular compartments (A), biological processes (B), pathways (C), and protein domains (D) for proteins identified by MS in the C-CAP elution fractions. The top 10 most significantly enriched terms are shown for each analysis.

In conclusion, the new chromatin preparation approach considerably reduced the share of abundant contaminant proteins. This helped me to improve the enrichment of CGI chromatin by ZF-CXXC affinity purification. Approximately 50% of the proteins detected in my CGI proteome are known to associate with CGIs / gene promoters.

4.3 Discussion

Previously, ChIP-MS approaches have been used to uncover the proteome of histone modifications associated with promoters, enhancers, transcribed gene bodies, and heterochromatin (Engelen et al., 2015; Ji et al., 2015; Soldi and Bonaldi, 2013). For

example, Engelen et al. (2015) and Ji et al. (2015) performed ChIP-MS in mouse embryonic stem cells (mESC) and both groups found similar sets of proteins enriched at H3K4me3 chromatin, a histone modification associated with transcribed promoters. Initially, I applied a similar approach but replaced the ChIP-MS antibody with a ZF-CXXC domain to target CGI-containing regions of the genome. However, this approach did not yield sufficient enrichment of CGI- and promoter-binding factors, presumably due to the presence of non-specifically bound proteins. One possible reason for the abundance of contaminants could be that the enrichment of CGIs by the ZF-CXXC domain does not allow stringent wash conditions, because they would disrupt the interaction between the affinity matrix and chromatin. In contrast, the histone antibodies in ChIP-MS can often withstand stringent ChIP wash conditions which involve high detergent (1% Triton X-100, NP-40, deoxycholate) and high salt (500 mM NaCl, LiCl) concentrations. Therefore, stringent washing might have helped to remove abundant contaminants in ChIP-MS, thereby increasing the enrichment of target proteins in above mentioned studies.

Since the ZF-CXXC/DNA-interaction does not allow stringent washing, I attempted to reduce non-crosslinked contaminants after performing the CGI affinity capture. To achieve this, I combined C-CAP with SICAP (Rafiee et al., 2016), a method which involves end-labelling of chromatin with biotinylated nucleotides and then immobilising the labelled chromatin on streptavidin beads (Figure 21). In theory, the strong interaction between biotin and avidin would allow stringent wash conditions that could effectively reduce non-crosslinked contaminants. However, MS data showed that SICAP considerably reduced total protein identifications (Figure 24), without having any specific impact on the enrichment of chromatin-bound factors (Figure 25). A possible cause could be that the wash conditions outlined in the published protocol are actually too harsh. Although the biotin/avidin-interaction is the strongest known non-covalent bond, it is not entirely resistant to disruption (Holmberg et al., 2005; Michael Green, 1990). In the SICAP protocol, the wash conditions include 1% SDS, 2 M NaCl, 20% isopropanol, and 40% acetonitrile. The washes seem too harsh, and they might have rinsed away not only contaminants but also the biotinylated DNA fragments together with crosslinked factors. Another potential cause for the loss of global protein abundance in the SICAP sample lies in the terminal deoxynucleotidyl transferase (TdT)-mediated biotinylation of DNA. To be able to incorporate nucleotides into DNA, TdT requires ssDNA of at least 3 nucleotides with a free 5'-phosphate and 3'-OH (Motea and Berdis, 2010). The sonication process used to fragment chromatin creates random breaks, which do not necessarily meet these criteria. Therefore, only chromatin fragments with the right break pattern can serve as a template for TdT, and all other fragments will be washed away. This could have led to a considerable

loss of sample. To solve this issue, I tried enzymatic fragmentation of DNA with different restriction enzymes that generate the required single-stranded initiator DNA. However, it proved to be difficult to find a restriction endonuclease that efficiently digests formaldehyde-crosslinked chromatin.

Given that SICAP did not improve chromatin purity, I designed a new chromatin purification approach with the aim to obtain a clean chromatin fraction before performing the CGI affinity capture step (Figure 27). With the new approach, I removed the non-chromatin parts of the cell in a stepwise manner. First, I disrupted the cell membrane and separated the nuclei from the cytoplasm by centrifuging the cell lysate through a sucrose density barrier. Subsequently, I lysed purified nuclei and removed proteins that were not crosslinked to chromatin by precipitating chromatin and applying several rounds of stringent washes as described by the ChEP method (Kustatscher et al., 2014a). MS confirmed that each step of the new procedure significantly enhanced sample purity (Figure 32). In particular, ChEP was very powerful in removing contaminants.

Having optimised the chromatin purification process, I tested if the new chromatin preparation enhanced CGI enrichment by C-CAP. This revealed a considerable improvement in the selectivity and specificity of the CGI proteome (Figure 36). Importantly, approximately half of the identified proteins were known CGI- and promoter-binding factors. In conclusion, I have developed a new method to efficiently enrich CGIs from bulk chromatin for proteomic investigation.

5 C-CAP identifies known and potentially novel CGI-associated proteins

5.1 Introduction

A considerable challenge in proteomic studies based on prey-bait affinity interactions is to discriminate between target enrichment and non-specific background binding. The difficulty in identifying specific protein interactions is usually caused by contaminating proteins that mask the true signal either due to their high cellular abundance or their non-specific interaction with the affinity matrix. In the previous chapter, I have shown how my new chromatin purification approach has considerably reduced abundant cellular contaminants that were impairing the enrichment of CGI-binding proteins. In this chapter, I develop my approach further to include a mutant ZF-CXXC protein (ZF-CXXC^{mt}) as an additional control in my C-CAP assay to discern specific interactions with the bait protein from non-specific binding to the affinity matrix. I will first validate the loss of CGI-binding capacity for the ZF-CXXC^{mt}. Then, I will investigate the genomic regions enriched in the C-CAP eluates by next generation sequencing. Finally, I will detail how the ZF-CXXC^{mt} helped me to improve the specificity of my CGI proteome and I will present new CGI-binding protein candidates.

5.2 Results

5.2.1 The ZF-CXXC^{mt} protein is deficient in binding non-methylated CpGs

Structural investigation of the KDM2B ZF-CXXC domain has shown that residues K643 and Q644 are responsible for binding non-methylated CpG dinucleotides (Xu et al., 2018). Importantly, mutations of these residues cause loss of the CpG-binding capacity without affecting the domain's fold (Allen et al., 2006a; Cierpicki et al., 2010; Xu et al., 2011). I reasoned that a mutant ZF-CXXC protein could help me to identify proteins that bind non-specifically to the bait or the support matrix in my C-CAP assay. I could then exclude the contaminants from the downstream analysis, thereby enhancing the specificity of my CGI proteome. To test this, I purified chromatin as described before (see 4.2.3) and then performed C-CAP including the ZF-CXXC^{mt} as a negative control (Figure 37). I expected the elution fractions of the ZF-CXXC^{mt} to contain a uniform mixture of methylated and unmethylated DNA similar to the composition of the input.

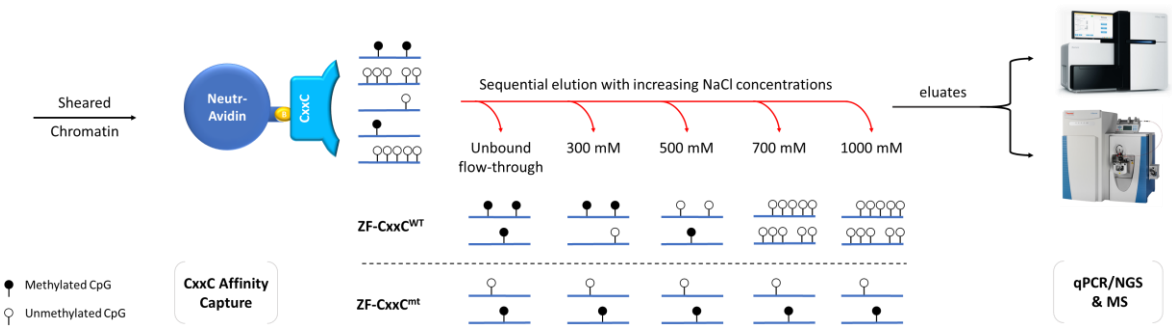


Figure 37: Schematic of CGI-chromatin affinity purification including the ZF-CXXC^{mt}
Schematic showing the ZF-CXXC affinity capture part of C-CAP. Sheared chromatin was prepared as in 4.2.3 and then loaded on NeutrAvidin beads bound by either ZF-CXXC^{WT} or ZF-CXXC^{mt}. Elution fractions were collected and analysed by qPCR, next generation sequencing (NGS), and MS.

First, I performed qPCR analysis of affinity-purified chromatin to verify that the ZF-CXXC^{mt} would no longer recover non-methylated chromatin. As done in previous experiments, I compared the PCR enrichment of the promoter and gene body regions of selected genes in the C-CAP fractions. The CGI-promoters of *Ncoa2*, *Bmp7*, and *Kmt5b* showed strong enrichment with the ZF-CXXC^{WT} but not with the ZF-CXXC^{mt} (Figure 38A-F). The non-CGI promoter of *Fgf7* is enriched neither in the ZF-CXXC^{WT}, nor the ZF-CXXC^{mt} eluates (Figure 38G-H). Furthermore, none of the ZF-CXXC affinity matrices enriched gene body DNA (Figure 38A-H). Thus, targeted gene-specific analysis confirmed the loss of CGI-binding affinity for the ZF-CXXC^{mt} protein.

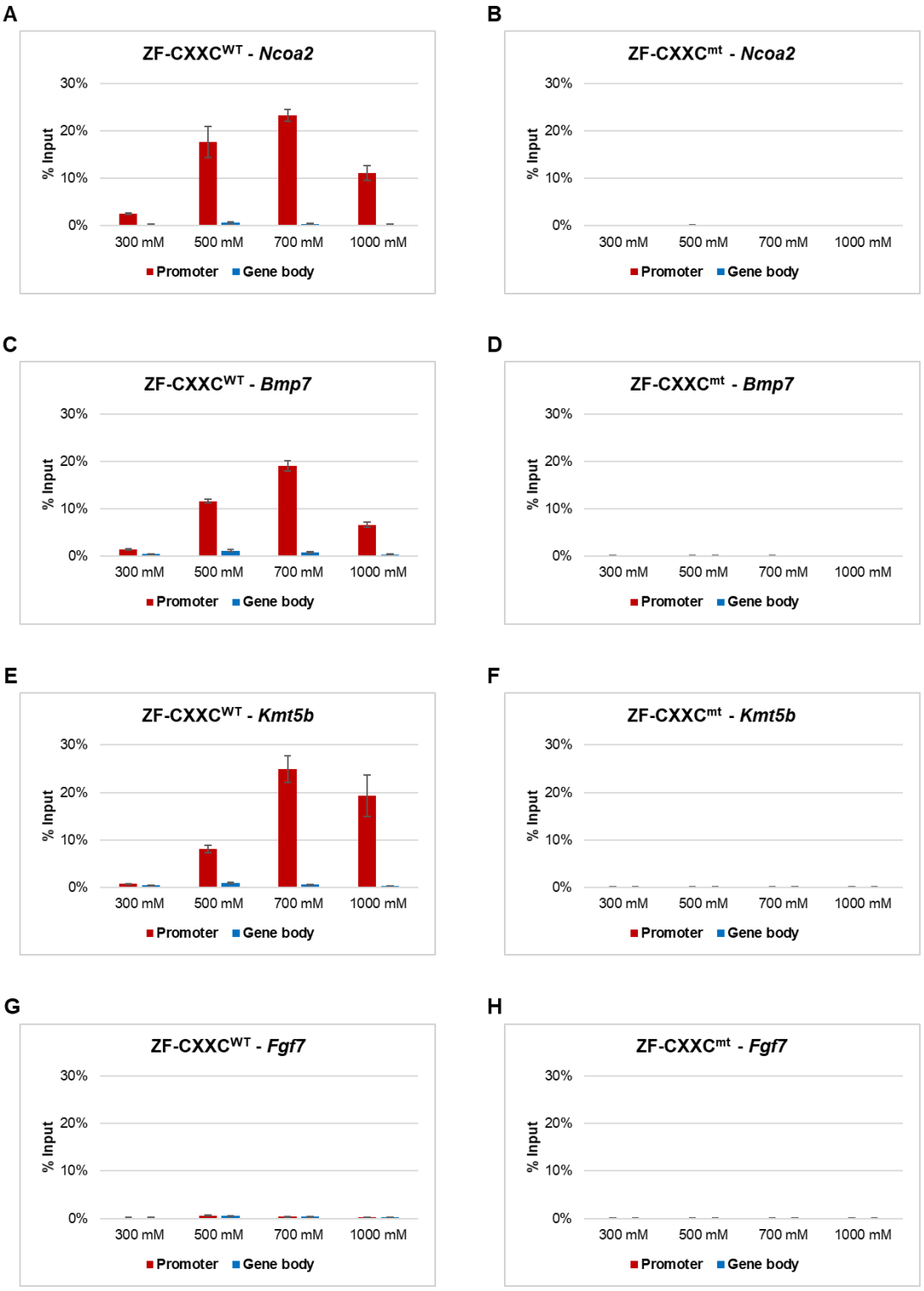


Figure 38: qPCR analysis for C-CAP elution fractions of ZF-CXXC^{WT} and ZF-CXXC^{mt}

Bar charts showing the qPCR analysis of C-CAP elution fractions of ZF-CXXC^{WT} (left charts) and ZF-CXXC^{mt} (right charts). CGI enrichment was evaluated by comparing amplification at promoter and gene body regions of three CGI-promoter genes (*Ncoa2, Bmp7, Kmt5b*) and one non-CGI promoter gene (*Fgf7*). CGI-promoters are enriched in the ZF-CXXC^{WT} eluates (A, C, E), but not in the ZF-CXXC^{mt} eluates (B, D, F). Eluates of both bait proteins are not enriched in gene body (A - H) or non-CGI-promoter regions (G, H). Enrichment is expressed as % of Input DNA. Some bars are not visible because the enrichment is below 1%. Error bars indicate the standard error, n = 3.

5.2.2 Next generation sequencing reveals enrichment of non-methylated islands and regions that contain transcription-associated histone modifications

To obtain a more comprehensive view of the genomic regions that are enriched by C-CAP, I performed massively parallel sequencing of my samples. First, I sought to validate that C-CAP enriches CGI regions uniformly. Since the ZF-CXXC^{WT} domain binds only non-methylated DNA, my reference to determine CGI enrichment was a dataset of non-methylated islands (NMIs) that were experimentally identified in mESCs by BioCAP carried out on purified naked DNA (Long et al., 2013b). This showed that enrichment of DNA in the 500 – 1000 mM elution fractions of the ZF-CXXC^{WT} (WT) strongly correspond to the location of NMIs identified in BioCAP (Figure 39). This is corroborated by an enrichment of the CG dinucleotide motif in these regions (Figure S4), which is a hallmark of CpG islands. In contrast, there was no enrichment of NMIs or the CG dinucleotide motif in the ZF-CXXC^{mt} (mt) samples, confirming at the genome-scale that this affinity matrix has lost its binding capacity to selectively enrich non-methylated chromatin.

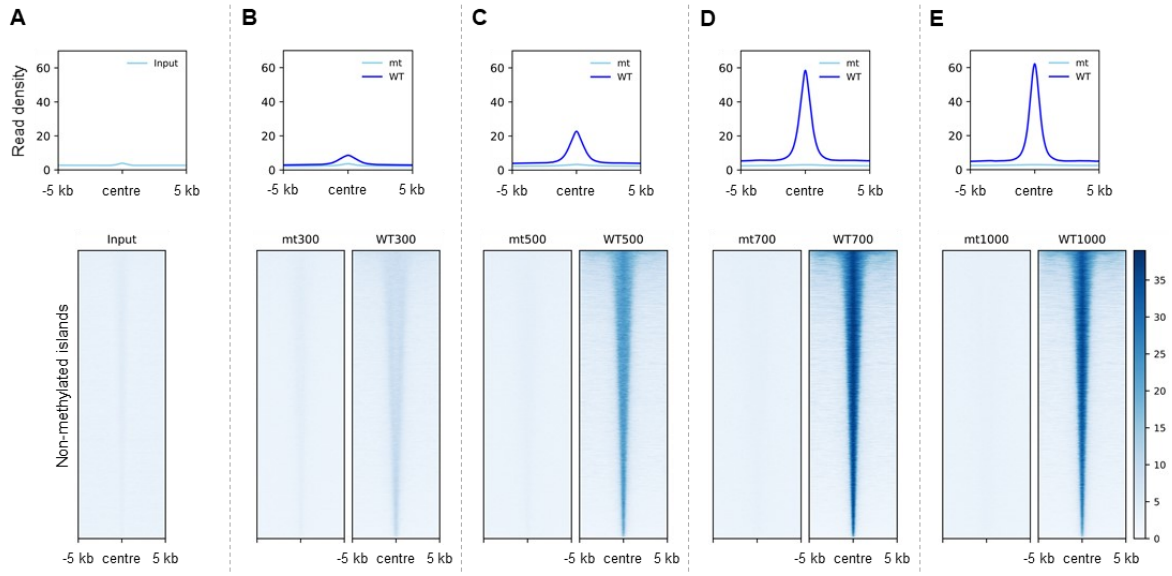


Figure 39: Genome-wide enrichment of NMIs in C-CAP eluates
Metaplots (top) and heatmaps (bottom) of C-CAP sequencing show genomic enrichment at non-methylated islands (NMIs) as identified by BioCAP. The same input sample (A) was loaded on ZF-CXXC^{mt} and ZF-CXXC^{WT} affinity matrices. Metaplots and heatmaps of the eluates (B – E) compare the enrichment of NMIs in the mt/WT pair of each salt elution fraction. NMI enrichment in the WT eluates begins in the 300 mM fraction (B) and reaches its maximum in the 700 and 1000 mM fractions (D, E). NMIs are not enriched in the mt eluates (B – E). n = 3.

Since CGIs are associated with ~70% of mammalian gene promoters (Saxonov et al., 2006), I set out to determine if C-CAP sequencing reads were enriched at the 5'-end of genes. I used a non-redundant mouse gene set (see 2.10) and analysed if C-CAP reads

localised within a region from -3 kb upstream of the transcription start site (TSS) up to the transcription end site (TES). This revealed that C-CAP reads from the 500 – 1000 WT fractions are markedly enriched around the TSS of 70 – 75% of the genes in the reference dataset (Figure 40). For a small portion of genes (<10%), C-CAP reads exhibit enrichment downstream of the TSS. This is in accordance with reports of intragenic CGIs that are particularly found in a subset of developmental genes (Gardiner-Garden and Frommer, 1987; Illingworth et al., 2010; Lee et al., 2017; Maunakea et al., 2010).

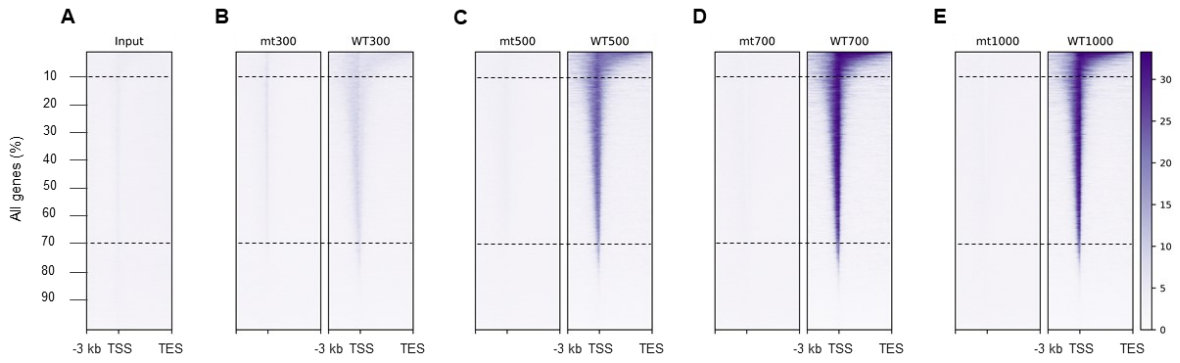


Figure 40: C-CAP sequencing reads are enriched at the TSS of genes
Heatmaps showing C-CAP sequencing reads enriched around the TSS of annotated mouse genes obtained from the RefSeq database. Dashed lines mark 10% and 70% of all genes, respectively. The same input sample (A) was loaded on ZF-CXXC^{mt} and ZF-CXXC^{WT} affinity matrices. Heatmaps of the eluates (B – E) compare the mt/WT pair of each salt elution fraction. Enrichment of transcription starts sites begins in the WT300 fraction (B) and reaches its maximum in the WT700 and WT1000 fractions (D, E). In all WT eluates, 70 – 75% of genes show enrichment of C-CAP reads around the TSS (B – E). For less than 10% of genes, WT eluates also enrich intragenic regions i.e., between TSS and TES. No enrichment is observed in mt eluates for the examined regions. n = 3.

Having validated that C-CAP specifically enriches NMIs (Figure 39) and transcription start sites (Figure 40) from chromatin samples, I set out to ensure my approach was not biased towards NMIs with particular chromatin states. Different genomic features (e.g. promoters and enhancers) and chromatin states (e.g. euchromatin vs. heterochromatin) are marked by distinctive histone modifications. Therefore, I used publicly available ChIP-sequencing data for histone modifications in mESCs and examined how my C-CAP signal was enriched at sites occupied by specific histone modifications. For this, I selected seven histone modifications which represent different regulatory features and chromatin states (Dunham et al., 2012; Ernst et al., 2011). H3K4me3 is found at active CGI-associated promoters, whereas H3K4me1 is mostly localised at enhancers, which have less dense regions of non-methylated DNA. Both, promoters and enhancers, are also marked by H3K9ac, which has a slight preference for active promoters, and H3K27ac, which is more pronounced at active enhancers. In contrast, H3K27me3 is enriched at silent CGI-associated promoters. As a

negative control, I also examined H3K9me3 and H3K36me3. H3K9me3 is associated with constitutive heterochromatin and H3K36me3 is found over actively transcribed regions such as gene bodies. Sites enriched for these modifications tend to lack non-methylated DNA and infrequently overlap with CGIs.

To perform this analysis, I first combined the C-CAP sequencing reads from the 500 – 1000 mM elution fractions as they showed the highest NMI enrichment in the WT condition. Then, I unbiasedly identified genomic sites of statistically significant enrichment (also known as peaks) for each histone modification and plotted C-CAP reads at these histone modification peaks. This revealed a strong enrichment of C-CAP signal at sites occupied by H3K4me3, H3K9ac, and H3K27ac which together mark active gene promoters where CGIs predominate (Figure 41A-C). C-CAP reads also co-localise with H3K4me1, however, to a lesser extent than with H3K4me3 (Figure 41D). This agrees with a previous study that found 17% of intergenic NMIs in mESC to be associated with the enhancer mark H3K4me1 (Long et al., 2013b). Importantly, when I plotted C-CAP reads over H3K27me3 peaks, there was also a clear enrichment of C-CAP signal indicating that silent chromatin is also efficiently recovered if it contains non-methylated DNA (Figure 41E). Together these observations indicate that C-CAP recovers chromatin containing non-methylated DNA independently of the underlying histone modification state. In contrast, there was little or no enrichment of C-CAP signal at sites containing H3K36me3 and H3K9me3 in agreement with the fact that sites with these histone modifications do not tend to have non-methylated DNA or overlap with CGIs (Figure 41F-G).

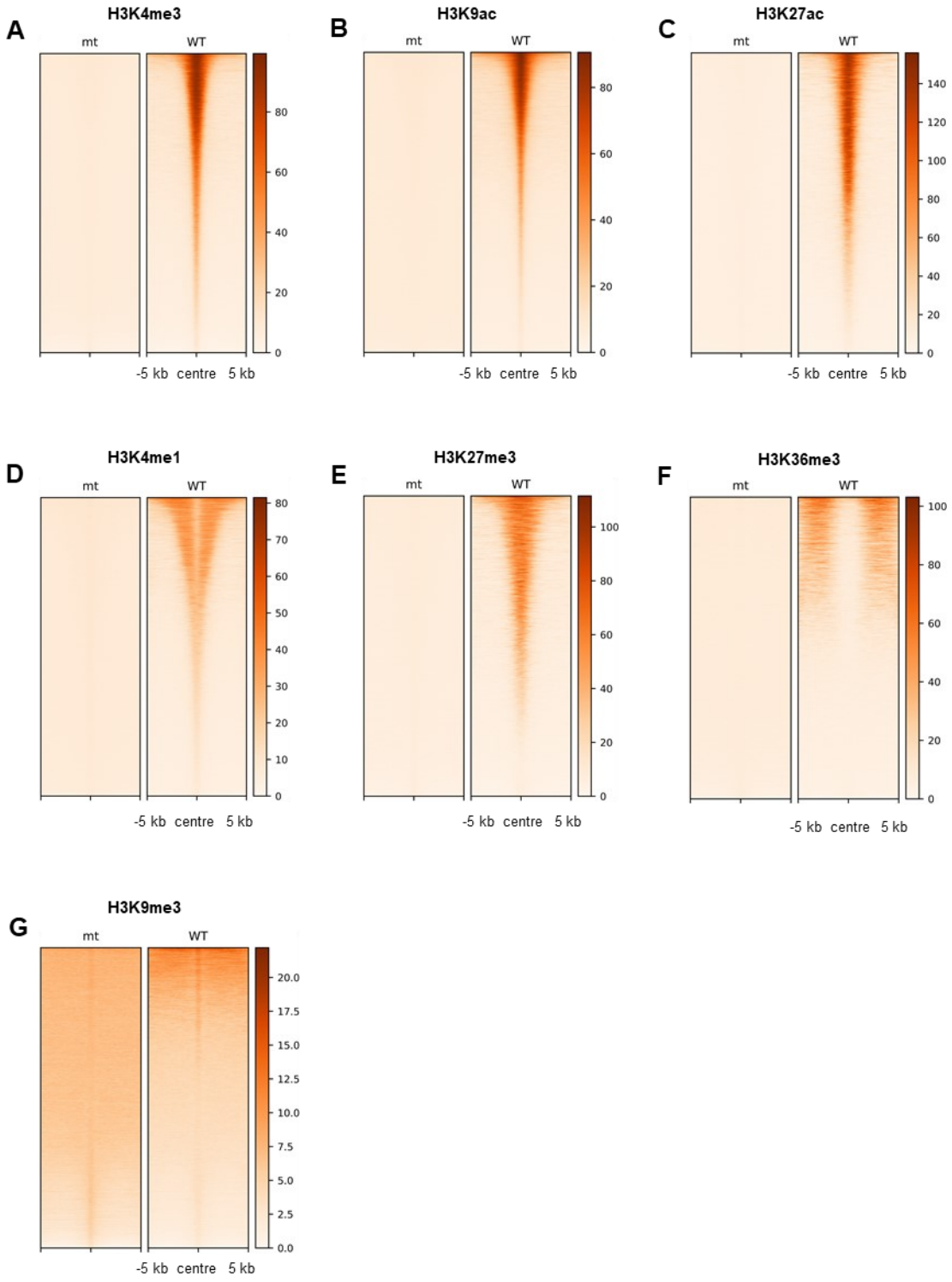


Figure 41: Analysis of C-CAP reads enriched at selected histone modification peaks

Heatmap analysis of C-CAP reads that are enriched at histone modification peaks which mark various gene regulatory elements and chromatin states. The sequencing reads from the 500 – 1000 mM elution fractions of each condition (mt, WT) have been combined and Input-normalised for the analysis. NMIs enriched by C-CAP correlate with histone modifications found at active promoters (A - C), inactive promoters (E), and active enhancers (B, C, D). In contrast, enriched NMIs are largely absent from regions occupied by histone modifications associated with actively transcribed gene bodies (F) and constitutive heterochromatin (G).

5.2.3 Sequencing analysis reveals enrichment of rDNA genes in C-CAP

The proteome analysis of the C-CAP eluates (see 4.2.4) showed enrichment of a considerable number of proteins involved in the transcription of ribosomal DNA (rDNA) and processing of ribosomal RNA (rRNA). I wondered if this enrichment was specific, or simply due to the abundance of factors involved in ribosome biogenesis. To answer this question, I focused on the rDNA gene locus in my C-CAP sequencing data. There are ~200 rDNA gene copies in the mouse genome which are arranged in arrays of tandem repeats across several chromosomes (Gaubatz and Cutler, 1978). The mouse rDNA gene spans ~45 kb and it is divided into two major parts (Figure 42A) (reviewed in McStay and Grummt, 2008). The coding region (~13 kb) is transcribed into a 45S pre-rRNA that contains the sequences of the 18S, 5.8S, and 28S rRNA. This is followed by a large intergenic spacer (IGS) that harbours several types of repeat elements (simple repeats, SINEs, LINEs) as well as a spacer promoter (SP) and a core promoter (CP). Importantly, the transcriptional activity of rDNA genes is regulated by methylation of conserved CpG sites in the promoter region (Santoro and Grummt, 2001). Additionally, the coding region is rich in CpG dinucleotides which are predominantly unmethylated (Yan et al., 2000).

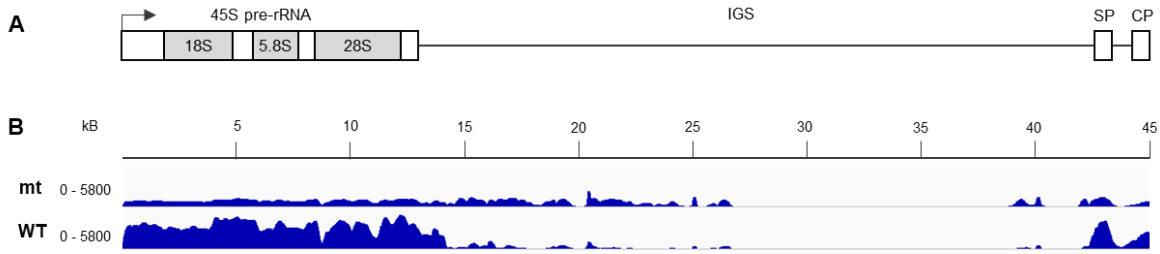


Figure 42: Specific regions of the rDNA gene locus are enriched by C-CAP
Schematic of the rDNA gene locus (A) and genome browser view (B) of mt and WT reads aligning to the rDNA gene locus. The entire locus is ~45 kb in length containing a ~13 kb coding region that is transcribed into the 45S pre-rRNA, and a large intergenic spacer (IGS) with a spacer promoter (SP) and a core promoter (CP). The mt and WT genome browser tracks represent the combined 500 – 1000 mM elution fractions. The arrow marks the transcription start site.

To examine if enrichment of rDNA was specific, I aligned the C-CAP reads to a single unit of a 45 kb rDNA gene copy and compared the combined 500 – 1000 mM elution fraction of mt and WT. This showed enrichment of WT reads in the promoter elements and the coding region of the 45S pre-rRNA (Figure 42B). Both are regions of the rDNA locus that contain non-methylated CpGs that can serve as potential binding sites for the ZF-CXXC^{WT} protein. The presence of non-methylated CpGs in combination with the large copy-number of rDNA loci seems to be the reason for the enrichment of proteins involved in rDNA transcription.

In summary, I have shown that the ZF-CXXC^{mt} protein has completely lost its ability to bind non-methylated CpGs (Figure 38). Massively parallel sequencing revealed that C-CAP enriches NMIs (Figure 39) in a manner similar to BioCAP on naked DNA, and predominantly recovers regions of DNA associated with annotated transcription start sites (Figure 40). Additionally, enriched chromatin is characterised by histone modifications that are associated with diverse expression states including active and silent promoters, and to a lesser degree CpG-rich enhancer elements (Figure 41) (Long et al., 2013b). Moreover, analysis of the rDNA gene locus showed specific enrichment at the promoter elements and the transcribed region, both of which are known to harbour non-methylated CpGs (Figure 42). Therefore, my genome-wide analysis of enriched DNA from C-CAP supports the conclusion that the ZF-CXXC WT and mt affinity matrices provide comparative tools to examine the CpG island proteome.

5.2.4 The ZF-CXXC^{mt} protein improves the specificity of the CGI proteome

Having shown that the ZF-CXXC^{WT} protein specifically enriches CGI chromatin but the ZF-CXXC^{mt} does not (Figure 39), I reasoned that this ZF-CXXC^{mt} could serve as an effective control in my C-CAP experiments to identify the *bona fide* CpG island proteome components. Therefore, I carried out MS on the ZF-CXXC^{WT} and ZF-CXXC^{mt} samples and then compared the resulting proteomes. I initially tested a two-step approach to determine the CGI proteome. In the first step, I tried to identify CGI-binding proteins by assessing their enrichment in the WT eluates relative to the initial chromatin preparation (input). This was the same approach that I used in my previous C-CAP analysis and it proved to be efficient in finding many CGI-binding proteins (see 4.2.4). However, this WT/Input analysis is not able to discriminate between those proteins that are enriched because they bind non-specifically to the affinity matrix and those that are enriched because they localise on CGIs. To overcome this shortcoming, I needed a second step in my analysis that allowed me to pinpoint the contaminants in my preliminary CGI proteome. Therefore, I tried to identify the contaminants by assessing the enrichment of proteins in the WT eluates relative to the mt eluates. Subsequently, I could exclude those proteins from my preliminary CGI proteome that showed enrichment in the mt eluates. The resulting CGI proteome would then have a higher specificity. To test if this two-step analysis improved specificity of the CGI proteome, I first combined the 500 – 1000 mM fractions as these showed the highest enrichment of CGIs in MS results (see 4.2.4) as well as the next generation sequencing analysis (see 5.2.2). Then, I quantified the abundance fold change of all proteins in the WT/Input and WT/mt pairs. The abundance fold changes in relation to their normalised log2 intensities

were visualised in MA plots to inspect and compare their distribution between both approaches (Figure 43). Additionally, components of PcG and TrxG complexes were highlighted as proxies for known CGI-binding proteins to obtain an initial indication of the enrichment. The MA plot of the WT/mt analysis showed a relatively sharp separation of proteins along the midline (log2 FC = 0) (Figure 43B). In contrast, proteins are more uniformly dispersed on both sides of the midline in the WT/Input analysis (Figure 43A). The sharper separation of proteins in the WT/mt compared to the WT/Input analysis possibly indicated an efficient identification of background proteins.

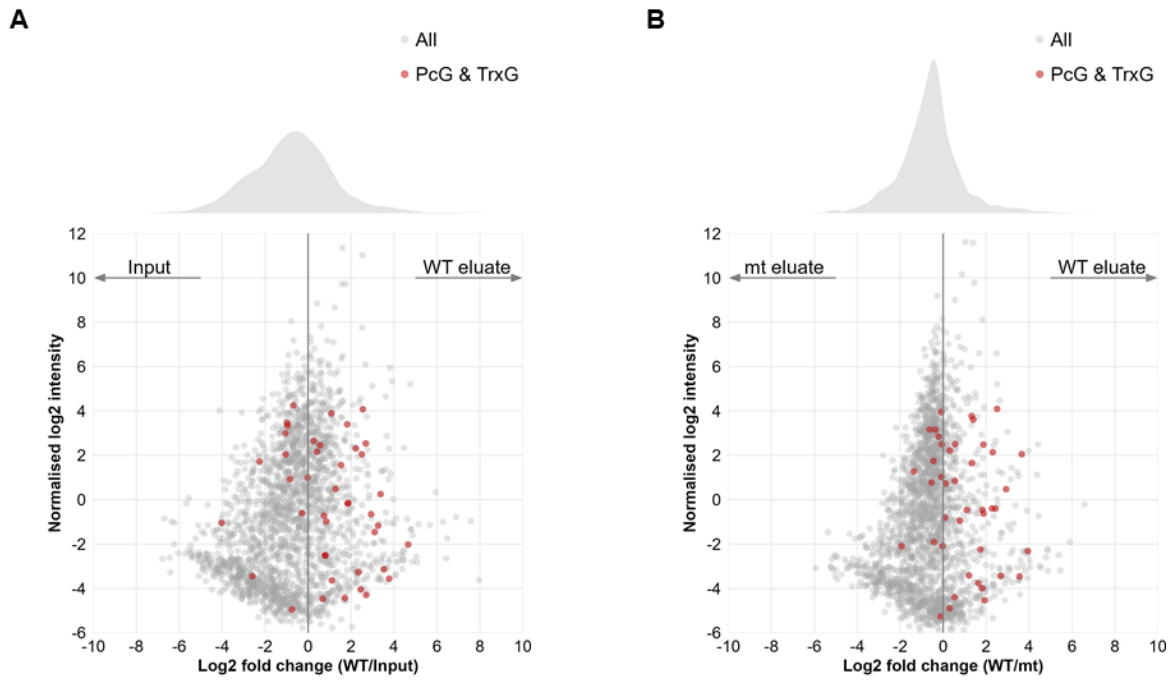


Figure 43: MA plots of WT/Input and WT/mt analyses
MA scatter plots comparing the distribution of abundance fold changes (log2 FC) of proteins in the WT/Input (A) and WT/mt (B) pairs. In (A), the enrichment in WT eluates is assessed relative to the input sample. In the second analysis (B), enrichment of potential background proteins in mt eluate is evaluated relative to the WT eluate. For both analyses, the 500 – 1000 mM elution fractions were combined. The density on top of the MA plots shows the distribution of log2 FC values for all proteins. Components of PcG and TrxG complexes have been highlighted in red to visually obtain a first indication of the enrichment of CGI-binding proteins. n = 3.

To examine in more detail if background proteins identified in the WT/mt analysis could help to improve the specificity of the CGI proteome, I performed statistical tests to focus on proteins that showed statistical significance in their enrichment. This resulted in 738 potential contaminants enriched in the mt eluate of the WT/mt analysis (Figure 44A) and 359 proteins enriched in the WT eluate of the WT/Input analysis (Figure 44B, Table S2). To exclude the contaminants from my preliminary CGI proteome of 359 proteins, I calculated

the overlap with the 738 background proteins identified in the mt eluates. Surprisingly, merely 39 proteins were shared between both datasets (Figure 44C). A qualitative examination revealed that the vast majority of the 39 proteins indeed seemed to be contaminants, however, many other contaminants were missed. For example, among the 39 proteins were only 10 out of 32 ribosomal proteins that were enriched in the preliminary CGI proteome. Additionally, the 39 proteins contained a few *bona fide* transcription-associated proteins such as LEO1 (PAF1 complex) and TAF2 (TFIID complex). Although the mt control slightly improved the specificity of the WT/Input dataset, the results indicated that the identification of contaminants in the preliminary CGI proteome was incomplete using the two-step analysis.

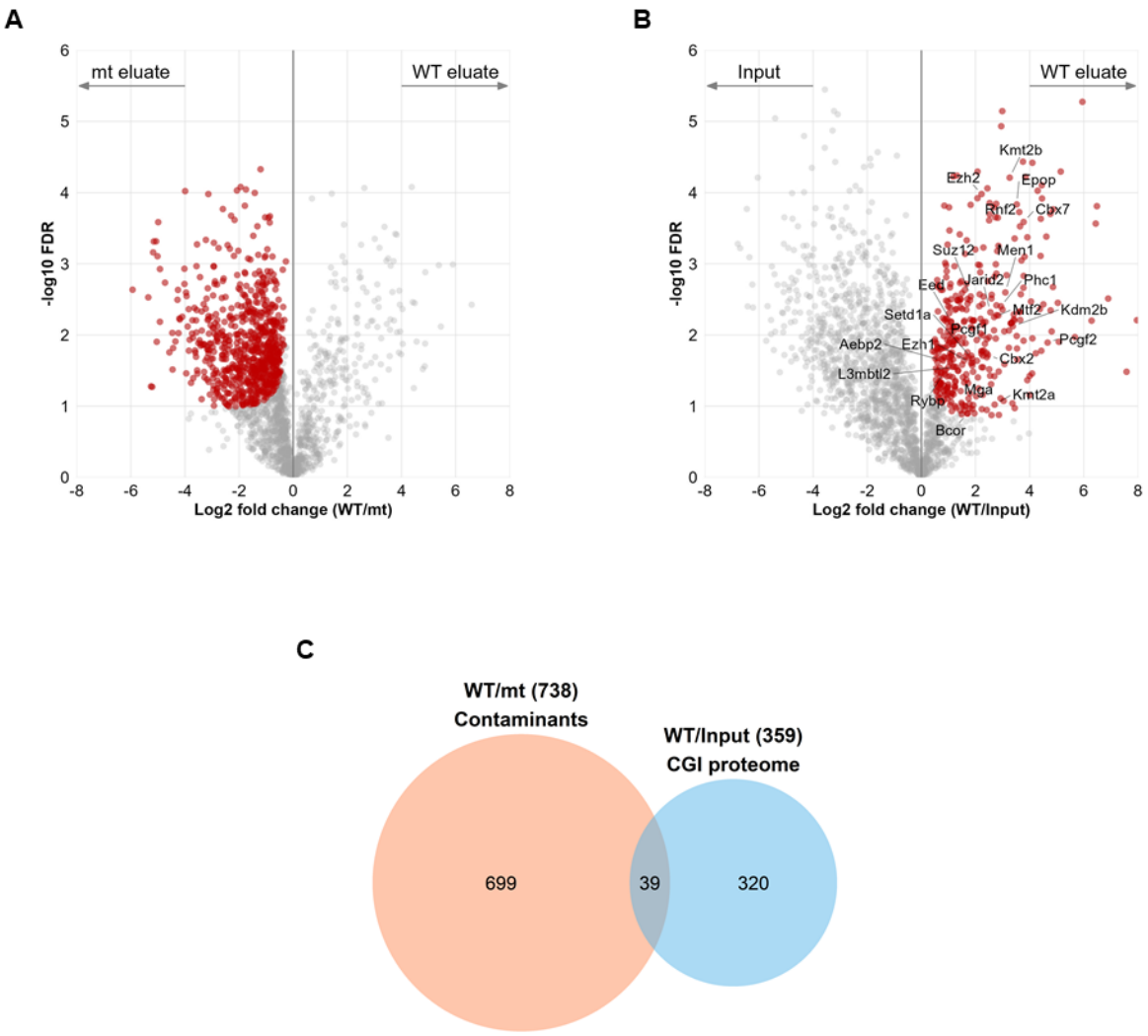


Figure 44: Volcano plots and Venn diagram of significantly enriched proteins in the mt eluate of WT/mt and in the WT eluate of the WT/Input analyses

(A) Volcano plot showing significantly enriched proteins in the WT/mt analysis. Proteins that are significantly enriched in the mt eluate are highlighted in red. These proteins represent potential contaminants. n = 3.

(B) Volcano plot showing significantly enriched proteins in the WT/Input analysis. Proteins that are significantly enriched in the WT eluate are highlighted in red, and additionally, components of PcG and TrxG complexes have been labelled. Highlighted proteins represent potential components of the CGI proteome. n = 3.

For (A) and (B) Welch's t-tests were performed to determine statistical significance with an FDR ≤ 0.05 considered significant. n = 3.

(C) Venn diagram illustrating the overlap of significantly enriched proteins between the mt eluate of (A) and the WT eluate of (B).

In my MA plots, I had noticed that PcG and TrxG components were similarly enriched in the WT eluates of the WT/Input and WT/mt approaches (Figure 43). Therefore, I wondered if I could use the WT eluate of the WT/mt approach directly as a CGI proteome instead of

performing the two-step analysis described above. To answer this question, I determined the significantly enriched proteins in the WT eluate of the WT/mt approach and compared them in terms of CGI-specificity to the 359 proteins that were identified in the WT/Input approach. This resulted in 178 significantly enriched proteins in WT/mt pair, with 132 proteins being shared with WT/Input pair (Figure 45, Table S3). As mentioned above, the significantly enriched proteins in WT/Input analysis contained 32 ribosomal proteins that were almost exclusively structural components of the ribosome. In contrast, the WT/mt dataset contained only one ribosomal protein. This suggested that the WT/mt approach was better at excluding abundant contaminants from the list of significantly enriched proteins, thereby improving the specificity of the CGI proteome.

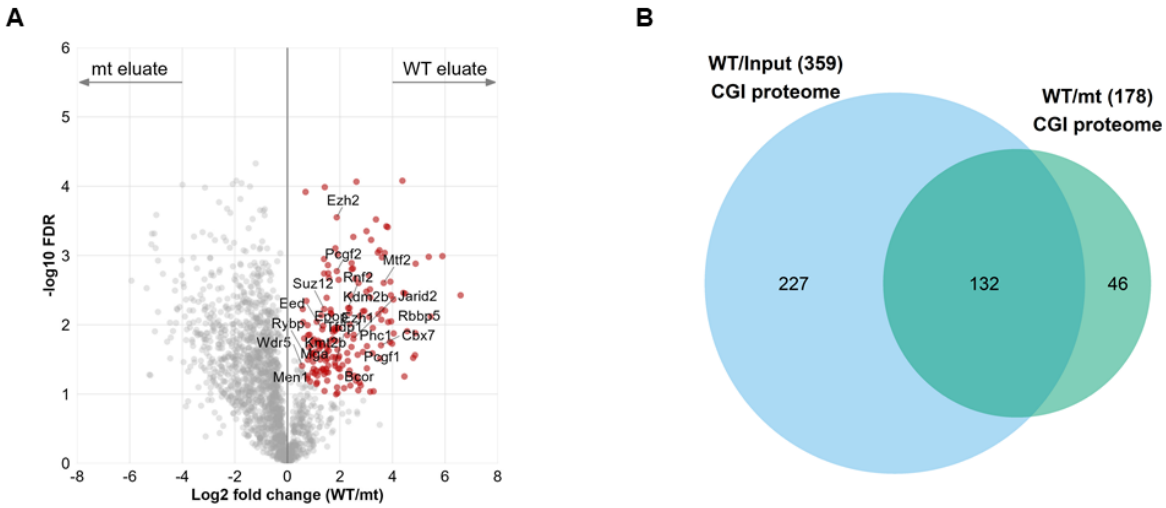


Figure 45: Volcano plot and Venn diagram of significantly enriched proteins in the WT eluate of the WT/mt analysis

(A) Volcano plot showing significantly enriched proteins in the WT eluate of the WT/mt analysis. Proteins that are significantly enriched in the WT eluate are highlighted in red, and additionally, components of PcG and TrxG complexes have been labelled. Highlighted proteins represent potential components of the CGI proteome. Welch's t-test was performed to determine statistical significance with an FDR ≤ 0.05 considered significant. n = 3.
(B) Venn diagram illustrating the overlap of significantly enriched proteins between the WT eluate of Figure 45A and the WT eluate of Figure 44B.

To investigate the specificity of the two approaches more comprehensively, I performed GO analysis on the two datasets of significantly enriched proteins. I hypothesised that the approach with less specificity would show GO terms in the top 10 list that are unrelated to transcription or the epigenetic regulation of gene expression. Indeed, the WT/Input analysis revealed in the top 10 list of all three GO categories (BP, CC, MF) terms that are associated with ribosomes and translation (red boxes in Figure 46A, C, E). It is important to differentiate here between factors that are involved in the transcription of ribosomal DNA and those that

are structural constituents of the ribosome, or other factors that participate in translation of poly(A) RNA in the cytosol. As I have shown in 5.2.3, parts of the rDNA gene locus are specifically enriched in the ZF-CXXC^{WT} eluates. Hence, the enrichment of proteins directly participating in rDNA transcription can be regarded as specific. This is also reflected in the GO analysis, where those factors are summarised e.g. under the terms “transcription initiation from RNA polymerase I promoter” and “chromatin silencing at rDNA”. Both of these GO terms are equally enriched in the WT/Input and WT/mt approaches (green boxes in Figure 46A, B). In contrast, proteins that are further downstream in the ribosome biogenesis pathway i.e., those that play a role after rDNA transcription, need to be categorised as contaminants. They feature in the list of significant proteins of the WT/Input approach due to their high cellular abundance and possibly stickiness to the affinity matrix. Importantly, by using the ZF-CXXC^{mt} eluates as my reference sample in the statistical analysis, I can differentiate between specific enrichment of rDNA transcription-associated proteins and non-specific ribosomal contaminants.

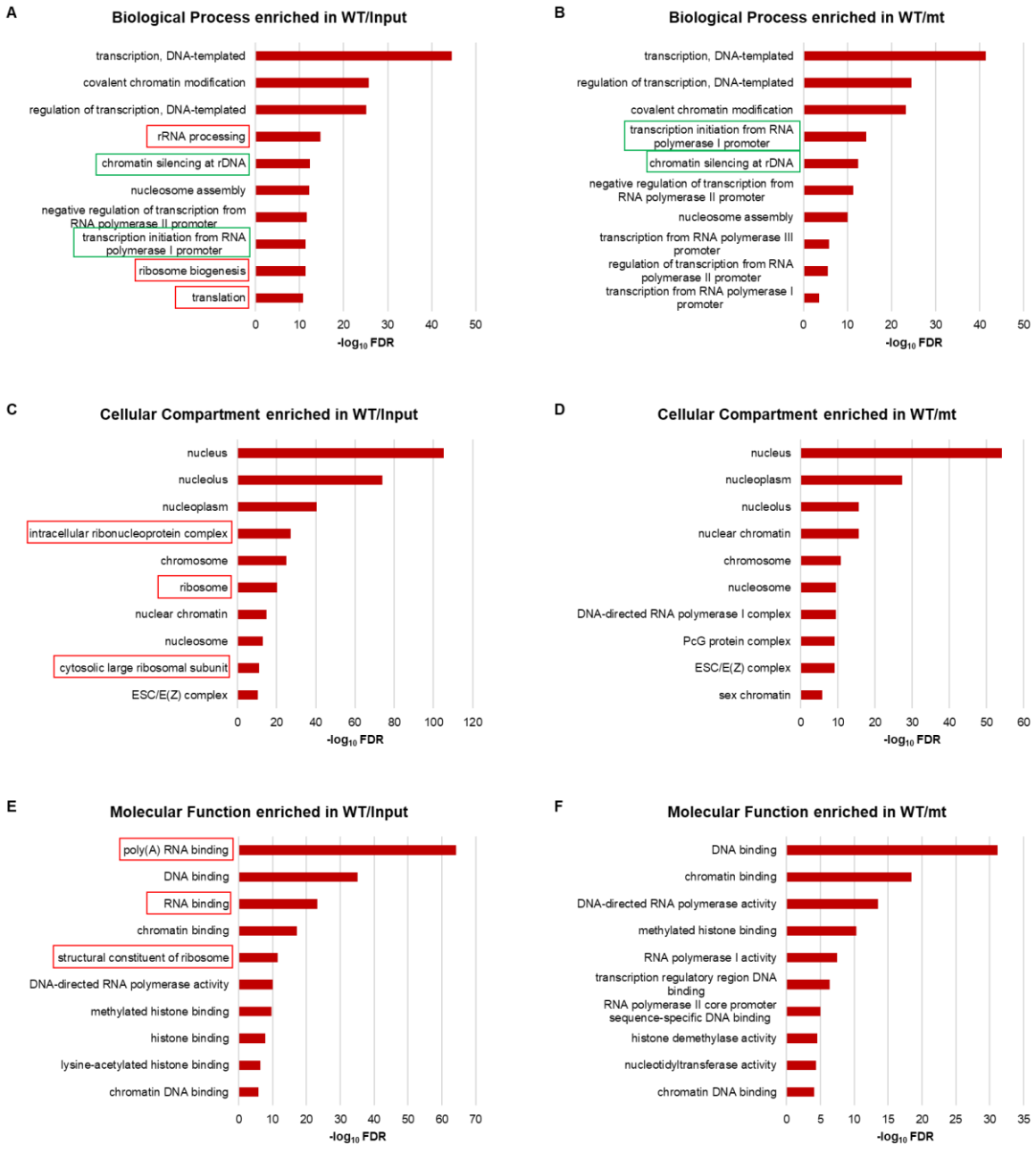


Figure 46: GO analysis of significantly enriched proteins in the WT/Input and WT/mt approach
Bar charts showing the top 10 enriched GO terms for the GO categories Biological Process (A - B), Cellular Compartment (C - D), and Molecular Function (E - F). The left side (A, C, E) shows significantly enriched GO terms for the WT/Input approach, while the right side (B, D, F) shows the same for the WT/mt approach. Red boxes highlight GO terms that are related to ribosomes and translation. Green boxes emphasise specifically rDNA transcription-associated GO terms.

In conclusion, the use of the ZF-CXXC^{mt} as a negative control in C-CAP helped to identify proteins that are abundant contaminants and/or interact non-specifically with the affinity matrix. Quantifying the enrichment of proteins relative to the mt eluates instead of the input, improved the overall specificity of the CGI proteome.

5.2.5 The identification of new CGI-associated proteins

The motivation for developing the C-CAP method was to investigate the CGI proteome in an unbiased and comprehensive manner, and possibly identify novel CGI-binding proteins. One way that can help to identify potentially new CGI interactors in a set of proteins is to investigate their interrelations by performing a network analysis. This is based on the hypothesis that proteins localising on CGIs would be functionally interconnected. To build a protein-protein interaction network, I retrieved information on protein associations from the STRING database (Doncheva et al., 2019; Szklarczyk et al., 2019) and combined this with an enrichment analysis to group functionally related proteins (Figure 47).

RNA polymerase I – III subunits and transcription initiation factors represent protein groups that are at the core of the transcription process. Therefore, these groups show the highest connectivity with each other and to other groups in the network (Figure 47). It is interesting to note that several subunits of rDNA-specific RNA polymerase I and pre-initiation complex are enriched in C-CAP, providing further evidence for specific enrichment of the rDNA gene locus.

Strongly connected to the core transcription machinery are protein groups that take up a regulating role. These are PcG and TrxG complexes, histone demethylases, transcription factors, and chromatin remodellers. Interestingly, network analysis also identified the Moz/Morf histone acetyltransferase complex among the transcription-regulating groups. Three out of four subunits of the Moz/Morf complex were significantly enriched, while the fourth subunit (MEAF6) was just below the significance cut-off. So far, the Moz/Morf complex has not been shown to be associated with CGIs in mESCs, and therefore, it might present a new component of the CGI proteome.

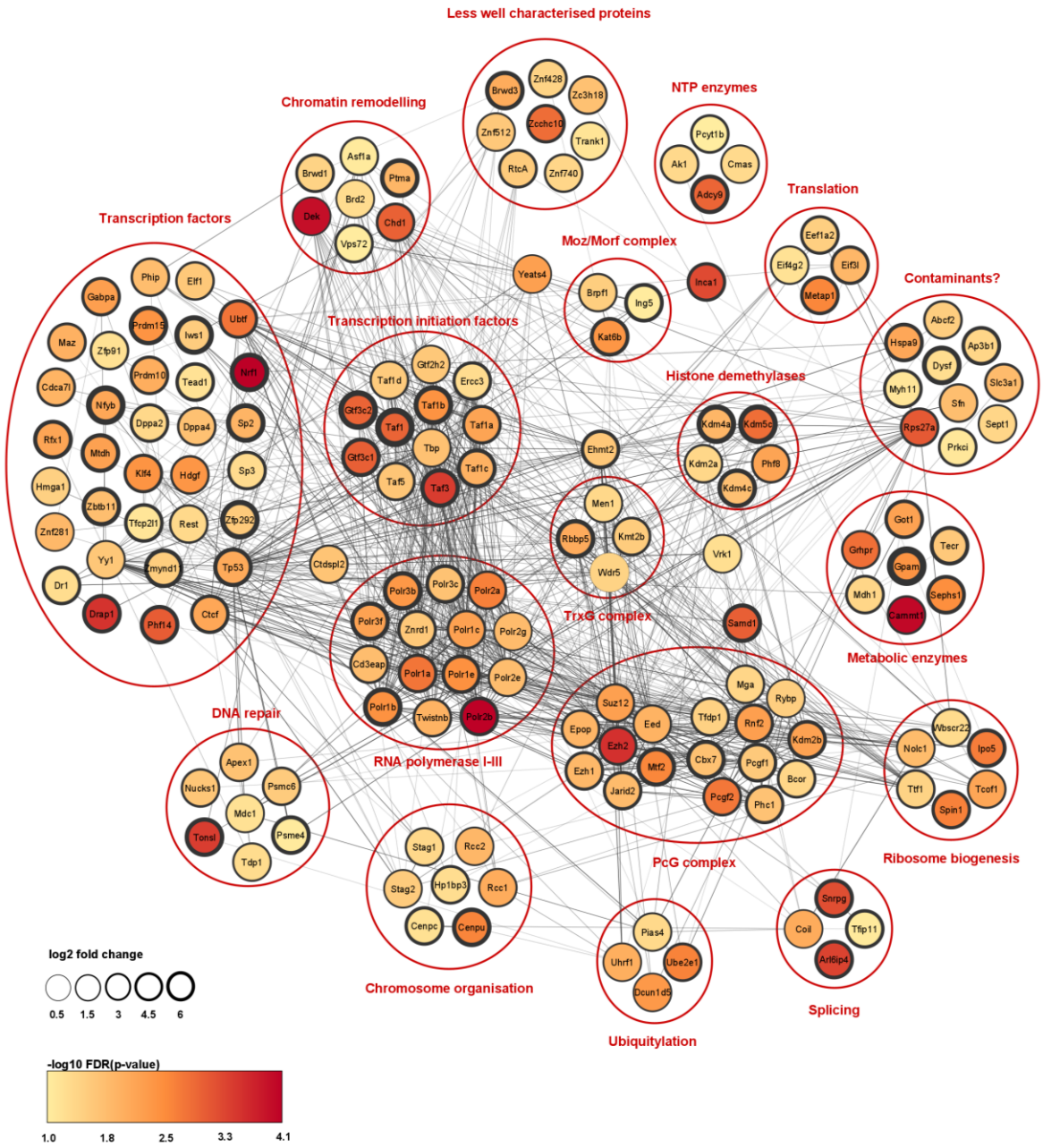


Figure 47: Network analysis of significantly enriched proteins in C-CAP
Network analysis showing significantly enriched proteins in the ZF-CXXC^{WT} eluates of C-CAP. Interaction data was obtained from the STRING database. Log2 FC and FDR values are taken from the analysis in Figure 45A. Proteins were grouped based on GO enrichment analysis. The darker the shade of the connecting lines, the higher the confidence of the interaction.

To narrow down my search for potentially new CGI-binding proteins, I looked for certain features in the enriched proteins which I considered to be indicative of a potential role in binding CGIs. The two most important features I looked for were nuclear localisation and a chromatin-binding domain. Since CGIs are an attribute of vertebrate genomes, I also searched phylogenomic databases to find out if the protein was conserved among

vertebrates, which could have indicated biological importance of the protein. Moreover, I gathered information about knockout and mutation phenotypes which could have given clues about the biological role in development or cellular function. Additionally, I reasoned that proteins associated with CGIs would likely co-localise with histone marks found at promoters or enhancers, and they would have a high affinity for non-methylated CpGs if they bound directly to DNA. There are several proteomic studies on proteins bound at various histone marks or at CpG dinucleotides with different modifications (Bartke et al., 2010; Eberl et al., 2013; Engelen et al., 2015; Ji et al., 2015; Spruijt et al., 2013a). I filtered the proteomics data from these studies for proteins that bound promoter- and enhancer-associated histone marks, or unmethylated CpGs. Afterwards, I searched the filtered list for proteins that were also found to be enriched in C-CAP. Using these various criteria to search for potentially new CGI-binding factors resulted in a list of 17 proteins (Table 12).

Table 12: Shortlist of potentially new CGI-binding proteins

BRPF1	PHF14	ZFP91
BRWD3	VRK1	ZMYND11
CDCA7L	ZC3H18	ZNF428
CDCA7	ZCCHC10	ZNF512
CMAS	ZFA	ZNF740
KAT6B	ZFP292	

Having narrowed down candidates that might be novel CGI-associated proteins, I then set out to verify their genomic localisation using ChIP-sequencing as an orthogonal approach. Since many of the proteins I was interested in were uncharacterised, I also had to take into account the availability of ChIP-grade antibodies which substantially limited the number of candidates I could investigate. In the end, I decided to examine CGI-binding by ChIP for the proteins KAT6B, CDCA7, CDCA7L. Although not proven to work in ChIP, the available antibodies for these proteins were validated for antigen-specificity by the manufacturers using e.g., western blot, co-immunoprecipitation, and immunofluorescence microscopy.

KAT6B is the catalytic subunit of the Moz/Morf complex. It was reported to bind to the promoter marks H3K4me3 and H3K27ac as well as to non-methylated cytosine in mESCs (Engelen et al., 2015; Ji et al., 2015; Spruijt et al., 2013a). Regarding CDCA7, studies have shown that it acts as a transcription regulator either on its own or through association with MYC (Gill et al., 2013; Goto et al., 2006). Furthermore, a complex comprising CDCA7 and the helicase HELLS showed nucleosome remodelling activity *in vitro* (Jenness et al., 2018). The CDCA7-HELLS complex might be involved in regulating CpG methylation at

centromeric and pericentromeric repeats (Unoki et al., 2019). CDCA7L is closely related to CDCA7, and just as CDCA7, it associates with MYC to function as a transcriptional regulator (Huang et al., 2005). Moreover, CDCA7L interacts with the transcriptional co-activator PSIP1 on chromatin (Hendrix et al., 2014). PSIP1 in turn has been shown to interact with the TrxG proteins KMT2A and MEN1 (Čermaková et al., 2014; Huang et al., 2012; Murai et al., 2014).

To increase the chances of obtaining a sufficiently high signal/noise ratio for ChIP sequencing, I tested three different crosslinking conditions. Two of the ChIP experiments were performed with different effective formaldehyde concentrations and the third included a long-range crosslinking agent in addition to formaldehyde. Experience from the Klose laboratory has shown that a double crosslinking approach can increase the signal/noise ratio considerably for some chromatin-binding proteins. After performing the ChIP assays, I examined the enrichment at various genomic locations by qPCR. This included the promoter and gene body regions of three CGI-promoter genes (*Ncoa2*, *Kmt5b*, *Bmp7*), one non-CGI promoter gene (*Fgf7*), and a region without any nearby genes (gene desert). A requirement for subsequent ChIP sequencing of the samples is at least a 10-fold enrichment of promoter over gene body (or gene desert) in qPCR. If the qPCR fold enrichment is below this threshold, the low signal/noise ratio will considerably impair identification of genomic binding sites in ChIP sequencing. Quantitative PCR analysis revealed that none of the antibodies achieved an acceptable fold enrichment in any of the tested conditions (Figure 48, 49, 50). This could indicate that the antibodies are not suitable for ChIP. Alternatively, these proteins may not bind to CGIs or the examined CGIs. Thus, CGI-binding could not be confirmed for the three protein candidates using the available antibodies.

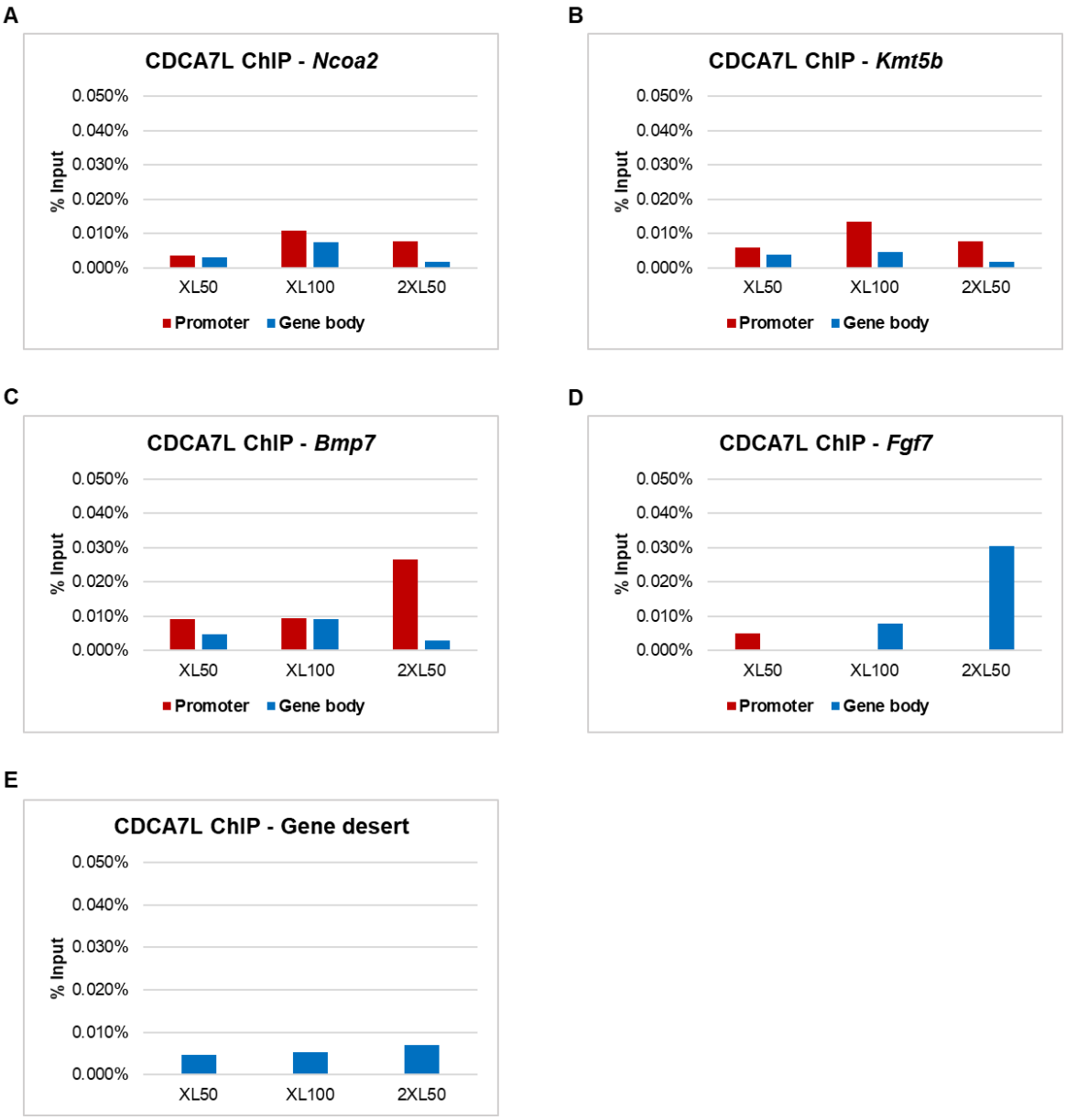


Figure 48: ChIP-qPCR analysis for CDCA7L

Bar charts showing the qPCR analysis of CDCA7L ChIP. Three conditions were tested: single crosslinking of 50 x 10^6 (XL50) or 100 x 10^6 (XL100) cells with formaldehyde, and double crosslinking of 50 x 10^6 cells with formaldehyde and DSG (2XL50). ChIP signal relative to Input DNA was evaluated at promoter, gene body, and gene desert regions. Fold-enrichment of promoter vs gene body (or gene desert) is less than 10 at CGI-promoter genes (*Ncoa2, Kmt5b, Bmp7*) and the non-CGI promoter gene (*Fgf7*). Some bars are not visible because the enrichment is too low.

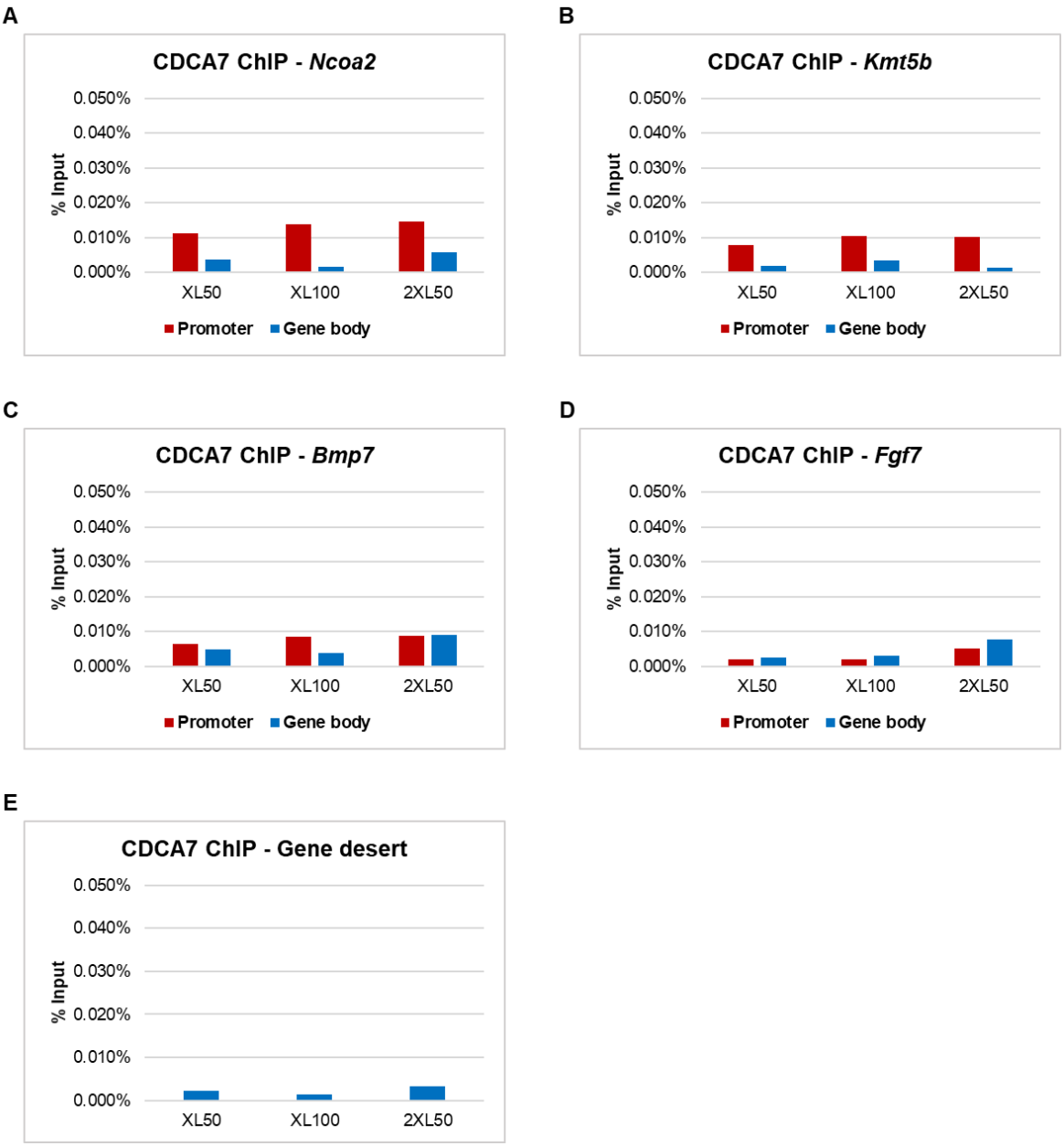


Figure 49: ChIP-qPCR analysis for CDCA7

Bar charts showing the qPCR analysis of CDCA7 ChIP. Three conditions were tested: single crosslinking of 50 x 10^6 (XL50) or 100 x 10^6 (XL100) cells with formaldehyde, and double crosslinking of 50 x 10^6 cells with formaldehyde and DSG (2XL50). ChIP signal relative to Input DNA was evaluated at promoter, gene body, and gene desert regions. Fold-enrichment of promoter vs gene body (or gene desert) is less than 10 at CGI-promoter genes (*Ncoa2, Kmt5b, Bmp7*) and the non-CGI promoter gene (*Fgf7*). Some bars are not visible because the enrichment is too low.

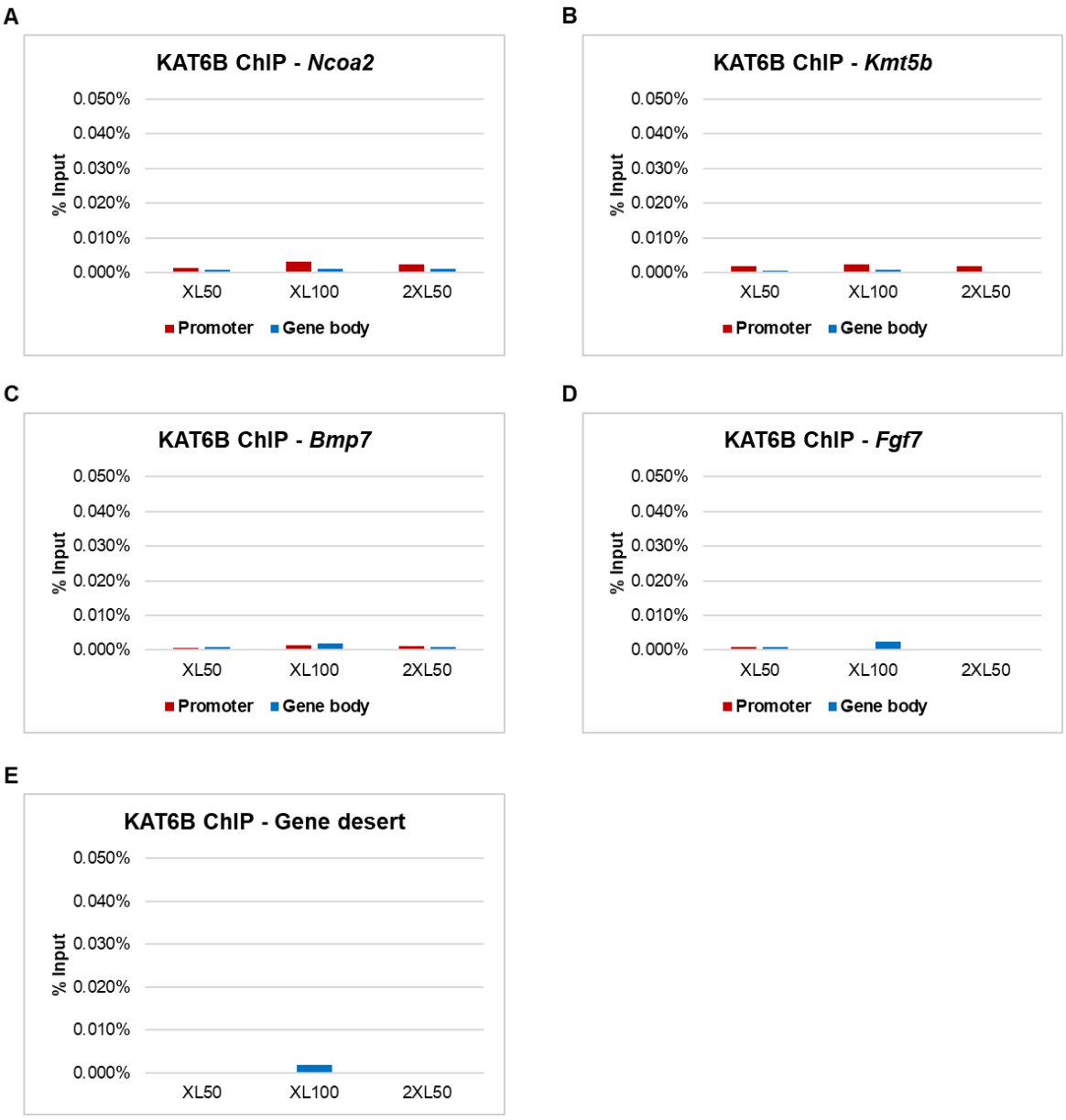


Figure 50: ChIP-qPCR analysis for KAT6B
Bar charts showing the qPCR analysis of KAT6B ChIP. Three conditions were tested: single crosslinking of 50 x 10^6 (XL50) or 100 x 10^6 (XL100) cells with formaldehyde, and double crosslinking of 50 x 10^6 cells with formaldehyde and DSG (2XL50). ChIP signal relative to Input DNA was evaluated at promoter, gene body, and gene desert regions. Fold-enrichment of promoter vs gene body (or gene desert) is less than 10 at CGI-promoter genes (*Ncoa2, Kmt5b, Bmp7*) and the non-CGI promoter gene (*Fgf7*). Some bars are not visible because the enrichment is too low.

While I was developing the C-CAP method, three studies were published that reported new CGI-binding properties for ZBTB11 (Wilson et al., 2020), SAMD1 (Stielow et al., 2021), and BANP (Grand et al., 2021). The authors of these studies performed ChIP sequencing in mESCs using either endogenously tagged targets or custom-made antibodies to determine the genomic distribution of the proteins. All three proteins were also identified in the CGI

proteome of my combined C-CAP experiments. Therefore, I wondered if the ChIP sequencing peaks of these proteins coincided with the location of NMIs as determined by the reference dataset (Long et al., 2013b). To investigate this, I analysed the publicly available sequencing data of SAMD1, BANP, and ZBTB11. The analysis revealed that peaks of all three examined proteins correlate well with NMI regions (Figure 51). Although I was not able to validate new CGI-binding proteins, evidence from aforementioned studies shows that C-CAP is capable of identifying novel CGI interactors.

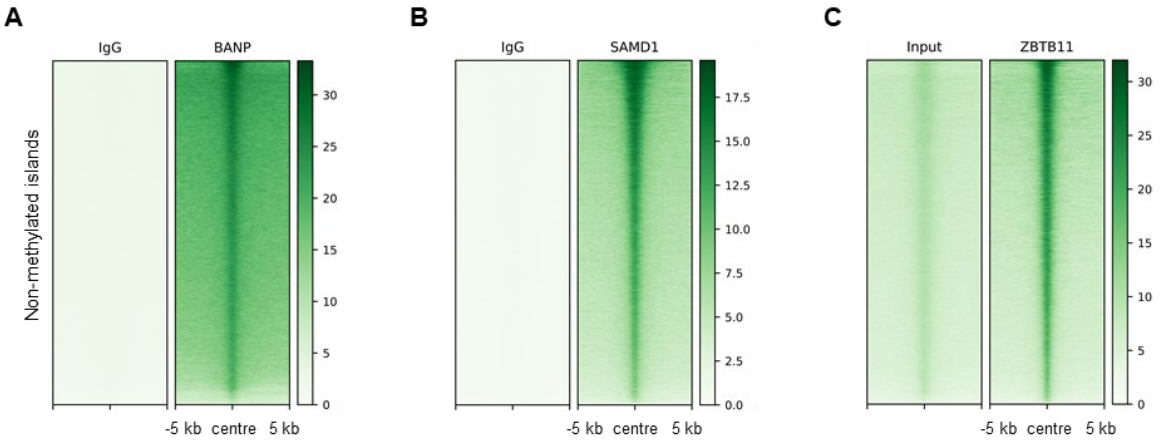


Figure 51: BANP, SAMD1, and ZBTB11 localise at NMIs in mESCs
Heatmap analysis showing ChIP-sequencing peaks of SAMD1 (A), BANP (B), and ZBTB11 (C) localising at NMIs. Peaks of all three proteins (A - C) are enriched at NMIs in the IP sample, but not in the respective control sample.

In conclusion, I demonstrated that C-CAP identifies many known CGI-binding proteins, among which are components of the core transcription machinery, PcG and TrxG complexes, and several other transcription regulators. Additionally, my analysis pointed to novel CGI interactors such as the Moz/Morf complex. Due to a lack of commercial ChIP-grade antibodies, I was not able to confirm CGI localisation for protein candidates. However, recent publications from others revealed that C-CAP has sufficient specificity to identify novel CGI interactors.

5.3 Discussion

Several methods have been developed in the past ~15 years to investigate the chromatin proteome by mass spectrometry (reviewed in Gauchier et al., 2020; van Mierlo and Vermeulen, 2021). However, a method to examine the CGI proteome in an unbiased and comprehensive manner was missing. CGIs exert an important role in gene regulation that

is mechanistically not well understood because we do not know the complete complement of proteins that contribute to their function. To overcome this fundamental lack of understanding, I have developed C-CAP. Using next generation sequencing, I have shown that C-CAP enriched non-methylated islands in the mouse genome (Figure 39). The enriched regions predominantly overlap with the 5' end of annotated genes (Figure 40), and they feature histone modifications associated with active and silent promotors as well as gene enhancers (Figure 41). Additionally, C-CAP enriched the promoter and coding regions of rDNA genes due to the presence of non-methylated CpGs in these regions (Figure 42). Therefore, C-CAP could also be applied to investigate transcription-related steps of ribosome biogenesis. This might help to better understand how defects in pre-rRNA synthesis cause certain ribosomopathies such as Treacher-Collins syndrome (Mills and Green, 2017; Sakai and Trainor, 2009).

To increase the specificity of my CGI proteome, I aimed to identify and exclude the contaminants from downstream analysis. For this, I used a mutant ZF-CXXC protein that had lost its binding capability for non-methylated CpGs. Using the ZF-CXXC^{mt} as a control in quantifying significantly enriched proteins considerably improved the overall specificity of the CGI proteome, however, two issues remained. First, when using the ZF-CXXC^{mt} eluates as the reference instead of the input some *bona fide* CGI-binding proteins (e.g. SETD1A, KMT2A, CBX2) dropped out of the list of significantly enriched proteins. Second, several obvious contaminants (e.g. myosin, translation initiation factors) were not excluded from the CGI proteome (Figure 47). These issues might have been caused by the quantification shortcomings of label-free proteomics and possibly they could be overcome by using stable isotope labelling in the future.

To identify new CGI-binding proteins in my eluates, I applied network analysis combined with functional enrichment analysis. This helped me to visualise the interrelations between the proteins and to group those that share common functions (Figure 47). Using network analysis, I identified the Moz/Morf histone acetyltransferase complex as a potentially new component of the CGI proteome in mESCs. Other new CGI interactors were less obvious and required further investigation by mining the literature for evidence. A general limitation of network analysis is that it relies on protein annotation data from public databases. Therefore, network analysis has a bias for well characterised proteins. To overcome this limitation, I tested Significance Analysis of Interactome (SAINT), which is not dependent on any public databases. Instead, it employs the protein abundance information provided by spectral counts to compute interaction probabilities (Teo et al., 2014). However, SAINT

missed many known CGI-binding proteins (e.g. EED, SUZ12, RNF2, RNAPII subunits) which made it unreliable for analysing the CGI proteome.

I aimed to determine the genomic localisation for a few selected CGI protein candidates by ChIP sequencing. The final choice of proteins had to be made based on the availability of antibodies. A ChIP-validated antibody was not available for any of the selected targets. In the end, I decided to chromatin-immunoprecipitate KAT6B, CDCA7, and CDCA7L because the antibody manufacturers at least provided validation data for antigen-specificity in other assay types. However, the ChIP experiments did not confirm CGI-binding for any of the proteins (Figure 48, 49, 50), possibly because the antibodies were not ChIP-suitable or alternatively these may be false positives in my analysis. To overcome the limitation of ChIP-grade antibody availability, one would have to make cells lines with endogenously tagged targets. For example, ChIP-grade antibodies are available for FLAG, HA, and T7 tags. However, a systematic production of such cell lines could not be pursued in this study due to time constraints and the COVID-19 pandemic.

Three recent studies presented SAMD1 (Stielow et al., 2021), BANP (Grand et al., 2021), and ZBTB11 (Wilson et al., 2020) as new CGI-binding proteins. All three proteins have also been detected in the CGI proteome of my combined C-CAP experiments. Strikingly, at the time of examining my C-CAP data for new CGI protein candidates, SAMD1 was still poorly characterised. It did not have any obvious chromatin-interaction domain, and it was annotated as a cytoplasmic and extracellular protein in public databases. However, proteomic studies suggested that it binds CpG-rich and non-methylated DNA (Bartke et al., 2010; Perino et al., 2018; Viturawong et al., 2013; Xiong et al., 2016). Following this clue, Stielow et al. identified a new type of winged-helix (WH) domain in SAMD1, which recognises CGIs in a completely different way than the WH domain of Polycomb protein MTF2 (Perino et al., 2018). Intriguingly, SAMD1-like WH domains can be found also in KAT6A, KAT6B, and ZMYND11 (Stielow et al., 2021). KAT6B and ZMYND11 are putative CGI-binding proteins according to my C-CAP dataset (Figure 47). In conclusion, although I could not validate CGI-binding for selected candidates due to a lack of ChIP-grade antibodies, evidence comes from other studies that C-CAP has the potential to identify novel CGI interactors.

6 C-CAP in PRC1 and PRC2 CKO cells provides new insight into chromatin composition and function of CGIs

6.1 Introduction

Chromatin is a heterogeneous assembly comprising DNA, histones, DNA- and histone-binding proteins, and non-coding RNAs (Bernstein et al., 2010). Its composition changes dynamically through the modification and exchange of bound factors. In particular, on gene regulatory elements such as promoters and enhancers, chromatin composition plays a pivotal role because it helps to regulate gene activity. A prerequisite for developmental processes and normal cell physiology is the ability to adapt gene activity according to environmental cues and metabolic requirements. The adaption of gene activity is in part regulated through dynamic changes of the chromatin composition at gene regulatory elements that help to control whether genes are actively transcribed, silenced, or kept in a transitional state. As recruitment sites for transcription-regulating proteins, CGIs exert an important function in determining the chromatin composition at gene promoters. Capturing the CGI chromatin composition is essential for understanding the role that CGIs play in gene regulation and in guiding developmental processes and maintaining normal cell physiology.

Typically, the proteome around gene promoters and enhancers has been examined with approaches such as ChIP-MS. This relies on using antibodies that bind to histone modifications or proteins that are known to bind to these gene regulatory elements in order to isolate and characterise the surrounding chromatin composition. However, ChIP-MS cannot be easily coupled to perturbation approaches to examine how factors influence the chromatin composition at gene regulatory elements. This is because a given perturbation may also influence the occupancy or distribution of the protein or modification used to isolate target site chromatin. Therefore, it currently remains challenging to unbiasedly examine how the chromatin composition at regulatory sites such as CGIs is affected, when factors that function at these elements are perturbed.

C-CAP applies a DNA-mediated chromatin pulldown which enables an unbiased investigation of the CGI proteome. Since C-CAP does not depend on the presence of a particular protein to isolate target site chromatin, it should in theory be capable of examining the CGI proteome under conditions where proteins that function at CGIs have been perturbed. In this chapter, I will investigate how the CGI proteome changes if I disrupt gene regulatory processes. To achieve this, I have used two different mESC lines in which I can

conditionally knockout (CKO) the genes for core subunits of Polycomb repressive complex (PRC) 1 and 2. I have chosen these complexes because they represent relatively well characterised CGI-binding factors in mESCs. The extensive literature on PcG proteins in mESCs allows me to form hypotheses about the effects I might observe in the CGI proteome when I knockout core subunits of PRC1 or PRC2. For example, structural and biochemical studies provide information about distinctive complex variants, the interactions within the complexes, and how they recognise their chromatin substrates (reviewed in Blackledge and Klose, 2021). Moreover, we have information about their genomic localisation and their function in gene regulation in wild-type cells and cells with perturbation of complex subunits. Beyond verifying the known effects of PRC1/2 knockouts, the unbiased nature of C-CAP can help to uncover novel synergistic or antagonistic relationships with PcG complexes.

To examine the CGI proteome changes in these CKO cells, I will first validate that the perturbation does not affect the enrichment of CGIs by C-CAP. I will then examine how the loss of core subunits affects other complex components and proteins localised on CGIs. Finally, I will use calibrated ChIP-sequencing as an orthogonal approach to further explore some of the proteome changes that have been detected by MS.

6.2 Results

6.2.1 C-CAP equally recovers CGIs in PRC1 and PRC2 CKO cells

PRC1 and PRC2 are essential transcription repressors in mESCs that guide cell differentiation programmes and help to maintain cell identity (Boyer et al., 2006; Schuettengruber and Cavalli, 2009). Each complex can counteract transcription on its own, however, their convergence on the same genomic sites amplifies the repressive effect on transcription. Several studies have shown that genetic knockout and catalytic inactivation of PRC1 or PRC2 has vast effects on transcription, leading to the derepression of hundreds of genes (Blackledge et al., 2020; Fursova et al., 2019; Loh et al., 2021). Since PRC1 and PRC2 regulate transcription through their association with CGI promoters, I reasoned that a knockout of these complexes could manifest itself in marked changes of the CGI proteome. These chromatin composition changes in turn, could provide new leads to explain how transcription is regulated at CGIs.

The PRC1 and PRC2 cell lines that I used to investigate the CGI proteome changes were previously developed and validated by members of the Klose laboratory (Fursova et al.,

2019 and Figure S1). The catalytic core of PRC1 comprises the E3 ubiquitin ligase RING1A or its paralogue RING1B (also known as RNF2), and one of six PCGF proteins (PCGF1-6). In mESCs, RING1A/B have overlapping functions in catalysing H2A monoubiquitylation and repressing transcription, and only knockout of both paralogues causes global gene derepression (Endoh et al., 2008; de Napoles et al., 2004). Therefore, I used a mESC line in which RING1B could be conditionally removed while RING1A was constitutively deleted (termed herein PRC1 CKO) (Figure 52A). In this cell line, the first coding exon of the *Ring1b* gene is flanked by LoxP recombination sites. Additionally, the cells also encode Cre recombinase fused to the oestrogen receptor (Cre-ERT2). Treatment with 4-hydroxytamoxifen (TAM) induces translocation of Cre-ERT2 into the nucleus where it catalyses recombination of the LoxP sites and excision of the first exon of *Ring1b*. Excision of the exon causes a frameshift in the mRNA which effectively knocks out *Ring1b*.

PRC2 has a tetrameric core formed by the H3K27 methyltransferase EZH1 or its paralogue EZH2, EED, SUZ12, and RBBP4 or RBBP7. Both RBBPs are not exclusive to PRC2 as they associate with multiple chromatin-binding complexes as stabilising factors (van Mierlo et al., 2019). To render PRC2 inactive, I used a mESC line in which SUZ12 can be conditionally knocked out (termed herein PRC2 CKO) using the same Cre/LoxP-mediated exon deletion method described above for PRC1 CKO (Figure 52B). Although SUZ12 is not the catalytic subunit, its deletion disrupts the entire complex and leads to genome-wide loss of the repressive H3K27me3 histone modification (Pasini et al., 2004b, 2007). A more comprehensive description of PcG complexes is given in 1.3.1.

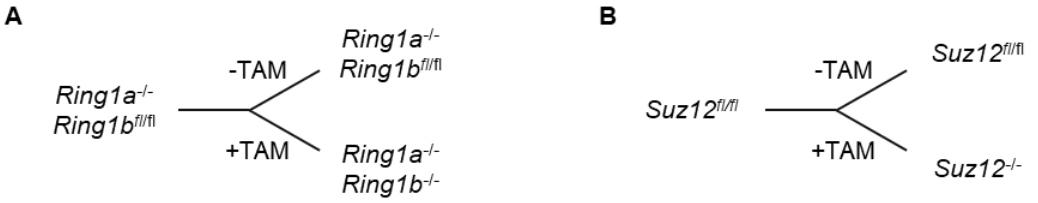


Figure 52: Conditional deletion systems for PRC1 and PRC2
(A) Schematic depicting the *Ring1a*$^{-/-}$ *Ring1b*$^{fl/fl}$ deletion system of the PRC1 CKO cell line. Addition of 4-hydroxytamoxifen (TAM) leads to genetic deletion of *Ring1b*.
(B) Schematic depicting the *Suz12*$^{fl/fl}$ deletion system of the PRC2 CKO cell line. Addition of 4-hydroxytamoxifen (TAM) leads to genetic deletion of *Suz12*.

To ensure that any changes I might observe in the CGI proteome are not caused by differential chromatin recovery between untreated (UNT) and TAM-treated cells, I first sought to verify that CGI enrichment was comparable in both conditions. Therefore, I

performed massively parallel sequencing of my C-CAP eluates and analysed the enrichment at non-methylated islands (NMIs). DNA binding motif analysis showed an equally strong enrichment of the CG dinucleotide motif in both conditions (Figure S5). Heatmap analyses of PRC1 and PRC2 samples visually suggested an almost identical level of NMI enrichment in both conditions (Figure 53A-B). However, metaplots indicated a slightly higher enrichment at the peak centre of the UNT condition, which was more pronounced in PRC1 than in PRC2 CKO cells (Figure 53A-B). To examine if this difference was statistically significant, I performed differential binding analysis. This revealed that none of the NMIs in the reference dataset was differentially bound (Figure 53C-D).

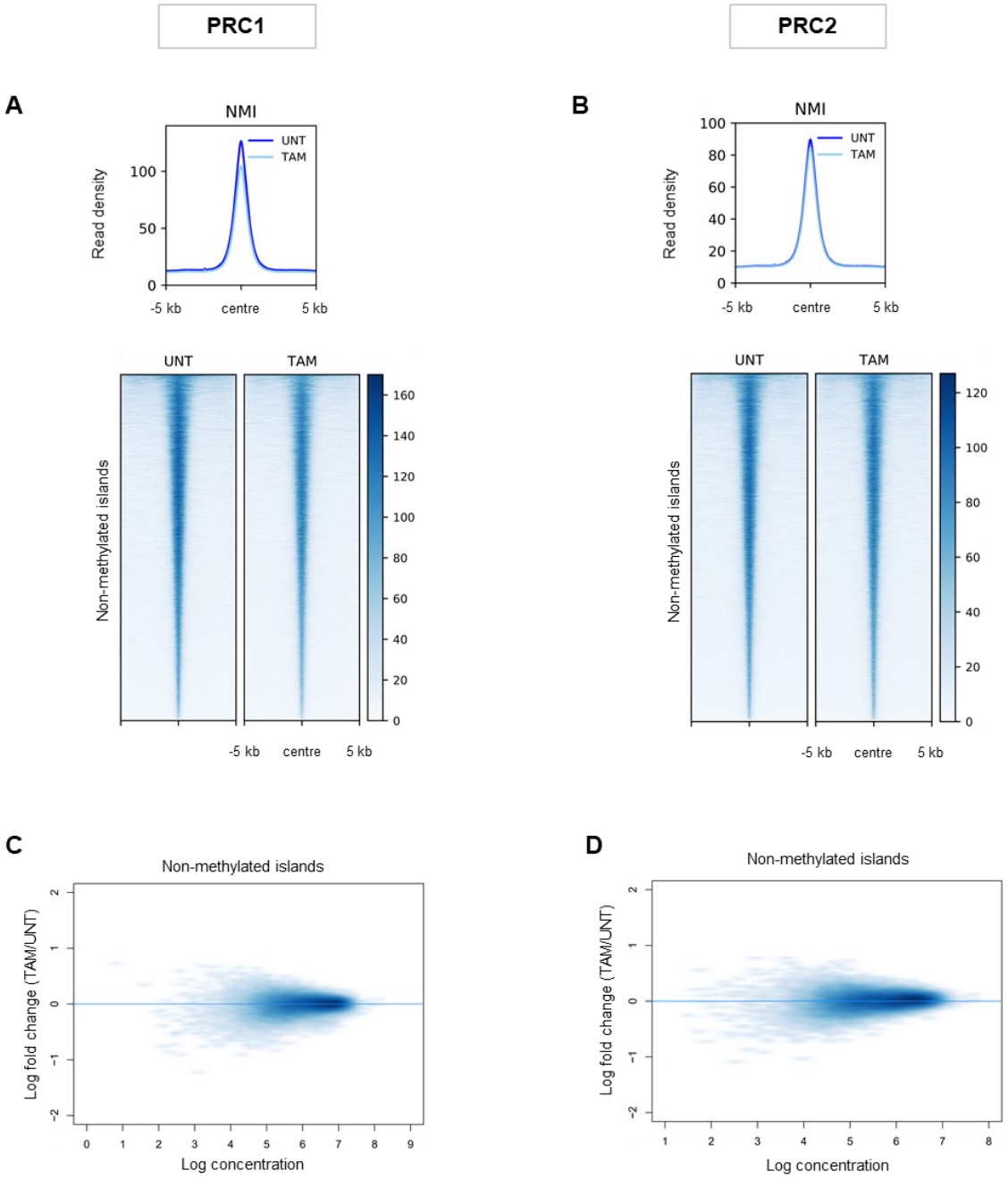


Figure 53: CGI chromatin enrichment in PRC1 and PRC2 CKO cells is comparable between untreated and TAM-treated cells

(A) Metaplot (top) and heatmap (bottom) of C-CAP sequencing showing genomic enrichment of non-methylated islands (NMIs) in the untreated (UNT) and TAM-treated condition of PRC1 CKO cells. The 500 – 1000 mM elution fractions of each condition were combined and Input-normalised for the analysis. The metaplot suggests slightly more enrichment at the peak centre of the UNT condition.
(B) As in (A) for PRC2.
(C) MA plot showing the differential binding analysis at NMIs of PRC1 CKO cells. Differential binding was assessed between the UNT and TAM condition. The log fold-change (Log FC) for virtually all sites lies between +1 and -1. None of the NMIs is differentially bound with an FDR ≤ 0.05 considered significant. n = 3.
(D) As in (C) for PRC2.

In summary, next generation sequencing has shown that loss of PRC1 or PRC2 does not affect the enrichment of CGIs by C-CAP. A comparable level of CGI chromatin enrichment will facilitate the interpretation of chromatin composition changes detected by MS.

6.2.2 Conditional knockout of PRC1 in mESCs causes extensive changes in the CGI proteome

Having verified that CGI enrichment is comparable between UNT and TAM conditions in both CKO cell lines, I set out to examine the chromatin composition changes at CGIs starting with PRC1 CKO cells. For this, I performed MS analysis of my C-CAP samples to identify proteins that were differentially enriched in the CGI proteome. Since loss of PRC1 leads to an extensive disruption of gene regulation in mESCs, I considered that some of the changes seen in the CGI proteome could be caused by gene expression alterations. To account for protein abundance changes caused by gene expression alterations, one can either analyse changes in the mRNA levels, or changes in nuclear protein abundances. Although mRNA is required for protein biosynthesis, a shift in mRNA levels does not lead to a direct and proportional shift in protein levels. This disconnect is caused by post-transcriptional and post-translational regulation, which induce decisive alterations in protein abundances that cannot be seen at the mRNA level (reviewed in Buccitelli and Selbach, 2020). This is evidenced by rather modest correlations between mRNA and protein levels, which has been reported in several studies (Gygi et al., 1999; Liu et al., 2016; Maier et al., 2009; de Sousa Abreu et al., 2009; Vogel and Marcotte, 2012), and also applies to my PRC1 and PCR2 CKO experiments (Figure S2Figure S3). Therefore, I analysed the nuclear proteomes of UNT and TAM cells by MS to account for differential nuclear protein abundances between these conditions. I then compared the CGI proteome and the nuclear proteome to identify those proteins that were differentially enriched in both datasets (Figure 54, marked in purple). These proteins need to be approached with caution as their differential enrichment on CGIs could have been caused by differential gene expression. To further increase the specificity of my CGI proteome, I created a list of high-confidence contaminants that were enriched in both the input and the ZF-CXXC^{mt} eluate of my previous C-CAP experiment (see 5.2.4 and Table S4). I then excluded the contaminants from the list of significantly changing proteins in the CKO dataset. The analysis yielded 460 proteins that increased and 172 proteins that decreased in abundance upon loss of PRC1 (Figure 54, marked in red).

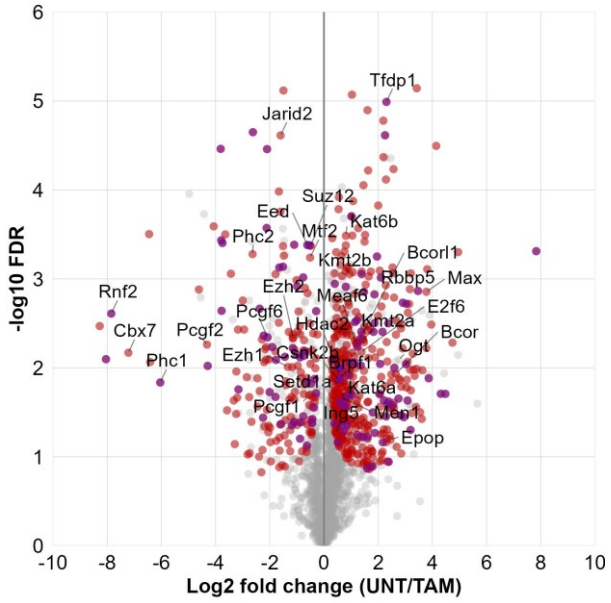


Figure 54: Volcano plot of differentially enriched proteins in C-CAP of PRC1 CKO
Volcano plot showing differentially enriched proteins in C-CAP eluates of PRC1 CKO cells. Differentially enriched proteins are highlighted in red, except those that overlap with a custom contaminant list (Table S4). Proteins that are differentially enriched in the CGI proteome, but also show differential enrichment in the nuclear proteome are marked in purple. Components of PcG, TrxG, and Moz/Morf complexes are labelled. Deletion of RING1A/B leads to a decrease in the abundance of 172 proteins and an increase of 460 proteins in the CGI proteome. Welch's t-test was performed to determine statistical significance with FDR ≤ 0.05 considered significant. n = 3.

To understand in more depth how PRC1 CKO impacted the CGI chromatin composition, I examined the proteome changes initially with a focus on PcG proteins. I expected PcG proteins to be most directly affected by the perturbation because deletion of the PRC1 core disrupts genomic binding of both PRC1 and PRC2 components (Blackledge et al., 2014; Scelfo et al., 2019). In mESCs, approximately half of RING1B-bound promoters are co-occupied by PCGF1/2/6, while PCGF3/5 have a relatively small overlap with RING1B occupancy, and PCGF4 is only expressed upon differentiation (Kloet et al., 2016; Scelfo et al., 2019). Moreover, ChIP-sequencing of PCGF proteins in PRC1 CKO cells has shown a widespread displacement of PCGF1/2 from chromatin, and a moderate reduction in genomic binding for PCGF3/6 (Scelfo et al., 2019). Therefore, I expected to see a decrease in protein levels on CGIs for PCGF1/2 and possibly also for PCGF6. I also anticipated a depletion of the auxiliary factors that are part of the complexes formed by PCGF1/2/6 core proteins. Indeed, proteomic analysis of my C-CAP samples revealed a significant reduction of PCGF1/2/6 in the CGI proteome (Figure 55A). Additionally, auxiliary factors CBX7 and PHC1/2 of canonical PRC1.2, which has PCGF2 at its core, also showed reduced protein levels upon loss of RING1A/B. However, facultative subunits of variant PRC1.1 and PRC1.6 increased in TAM-treated cells. These proteins included BCOR and BCORL1 for PRC1.1

and E2F6, TFDP1, and MAX for PRC1.6 (Figure 55A). Furthermore, I expected the loss of RING1A/B catalytic activity to have knock-on effects on PRC2 occupancy at CGIs. In PRC1 CKO cells, it has been shown that deletion of RING1A/B is accompanied by a decrease in PRC2 genomic occupancy and H3K27me3 (Blackledge et al., 2014). Since PRC1 CKO cells cannot catalyse H2AK119ub anymore, I suspected that the JARID2 subunit of PRC2, which binds this histone modification, might also be depleted in the CGI proteome (Cooper et al., 2014). Moreover, reduced JARID2 levels have been shown to correlate with a decrease of PRC2 occupancy at its target genes (Pasini et al., 2010b; Peng et al., 2009). Therefore, I anticipated to see a loss of several PRC2 core and facultative subunits at CGIs. In line with my hypotheses, proteomics analysis revealed depletion of PRC2 core proteins EZH1/2, EED, SUZ12, and facultative subunits JARID2 and MTF2 (Figure 55B). The only PRC2 subunit that gained in abundance on CGIs was EPOP. However, the association of EPOP with PRC2 is not exclusive as it also interacts with other proteins involved in transcription regulation (Beringer et al., 2016; Liefke et al., 2016).

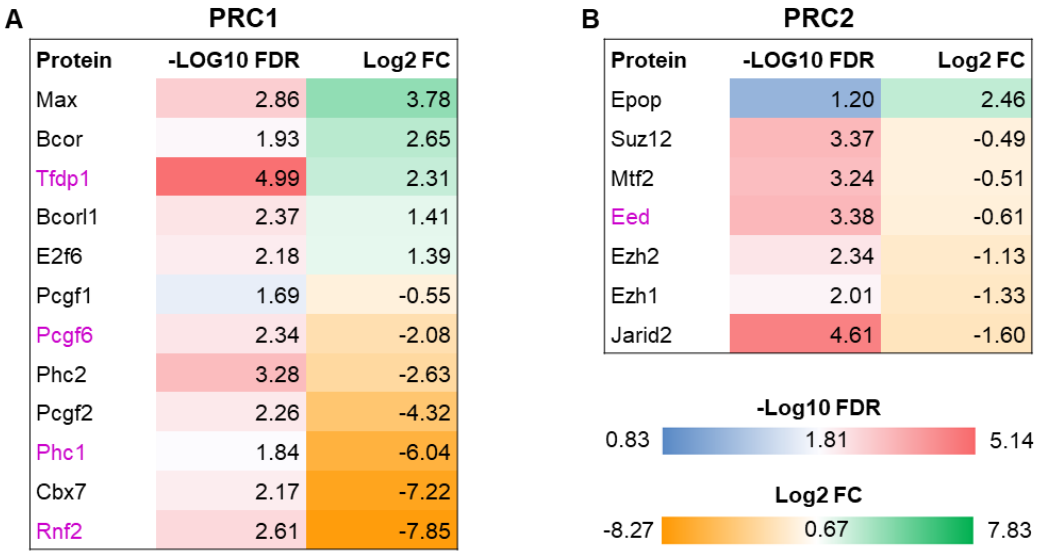


Figure 55: Differentially enriched PcG proteins in the CGI proteome of PRC1 CKO
(A-B) Tables showing differentially enriched PcG proteins in the CGI proteome of PRC1 CKO cells. PcG proteins were grouped based on their association with PRC1 (A) or PRC2 (B). Proteins are sorted in descending order of their log2 FC. Similar to Figure 54, protein names that are differentially enriched in the CGI proteome, but also show differential enrichment in the nuclear proteome are highlighted in purple. The colour bars mark the lowest, median, and highest -log10 FDR and log2 fold change, respectively. Statistical significance and fold change values are based on Figure 54.

The changes in PcG proteins that I observed after PRC1 removal generally correspond to those expected based on previous candidate-based ChIP approaches. This gave me confidence that my approach would also allow me to unbiasedly determine how PRC1

removal affects the CGI proteome more broadly. To achieve this, I performed GO analysis for the differentially enriched proteins and then grouped proteins with shared annotations. This revealed that loss of PRC1 affected several protein groups involved in chromatin modification, remodelling, and transcription regulation (Figure 56). Loss of PcG activities leads to a depletion of repressive histone modifications and concomitant gain in H3K4me3 and transcription activity (Dobrinić et al., 2021; Loh et al., 2021). In agreement with this, I observed an increase in the protein levels of the H3K4 methyltransferases SETD1A, KMT2A, KMT2B, and their interactors MEN1 and RBBP5 (Figure 56A). Furthermore, the abundance of several RNAP subunits increased (Figure 56B) which is supported by a recent study that found new binding of RNAPII at TSSs after rapid loss of PRC1 catalytic activity (Dobrinić et al., 2021). Coherent with the increase in RNAP occupancy and transcription activity, the analysis also showed a gain in transcription initiation factors on CGIs (Figure 56C).

Among the chromatin modifiers, several histone demethylases were affected by the PRC1 CKO (Figure 56D). KDM5A, KDM6B, and JMJD6 protein levels increased at CGIs, whereas another five demethylases exhibited a loss of protein abundance. KDM6B removes the repressive histone mark H3K27me2/3 (Hong et al., 2007; Lan et al., 2007), and JMJD6 cooperates with BRD4 to recruit the transcription-activating pTEFb complex to promoters (Liu et al., 2013). In contrast, KDM5A has a transcription-inhibiting role as a H3K4me2/3 demethylase (Klose et al., 2007; Tu et al., 2008). The histone demethylases that were depleted at CGIs (KDM4A/B/C, KDM7A, JMJD1C) catalyse the demethylation of histone 3 at residues K9, K27, and K36 (Cloos et al., 2006; Fodor et al., 2006; Horton et al., 2010; Izaguirre-Carbonell et al., 2019; Klose et al., 2006; Pedersen et al., 2016; Whetstine et al., 2006). Methylation of H3K9 and H3K27 is often associated with transcription repression whereas H3K36me3 is a mark of actively transcribed gene bodies (Dunham et al., 2012; Ernst et al., 2011). Another group of chromatin modifiers whose constituents were enriched at CGIs are histone acetyltransferases (HATs) (Figure 56E). Notably, this included all subunits of the Moz/Morf HAT complex (KAT6A/B, ING5, BRPF1, MEAF6), and six of twelve subunits of the NuA4 HAT complex. Histone acetylation is a mark of active gene regulatory elements (promoters, enhancers), and it generally anticorrelates with H3K27me3 (Dunham et al., 2012; Papin et al., 2021; Taylor et al., 2013; Tazi and Bird, 1990; Wang et al., 2008).

A TrxG

Protein	-LOG10 FDR	Log2 FC
Rbbp5	2.93	2.28
Kmt2a	2.31	1.45
Kmt2b	3.05	0.95
Men1	1.67	0.62
Setd1a	1.82	0.56

B RNAPI-III

Protein	-LOG10 FDR	Log2 FC
Polr3c	1.59	2.34
Polr3g	1.03	0.91
Polr2j	0.99	0.82
Polr3f	2.91	0.80
Polr3b	1.86	0.71
Polr1a	1.56	0.53
Polr1c	1.66	0.51
Cd3eap	2.07	0.48
Polr2e	1.70	0.32

C Transcription initiation

Protein	-LOG10 FDR	Log2 FC
Brf1	2.49	2.86
Gtf2h4	1.17	1.03
Taf1a	2.36	1.02
Taf1d	1.41	0.97
Med14	1.65	0.94
Taf1b	2.70	0.92
Cdk7	2.74	0.90
Med19	1.03	0.85
Taf1c	1.65	0.79
Gtf3c4	1.58	0.70
Gtf2f2	1.44	0.56
Gtf3c1	3.92	0.55
Gtf3c2	1.22	0.52
Gtf2e2	2.18	0.47
Ubtf	1.27	0.45
Taf1	2.50	-0.51

D Histone demethylation

Protein	-LOG10 FDR	Log2 FC
Kdm5a	1.20	1.61
Kdm6b	1.38	1.49
Jmjd6	2.11	0.34
Kdm4a	1.89	-0.63
Kdm4b	2.47	-0.89
Kdm4c	2.22	-1.15
Kdm7a	2.19	-1.79
Jmjd1c	4.65	-2.62

E Histone acetylation

Protein	-LOG10 FDR	Log2 FC
Tada3	1.25	2.10
Actr6	1.64	1.22
Meaf6	2.28	1.17
Morf4l2	2.31	1.13
Morf4l1	1.70	1.03
Kat6b	3.48	0.81
Ep400	1.50	0.69
Kat6a	1.67	0.62
Brpf1	1.89	0.60
Ing5	1.27	0.43
Dmap1	1.89	0.32

F Chromatin Remodelling

Protein	-LOG10 FDR	Log2 FC
Pole3	2.53	1.21
Baz1a	2.38	0.70
Chrac1	1.48	0.58
Baz1b	1.53	0.53
Brd9	2.20	-1.08
Cecr2	1.65	-1.10
Chd2	2.38	-1.15
Arid2	1.37	-1.58
Chd9	1.16	-1.97
Smarca1	4.46	-3.81

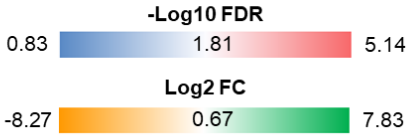


Figure 56: Differentially enriched protein groups in the CGI proteome of PRC1 CKO
(A-F) Tables showing differentially enriched protein groups in the CGI proteome of PRC1 CKO cells. Proteins were grouped based on shared annotations in the GO database. Proteins are sorted in descending order of their log2 FC. Similar to Figure 54, protein names that are differentially enriched in the CGI proteome, but also show differential enrichment in the nuclear proteome are highlighted in purple. The colour bars mark the lowest, median, and highest -log10 FDR and log2 fold change, respectively. Statistical significance and fold change values are based on Figure 54.

Finally, the analysis also revealed differential enrichment of chromatin remodeller components, which included the SWI/SNF, CHD, and ISWI subfamilies. Among them, only the ISWI subtype CHRAC complex (CHRAC1, POLE3, BAZ1A/B) (Poot et al., 2000) was enriched at CGIs upon PRC1 loss. The effect that these complexes have on DNA accessibility and transcription activity is largely contextual and depends on various factors such as the nucleosome substrate and genomic location, or their interaction with transcription activators and repressors (Clapier et al., 2017; Goodwin and Picketts, 2018).

In summary, my analysis has shown that the loss of RING1A/B have extensive effects on the CGI proteome (Figure 54) that go beyond the disruption of interactions between PcG proteins (Figure 55). In agreement with the loss of Polycomb-mediated gene repression, RNAPs and transcription initiation factors increased at CGIs (Figure 56B-C). This was accompanied by a gain in other transcription activators such as TrxG proteins and histone acetyltransferases (Figure 56A, E). Moreover, my results suggest that in the absence of PcG proteins, chromatin modifiers and remodellers change the chromatin architecture in way that promotes transcription activity. Taken together, these results support a pivotal role for PRC1 in regulating gene transcription at CGIs.

6.2.3 Conditional knockout of PRC2 in mESCs causes less pronounced CGI proteome changes than loss of PRC1

Having analysed the PRC1 CKO, I set out to examine how the loss of PRC2 affected the CGI proteome by performing the same analysis steps as before (see 6.2.2). This revealed that PRC2 CKO overall caused a less pronounced change of the CGI chromatin composition than the PRC1 CKO. The loss of the PRC2 core component SUZ12 led to a decrease in the abundance of 68 proteins and an increase of 129 proteins at CGIs (Figure 57).

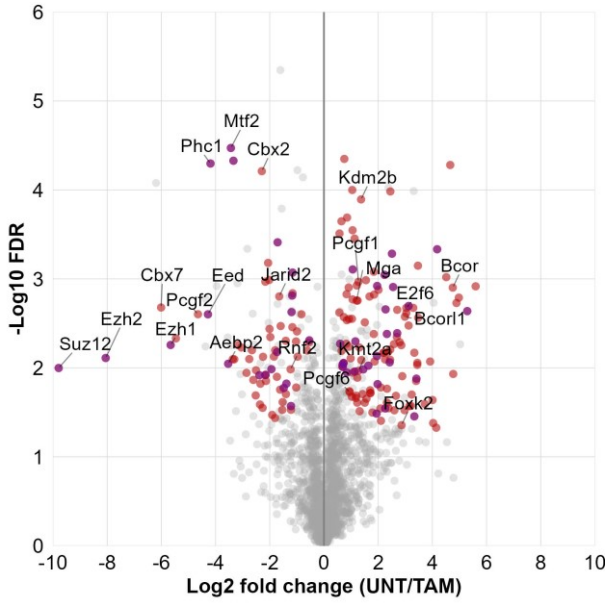


Figure 57: Volcano plot of differentially enriched proteins in C-CAP of PRC2 CKO
Volcano plot showing differentially enriched proteins in C-CAP eluates of PRC2 CKO cells. Differentially enriched proteins are highlighted in red, except those that overlap with a custom contaminant list (Table S4). Proteins that are differentially enriched in the CGI proteome, but also show differential enrichment in the nuclear proteome are marked in purple. Components of PcG and TrxG complexes are labelled. Deletion of SUZ12 leads to a decrease in the abundance of 68 proteins and an increase of 129 proteins in the CGI proteome. Welch's t-test was performed to determine statistical significance with FDR ≤ 0.05 considered significant. n = 3.

Similar to the PRC1 CKO, I approached the proteome analysis of PRC2 CKO by initially concentrating on the changes related to PcG proteins because I expected them to be most severely affected. Several studies have demonstrated that the catalytic core of PRC2 consisting of SUZ12, EZH1/2, and EED cannot be formed in the absence of any of the three core proteins (Cao and Zhang, 2004; Pasini et al., 2004c, 2007). Moreover, without the core, PRC2 accessory subunits would not be able to assemble into a complex either. As expected, EZH1/2, EED, JARID2, MTF2, and AEBP2 were depleted in the CGI proteome upon loss of SUZ12 (Figure 58A). Considering the reciprocal relationship between PRC1 and PRC2, I also looked for changes in PRC1 protein levels on CGIs. Loss of PRC2 catalytic activity has been shown to cause displacement of canonical but not variant PRC1 from chromatin (Tavares et al., 2012). Specifically, this affected CBX7-containing PRC1.2 assemblies because CBX7 is the predominant CBX-subunit in mESCs which is also highly dependent on H3K27me3 for its localisation on CGIs (Morey et al., 2012, 2013). Indeed, MS analysis revealed reduction of PRC1.2 core components RING1B and PCGF2, as well as accessory subunits CBX2/7 and PHC1 (Figure 58B). Moreover, seven PRC1 proteins increased in the CGI proteome when SUZ12 was deleted. These were all part of the variant complexes PRC1.1 (PCGF1, KDM2B, BCOR, BCORL1) and PRC1.6 (PCGF6, MGA,

E2F6). The finding that cPRC1 is lost while vPRC1 is retained at CGIs in the absence of PRC2 catalytic activity agrees with a previous study in mESCs (Tavares et al., 2012). One observation that strikingly diverged from results in the PRC1 CKO was that none of the transcription initiation factors were enriched, while only one RNAP subunit (POLR3B) and one TrxG protein (KMT2A) gained in abundance.

A

PRC2

Protein	-LOG10 FDR	Log2 FC
Jarid2	2.80	-1.65
Mtf2	4.47	-3.43
Aebp2	2.05	-3.53
Eed	2.60	-4.28
Ezh1	2.26	-5.66
Ezh2	2.11	-8.05
Suz12	2.00	-9.79

-Log10 FDR

1.33 2.18 4.47

Log2 FC

-9.79 1.13 5.60

B

PRC1

Protein	-LOG10 FDR	Log2 FC
Bcor	2.90	4.75
E2f6	2.67	3.04
Bcorl1	2.57	2.99
Kdm2b	3.89	1.37
Pcgf1	2.97	1.27
Mga	2.76	1.20
Pcgf6	1.73	0.95
Rnf2	1.98	-1.23
Cbx2	4.21	-2.29
Phc1	4.30	-4.19
Pcgf2	2.60	-4.65
Cbx7	2.68	-6.01

Figure 58: Differentially enriched PcG proteins in the CGI proteome of PRC2 CKO
(A-B) Tables showing differentially enriched PcG proteins in the CGI proteome of PRC2 CKO cells. PcG proteins were grouped based on their association with PRC2 (A) or PRC1 (B). Proteins are sorted in descending order of their log2 FC. Similar to Figure 57, protein names that are differentially enriched in the CGI proteome, but also show differential enrichment in the nuclear proteome are highlighted in purple. The colour bars mark the lowest, median, and highest -log10 FDR and log2 fold change, respectively. Statistical significance and fold change values are based on Figure 57.

As described for PRC1 CKO (see 6.2.2), I expanded my proteome analysis to other proteins in order to identify functionally related groups based on shared GO annotations. This revealed several histone modifiers and chromatin remodellers among the differentially enriched protein groups (Figure 59). Some of these proteins were in common between PRC2 and PRC1 CKO experiments but did not always follow the same direction of enrichment. The H3K4me2/3 demethylase KDM5A was enriched at CGIs in both experiments (Figure 59A, 56D). In contrast, the H3K9 and H3K36 demethylases KDM4A/B were enriched in PRC2 CKO (Figure 59A) but decreased in abundance in PRC1 CKO (Figure 56D). A similar divergence of enrichment was also observed for the H3K27me2/3 demethylase KDM6B. It increased in PRC1 CKO (Figure 56D) but is lost from CGIs in PRC2 CKO cells (Figure 59A). Interestingly, components of the NuA4 histone acetyltransferase complex and the CHRAC chromatin remodelling complex were enriched in PRC2 and

PRC1 CKO cells (Figure 56F, 59B), possibly indicating common pathways that lead to transcription activation when PcG catalytic activities are absent.

Finally, the analysis revealed enrichment of proteins involved in H3K79 dimethylation (Figure 59D). This transcription-associated histone modification tends to be present near the 5' end of active genes and plays an important role in enhancing transcription elongation (Dunham et al., 2012; Ljungman et al., 2019). The enriched proteins MLLT3, MLLT10, and AFF1 are part of the super elongation complex and have been shown to recruit the H3K79 methyltransferase DOT1L to chromatin (Luo et al., 2012). In their function as DOT1L recruiters, MLLT3 binds sites with acetylated histones, in particular H3K9ac, whereas MLLT10 binds unmodified H3K27 (Chen et al., 2015; Li et al., 2014).

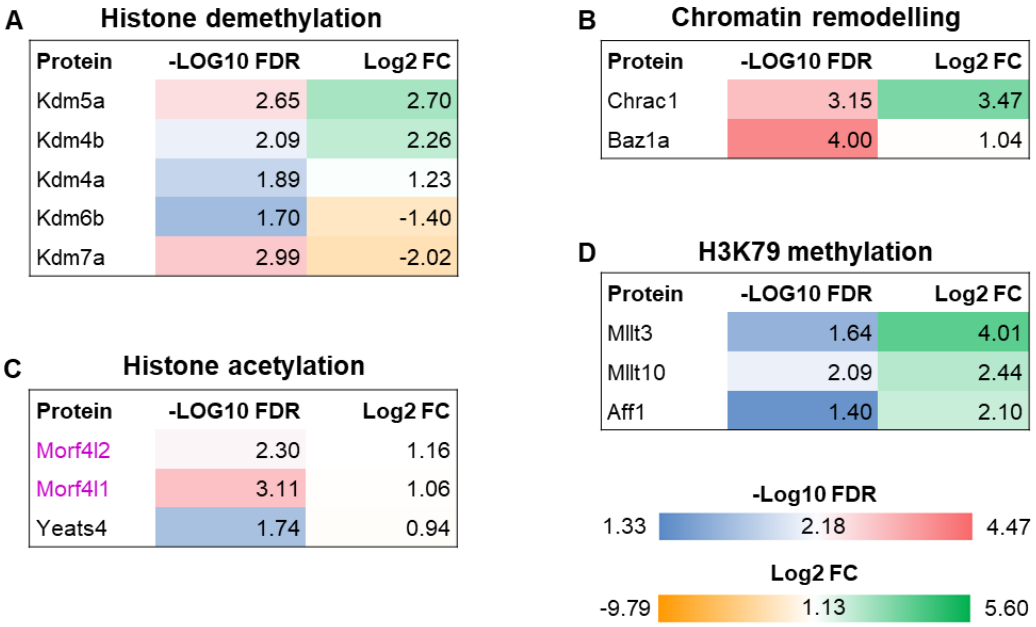


Figure 59: Differentially enriched protein groups in the CGI proteome of PRC2 CKO
(A-D) Tables showing differentially enriched proteins groups in the CGI proteome of PRC2 CKO cells. Proteins were grouped based on shared annotations in the GO database. Proteins are sorted in descending order of their log2 FC. Similar to Figure 57, protein names that are differentially enriched in the CGI proteome, but also show differential enrichment in the nuclear proteome are highlighted in purple. The colour bars mark the lowest, median, and highest -log10 FDR and log2 fold change, respectively. Statistical significance and fold change values are based on Figure 57.

In conclusion, loss of PRC2 in mESCs affected fewer proteins in the CGI proteome compared to PRC1 CKO. For example, RNAP subunits, transcription initiation factors, and TrxG proteins were hardly enriched in PRC2 CKO whereas these three protein groups gained considerably in abundance in PRC1 CKO. A similar set of histone demethylases was differentially enriched in both CKO cells, however, the direction of enrichment diverged

between the experiments for some of the demethylases. Strikingly, components of the same histone acetyltransferase and chromatin remodelling complexes were enriched at CGI, possibly indicating an antagonistic relationship with PcG proteins. Finally, super elongation complex and H3K79 methyltransferase complex components were exclusively enriched in PRC2 CKO, which could suggest a distinctive pathway of transcription activation when PRC2-mediated repression is lost. In summary, these results suggest that loss of PRC1 and PRC2 impacted some transcription regulatory pathways that are common to both Polycomb complexes, however, the results also pointed to pathways that seem to be unique to each complex.

6.2.4 ChIP-sequencing of differentially enriched proteins in PRC2 CKO cells corroborates MS results

The proteomic analysis of PRC1 and PRC2 CKO revealed several changes in the CGI chromatin composition that were in common between both experiments. One of them was the enrichment of variant PRC1.1 and PRC1.6 subunits upon loss of RING1A/B and SUZ12 (Figure 55A, 58B). To follow up on this, I sought to verify for selected subunits that the increase in protein levels in the proteome was caused by a gain in occupancy at CGIs. Moreover, I was interested in quantifying the genome-wide extent of the occupancy changes. For this, I performed calibrated ChIP-sequencing for KDM2B and PCGF1 (PRC1.1 subunits), and PCGF6 (PRC1.6 subunit) in PRC2 CKO cells. In contrast to conventional ChIP-seq, in calibrated ChIP-seq a small amount of a foreign genome is added to the samples that serves as an internal reference for normalisation and enables quantitative comparison of samples. Metaplot and heatmap analyses of the ChIP-seq data confirmed increased binding at NMIs for all three proteins upon deletion of SUZ12 (Figure 60). Notably, the gain in binding was higher for KDM2B and PCGF1 than it was for PCGF6. This echoed the differences in log2 abundance fold-changes observed in the MS data for the three proteins (Figure 58B). Moreover, a DNA binding motif analysis of all significant ChIP peaks of KDM2B, PCGF1, and PCGF6 revealed equal enrichment of the CG dinucleotide motif for all three proteins and in both conditions (Figure S6). This could indicate that the three proteins do not relocate randomly in the genome, but do so in a targeted manner to additional CGIs.

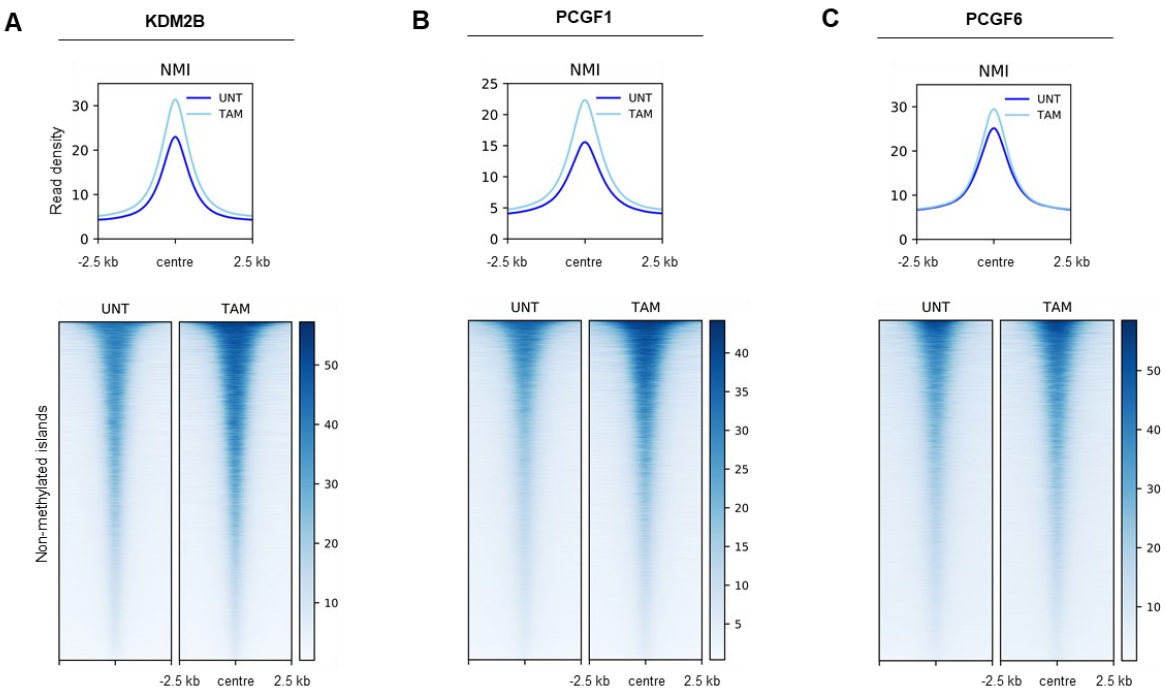


Figure 60: Calibrated ChIP-seq analysis of KDM2B, PCGF1, and PCGF6 shows increased binding at NMIs in PRC2 CKO

(A) Metaplot (top) and heatmap (bottom) of calibrated KDM2B ChIP-sequencing in PRC2 CKO cells. Both plots compare the enrichment of KDM2B at NMIs between untreated (UNT) and TAM-treated cells. TAM treatment leads to genetic deletion of SUZ12. KDM2B occupancy at NMIs increases upon loss of SUZ12. n = 3.
(B) As in (A) for PCGF1. PCGF1 occupancy at NMIs increases upon loss of SUZ12.
(C) As in (A) for PCGF6. PCGF6 occupancy at NMIs increases upon loss of SUZ12.

I then performed differential binding analysis to quantify the number of genomic sites at which these proteins gained or lost occupancy. For KDM2B and PCGF1 (both part of PRC1.1), this revealed a statistically significant increase of at least 1.5-fold at 12,380 and 10,094 sites, respectively (Figure 61A, C). A gain of at least 1.5-fold in PCGF6 genomic occupancy was limited to 437 sites, while 86 sites lost PCGF6 binding (Figure 61E). The differential binding analysis showed a similar increase in DNA occupancy on a genome-wide and on a single gene level for the PRC1.1 subunits KDM2B and PCGF1 (Figure 61A-D). Therefore, I wondered if they gained binding at the same genomic sites. To answer this, I determined the overlap of the binding sites, and this showed that the majority of sites with increased binding are occupied by both KDM2B and PCGF1 (Figure 62A). Interestingly, the overlap analysis also revealed that the majority of sites with increased PCGF6 occupancy are shared with KDM2B and PCGF1 (Figure 62B, C).

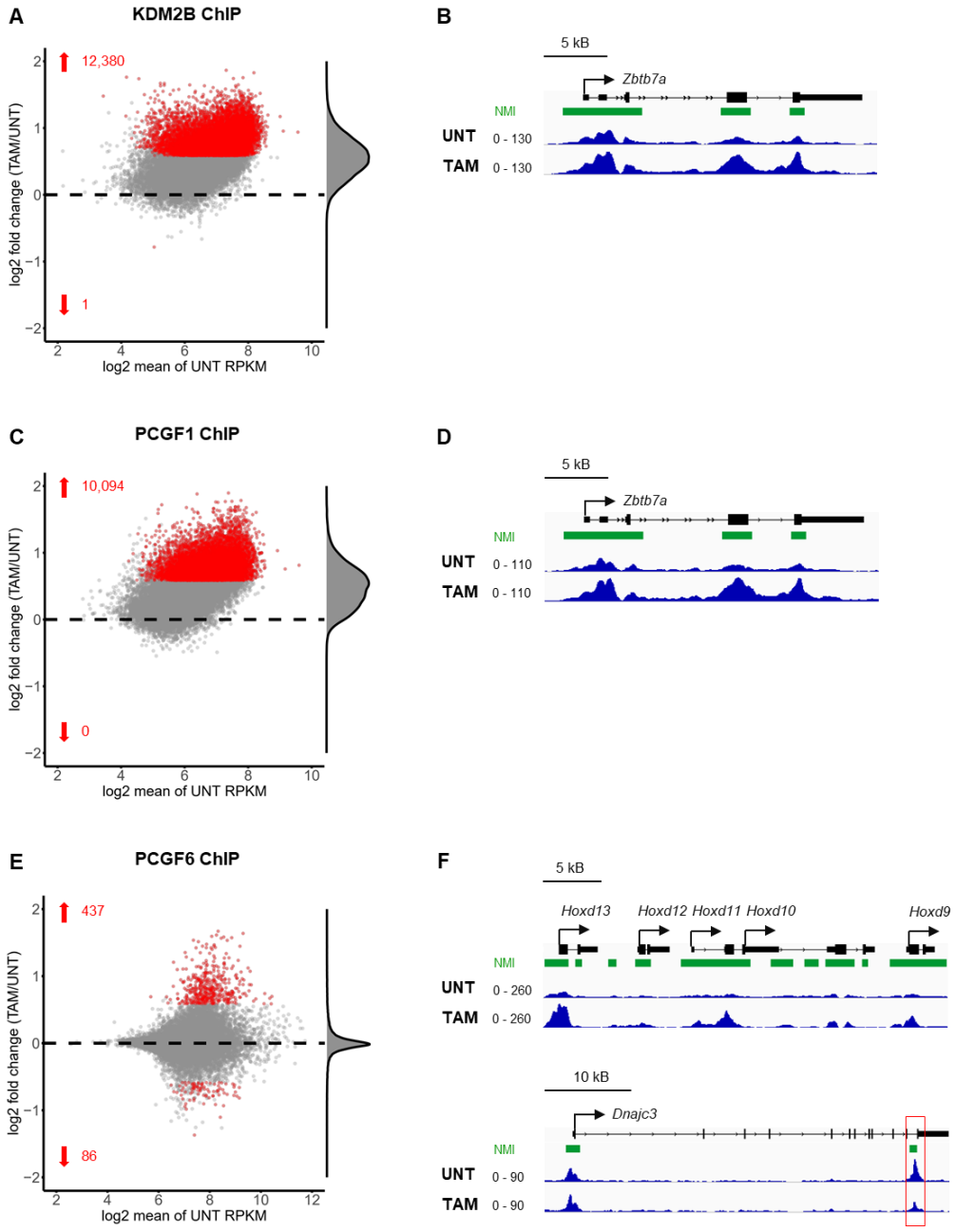


Figure 61: Differential binding analysis of KDM2B, PCGF1, and PCGF6 in PRC2 CKO

(A) MA plot showing the differential binding analysis of calibrated KDM2B ChIP-sequencing in PRC2 CKO cells. Differentially bound genomic sites are highlighted in red and the count is shown. Loss of PRC2 causes a global increase in KDM2B genomic occupancy. The statistical significance threshold for differential binding was set at FDR ≤ 0.05 and fold-change ±1.5. n = 3.
(B) Genome browser snapshot of *Zbtb7a* which was among the enriched sites with the highest fold-change. The location of NMIs was obtained from Long et al., 2013.
(C) As in (A) for PCGF1. Loss of PRC2 causes a global increase in PCGF1 genomic occupancy.
(D) As in (B) for PCGF1. *Zbtb7a* was among the enriched genes with the highest fold-change.
(E) As in (A) for PCGF6. Loss of PRC2 causes an increase in PCGF6 occupancy at 437 sites and a decreased at 86 sites.
(F) As in (B) for PCGF6. Genome browser snapshot of *Hox* and *Dnajc3* gene loci. Peaks at the *Hox* genes where among the sites with the highest gain in PCGF6 occupancy, whereas the peak at the 3'-end of *Dnajc3* (red box) was one of the sites with the highest loss of PCGF6 occupancy.

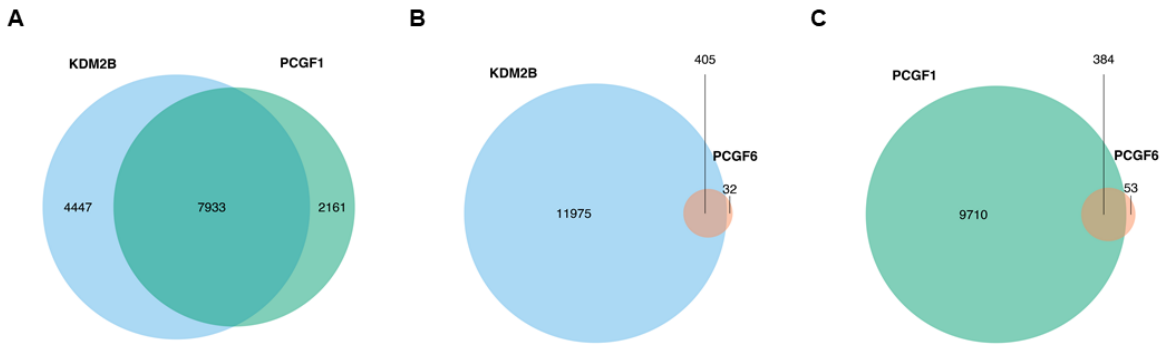


Figure 62: Overlap of genomic sites that gained occupancy of KDM2B, PCGF1, and PCGF6 in PRC2 CKO
(A) Venn diagram showing the overlap of genomic binding sites that increased in KDM2B and PCGF1 occupancy upon loss of SUZ12. The number of binding sites with a significant increase of at least 1.5-fold was taken from the analysis in Figure 61.
(B) As in (A) for KDM2B and PCGF6.
(C) As in (A) for PCGF1 and PCGF6.

In summary, using calibrated ChIP-sequencing as an orthogonal approach, I have verified for selected proteins that the CGI proteome alterations are caused by changes in genomic occupancy. Furthermore, I have shown that PRC2 CKO leads to a global increase of KDM2B and PCGF1 binding at their respective target sites. In contrast, the increase in binding for PCGF6 is limited to ~400 sites only, while less than 100 sites show decreased PCGF6 binding. Looking at the overlap of genomic sites with increased occupancy for the three proteins revealed that the majority of sites are in common between them.

6.3 Discussion

The first step in understanding how cellular processes are initiated and regulated at the genome level, is to identify the factors that localise at gene regulatory elements – most importantly at gene promoters. In a further step, capturing the proteome changes at promoters can point to novel synergistic and antagonistic interactions between protein groups that are involved in gene regulation. To this end, I have used a conditional knockout (CKO) system to induce changes of the CGI proteome which I could then investigate using C-CAP. To perturb the CGI proteome, I chose core subunits of PRC1 and PRC2, which are relatively well characterised CGI-associated complexes. This allowed me to anticipate some of the proteome changes caused by the perturbation. More importantly, C-CAP enabled me to identify for the first time in an unbiased and comprehensive manner the factors that changed after perturbating PcG complex interactions. Those proteome changes provided new hints as to how PcG proteins might regulate gene expression at CGIs.

First, I verified that CKO of PRC1 and PRC2 did not alter the enrichment of CGIs by C-CAP (Figure 53). Equal levels of CGI enrichment was a requirement for reliable quantification of changes in the protein abundances between untreated and TAM-treated cells. Subsequent MS analysis revealed that many of the effects of PRC1 and PRC2 CKO on the CGI proteome agreed with our current knowledge about their interactions at gene promoters. For instance, loss of the transcription repressing activity of PRC1 causes a widespread activation of gene expression (Blackledge et al., 2020; Fursova et al., 2019). On the CGI proteome level, this was reflected in an increase in transcription-promoting factors such as TrxG proteins, RNAPs, and transcription initiation factors (Figure 56A-C). For PRC2 CKO, in contrast, only few of these transcription-promoting factors were gained at CGIs. Additionally, the overall impact of PRC2 CKO on the CGI proteome seemed to be less severe than for PRC1 CKO (Figure 54, 57). This was somewhat surprising and might indicate a smaller magnitude of gene derepression upon loss of PRC2 activity (discussed further below).

Three themes emerged in the proteome analysis of PRC1/2 CKO. The themes were differential enrichment of histone demethylases, chromatin remodellers, and histone acetyltransferases. The results suggest that these three protein groups contribute to creating a chromatin architecture that is permissive to gene transcription when the repressive influence of PcG complexes is disrupted. Considering the proteome changes observed in PRC1 CKO (Figure 56), a hypothetical mechanism for gene activation could involve the following steps. KDM6B demethylates H3K27me2/3 at Polycomb target sites (Agger et al., 2007; Min et al., 2007; de Santa et al., 2007), which makes H3K27 available for acetylation by histone acetyltransferases (HATs). At the same time, KDM4A/B/C and KDM7A, are evicted from genes, where they would otherwise demethylate the transcription elongation mark H3K36me3. Other than demethylating H3K36me2/3, KDM4A/B/C and KDM7A can also remove heterochromatin-associated H3K9me2/3 from gene regulatory elements. Since KDM4A/B/C and KDM7A decreased in the CGI proteome, this could indicate that at the time of cell harvest, i.e. 72 h after TAM-treatment was commenced, the repressive mark H3K9me3 might have been already erased. This could have allowed acetylation of H3K9, which is supported by the increase in HATs in the CGI proteome. Additionally, JMJD6 in cooperation with BRD4 might recruit the pTEFb complex to promoters to facilitate transcription (Liu et al., 2013). Initially surprising was the enrichment of the H3K4me2/3 demethylase KDM5A in the CGI proteome. Removing methylation from H3K4 would counteract active transcription and indeed, KDM5A has been generally implicated in transcription repression (Kirtana et al., 2020). However, there are also reports that KDM5A can act as a transcription activator, in particular, on CGI promoters (Brier et

al., 2017). For example, it was shown that KDM5A can indirectly act as a transcription co-activator by inhibiting histone deacetylase activity (DiTacchio et al., 2011).

The differential enrichment of many chromatin remodellers (Figure 56F) further supports the notion that alterations of the chromatin architecture may play an important role in gene activation after loss of PRC1. Interestingly, three of four CHRAC complex subunits (CHRAC1, POLE3, BAZ1A/B) were enriched in RING1A/B depleted cells (Figure 56F). This complex catalyses nucleosome sliding, thereby modulating the spacing of nucleosomes (Kukimoto et al., 2004; Poot et al., 2000). Loss of RING1A/B is associated with changes in nucleosome spacing at Polycomb target sites (King et al., 2018) and the CHRAC complex might be responsible for this.

Active gene regulatory regions are often marked by acetylated histone residues such as H3K27ac, H3K9ac, and H4K16ac (Dunham et al., 2012; Papin et al., 2021; Taylor et al., 2013; Tazi and Bird, 1990; Wang et al., 2008). In line with an increase in gene expression in PRC1/2 CKO, several HAT complex subunits gained in abundance in the CGI proteome (Figure 56E, 59C). For example, all subunits of the Moz/Morf histone acetyltransferase complex were enriched upon deletion of RING1A/B. In the previous chapter (see 5.2.5), the Moz/Morf complex was identified as a potentially new component of the CGI proteome in mESCs. In HEK293T cells, KAT6A (also known as MOZ) has been recently shown to be part of a transcriptional activation system on CGI promoters together with KMT2A (also known as MLL1) (Miyamoto et al., 2020). The authors reported that KMT2A recruits KAT6A to CGI promoters of actively transcribed genes where KAT6A preferentially binds RNAPII that is not yet committed to transcription (i.e. lacking Ser5p). Moreover, KAT6A was excluded at regions bound by canonical PRC1.4. Similarly, an opposing relationship between PCGF4 and KAT6A was reported in mice (Sheikh et al., 2015). Moreover, there is evidence from zebrafish that Moz/Morf components act in concert with TrxG members to counteract PcG proteins (Laue et al., 2008). If the same mechanisms hold true in mESCs, the gain of Moz/Morf complex following RING1A/B deletion could indicate an antagonistic relationship with PRC1 similar to the TrxG proteins.

The proteome changes observed in PRC2 CKO suggest that gene derepression follows slightly distinctive mechanism compared to PRC1 CKO. For example, the H3K27me2/3 demethylase KDM6B is depleted in PRC2 CKO (Figure 59A). The reason is most likely the absence of H3K27me2/3 substrate at PcG target sites without the catalytic activity of PRC2. Furthermore, H3K9/K36 di-/tri-demethylases KDM4A/B were enriched in PRC2 CKO, indicating that the repressive H3K9me3 modification was still present at promoters. This

could suggest that gene derepression is incomplete and would support the observation that the CGI proteome disruption seems to be less severe in PRC2 CKO. Moreover, some components of the super elongation complex (SEC), which also interact with the H3K79 methyltransferase DOT1L, were enriched only in PRC2 CKO (Figure 59D) (Bitoun et al., 2007; Chen et al., 2015; Li et al., 2014; Luo et al., 2012; Mahmoudi et al., 2010). The SEC and H3K79me2 promote transcription elongation. Thus, loss of PRC2 might have more of an effect on the transcription elongation phase whereas loss of PRC1 seems to affect transcription initiation (Dobrinić et al., 2021).

To follow up on some of the proteome changes discussed above, one could perform C-CAP at different time points, for example at 24, 48, 72 h post TAM-treatment in PRC1 CKO. Combined with ChIP-seq for selected chromatin modifiers and histone modifications, one could get more detailed temporal resolution about the factors that become enriched or depleted in the CGI proteome. This could provide mechanistic insights about the order of gene derepression steps that take place when PRC1 catalytic activity is lost.

Another interesting finding related to PcG proteins was the enrichment of variant PRC1.1 and PRC1.6 components in TAM-treated cells. Although it was previously reported that variant PRC1 remains at Polycomb target sites when PRC2 is knocked out, the study did not specify which variant complexes were affected (Tavares et al., 2012). In my study, I found that in PRC1 CKO the enrichment of variant PRC1 comprised only auxiliary subunits (Figure 55A) whereas in PRC2 CKO it also included core subunits PCGF1/6 (Figure 58B). This observation might be related to the distinctive DNA recognition modes of these two variant complexes compared to canonical PRC1. The ZF-CXXC domain of KDM2B targets PRC1.1 to CGIs and KDM2B does not require RING1A/B for this targeting (Blackledge et al., 2014; Farcas et al., 2012). Although loss of RING1A/B causes displacement of PCGF1 from chromatin (Scelfo et al., 2019), the modularity reported for PRC1.1 could still allow other subunits to localise on CGIs (Rose et al., 2016). Similarly, PRC1.6 is recruited to chromatin via DNA sequence-specific binding of the E2F6-TFDP1 and MAX-MGA dimers of the complex (Scelfo et al., 2019; Stielow et al., 2018). PCGF6 itself contributes little to genomic loading of PRC1.6 and RING1A/B is not required for complex assembly (Scelfo et al., 2019; Stielow et al., 2018).

The gain of variant PRC1 components in the CGI proteome of TAM-treated cells prompted me to select PRC1.1 and PRC1.6 subunits to validate their enrichment using calibrated ChIP-seq as an orthogonal approach. The analysis confirmed that KDM2B and PCGF1/6 occupancy increased at NMIs in PRC2 CKO (Figure 60). Additionally, it revealed that

binding of KDM2B and PCGF1 was significantly increased at more than 10,000 sites, while the increase in binding for PCGF6 was limited to less than 500 sites (Figure 61). Interestingly, the sites at which they gained occupancy largely overlapped for all three proteins (Figure 62). Why do these two variant PRC1 complexes gain in binding and what is the functional implication of this? Breaking down the known interaction dependencies could shed light on this question. In the PRC1 CKO proteome, only the auxiliary subunits MAX, E2F6, TFDP1 (PRC1.6) and BCOR/L1 (PRC1.1) were enriched (Figure 55A). As described above, the MAX-MGA and E2F6-TFDP1 dimers do not require PCGF6 or RING1A/B for localisation at CGIs because they have their own DNA sequence-specific binding modules (Scelfo et al., 2019; Stielow et al., 2018). Without the physical presence and gene repressive activity of PRC1, target DNA sequences of MAX-MGA and E2F6-TFDP1 might be more accessible, which could explain the gain in these proteins. Also enriched in the PRC1 CKO proteome were BCOR/L1, whereas KDM2B was enriched but below the significance threshold, and PCGF1 was depleted (Figure 55A). Crosslinking mass spectrometry has shown that BCOR and KDM2B form direct interactions and at least part of the recruitment of BCOR/1L to CGIs might be attributed to KDM2B (Rose et al., 2016). Additionally, it is possible that BCOR/1L is recruited to promoters through interaction with NCOR2 (Hatzi et al., 2013), which was also significantly enriched. PCGF1 requires RING1A/B for efficient recruitment to Polycomb target sites (Scelfo et al., 2019), and deletion of RING1A/B likely accounts for PCGF1 depletion in the CGI proteome. Considering the global derepression of genes in PRC1 CKO (Blackledge et al., 2020; Fursova et al., 2019), enrichment of described auxiliary vPRC1 subunits seems unlikely to have an impact on gene repression without any of the core subunits (RING1A/B, PCGF1/6).

In PRC2 CKO, the core subunits PCGF1/6 were enriched in addition to the auxiliary subunits (Figure 58B). Moreover, RING1B was only moderately depleted in the CGI proteome (Figure 58B). Since PCGF1 cannot be recruited efficiently to chromatin without RING1B (Scelfo et al., 2019), the enrichment of PCGF1 in PRC2 CKO cells indicates that it did interact with RING1B and likely formed complete and functional vPRC1.1 at CGIs. The same is possibly true for vPRC1.6 considering the enrichment of PCGF6 and auxiliary subunits at CGIs. Others have shown that in the absence of PRC2, RING1B is lost at sites occupied by canonical PRC1 but not at sites bound by variant PRC1 (Tavares et al., 2012). This is consistent with a modest loss of RING1B but a strong depletion of the cPRC1 components PCGF2, PHC1, and CBX2/7 in the CGI proteome. Importantly, vPRC1.1/1.6 contribute most to PRC1-mediated gene repression in mESCs (Fursova et al., 2019). Hence, even when PRC2 is lost, vPRC1.1/1.6 still seem to perform their transcription-repressing function, which would explain why hardly any factors associated with active

transcription (e.g. RNAPs, transcription initiation factors) were enriched in PRC2 CKO. To elucidate the contribution of PCGF1/6 in maintaining gene repression when PRC2 catalytic activity is lost, one could conditionally knockout PCGF1/6 in PRC2 CKO cells and perform gene expression analysis using RNA-seq.

In conclusion, I have demonstrated that C-CAP can unbiasedly and comprehensively identify chromatin composition changes at CGIs. In PRC1/2 CKO, this has revealed that several histone modification and chromatin remodelling complexes may have pivotal roles in re-activating gene expression in the absence of Polycomb repressive activities. My analysis provided new leads for further investigation of the CGI protein interaction network. Thus, C-CAP is a novel tool to investigate the CGI proteome which can advance our knowledge of how these genomic elements contribute to gene regulation in developmental processes and pathogenesis.

7 Discussion

CpG islands (CGIs) are GC-rich regions with clusters of unmethylated CpG dinucleotides (Bird et al., 1985; Cooper et al., 1983; Tykocinski and Max, 1984). In vertebrate genomes, CGIs are associated with the majority of gene promoters where they play an important role in gene expression by recruiting regulatory proteins such as transcription factors, chromatin modifiers and remodellers to promoters (Deaton and Bird, 2011). A considerable amount of effort has been made in the past ~40 years to understand the mechanisms by which they contribute to gene regulation. However, a major obstacle in uncovering the function of CGIs in gene regulation, and possibly beyond, has been our lack of knowledge about the complement of proteins that localise at these crucial genomic elements.

Here, I have developed a novel method to investigate the CGI proteome in an unbiased manner using mass spectrometry. To achieve this, I based my approach on BioCAP, which is a method to enrich CGIs from naked DNA (Blackledge et al., 2012). BioCAP utilises a ZF-CXXC-containing affinity matrix to selectively capture and enrich regions of non-methylated DNA which can then be profiled genome-wide using next generation sequencing. Building on this tried and tested method to capture CGIs, I developed C-CAP which enriches CGIs from chromatin instead of naked DNA, thereby allowing genomic as well as proteomic profiling of CGIs (Chapter 4). I demonstrated that C-CAP can identify many known and potentially novel CGI-associated proteins (Chapter 5). Moreover, I showed that C-CAP is sufficiently sensitive to detect dynamic changes of the CGI proteome when gene regulatory processes are disrupted (Chapter 6). This last feature is crucial for uncovering the chromatin composition changes that take place at CGIs, for example, during cell differentiation or carcinogenesis. Thus, C-CAP can provide new leads for the investigation of proteins that regulate transcription in diverse cellular processes.

In the following sections, I want to give some insights beyond the discussions at the end of each results chapter. I first present the rationale behind the design of my chromatin purification and CGI enrichment workflow. I then discuss the strength of an unbiased MS-based approach in identifying potentially new CGI-associated proteins. Finally, I highlight how C-CAP could be applied to enhance our understanding of developmental processes and pathogenesis.

7.1 The choice of the chromatin purification approach

A major challenge in developing C-CAP was the chromatin purification step. My initial attempt to examine the CGI proteome was hampered by contaminating proteins that masked the CGI-associated protein signal in my MS analysis (see 4.2.1). This highlighted the need to improve the purity of my chromatin preparation. To this end, I tried to remove the contaminants after the CGI affinity capture step using SICAP (Selective Isolation of Chromatin-Associated Proteins, Rafiee et al., 2016) (see 4.2.2). SICAP involves immobilising biotin-labelled DNA fragments on an avidin-based affinity matrix and rinsing off contaminating proteins through a series of stringent washes. However, SICAP did not yield the aspired improvement in chromatin purity for reasons which were discussed in 4.3. Here, I briefly present alternative methods to purify chromatin and discuss the reasons for the approach I have ultimately chosen.

Chromatin purification strategies can be summarised in three groups (see 1.4): (1) single-locus isolation, (2) proximal proteome enrichment, and (3) whole-chromatin purification. Single-locus strategies will not be discussed here as they are applicable only for genomic elements with many copies such as telomeres or rDNA. For CGIs, which all have unique sequences, a single-locus approach would not provide sufficient enrichment (Vermeulen and Déjardin, 2020).

A candidate-based strategy is one option to examine the proximal proteome of a CGI-bound factor. In ChIP followed by MS (ChIP-MS), the protein of interest together with the crosslinked chromatin fragment is enriched using an antibody, and the proteins that are bound to the enriched chromatin fragments are then subjected to MS analysis (Engelen et al., 2015; Ji et al., 2015). One possibility to investigate the CGI proteome would have been ChIP-MS for KDM2B as it is found *in vivo* at virtually all CGIs in mESCs (Farcas et al., 2012). However, this would not have given an unbiased and comprehensive view of the CGI proteome because it is dependent on the genomic occupancy of KDM2B which can change, for example, in differentiated or diseased cells. An experiment like the PRC2 CKO would not have allowed the comparison of native and perturbed CGI proteomes with a KDM2B ChIP-MS approach because the genomic occupancy of KDM2B changes upon loss of SUZ12 (see 6.2.4). Considering this critical disadvantage of a candidate-based strategy highlighted that my CGI enrichment method had to be based on the direct interaction of the ZF-CXXC domain with non-methylated DNA. This DNA-mediated enrichment would be the basis for an unbiased investigation of CGIs because it is not dependent on the localisation of CGI-bound proteins.

Another possibility to enrich the CGI proteome was proximity biotinylation (reviewed in Gingras et al., 2019). This approach genetically fuses a protein of interest to a promiscuous biotin ligase or an engineered peroxidase, and the fusion construct is stably expressed inside cells. Upon initiation of the labelling reaction, the respective enzyme catalyses the biotinylation of proteins in its vicinity, thereby marking proteins that share a specific localisation. Biotinylated proteins can then be purified on streptavidin beads and subjected to MS for identification. Recently, in a new method termed ChromID, a promiscuous biotin ligase was genetically fused to different chromatin reader domains in mESCs to identify the proteome at genomic sites marked by distinctive epigenetic modifications (Villaseñor et al., 2020). For my project, I briefly considered a similar strategy with a ZF-CXXC/promiscuous biotin ligase fusion protein to target non-methylated CpGs. In theory, the fusion protein would bind to CGIs *in vivo*, biotinylate proteins in its proximity, which can then be purified and subjected to MS. However, proximity biotinylation comes with several challenges which are detailed in 1.4.3. Here, I want to point out only one disadvantage of proximity biotinylation which is relevant in the context of chromatin-associated proteins. C-CAP allows simultaneous genomic and proteomic profiling from the same sample. I used this dual-purpose feature of C-CAP to validate the unbiased and genome-wide enrichment of NMIs (see 5.2.2), and to verify equal NMI enrichment between different conditions in PRC1/2 CKO experiments (see 6.2.1). In contrast, using aliquots from the same sample for NGS and MS is not possible with proximity biotinylation. The developers of ChromID solved this issue by generating one cell line to validate the genomic localisation of a chromatin reader domain by ChIP-seq, and another cell line to examine the proximity proteome of the same chromatin reader domain by MS. Using two different cell lines is not essential, however, it is necessary to perform a ChIP assay in addition to the proximity biotinylation experiment with a separate aliquot of cells. Therefore, proximity biotinylation considerably increases the time and effort needed to obtain the same information compared to C-CAP. Considering this and other disadvantages (see 1.4.3), proximity biotinylation was dismissed as an option to investigate the CGI proteome.

Click chemistry-based and centrifugation-based methods can be employed to isolate the entire chromatin fraction from cells. For example, the Di Croce laboratory isolated chromatin from mESCs by biotinylating the entire cellular DNA molecule *in vivo* using click chemistry and subsequent streptavidin affinity purification (Aranda et al., 2019). A click chemistry-based approach is indeed useful for whole-chromatin purification, however, technically it would not be possible to continue with the enrichment of CGIs using a NeutrAvidin/ZF-CXXC affinity matrix. Since chromatin itself is biotinylated, it would compete with the biotinylated ZF-CXXC domain for binding sites on the NeutrAvidin resin which would lead

to displacement of the ZF-CXXC domain. One way to bypass the issue is to change the affinity resin and affinity tag on the ZF-CXXC domain. However, most affinity interactions other than biotin/avidin would not reliably withstand salt concentrations beyond 500 mM. The stepwise salt gradient from 300 – 1000 mM is a crucial part of C-CAP because fractionation decreases sample complexity which benefits MS measurement of proteins in the low-abundance range. Moreover, using aforementioned salt gradient on affinity matrices that are not based on the biotin/avidin interaction would leach out the bait protein and lead to a bait contamination in the high-salt fractions. Bait-derived peptides could then interfere with the detection of low abundant proteins during MS acquisition.

Another way to purify chromatin is fractionation of the cellular components by centrifugation. This can be achieved, for example, using a caesium chloride (CsCl) gradient (Ginno et al., 2018). Through extensive ultracentrifugation, crosslinked chromatin is separated from free DNA and free proteins based on its distinct buoyant density within the CsCl gradient. I briefly explored this option to improve the purity of my chromatin preparation but did not pursue it further because of two major drawbacks. First, generating a stable isopycnic CsCl density gradient required 60 h of ultracentrifugation and extended the processing time of my chromatin preparation considerably. Second, to recover the chromatin band from the gradient, a needle had to be pierced through the wall of the ultracentrifuge tube at the right gradient position. This process easily disrupted the gradient and also led to sample loss. Therefore, from a pragmatic standpoint, CsCl gradient ultracentrifugation was a less attractive option to purify chromatin.

A whole-chromatin purification approach that seemed to be compatible with CGI enrichment by the ZF-CXXC affinity matrix was ChEP (Chromatin Enrichment for Proteomics, (Kustatscher et al., 2014a). This method addressed the issue I had with contaminants in my chromatin preparation, and it was easy to implement in front of the CGI affinity capture step. ChEP exploits the distinct biochemical properties of crosslinked chromatin to precipitate it under defined buffer conditions. Non-crosslinked contaminants are removed by repeatedly precipitating and washing chromatin under denaturing conditions using 4% SDS and 8 M urea. I implemented this method in my workflow with some buffer composition changes that improved the reproducibility of the chromatin precipitation step and enabled me to perform the procedure in a cooled centrifuge (see 2.4.3). However, when I looked at the MS data from the first publication of this method, I noticed that the purified chromatin fraction still contained many cytoplasmic contaminants despite the stringent wash conditions (Kustatscher et al., 2014b). This prompted me to include another purification step designed to reduce cytoplasmic proteins. To achieve this, I isolated intact nuclei from

cells using a gentle lysis procedure and purified the nuclei on a sucrose density barrier by ultracentrifugation. In contrast to a CsCl gradient, this ultracentrifugation step is easy to set up and can be done in just one hour. Nuclei pass through the highly viscous sucrose density barrier and pellet at the bottom of the ultracentrifuge tube while lighter, cytoplasmic cell components cannot penetrate the barrier and therefore remain at the top. The sucrose density barrier in combination with ChEP improved the chromatin purity significantly (see 4.2.3), which later enhanced the enrichment of the CGI proteome (see 4.2.4).

In conclusion, when I was developing C-CAP, I systematically assessed the suitability of different techniques to improve chromatin purity and CGI enrichment. In the end, I decided to use a 2-step approach with sucrose density barrier ultracentrifugation for nuclei purification, and ChEP for removing non-crosslinked contaminants from chromatin. Importantly, this approach did not compromise on the strengths of the CGI affinity capture step, which are, direct binding of the ZF-CXXC domain to DNA, fractionation of CGI chromatin, and being able to perform NGS and MS analysis from the same sample.

7.2 The strength of MS-based techniques in identifying chromatin interactions *de novo*

MS-based techniques for investigating the chromatin proteome have the advantage that they are not dependent on any pre-existing knowledge about protein-DNA or protein-protein interactions. Such an unbiased approach enables us to discover novel chromatin-associated proteins, including those that are not expected to have a chromatin-related function. Here I want to highlight how mining proteomic datasets can provide new leads for the discovery of novel promoter and CGI-associated proteins.

SAMD1 has been recently validated as a CGI-binding protein (Stielow et al., 2021), but at the time when I was analysing my C-CAP data, it was essentially uncharacterised. There was no indication of a chromatin-related function in public databases. On the contrary, in UniProt it was described as an extracellular protein that binds low-density lipoprotein (also known as “bad” cholesterol) (Lees et al., 2005). However, multiple proteomic studies pointed to a potentially chromatin-related function. The first clue for a chromatin-association came from pulldown assays with reconstituted modified nucleosomes where SAMD1 was found to be enriched on H3K4me3 and H3K27me3 (Bartke et al., 2010). Moreover, IP-MS of HDAC1/2 and SFMBT1/2 suggested that SAMD1 might be part of transcription repression complexes (Malovannaya et al., 2010; Zhang et al., 2013), and ChIP-MS studies

supported a localisation on H3K4me3 *in vivo* (Engelen et al., 2015; Ji et al., 2015; Soldi and Bonaldi, 2013). Additionally, SAMD1 was found to be repelled by methylated DNA and enriched on CpG-rich DNA which indicated a function on CGIs (Perino et al., 2018; Xiong et al., 2016). Mining the literature and mass spectrometry datasets led Liefke et al. to speculate that SAMD1 might be a novel CGI-binding protein. Eventually, they uncovered a new type of WH domain in SAMD1 which enables it to bind non-methylated CGI promoters where it acts as a transcription repressor.

As exemplified by SAMD1, proteomic datasets are valuable resources for providing leads to identify novel chromatin-associated proteins. In section 5.2.5, I tried verifying CGI-binding for CDCA7/L and KAT6B, and part of the evidence that led me to follow up on these proteins was found in proteomic datasets. Similar to SAMD1, CDCA7L was enriched on H3K4me3 and H3K27me3 in pulldown assays with reconstituted modified nucleosomes (Bartke et al., 2010). ChIP-MS experiments in mESCs reported enrichment of CDCA7 and KAT6B on H3K4me1/3 and H3K27ac (Engelen et al., 2015; Ji et al., 2015). Additionally, KAT6B bound non-methylated cytosine in DNA pulldown assays with nuclear extracts from mESCs (Spruijt et al., 2013b). My CGI proteome contains several other proteins with evidence in the literature and MS datasets that point to CGI-related functions. To give some examples, ZNF512 is a largely uncharacterised Zinc finger protein that might interact with CGI-binding protein BANP, and ZNF512 was found to be enriched at active (H3K4me1/3, H3K27ac) and repressive (H3K9me3) histone marks in mESCs using ChIP-MS (Engelen et al., 2015; Ji et al., 2015; Luck et al., 2020). Another example of a poorly characterised protein is BRWD3 which has multiple bromodomains and WD-40 repeats. Bromodomains can bind acetylated histones, whereas WD-40 repeats mediate DNA-protein and protein-protein interactions or serve as platforms for complex assembly (Jain and Pandey, 2018; Josling et al., 2012). Consistent with its domain composition, BRWD3 can recruit other proteins to acetylated histones in mammalian cells, and ChIP-MS showed enrichment on H3K4me3 and H3K27ac (Ji et al., 2015; Wang et al., 2021b). Although the cellular roles of ZNF512 and BRWD3 are still unexplored, their chromatin-binding domains, and their co-precipitation with histone modifications found at gene regulatory elements indicate functions in gene expression pathways.

Metabolic enzymes are at first glance not considered to be involved in gene regulation. When they are enriched in chromatin proteomics studies, they are usually not the focus of interest, and therefore tend to be dismissed as contaminants. Interestingly, moonlighting functions are often found in metabolic enzymes (reviewed in Li et al., 2018). Moonlighting refers to additional, unexpected, and sometimes unrelated functions outside of the

canonical cellular environment of the protein. For example, the pyruvate kinase isoform PKM2 possesses both metabolic and non-metabolic functions (Zhang et al., 2019b). Apart from producing pyruvate, which is one of the substrates for acetyl-CoA production, PKM2 can also act as a protein kinase. In its non-metabolic role, PKM2 participates in transcription regulation by phosphorylating histone H3T11 and the transcription factor STAT3 (Gao et al., 2012a; Yang et al., 2012). The power of MS-based approaches to uncover metabolic enzymes with moonlighting functions has been recently demonstrated by the Di Croce Lab. Performing chromatin purification followed by MS, they found Adenosylhomocysteinase (AHCY) to be associated with chromatin in mESCs (Aranda et al., 2019). Genomic profiling and functional assays revealed that AHCY is localised around the TSS of several hundred highly expressed genes where it impacts their transcriptional state.

Seven metabolic enzymes were enriched in my CGI proteome (see 5.2.5). These include an aminotransferase (GOT1), two reductases (TECR, GRHPR), a kinase (SEPHS1), a methyltransferase (CARNMT1), a dehydrogenase (MDH1), and an acyltransferase (GPAM). Similar to the PKM2 example given above, it is conceivable that some of them perform catalytic reactions on substrates other than their canonical metabolites. Interestingly, for two enzymes (SEPHS1, GRHPR), ChIP-MS showed co-precipitation on chromatin marked by H3K4me3 and H3K79me2 (Engelen et al., 2015; Ji et al., 2015). Just as AHCY, the seven metabolic enzymes enriched in my CGI proteome do not have any annotated chromatin-binding domains. However, that does not exclude the possibility of a yet undiscovered moonlighting function on chromatin.

In conclusion, the major strength of MS-based techniques is that they are unbiased, which makes them well-suited for identifying novel interactions. The recent examples of SAMD1 (Stielow et al., 2021) and AHCY (Aranda et al., 2019) demonstrate that clues provided by proteomic studies can help in identifying transcription regulators *de novo*. The CGI proteome generated in this work provides an inventory of new CGI-binding candidates. Some of these proteins, such as ZNF512 and BRWD3, seem more likely to be CGI interactors based on their chromatin-binding domains whereas metabolic enzymes which lack any obvious chromatin connection are more risky bets. Going forward, candidate-based experiments are required to systematically validate potentially novel CGI interactors.

7.3 Application of C-CAP to understand CGI function in development and disease

Through their co-localisation with promoters and enhancers, CGIs play a pivotal role in transcription regulation (Deaton and Bird, 2011). To understand how CGIs exert their gene regulatory function, it is essential to elucidate the identity and interrelations of CGI-associated proteins. PcG and TrxG complexes are master regulators of transcription that associate with CGIs, and understanding the role of these complexes is key to understanding CGI biology. Transcription regulation by PcG and TrxG complexes has been investigated for over 40 years in several model organism and cell types, yet, we still have an incomplete understanding how this is mechanistically achieved (reviewed in Schuettengruber et al., 2017). Although it is clear that PcG and TrxG proteins are crucial for modulating transcriptional programmes, for example during cell differentiation, we do not know how those transcriptional changes are brought about on the proteome level. This is largely due to the lack of molecular tools that would enable us to capture how disruption of PcG and TrxG functions impact the CGI proteome. With C-CAP I have developed a novel tool that can for the first time provide an unbiased and comprehensive picture of the chromatin composition changes at CGIs. Using C-CAP, I have shown in mESCs that loss of PRC1 and PRC2 affected a multitude of transcription factors and proteins that shape the chromatin architecture at CGIs (see 6.2). Thus, the C-CAP experiments of PRC1/2 CKO provided a comprehensive inventory of proteins that are components of a transcriptional regulation system at CGIs. Importantly, these experiments provided new leads how the balance between PcG and TrxG proteins might be regulated.

PcG and TrxG proteins were first identified as key regulators of developmental processes in *Drosophila* (Ingham, 1983; Ingham and Whittle, 1980; Lewis, 1978) . Later studies have shown that these complexes are functionally conserved also in other animals, including in mammals (Holdeman et al., 1998; Lee and Skalnik, 2005; Robert et al., 1996). A hallmark phenotype of PcG and TrxG mutations across animal species are morphological abnormalities in body axis and pattern formation. The correct temporal and spatial execution of transcriptional programmes is crucial for forming a normal body plan, and it is the balance of PcG and TrxG proteins at gene promoters that guides these programmes in the right direction (Poynter and Kadoch, 2016). Interestingly, PcG and TrxG proteins often co-occupy the same genomic sites and many developmental gene promoters in pluripotent cells are marked with the respective histone modifications of both complexes. These domains where silent H3K27me3 and active H3K4me3 coexist on the same histones have been described as bivalent chromatin (Azuara et al., 2006; Bernstein et al., 2006a). Bivalent domains are thought to keep the gene promoter in a poised state that can be readily switched into an active or silent state during cell differentiation. Keeping the promoter

uncommitted in pluripotent cells would allow development to progress through distinctive transcriptional programmes by tipping the balance in favour of PcG or TrxG proteins.

An intriguing question is which factors tip the balance at promoters towards activation or repression? My C-CAP experiments of PRC1/2 CKO suggested histone modifiers as strong candidates (see 6.2). Several lines of evidence have indicated an antagonistic relationship of PcG proteins with histone acetylation, whereas histone acetyltransferases (HATs) often act in concert with TrxG proteins (Laue et al., 2008; Miyamoto et al., 2020; Paggetti et al., 2010; Pasini et al., 2010c; Sheikh et al., 2015). One possibility is that PcG complexes counteract HATs, thereby keeping acetylation at promoters below a threshold at which transcription-promoting activities would dominate (Kuroda et al., 2020). Acetylation and methylation on the same histone residue are mutually exclusive. In this context, it is interesting to note that many histone demethylases (HDMs) were differentially enriched in PRC1/2 CKO cells (see 6.2). When PcG protein levels at CGIs are reduced, certain HDMs possibly remove repressive methylation marks. Some of the demethylated histone residues could then serve as substrates for HATs. Once transcription-activating acetylation marks prevail, other HDMs might help to maintain the active state. Thus, HATs and HDMs could have a crucial role in the balance between PcG and TrxG activities.

To fully understand how the balance between PcG, TrxG, HATs, and HDMs regulates the transcriptional state of genes, it will not be sufficient to quantify the protein levels of these groups at CGI promoters. Additionally, we need to track the products of their catalysis, i.e. the histone modifications that they place or remove. To achieve this, it will be necessary to extend C-CAP to PTM analysis of histones. This can be easily accomplished by taking an aliquot of the C-CAP eluates and preparing them for MS using protocols that are specifically designed for histone proteomics (Sidoli et al., 2012). Elucidating the histone modifications at CGI promoters will be crucial in understanding how they influence the chromatin architecture, and how they act as recruitment sites for transcription activators or repressors. Furthermore, C-CAP in combination with ChIP-seq could identify novel histone modifications that might be specific to CGIs, and therefore might explain how CGI promoters uniquely differ from other gene regulatory regions.

Using C-CAP (including histone PTM analysis) to compare the CGI proteomes of pluripotent and differentiated cells could help us identify the protein factors that guide development through transcriptional programmes. Performing C-CAP in a time-resolved manner, i.e. sampling at different time points, could provide an indication of the sequence of processes that take place at CGI promoters during cell differentiation. For example, one

could determine in which temporal order PcG and TrxG complexes, histone acetyltransferases and demethylases are gained or lost at CGI promoters. Obviously, this can also be done in reverse order by inducing pluripotency in somatic cells. One could then check if the sequence of proteome changes is reversed, and it would allow comparison between the CGI proteomes of native and induced pluripotent cells. This could give further insights about unique CGI proteome features of pluripotency.

Disruption of the balance between PcG and TrxG proteins can lead to aberrant development and cause cancer (Piunti and Shilatifard, 2021). Several PcG and TrxG components such as EZH2 and MLL proteins have well documented roles in carcinogenesis. Therapeutic interventions usually aim to block the catalytic activity of misregulated EZH2 or MLL proteins using small molecule inhibitors. C-CAP could be performed with cancer cells to enhance our understanding how PcG and TrxG components cause changes of the CGI proteome that lead to cancer. This could possibly identify new pharmacological targets for the treatment of cancer. In summary, the application of C-CAP in different cell types and physiological conditions will help to elucidate the mechanisms by which CGIs and associated proteins contribute to gene regulation in normal developmental processes and pathogenesis.

8 Summary and perspectives

I have developed a novel CGI chromatin enrichment method that allows an unbiased and comprehensive investigation of the CGI proteome by mass spectrometry. Using this new biochemical tool termed C-CAP, I have examined the CGI proteome of mESCs and identified potentially novel CGI-associated proteins. Further ChIP-seq experiments will be needed to verify the genomic localisation for new CGI-binding candidates. Once CGI-binding is confirmed, additional assays can unravel how they contribute to CGI function. Furthermore, I examined how the CGI proteome changes if I disrupt gene regulatory processes. To achieve this, I conditionally knocked out PRC1 and PRC2 in mESCs and used C-CAP to investigate the proteome changes. The proteome analysis uncovered new interrelations of PcG proteins with other components of the gene regulation network at CGI promoters. Going forward, one could include histone PTM analysis and several sampling time points to obtain a more granular view of the proteome changes. This can be complemented by calibrated ChIP-seq for selected factors to verify and quantify the changes on the genome level. For proteins that are significantly affected by loss of PcG complexes, one could then create conditional knockout mESCs and characterise the transcriptome and proteome changes for these cells. Ultimately, complementing C-CAP with other types of assays will enable us to dissect the processes that regulate gene expression at CGI promoters.

Appendix: Supplementary data

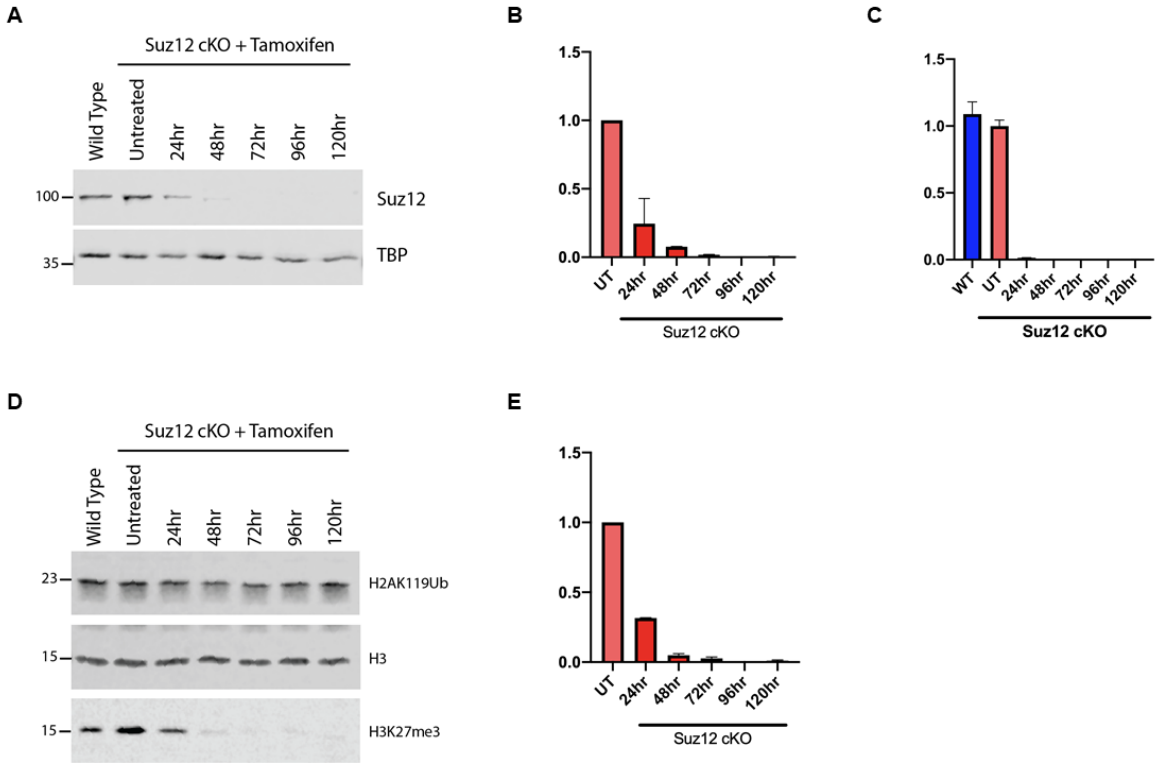


Figure S1: Validation of the SUZ12 CKO cell line (courtesy of Iona Manley)
(A) Western blot showing the time course of nuclear SUZ12 protein level depletion during TAM treatment of SUZ12 CKO mESCs. Nuclear SUZ12 protein is undetectable after 48 h of TAM treatment. TBP (TATA box binding protein) is the loading control.
(B) Bar chart showing the quantification of nuclear SUZ12 protein level depletion at different time points during TAM treatment of SUZ12 CKO mESCs. Treatment time points were normalised to untreated (UT).
(C) Bar chart showing SUZ12 RNA level depletion at different time points during TAM treatment of SUZ12 CKO mESCs. Reverse transcription followed by qPCR was performed to measure RNA levels. Treatment time points were normalised to untreated (UT).
(D) Western blot showing the time course of H3K27me3 protein level depletion during TAM treatment of SUZ12 CKO mESCs. H3K27me3 protein is almost undetectable after 48 h of TAM treatment. H3 is the loading control.
(E) Bar chart showing the quantification of H3K27me3 protein level depletion at different time points during TAM treatment of SUZ12 CKO mESCs. Treatment time points were normalised to untreated (UT).

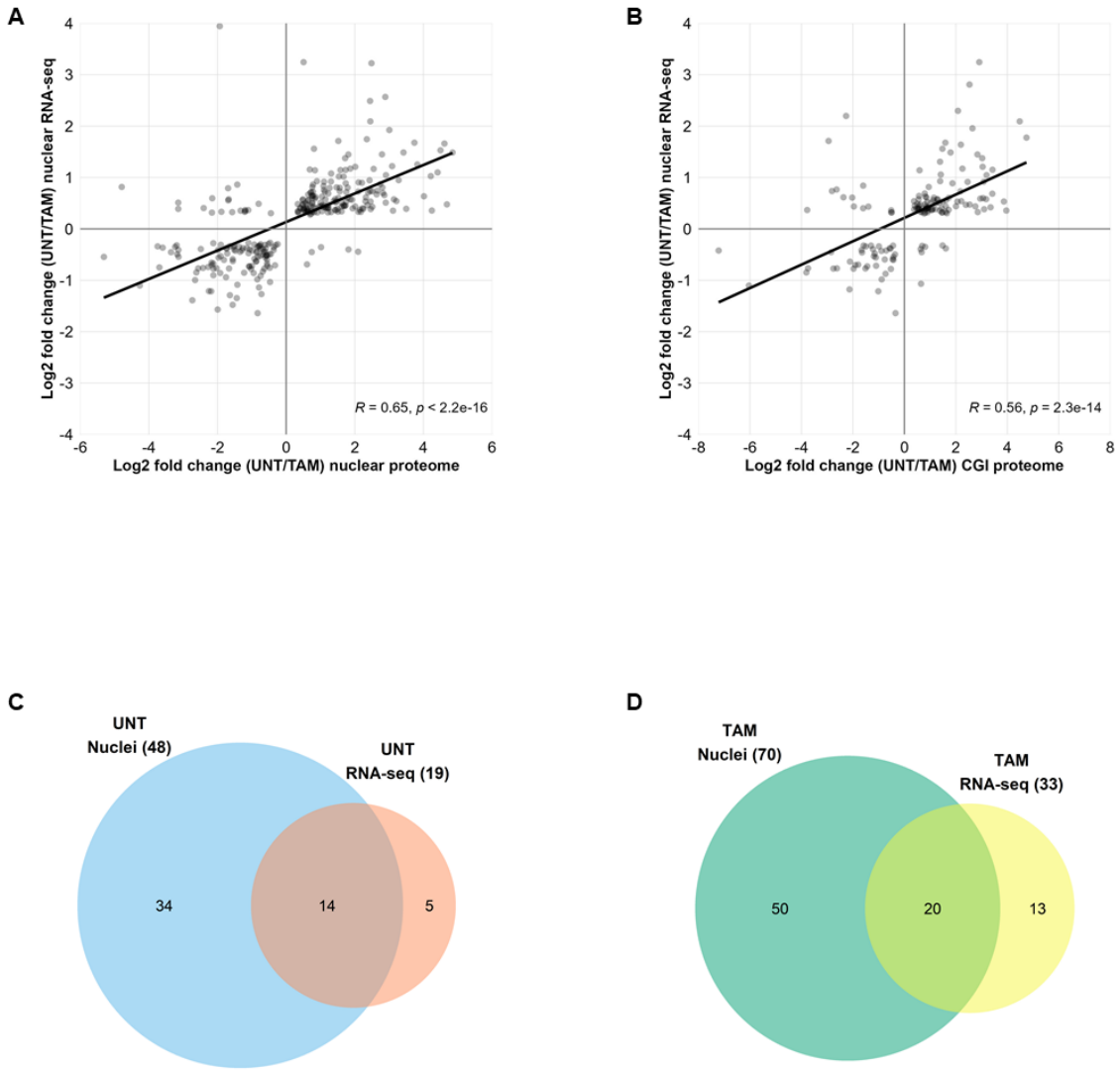


Figure S2: Correlation between mRNA and protein log2 fold changes in PRC1 CKO

(A) Scatter plot showing the Pearson correlation between nuclear mRNA and nuclear protein log2 fold changes in the PRC1 CKO experiment. Shown are only proteins with a statistically significant log2 fold change in both the MS and RNA-seq datasets. Of the 945 proteins that have changed significantly in the nuclear proteome, only 301 have also changed significantly on the mRNA level. The Pearson correlation of log2 fold changes of these 301 common proteins is R=0.65. RNA-seq data courtesy of Nadya Fursova.

(B) As in (A) for the CGI proteome. Of the 632 proteins that have changed significantly in the CGI proteome, only 155 have also changed significantly on the mRNA level. The Pearson correlation of log2 fold changes of these 155 common proteins is R=0.56, which is lower than the correlation with the nuclear proteome. This is expected as the RNA-seq data cannot specifically reflect changes at CGIs.

(C) Venn diagram showing the overlap of nuclear proteome and RNA-seq datasets for proteins that decrease in abundance at CGIs in PRC1 CKO cells. A total of 172 proteins were depleted at CGIs after tamoxifen treatment. Of these, 48 also showed depletion in the nuclear proteome, and 19 were depleted on mRNA level. Only 14 proteins are depleted in both the nuclear proteome and on mRNA level.

(D) As in (C) for proteins that increase in abundance at CGIs in PRC1 CKO cells. A total of 460 proteins were enriched at CGIs after tamoxifen treatment. Of these, 70 also showed enrichment in the nuclear proteome, and 33 were enriched on mRNA level. Only 20 proteins are depleted in both the nuclear proteome and on mRNA level.

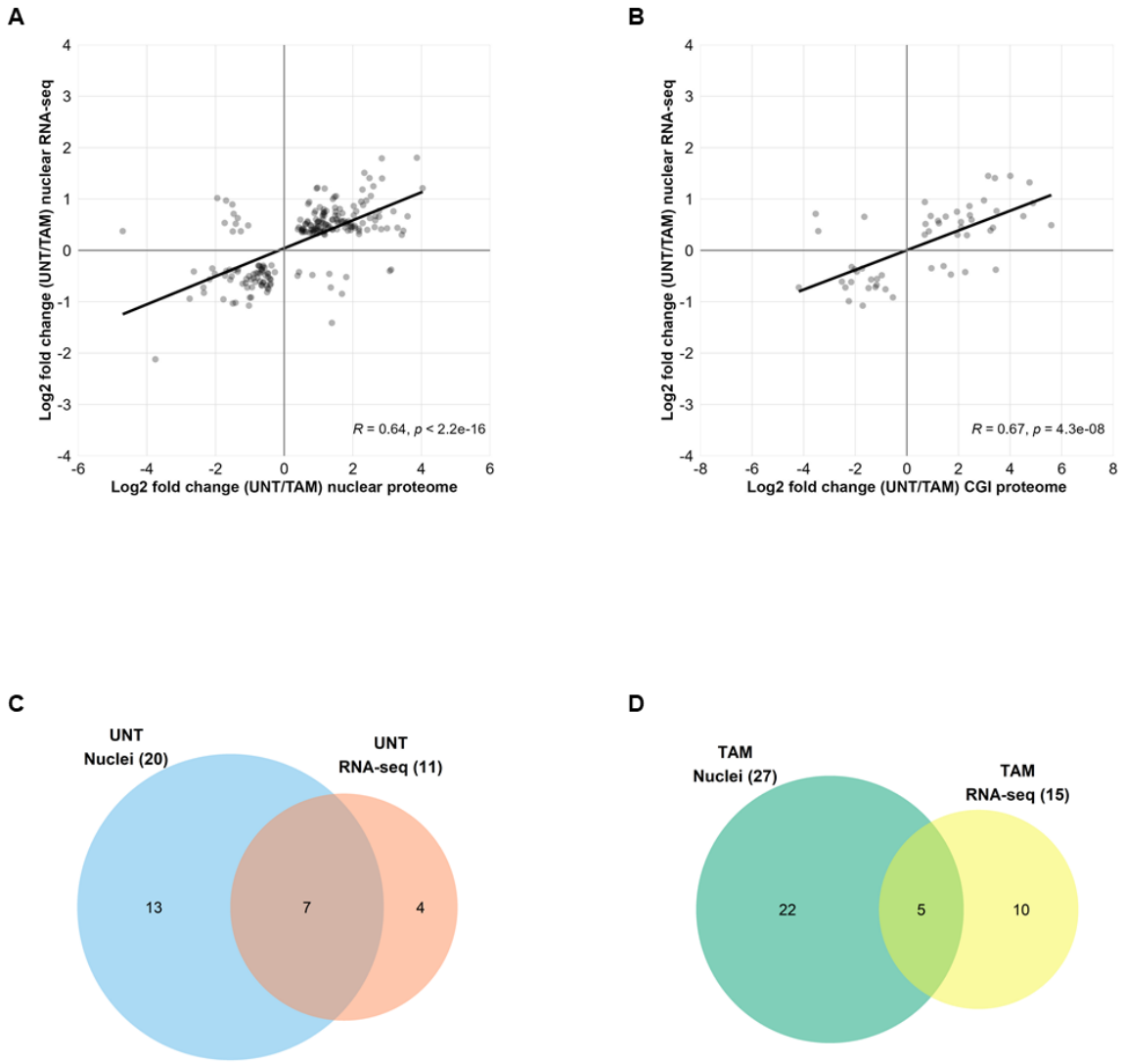


Figure S3: Correlation between mRNA and protein log2 fold changes in PRC2 CKO

(A) Scatter plot showing the Pearson correlation between nuclear mRNA and nuclear protein log2 fold changes in the PRC2 CKO experiment. Shown are only proteins with a statistically significant log2 fold change in both the MS and RNA-seq datasets. Of the 944 proteins that have changed significantly in the nuclear proteome, only 217 have also changed significantly on the mRNA level. The Pearson correlation of log2 fold changes of these 217 common proteins is R=0.64. RNA-seq data courtesy of Paula Dobrinić.

(B) As in (A) for the CGI proteome. Of the 281 proteins that have changed significantly in the CGI proteome, only 53 have also changed significantly on the mRNA level. The Pearson correlation of log2 fold changes of these 53 common proteins is R=0.67, which is surprisingly higher than the correlation with the nuclear proteome. This might be due to the comparatively small size of the dataset.

(C) Venn diagram showing the overlap of nuclear proteome and RNA-seq datasets for proteins that decrease in abundance at CGIs in PRC2 CKO cells. A total of 68 proteins were depleted at CGIs after tamoxifen treatment. Of these, 20 also showed depletion in the nuclear proteome, and 11 were depleted on mRNA level. Only 7 proteins are depleted in both the nuclear proteome and on mRNA level.

(D) As in (C) for proteins that increase in abundance at CGIs in PRC2 CKO cells. A total of 129 proteins were enriched at CGIs after tamoxifen treatment. Of these, 27 also showed enrichment in the nuclear proteome, and 15 were enriched on mRNA level. Only 5 proteins are depleted in both the nuclear proteome and on mRNA level.

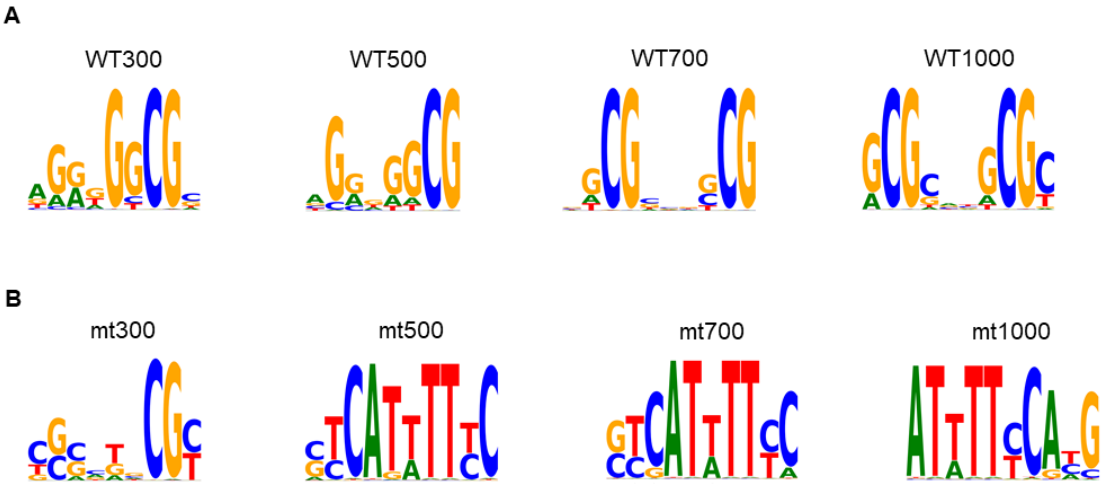


Figure S4: DNA binding motifs enriched in ZF-CXXC^{WT} and ZF-CXXC^{mt} C-CAP eluates
(A) DNA binding motifs enriched in the WT eluates of the C-CAP assay performed in Chapter 5. All four motifs show an enrichment of the CG dinucleotide motif, which is a landmark of CpG islands, and the binding motif of the ZF-CXXC^{WT} bait domain derived from KDM2B.
(A) DNA binding motifs enriched in the mt eluates of the C-CAP assay performed in Chapter 5. A CG dinucleotide motif is not consistently enriched in the mt samples, indicating that unmethylated CpGs were not bound by the ZF-CXXC^{mt} bait protein.



Figure S5: DNA binding motifs enriched in C-CAP eluates of PRC1 and PRC2 CKO
DNA binding motifs enriched in the UNT and TAM eluates of the C-CAP assays performed with PRC1 and PRC2 CKO cells in section 6.2.1. All four motifs show an enrichment of the CG dinucleotide motif, which is a landmark of CpG islands, and the binding motif of the ZF-CXXC^{WT} bait domain derived from KDM2B.

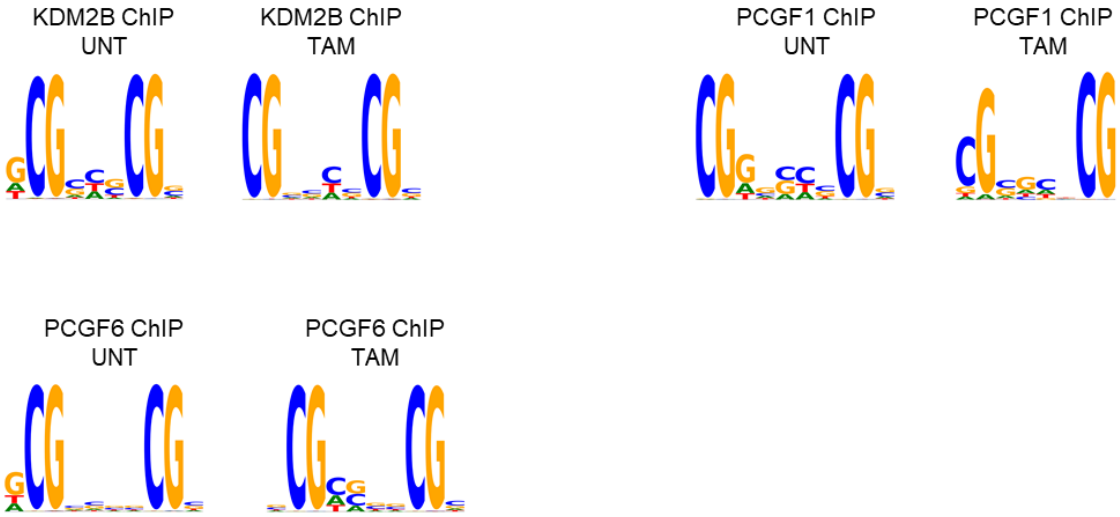


Figure S6: DNA binding motifs enriched in KDM2B, PCGF1, and PCGF6 ChIP assays
DNA binding motifs enriched in the UNT and TAM eluates of KDM2B, PCGF1, and PCGF6 ChIP assays performed with PRC2 CKO cells in section 6.2.4. All motifs show an enrichment of the CG dinucleotide motif, which is a landmark of CpG islands, and the binding motif of the ZF-CXXC domain of KDM2B. PCGF1 can be recruited to CGIs as part of PRC1.1, which comprises KDM2B as its CGI-recruitment module (Farcas et al., 2012). Similarly, the MAX/MGA and E2F6/TFDP1 dimers have the capacity to bind the CG dinucleotide motif and recruit PRC1.6, with PCGF6 at its core, to CGIs (Stielow et al., 2018).

Table S1: List of enriched proteins from Figure 35

Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC
Acot7	1.5412	2.7807	Ecm29	1.7262	2.3742	Lire1	1.9868	1.5490	Rad18	1.2624	3.1334	Terf2	1.2546	1.8594
Aff4	2.6269	3.1218	Eed	1.8076	1.2280	Maz	1.2243	3.5814	Rbbp5	1.5445	1.8041	Tfap2d	1.4485	3.2667
Ap3b1	1.7709	4.9465	Eefsec	1.5083	2.0558	Meaf6	1.3758	2.0352	Rcc1	2.3559	0.6342	Tfcp2l1	2.0561	4.5283
Ap3d1	2.3235	3.0938	Elf1	1.4124	1.9729	Mecp2	2.4896	3.5176	Rfx1	1.4643	2.3668	Tfdp1	1.3706	1.4063
Ap3m1	1.3568	3.3746	Elf2	1.3057	2.7732	Mfap1b	3.2913	0.3273	Rnf2	1.2800	2.3887	Tgm3	1.4036	1.8966
Apex1	1.6415	1.1368	Ell	2.4707	2.3504	Mga	1.5548	1.6840	Rnpc3	1.2972	1.7798	Thbs2	2.7018	6.5538
Banp	1.8161	1.9465	Ezh1	1.9419	2.0708	Mnt	1.9684	1.9547	Rpl11	2.2629	1.0509	Top1	1.9624	1.3449
Bax	2.2882	3.0760	Fancd2	1.5566	1.7737	Morf4l1	3.2919	0.9096	Rpl13a	1.4459	0.9741	Trim44	2.3273	2.4992
Baz2a	2.2732	0.6814	Fbxo28	1.9222	0.8387	Morf4l2	1.7972	0.9852	Rpl22	1.3504	1.2539	Ttc27	3.7196	8.0923
Bcor	2.4059	5.0468	Fmr1	2.8202	2.0847	Mpg	3.3639	9.8445	Rpl35	2.0004	1.2918	Ttf1	1.3655	3.4434
Bcorl1	1.7308	1.9124	Gabpa	1.4678	1.5511	Mtf2	2.2557	2.5560	Rpl7a	1.6606	0.9181	Ubp1	1.8876	1.7642
Bop1	1.7155	1.2772	Gpalpp1	1.5333	3.3342	Mum1	1.3764	1.8738	Rps27a	1.6401	0.9307	Ubr4	1.6965	4.2428
Brat1	2.7633	4.7495	Grhl2	3.1652	4.9700	Naf1	2.0302	3.8374	Rps7	1.3622	1.5362	Ubtf	1.8029	2.1341
Brd2	3.3465	1.5147	Gtf2b	1.5746	1.2940	Ncl	1.8548	0.8593	Rtf1	2.0711	2.3266	Uhrf1	1.7887	0.7451
Brd3	2.1486	1.6036	Gtf2f1	3.4027	0.7023	Nfyb	2.0920	3.8818	Samd1	1.6360	3.3639	Upp2	1.4651	4.7361
Brd4	1.5964	0.7690	Gtf2f2	1.8332	1.3493	Nhp2	1.7611	1.9698	Sap30bp	2.1809	1.1397	Usf1	1.7982	1.8979
Brdt	2.1263	2.2572	Gtf2h4	1.3571	1.6727	Nktr	2.0768	4.5999	Setd1a	2.1398	4.3071	Utf1	1.8537	0.7758
Brwd1	2.6049	2.6273	H1f0	1.7031	2.0706	Nolc1	2.0647	4.2120	Setd1b	1.2783	2.6593	Wars	1.2873	2.2740
Brwd3	1.9675	4.5473	H2afv	3.7136	1.1053	Nrf1	2.2306	3.0589	Sirt1	2.3670	2.7347	Wdr5	2.2541	0.4649
C18orf25	2.1418	2.8068	H2afy	1.9377	1.6633	Nucks1	1.9087	3.9490	Slc7a6os	2.0624	2.4384	Wdr89	1.5493	2.0612
Ccdc15	1.5558	3.2380	H2afy2	1.4728	1.2122	Paf1	3.8102	3.4255	Smarca5	2.2455	0.5698	Ywhag	1.9571	5.0711
Ccnl1	1.6077	1.2899	Hars	1.5228	1.3876	Parp1	1.8114	0.6952	Smc1b	1.4579	0.9347	Yy1	1.3104	1.8228
Cd3eap	2.0991	1.9243	Hdgfrp2	2.5459	3.5443	Pbx2	1.6370	3.3522	Snapc4	1.3848	2.5924	Zbed6	2.5521	3.5473
Cdca7l	1.4997	2.1324	Hdgfrp3	1.3969	2.9048	Pcbp2	3.1645	2.7662	Sntb2	2.2226	2.7737	Zbtb11	1.6012	3.7284
Cdk11b	2.9507	2.9512	Hirip3	1.4660	1.9445	Pcgf1	1.2866	3.5722	Sp2	2.3286	2.5794	Zc3h18	2.0423	2.1949
Cfap20	2.0210	2.0631	Hist1h2ah	1.4258	1.5771	Pcgf2	1.3698	2.9665	Spin1	1.3513	3.8877	Zdbf2	2.2673	3.6994
Cfdp1	1.7511	2.3549	Hist1h2ba	1.9559	1.5737	Pds5a	1.8693	0.6033	Ssrp1	1.7528	1.3391	Zfa	1.2459	2.9596
Chd1	4.5835	2.6223	Hist1h2bp	2.6405	1.8343	Phf23	2.1342	1.2370	Stag1	1.6431	1.3687	Zmynd11	1.3088	2.8043
Chd9	1.5043	3.5006	Hist2h2aa1	1.6913	3.8873	Phf8	1.5212	3.0036	Stag2	1.4185	1.2351	Znf143	2.8713	2.9150
Cmas	1.2393	3.5012	Hist2h2bb	1.2972	2.4372	Phip	2.0449	1.5621	Suds3	2.0176	1.4694	Znf428	1.2100	3.0390
Coil	2.4479	2.2046	Hist3h2ba	2.1328	1.1157	Pinx1	1.4783	2.1013	Supt16h	3.2526	1.8365	Znrd1	1.4459	1.2677
Col18a1	1.4516	3.6917	Hmga1	2.0175	1.6052	Pknox1	1.7842	3.4480	Supt5h	1.7575	1.7796	Zranb2	2.7118	3.0882
Crebrf	4.2358	9.2884	Hp1bp3	1.5437	2.6472	Pkp1	1.7148	1.7782	Suv39h2	2.2182	1.5147			
Csnk2a1	2.2016	1.4906	Htatsf1	2.9334	4.8148	Pnisr	1.2377	4.2449	Suz12	1.4815	1.4262			
Csnk2b	1.7148	1.0772	Inca1	2.7465	1.7070	Pnn	2.5664	0.9330	Synm	1.3663	2.0225			
Ctdspl2	2.2661	1.4075	Ing3	1.7770	1.1263	Polr1a	1.4334	2.0774	Tacc3	1.5213	2.2794			
Custos	1.7784	2.7147	Jarid2	2.0080	2.5493	Polr1b	1.4407	1.6597	Taf1	1.4208	3.8934			
Cwc22	1.7582	2.0607	Jmjd6	1.4550	2.9013	Polr1c	1.4835	1.3399	Taf1a	2.0898	3.7807			
Dcun1d1	1.4261	0.9747	Kat6a	1.3105	3.2435	Polr1e	1.2818	2.7465	Taf1b	2.2963	2.8033			
Dek	2.1160	1.5455	Kat6b	2.5196	2.2325	Polr2a	2.4865	1.2489	Taf1c	1.6127	2.6779			
Dkc1	2.1141	0.8805	Kdm2a	1.5945	1.8151	Polr2b	2.0575	1.8893	Taf1d	1.9883	3.1630			
Dmap1	1.7672	1.6288	Kdm2b	1.6253	3.2108	Polr2c	1.6324	1.8983	Taf2	1.5730	3.9211			
Dnajc9	3.1513	0.5570	Kdm4a	1.3411	2.7713	Polr2e	2.0838	0.9970	Taf3	1.7184	2.9889			
Dohh	1.2409	3.1917	Kdm4c	1.3373	2.7470	Polr2h	1.4434	2.3947	Taf4b	2.2587	2.9168			
Dppa2	1.7648	1.5764	Kif17	3.6080	4.3698	Polr3b	1.3472	3.3913	Tbc1d15	2.0900	4.6221			
Dppa4	1.8488	1.8158	Kif22	1.7805	1.0553	Polr3c	1.3339	2.3609	Tbp	2.0935	1.8574			
Dr1	2.1391	2.8268	Kmt2a	1.2589	2.3683	Polr3f	1.4914	3.4922	Tceb3	2.1089	1.9150			
Drap1	2.9109	4.2980	L1td1	1.9317	3.2067	Ppm1g	1.5194	2.2222	Tcof1	3.5189	3.7934			
Dsg1a	2.4643	3.3910	Larp7	2.1436	1.3830	Prdm10	1.4106	2.5592	Tead1	1.3432	1.5005			
Eaf1	1.9639	5.1099	Leo1	3.6870	3.2545	Psip1	2.1999	0.8807	Terf1	1.5073	1.4844			

Table S2: List of enriched proteins from Figure 44B

Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC
Adcy9	2.6633	3.7688	Epop	3.8345	3.5223	Kmt2b	4.2106	3.2612	Prdm10	1.4209	2.6974	Supt5h	4.2253	1.1729
Aebp2	1.6390	0.7928	Ercc3	1.1798	1.7998	Kri1	1.9055	1.1678	Prdm15	4.1035	4.4486	Suz12	2.5420	1.8236
Ap3b1	2.2077	7.9633	Esf1	3.7931	1.0206	Krr1	1.3237	1.7244	Prpf40a	1.4188	0.9445	Taf1	2.3462	4.7738
Apex1	1.9311	0.9508	Etf1	1.6619	0.7516	Ktn1	1.4592	4.0897	Psip1	2.3660	1.5092	Taf1a	1.6546	3.5362
Arglu1	1.9263	1.2330	Ezh1	1.6061	1.8344	L1td1	2.1840	2.3709	Psma1	0.9407	1.4408	Taf1b	1.9527	4.1009
Arl6ip4	1.6113	1.8883	Ezh2	3.9800	2.2226	L3mbtl2	1.5497	1.2774	Psme4	3.5652	6.4388	Taf1c	1.7424	4.2262
Asf1a	1.7563	1.7039	Fam133b	1.5465	2.2234	Larp7	4.2395	1.3162	Pum2	2.2012	6.2864	Taf1d	3.0463	3.6961
Atp5s	1.0474	1.4792	Fbl	2.1132	1.0204	Lbr	2.1849	0.9351	Pus1	2.0909	1.8884	Taf2	2.4045	2.1392
Aurkb	1.1372	1.0939	Fbll1	1.4948	1.3514	Leo1	1.9351	1.2817	Pus10	1.9498	2.7464	Taf3	3.6326	4.4103
Banf1	1.0005	1.2492	Frg1	1.1337	0.6168	LRWD1	1.9034	0.4985	Raly	1.2187	0.5675	Taf7	2.2097	0.8539
Baz1b	1.4210	0.7039	Gabpa	3.0123	0.8797	Lyar	1.9903	1.1405	Rbm19	1.2577	0.7287	Tbp	2.7565	2.4629
Baz2a	2.0597	1.1979	Gltscr2	1.3061	0.5628	Mbtd1	1.1592	2.2700	Rcc1	2.0860	1.4024	Tcof1	5.1453	2.9875
BC027231	2.4727	0.9730	Gnl3	2.6340	0.7826	Mdc1	3.8286	1.8203	Rest	1.9283	2.0447	Tead1	1.8094	3.5306
Bcor	0.8978	1.7088	Got1	1.4271	2.5400	Mecp2	2.1791	2.2759	Rfx1	1.0741	3.0372	Terf1	1.0500	0.7676
Bnip1	2.0501	4.8120	Gpatch4	1.2506	0.9336	Men1	2.5978	3.1003	Rlim	2.5381	2.2921	Tfap2d	1.8198	3.2307
Brd2	2.5713	2.1526	Gpatch8	1.7063	2.2733	Mepce	0.8766	2.8624	Rnf2	3.6108	2.5033	Tgm3	1.9945	2.2645
Brd3	1.2455	1.2376	Gpx4	1.3251	0.6994	Mga	1.4068	1.5339	Rpf2	2.0972	1.1743	Thumpd1	2.1127	1.8337
Brms1l	0.9574	2.2341	Gsg2	1.0330	1.7787	Mki67	2.8100	0.9190	Rpl10	1.2589	0.9701	Tigar	1.0165	1.0596
Brpf1	3.7647	4.8670	Gtf2b	1.6132	1.0665	Morf4l2	1.1596	0.9349	Rpl10a	1.2737	0.6094	Tma16	2.5009	1.0935
Brwd1	2.4304	4.5061	Gtf2f2	1.1832	1.1295	Mrto4	1.7554	0.6177	Rpl11	2.5095	1.4414	Tmpo	1.6149	0.5903
Brwd3	2.6743	4.8612	Gtf3c1	2.2046	3.4334	Mtf2	2.2517	2.6878	Rpl14	2.3319	0.6114	Top1	2.4998	1.4487
Bzw2	1.6989	1.6363	Gtf3c2	1.1679	1.1931	Naf1	1.1395	2.0247	Rpl17	1.2771	1.2363	Tpx2	1.4726	0.7356
C18orf25	2.9879	2.0950	Gtpbp4	2.0581	0.6962	Ndnl2	2.3574	2.9943	Rpl18	2.6768	0.6068	Trmt10a	0.9725	3.4464
Sephs1	2.4396	1.7051	H1f0	1.7703	2.2496	Nfyb	1.4254	4.0006	Rpl19	2.2223	0.9588	Trmt112	0.9960	1.6357
Cggbp1	2.2883	3.4137	H2afv	3.7917	2.6765	Nhp2	1.9694	2.5030	Rpl22	2.6350	0.7161	Tsr1	2.7179	1.0742
C4orf27	2.9791	2.1591	H2afx	1.4830	7.5783	Nhp2l1	1.8197	0.7923	Rpl3	2.6688	1.0566	Ttf1	3.2528	2.8458
C7orf50	1.7088	1.8466	H2afy	1.8844	1.6338	Nifk	1.5385	0.6819	Rpl34	1.2057	0.5470	Twistnb	2.8628	2.7697
Casp3	2.8805	2.1747	H2afy2	2.4917	1.3372	Nktr	2.4515	5.0392	Rpl35a	2.1087	1.1742	Ube2l3	2.1438	2.2747
Cbx1	1.2940	0.8277	Hdgf	3.4133	1.4163	Nmt1	0.9772	1.2737	Rpl38	1.9237	1.3228	Ube2v1	2.0491	0.9762
Cbx2	1.7197	2.4559	Hdgfrp2	3.8429	2.7626	Nolc1	3.3732	3.9135	Rpl4	1.8018	0.5651	Ubtf	3.1002	3.8139
Cbx7	3.5874	3.7750	Hirip3	1.6673	0.5155	Nop10	2.3655	4.3777	Rpl6	1.8943	1.0379	Uhrf1	1.6955	0.8152
Cd3eap	4.4351	3.7503	Hist1h1a	3.1401	1.6150	Nop16	2.9661	1.2220	Rpl7	3.4687	1.0357	Upp2	1.1559	4.0013
Cdca7	1.5177	2.5056	Hist1h1b	1.6557	1.1744	Nop2	1.1125	0.6791	Rpl9	1.5576	1.2070	Usf2	1.6866	2.4195
Cdca7l	4.0266	4.2951	Hist1h1c	2.5876	1.4959	Nop56	1.8530	0.9650	Rpp30	2.1241	1.3237	Utf1	1.6375	2.0536
Cdca8	2.4740	1.3959	Hist1h1d	2.3536	1.2403	Nop58	1.6764	0.5963	Rpp38	1.8191	0.8909	Utp11l	2.2037	1.9058
Cdk11b	1.4225	1.0411	Hist1h1e	1.3556	1.2582	Nploc4	0.9298	1.1237	Rps10	2.2046	1.9474	Utp3	2.4019	0.9881
Cenpq	2.7305	1.4276	Hist1h1t	0.9684	1.7926	Nrf1	5.2760	5.9471	Rps13	1.3709	0.5399	Vps72	0.9455	2.4151
Cenpu	1.4836	1.9084	Hist1h2ah	1.6839	1.7440	Nucks1	2.8168	2.7004	Rps14	1.9281	1.0428	Vrk1	2.3549	1.6641
Chchd2	1.7474	2.4101	Hist1h2ba	1.5716	1.8767	Nusap1	1.0208	1.5032	Rps16	1.3242	0.6996	Wbscr22	2.3223	3.6355
Chd1	3.6535	2.7339	Hist1h2bp	3.8534	2.5217	Orc1	3.8169	0.8469	Rps19	1.1083	1.7405	Wdr46	2.0072	1.0515
Chd2	2.0629	3.2342	Hist1h3a	2.1827	1.5724	Paf1	1.5818	1.5053	Rps19bp1	1.1062	0.9588	Wdr74	1.0545	1.1025
Chd8	1.3285	0.6681	Hist1h4a	2.2002	1.5909	Parp1	1.4613	0.5660	Rps2	1.4632	0.7362	Yy1	1.8598	1.1457
Cmas	1.0104	1.5863	Hist2h2aa1	2.3293	3.6091	Pcbp2	1.3091	2.5745	Rps25	1.3499	0.6596	Zbtb11	4.2957	5.1470
Coil	3.3302	1.6735	Hist2h2ab	2.7563	1.4730	Pcgf1	1.8757	2.3457	Rps26	1.7427	1.0614	Zc3h11a	1.7608	0.4982
Crip2	1.2767	0.9876	Hist3h2ba	1.5511	1.1605	Pcgf2	1.8930	4.6625	Rps27a	3.6939	2.5205	Zc3h13	4.2194	3.8465
Ctcf	1.9417	2.4247	Hjurp	0.8962	1.6979	Pds5b	1.8649	1.1790	Rps3	1.6895	1.0029	Zc3h18	2.4363	2.2603
Ctdspl2	3.2287	2.2900	Hmga1	2.2275	2.1878	Pes1	1.8270	0.7016	Rps4x	1.0850	0.8368	Zc3h8	1.4486	1.6606
Cwc22	1.4072	1.1550	Hmga1	2.1645	3.3030	Phax	3.5263	3.6421	Rps6	2.2374	0.7916	Zcchc10	1.9705	5.6695
Cwc25	1.6845	0.8689	Hmgb3	0.9605	1.8801	Phc1	2.4037	2.9372	Rps8	2.3902	1.0579	Zdbf2	1.7777	4.4293
Dcun1d5	1.0659	0.6512	Hnrnpa3	2.9785	0.9119	Phf14	2.1633	3.3051	Rrp9	1.2248	0.9218	Zfa	2.0652	2.6831
Ddx21	2.2334	0.8971	Hnrnpab	0.8876	1.4929	Phf23	3.2011	1.9920	Rsbn1	2.4171	2.7721	Zfp120	1.5460	2.2770
Ddx31	1.6994	1.0902	Hnrnpc	1.9375	0.4995	Phf8	4.4196	4.0986	Rsl1d1	2.8953	0.9342	Zmynd11	1.2505	2.1096
Ddx39a	1.7708	0.4184	Hnrnpdl	1.5282	0.5339	Phip	2.5130	2.5882	Rsl24d1	1.8127	0.6175	Znf428	2.2745	2.8427
Ddx49	1.4994	0.5933	Hp1bp3	1.4863	2.8554	Pin4	2.1777	1.1070	Rtf1	2.4831	1.3316	Znf451	1.8308	0.5947
Ddx54	3.7014	4.7645	Hspa9	1.1102	2.9256	Pinx1	3.9210	2.0705	Rybp	1.4892	0.8499	Znf512	2.2175	1.8139
Dek	4.2954	2.0735	Htatsf1	1.3723	1.6665	Pkp1	2.4704	4.0484	Samd1	2.8399	3.1476	Znf740	2.3658	1.2871
Denr	1.1846	1.1184	Inca1	3.7267	3.6144	Plin3	2.5115	6.8989	Sdad1	2.5502	1.2696	Znf771	0.8992	1.9599
Dimt1	1.3272	0.7437	Incenp	1.5518	1.0086	Pnisr	3.1105	4.4026	Set	1.3088	0.9943	Znf800	1.0399	3.3626
Dkc1	1.6417	1.5762	Ing5	1.1714	0.6869	Pno1	1.5100	0.8948	Setd1a	1.6667	1.8744	Znrd1	2.2196	3.4128
Dmap1	1.7127	1.3984	Ino80	4.9330	2.9493	Pole3	1.0149	1.9125	Sgol1	2.1736	3.3353	Zranb2	2.7171	1.6585
Dnajc9	2.4829	0.9104	Iws1	0.8744	2.5898	Polr1a	2.4887	1.8003	Sirt1	1.7732	2.2114			
Dnttip1	1.1673	0.8673	Jade1	1.2134	0.8306	Polr1b	2.5719	3.6710	Slc3a1	0.9421	1.5912			
Dpf2	1.6394	1.0677	Jarid2	2.3192	2.5493	Polr1c	1.5046	1.3567	Smarca5	2.7011	0.6144			
Dppa2	2.2792	2.7736	Jmjd6	1.3652	3.9125	Polr1e	3.9174	4.4553	Snrpa1	1.7541	1.1275			
Dppa4	2.1993	1.9023	Kat6b	3.1794	2.8035	Polr2a	2.6323	1.5995	Snrpe	1.7178	1.1128			
Dr1	1.5916	3.0685	Kdm2a	2.3869	1.4465	Polr2b	4.0605	2.4391	Snrpg	3.3554	3.4434			
Drap1	3.3825	4.6150	Kdm2b	2.1235	3.3768	Polr2e	3.6468	2.8221	Sntg1	2.2124	3.6487			
Dysf	3.8100	6.4830	Kdm4a	2.0214	2.1904	Polr2g	2.3085	1.0304	Sp2	2.8299	3.7779			
Ebna1bp2	1.8818	0.5563	Kdm4c	1.9193	3.7733	Polr3b	1.9858	3.5194	Spin1	2.9938	2.7487			
Edf1	1.2119	1.4540	Kdm5c	1.8825	1.5060	Polr3c	1.4039	2.0873	Ssb	3.2699	0.9672			
Eed	2.1911	1.0902	Kiaa0020	1.2544	0.6638	Polr3f	2.0516	3.0623	Ssrp1	1.9699	1.3731			
Eif2s1	1.7997	0.7182	Kif17	1.9041	5.0781	Ppan	1.1977	0.9772	Stk26	1.7368	2.2161			
Eif3c	1.2749	1.2337	Kif22	2.5061	1.7025	Ppia	2.7715	0.5848	Sub1	1.9856	0.8228			
Eif3l	1.3337	1.5845	Klf4	2.5650	2.5969	Ppm1g	1.4499	0.8632	Sumo3	1.2998	0.6996			
Elf1	1.7823	2.3020	Kmt2a	1.0221	2.7059	Ppp1r10	1.3374	1.2487	Supt16h	1.9643	1.5500			

Table S3: List of enriched proteins from Figure 45A

Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC	Protein	-Log10 FDR	Log2 FC
Abcf2	1.5092	1.5167	H2afy	1.7374	1.1747	Polr1a	2.8026	2.4443	Tfcp2l1	1.1219	2.7893
Adcy9	2.9737	3.5963	H2afy2	2.7432	1.5525	Polr1b	2.4611	4.4123	Tfdp1	1.4480	1.8132
Ak1	1.3121	1.5520	Hdgf	2.3898	1.4929	Polr1c	2.2176	1.6424	Tfip11	1.0524	2.6973
Ap3b1	1.3708	3.0243	Hist1h1a	1.6781	0.8077	Polr1e	2.4785	3.0023	Tonsl	3.4094	3.8177
Apex1	1.6912	1.2489	Hist1h1c	1.7617	0.9794	Polr2a	2.6755	1.5828	Tp53	2.0730	3.5743
Arl6ip4	3.3510	3.0090	Hist1h2bc	1.9530	1.7860	Polr2b	4.0651	2.6251	Trank1	1.1439	1.1034
Asf1a	1.0434	1.4111	Hist1h2bp	2.9487	1.3919	Polr2e	1.7705	1.1659	Ttf1	1.4112	1.5527
Bcor	1.3704	1.9304	Hist1h3a	1.4858	1.4545	Polr2g	1.8037	0.6384	Twistnb	1.9337	1.7521
Brd2	1.2617	0.7715	Hist1h4a	1.5460	1.0364	Polr3b	2.3650	4.0324	Ube2e1	2.6492	1.9483
Brpf1	1.4978	1.2440	Hist2h2aa1	1.5160	1.9387	Polr3c	1.6910	3.0277	Ubtf	2.8116	2.4935
Brwd1	1.4180	2.0958	Hist2h2ab	1.9850	1.3591	Polr3f	2.2033	3.7172	Uhrf1	1.9940	0.7689
Brwd3	2.0387	3.8325	Hist3h2ba	1.7361	1.1969	Prdm10	2.0072	2.3726	Vps72	1.0150	1.9053
Calml3	1.0851	2.1560	Hmga1	2.7396	1.3956	Prdm15	2.4261	3.9367	Vrk1	1.2254	0.7644
Carnmt1	3.9861	1.4213	Hmga1	1.4664	0.9768	Prkci	1.0962	1.8895	Wbscr22	1.2847	1.4041
Cbx7	1.7030	3.5610	Hp1bp3	1.2474	2.0173	Psmc6	1.5344	1.7156	Wdr5	1.4068	0.5574
Cd3eap	1.7672	1.8488	Hspa9	2.1131	3.1084	Psme4	1.0393	3.2783	Yeats4	2.2266	0.5833
Cdca7l	1.9156	1.8237	Inca1	3.2679	2.5077	Ptma	1.8750	4.0416	Yy1	1.6132	1.0582
Cenpc	1.1908	2.6088	Ing5	1.0333	3.1505	Rbbp5	2.0485	3.9473	Zbtb11	1.7524	3.9018
Cenpu	2.6220	3.9165	Ipo5	2.7104	3.1035	Rcc1	2.0449	1.1180	Zc3h18	1.7048	1.3230
Chd1	3.0202	1.9096	Iws1	1.5213	4.7868	Rcc2	1.8569	0.8437	Zcchc10	2.8912	2.4411
Cmas	1.2940	1.0435	Jarid2	1.7978	2.5113	Rest	1.3167	1.4727	Zfa	1.6663	1.5032
Coil	2.0384	0.5954	Kat6b	2.3940	3.1535	Rfx1	2.1550	3.4582	Zfp292	1.5614	4.8548
Ctcf	2.1257	2.7488	Kdm2a	1.3525	1.3626	Rnf2	2.2417	2.3101	Zfp91	1.1932	1.5145
Ctdspl2	1.6539	1.2179	Kdm2b	2.1996	2.9311	Rps27a	3.1045	1.8242	Zmynd11	1.5153	3.5253
Dcun1d5	2.3442	0.7211	Kdm4a	1.8787	4.8709	RtcA	1.5470	1.9683	Znf281	1.8489	0.7826
Dek	3.9171	0.6900	Kdm4c	1.7278	3.9845	Rybp	1.4952	0.7800	Znf428	1.4077	1.8172
Dppa2	1.3382	2.3638	Kdm5c	2.8815	4.8777	Samd1	3.0373	3.7026	Znf512	1.5837	0.9713
Dppa4	1.7067	1.5504	Klf4	2.4256	2.3834	Sephs1	2.5111	3.1422	Znf740	1.3581	1.4717
Dr1	1.2509	2.6646	Kmt2b	1.5305	1.8287	Sept1	1.3657	1.0021	Znrd1	1.5924	2.2758
Drap1	3.5203	3.3698	Maz	1.9521	1.9768	Sfn	1.7461	1.3016			
Dysf	1.2538	4.4515	Mdc1	1.1823	0.8664	Slc3a1	1.7937	1.7213			
Eed	1.9335	1.3273	Mdh1	1.3596	1.9904	Snrpg	3.2248	3.1857			
Eef1a2	1.4784	2.2972	Men1	1.3389	1.1012	Sp2	1.9071	4.5561			
Ehmt2	1.6081	2.8762	Metap1	2.6027	2.7048	Sp3	1.2803	2.3633			
Eif3l	1.8560	2.5215	Mga	1.3647	1.3413	Spin1	2.6646	2.5808			
Eif4g2	1.1666	1.1264	Mtdh	2.2195	3.5874	Stag1	1.3309	1.2503			
Elf1	1.5924	1.4798	Mtf2	2.6034	3.6627	Stag2	1.5435	0.8567			
Epop	1.8961	1.7365	Myh11	1.1225	2.3924	Suz12	2.2270	1.4001			
Ercc3	1.2630	2.5347	Nfyb	2.1181	5.4520	Taf1	2.9909	5.8980			
Ezh1	1.8490	2.2769	Nolc1	1.7904	1.0060	Taf1a	2.1694	2.3575			
Ezh2	3.5505	1.8755	Nrf1	4.0789	4.3757	Taf1b	2.4480	4.4846			
Gabpa	2.2414	2.3633	Nucks1	1.6265	1.5723	Taf1c	1.9533	3.2487			
Got1	2.1960	2.8711	Pcgf1	1.5664	2.6726	Taf1d	1.5420	1.3399			
Gpam	2.4258	6.5892	Pcgf2	2.7733	1.8776	Taf3	3.4211	3.7652			
Grhpr	2.8605	1.5416	Pcyt1b	0.9964	1.8544	Taf5	1.6414	1.9049			
Gtf2h2	1.5942	0.9055	Phc1	1.6800	2.4366	Tbp	1.7030	0.9266			
Gtf3c1	3.0422	3.4410	Phf14	3.0754	3.5031	Tcof1	2.1352	1.2904			
Gtf3c2	2.9805	5.3772	Phf8	2.1694	1.6498	Tdp1	1.2665	1.1413			
H1f0	1.9528	1.9132	Phip	1.7708	1.6579	Tead1	1.1650	2.7456			
H2afv	2.1507	1.6749	Pias4	1.3136	0.9528	Tecr	1.5881	3.2341			

Table S4: List of background proteins (enriched in Input and ZF-CXXC^{mt})

Protein	Protein	Protein	Protein	Protein	Protein	Protein	Protein	Protein	Protein
Aars	Cct8	Dnaja1	Hk2	Mad1l1	Nup88	Prpf8	Rtcb	Ssu72	Uchl5
Abt1	Cd2bp2	Dnaja2	Hmmr	Mapk1	Nup93	Psma5	RTRAF	St13	Umps
Acin1	Cdc16	Dnajc7	Hnrnpf	Matr3	Nup98	Psma7	Ruvbl2	Stat3	Urb1
Acly	Cdc37	Drosha	Hnrnpm	Mbd3	Nxf1	Psmc1	Sall1	Steep1	Usp14
Actn4	Cdk1	Dtymk	Hsp90aa1	Mcm2	Ogt	Psmc2	Sall4	Stip1	Usp37
Adss	Celf1	Dync1h1	Hsp90ab1	Mcm3ap	Pabpc1	Psmc3	Samhd1	Suds3	Usp5
Ahctf1	Champ1	Ect2	Hsp90b1	Mcm6	Pabpn1	Psmd12	Sars	Sugt1	Vcp
Ahcy	Cherp	Eef1b	Hspa4	Mcm7	Paics	Psmd14	Sart1	Supt6h	Vim
Akr1b1	Chmp4b	Eef1g	Hspa5	Med17	Park7	Psmd3	Sart3	Swi5	Vps29
Aldoa	Chtop	Eef2	Hspa8	Memo1	Parn	Psme3	Scaf1	Sympk	Vps4b
Anln	Ckb	Eftud2	Hspbap1	Mettl3	Parp2	Pspc1	Scaf4	Taf6	Wapal
Anxa11	Clk1	Ehd1	Hsph1	Mfap1b	Paxbp1	Puf60	Seh1l	Tardbp	Wbp11
Ap2a1	Cmtr1	Ehmt1	Huwe1	Morc2a	Pcbp1	Pus7	Sept11	Tbcb	Wdhd1
Ap2a2	Cnn3	Eif3a	Igbp1	Mta1	Pcnp	Qki	Sept2	Tceb1	Wdr1
Ap2m1	Copb1	Eif3b	Ik	Mta2	Pdcd6ip	Rad50	Setd2	Tcerg1	Wrnip1
Ap2s1	Cops5	Eif3e	Ikbkap	Mta3	Pdlim1	Rae1	Setx	Tcp1	Wtap
Aqr	Cpsf1	Eif3f	Ilkap	Mthfd1	Pelp1	Ranbp1	Sf1	Tex10	Xpo1
Arcn1	Cpsf2	Eif3h	Ina	Myef2	Pfas	Ranbp2	Sf3a1	Thoc1	Xpo4
Arf4	Cpsf3	Eif4g1	Ints1	Naca	Pfdn2	Ranbp3	Sf3b1	Thoc2	Xpo5
Arid1a	Cpsf3l	Eif5	Ints4	Nap1l4	Pgd	Rangap1	Sf3b3	Thoc5	Xpo7
Arih1	Csde1	Ep300	Ints7	Napa	Pgk1	Rars	Sgol2	Thoc6	Xpot
Armc4	Csnk1a1	Erh	Ints9	Ncapd2	Phf2	Raver1	Skiv2l2	Thop1	Xrcc5
Asna1	Csrp2	Esd	Ipo4	Ncbp1	Pias2	Rbbp6	Sltm	Thrap3	Xrcc6
Atad2	Cstf2	Exosc3	Ipo5	Ncoa5	Picalm	Rbm10	Slu7	Tjp1	Xrn2
Atf7ip	Cstf3	Exosc4	Ipo7	Nedd4	Plaa	Rbm12	Smap	Tjp2	Ybx1
Atic	Ctdp1	Exosc7	Iqgap1	Nedd4l	Pmm1	Rbm14	Smarcad1	Tle3	Ylpm1
Atxn10	Ctnna1	Fam208a	Irf2bpl	Nelfb	Pnkp	Rbm17	Smarcc1	Tle4	Ywhae
Bclaf1	Ctnnb1	Fam50a	Isy1	Nfrkb	Pogz	Rbm22	Smarcd1	Tlk2	Ywhaq
Birc6	Ctnnd1	Fanci	Katnal2	Nmd3	Pola1	Rbm25	Smarce1	Tmem209	Zc3h14
Bod1l	Cul2	Farsb	Kdm1a	Nme1	Pola2	Rbm33	Smc2	Tnpo1	Zcchc8
Bub3	Cul3	Fasn	Kdm3b	Nmral1	Pold1	Rbm42	Smc4	Tpm3	Zfr
Bud13	Cul4b	Fip1l1	Kdm6a	Nol11	Pole	Rbm5	Smc5	Tpr	Zmym2
Caap1	Cul5	Fkbp4	Khsrp	Nono	Pom121	Rcor2	Smc6	Tra2a	Zmym4
Cad	Cwc15	Fkbp5	Kiaa1429	Nr0b1	Ppid	Rexo4	Smek1	Trim24	Znf207
Cand1	Cwc27	Fubp1	Kif20b	Nsfl1c	Ppie	Rfc2	Smpd4	Trmt1l	Znf532
Capza1	Cwf19l1	G3bp1	Kif4	Nsmce1	Ppil1	Rfc5	Snd1	Trmt2a	Znf830
Capzb	Dbr1	G3bp2	Kpna2	Nsrp1	Ppil2	Rnf138	Snrnp200	Trnt1	Zpr1
Carm1	Ddb1	Gars	Kpna3	Nudc	Ppm1a	Rnf20	Snrpc	Trrap	Zscan4f
Ccar1	Ddx1	Gatad2a	Kpna6	Nudcd1	Ppm1b	Rnf220	Snw1	Ttc33	
Ccar2	Ddx20	Gatad2b	Kpnb1	Nudt5	Ppp1r7	Rnf40	Snx5	Tuba1b	
Ccnb1	Ddx3y	Gemin5	Lactb2	Nup107	Ppp2ca	Rnh1	Son	Tubb4b	
Ccnh	Ddx41	Gfpt2	Lancl1	Nup133	Ppp2r1a	Rpa2	Spag7	Tubb5	
Ccnk	Ddx42	Gnb2	Lars	Nup155	Ppp2r2a	Rpl21	Spr	Txndc9	
Ccnt1	Ddx46	Grb2	Las1l	Nup160	Ppp4c	Rpl23	Sptbn1	Txnl4a	
Cct2	Ddx56	Gsta4	Lasp1	Nup188	Ppwd1	Rpl30	Srek1	U2surp	
Cct3	Ddx6	Gtf2e1	Lmnb2	Nup214	Prc1	Rpl32	Srpk1	Uba1	
Cct4	Dgcr8	Gtf3c5	Lrrc40	Nup35	Prim2	Rprd1a	Srrm2	Ube4b	
Cct5	Dhx8	Hdac2	Lsm14a	Nup54	Prpf31	Rprd1b	Srsf10	Ubr5	
Cct6a	Dis3	Hdlbp	Luc7l	Nup62	Prpf4b	Rps15a	Srsf4	Ubxn1	
Cct7	Dlgap5	Herc4	Luc7l2	Nup85	Prpf6	Rrp12	Srsf9	Ubxn7	

References

Aapola, U., Shibuya, K., Scott, H.S., Ollila, J., Vihinen, M., Heino, M., Shintani, A., Kawasaki, K., Minoshima, S., Krohn, K., et al. (2000). Isolation and initial characterization of a novel zinc finger gene, DNMT3L, on 21q22.3, related to the cytosine-5-methyltransferase 3 gene family. Genomics *65*, 293–298. https://doi.org/10.1006/GENO.2000.6168.

Agger, K., Cloos, P.A.C., Christensen, J., Pasini, D., Rose, S., Rappsilber, J., Issaeva, I., Canaani, E., Salcini, A.E., and Helin, K. (2007). UTX and JMJD3 are histone H3K27 demethylases involved in HOX gene regulation and development. Nature *449*, 731–734. https://doi.org/10.1038/nature06145.

Alfert, A., Moreno, N., and Kerl, K. (2019). The BAF complex in development and disease. Epigenetics and Chromatin *12*, 1–15. https://doi.org/10.1186/s13072-019-0264-y.

Allan, J., Hartman, P.G., Crane-robinson, C., and Aviles, F.X. (1980). The structure of histone H1 and its location in chromatin. Nature *288*, 675–679. https://doi.org/10.1038/288675a0.

Allen, M.D., Grummitt, C.G., Hilcenko, C., Min, S.Y., Tonkin, L.M., Johnson, C.M., Freund, S.M., Bycroft, M., and Warren, A.J. (2006a). Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. The EMBO Journal *25*, 4503–4512. https://doi.org/10.1038/SJ.EMBOJ.7601340.

Allen, M.D., Grummitt, C.G., Hilcenko, C., Min, S.Y., Tonkin, L.M., Johnson, C.M., Freund, S.M., Bycroft, M., and Warren, A.J. (2006b). Solution structure of the nonmethyl-CpG-binding CXXC domain of the leukaemia-associated MLL histone methyltransferase. EMBO J *25*, 4503–4512. https://doi.org/10.1038/SJ.EMBOJ.7601340.

Aloia, L., di Stefano, B., and di Croce, L. (2013). Polycomb complexes in stem cells and embryonic development. Development (Cambridge) *140*, 2525–2534. https://doi.org/10.1242/dev.091553.

Amemiya, H.M., Kundaje, A., and Boyle, A.P. (2019). The ENCODE Blacklist: Identification of Problematic Regions of the Genome. Scientific Reports *9*, 1–5. https://doi.org/10.1038/s41598-019-45839-z.

An, J., Rao, A., and Ko, M. (2017). TET family dioxygenases and DNA demethylation in stem cells and cancers. Experimental and Molecular Medicine *49*, e323–e323. https://doi.org/10.1038/emm.2017.5.

Andersen, K.R., Leksa, N.C., and Schwartz, T.U. (2013). Optimized E. coli expression strain LOBSTR eliminates common contaminants from His-tag purification. Proteins: Structure, Function, and Bioinformatics *81*, 1857–1861. https://doi.org/10.1002/prot.24364.

Aranda, S., Alcaine-Colet, A., Blanco, E., Borràs, E., Caillot, C., Sabidó, E., and di Croce, L. (2019). Chromatin capture links the metabolic enzyme AHCY to stem cell proliferation. Science Advances *5*, eaav2448. https://doi.org/10.1126/sciadv.aav2448.

Ardehali, M.B., Mei, A., Zobeck, K.L., Caron, M., Lis, J.T., and Kusch, T. (2011). Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription. EMBO J *30*, 2817–2828. https://doi.org/10.1038/EMBOJ.2011.194.

Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y., and Shirakawa, M. (2008). Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature *455*, 818–821. https://doi.org/10.1038/nature07249.

Arrigoni, R., Alam, S.L., Wamstad, J.A., Bardwell, V.J., Sundquist, W.I., and Schreiber-Agus, N. (2006). The Polycomb-associated protein Rybp is a ubiquitin binding protein. FEBS Lett *580*, 6233–6241. https://doi.org/10.1016/J.FEBSLET.2006.10.027.

van Attikum, H., Fritsch, O., and Gasser, S.M. (2007). Distinct roles for SWR1 and INO80 chromatin remodeling complexes at chromosomal double-strand breaks. EMBO J *26*, 4113–4125. https://doi.org/10.1038/SJ.EMBOJ.7601835.

Avvakumov, G. v., Walker, J.R., Xue, S., Li, Y., Duan, S., Bronner, C., Arrowsmith, C.H., and Dhe-Paganon, S. (2008). Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. Nature *455*, 822–825. https://doi.org/10.1038/nature07273.

Ayton, P.M., Chen, E.H., and Cleary, M.L. (2004). Binding to Nonmethylated CpG DNA Is Essential for Target Recognition, Transactivation, and Myeloid Transformation by an MLL Oncoprotein. Molecular and Cellular Biology *24*, 10470–10478. .

Azuara, V., Perry, P., Sauer, S., Spivakov, M., Jørgensen, H.F., John, R.M., Gouti, M., Casanova, M., Warnes, G., Merkenschlager, M., et al. (2006). Chromatin signatures of pluripotent cell lines. Nature Cell Biology *8*, 532–538. https://doi.org/10.1038/ncb1403.

Babu, A., and Verma, R.S. (1987). Chromosome Structure: Euchromatin and Heterochromatin. International Review of Cytology *108*, 1–60. https://doi.org/10.1016/S0074-7696(08)61435-7.

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

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308. https://doi.org/10.1016/0092-8674(81)90413-X.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell *33*, 729–740. https://doi.org/10.1016/0092-8674(83)90015-6.

Bannister, A.J., Schneider, R., Myers, F.A., Thorne, A.W., Crane-Robinson, C., and Kouzarides, T. (2005). Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes. J Biol Chem *280*, 17732–17736. https://doi.org/10.1074/JBC.M500796200.

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S.C., Mann, M., and Kouzarides, T. (2010). Nucleosome-Interacting Proteins Regulated by DNA and Histone Methylation. Cell *143*, 470–484. https://doi.org/10.1016/j.cell.2010.10.012.

Bashtrykov, P., Jankevicius, G., Smarandache, A., Jurkowska, R.Z., Ragozin, S., and Jeltsch, A. (2012). Specificity of Dnmt1 for methylation of hemimethylated CpG sites resides in its catalytic domain. Chem Biol *19*, 572–578. https://doi.org/10.1016/J.CHEMBIOL.2012.03.010.

Bateman, A., Martin, M.J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E.H., Britto, R., Bursteinas, B., et al. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research *49*, D480–D489. https://doi.org/10.1093/NAR/GKAA1100.

Baubec, T., Colombo, D.F., Wirbelauer, C., Schmidt, J., Burger, L., Krebs, A.R., Akalin, A., and Schübeler, D. (2015). Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature 2015 520:7546 *520*, 243–247. https://doi.org/10.1038/nature14176.

Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. Genome Res *10*, 1001–1010. https://doi.org/10.1101/GR.10.7.1001.

Bell, J.S.K., and Vertino, P.M. (2017). Orphan CpG islands define a novel class of highly active enhancers. Epigenetics *12*, 449–464. https://doi.org/10.1080/15592294.2017.1297910.

Beltran, M., Yates, C.M., Skalska, L., Dawson, M., Reis, F.P., Viiri, K., Fisher, C.L., Sibley, C.R., Foster, B.M., Bartke, T., et al. (2016). The interaction of PRC2 with RNA or chromatin s mutually antagonistic. Genome Research *26*, 896–907. https://doi.org/10.1101/GR.197632.115/-/DC1.

Beltran, M., Tavares, M., Justin, N., Khandelwal, G., Ambrose, J., Foster, B.M., Worlock, K.B., Tvardovskiy, A., Kunzelmann, S., Herrero, J., et al. (2019). G-tract RNA removes Polycomb repressive complex 2 from genes. Nat Struct Mol Biol *26*, 899–909. https://doi.org/10.1038/S41594-019-0293-Z.

Beringer, M., Pisano, P., di Carlo, V., Blanco, E., Chammas, P., Vizán, P., Gutiérrez, A., Aranda, S., Payer, B., Wierer, M., et al. (2016). EPOP Functionally Links Elongin and Polycomb in Pluripotent Stem Cells. Molecular Cell *64*, 645–658. https://doi.org/10.1016/J.MOLCEL.2016.10.018.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006a). A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell *125*, 315–326. https://doi.org/10.1016/J.CELL.2006.02.041.

Bernstein, B.E., Meissner, A., and Lander, E.S. (2007). The Mammalian Epigenome. Cell *128*, 669–681. https://doi.org/10.1016/J.CELL.2007.01.033.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). The NIH roadmap epigenomics mapping consortium. Nature Biotechnology *28*, 1045–1048. https://doi.org/10.1038/nbt1010-1045.

Bernstein, E., Duncan, E.M., Masui, O., Gil, J., Heard, E., and Allis, C.D. (2006b). Mouse Polycomb Proteins Bind Differentially to Methylated Histone H3 and RNA and Are Enriched in Facultative Heterochromatin. Molecular and Cellular Biology *26*, 2560–2569. https://doi.org/10.1128/mcb.26.7.2560-2569.2006.

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

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

Bird, A.P. (1980). DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Research *8*, 1499–1504. https://doi.org/10.1093/NAR/8.7.1499.

Bird, A.P. (1987). CpG islands as gene markers in the vertebrate nucleus. Trends in Genetics *3*, 342–347. https://doi.org/10.1016/0168-9525(87)90294-0.

Bird, A.P., and Southern, E.M. (1978). Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from Xenopus laevis. Journal of Molecular Biology *118*, 27–47. https://doi.org/10.1016/0022-2836(78)90242-5.

Bird, A.P., and Wolffe, A.P. (1999). Methylation-Induced Repression— Belts, Braces, and Chromatin. Cell *99*, 451–454. https://doi.org/10.1016/S0092-8674(00)81532-9.

Bird, A., Taggart, M., Frommer, M., Miller, O.J., and Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell *40*, 91–99. https://doi.org/10.1016/0092-8674(85)90312-5.

Birger, Y., Shemer, R., Perk, J., and Razin, A. (1999). The imprinting box of the mouse Igf2r gene. Nature *397*, 84–86. https://doi.org/10.1038/16291.

Bitoun, E., Oliver, P.L., and Davies, K.E. (2007). The mixed-lineage leukemia fusion partner AF4 stimulates RNA polymerase II transcriptional elongation and mediates coordinated chromatin remodeling. Human Molecular Genetics *16*, 92–106. https://doi.org/10.1093/HMG/DDL444.

Blackledge, N.P., and Klose, R.J. (2021). The molecular principles of gene regulation by Polycomb repressive complexes. Nature Reviews Molecular Cell Biology *22*, 815–833. https://doi.org/10.1038/s41580-021-00398-y.

Blackledge, N.P., Zhou, J.C., Tolstorukov, M.Y., Farcas, A.M., Park, P.J., and Klose, R.J. (2010). CpG Islands Recruit a Histone H3 Lysine 36 Demethylase. Molecular Cell *38*, 179–190. https://doi.org/10.1016/j.molcel.2010.04.009.

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

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

Blackledge, N.P., Fursova, N.A., Kelley, J.R., Huseyin, M.K., Feldmann, A., and Klose, R.J. (2020). PRC1 Catalytic Activity Is Central to Polycomb System Function. Molecular Cell *77*, 857-874.e9. https://doi.org/10.1016/J.MOLCEL.2019.12.001.

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

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. https://doi.org/10.3390/cells7030017.

Bogdanović, O., Smits, A.H., de La Calle Mustienes, E., Tena, J.J., Ford, E., Williams, R., Senanayake, U., Schultz, M.D., Hontelez, S., van Kruijsbergen, I., et al. (2016). Active DNA demethylation at enhancers during the vertebrate phylotypic period. Nat Genet *48*, 417–426. https://doi.org/10.1038/NG.3522.

Bondarenko, V.A., Steele, L.M., Újvári, A., Gaykalova, D.A., Kulaeva, O.I., Polikanov, Y.S., Luse, D.S., and Studitsky, V.M. (2006). Nucleosomes Can Form a Polar Barrier to Transcript Elongation by RNA Polymerase II. Molecular Cell *24*, 469–479. https://doi.org/10.1016/J.MOLCEL.2006.09.009.

Borgel, J., Guibert, S., Li, Y., Chiba, H., Schübeler, D., Sasaki, H., Forné, T., and Weber, M. (2010). Targets and dynamics of promoter DNA methylation during early mouse development. Nature Genetics *42*, 1093–1100. https://doi.org/10.1038/ng.708.

Bostick, M., Jong, K.K., Estève, P.O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science (1979) *317*, 1760–1764. https://doi.org/10.1126/science.1147939.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., et al. (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349–353. https://doi.org/10.1038/NATURE04733.

Bracken, A.P., Dietrich, N., Pasini, D., Hansen, K.H., and Helin, K. (2006). Genome-wide mapping of Polycomb target genes unravels their roles in cell fate transitions. Genes Dev *20*, 1123–1136. https://doi.org/10.1101/GAD.381706.

Brand, M., Nakka, K., Zhu, J., and Dilworth, F.J. (2019). Polycomb/Trithorax Antagonism: Cellular Memory in Stem Cell Fate and Function. Cell Stem Cell *24*, 518–533. https://doi.org/10.1016/j.stem.2019.03.005.

Brien, G.L., Gambero, G., O'Connell, D.J., Jerman, E., Turner, S.A., Egan, C.M., Dunne, E.J., Jurgens, M.C., Wynne, K., Piao, L., et al. (2012). Polycomb PHF19 binds H3K36me3 and recruits PRC2 and demethylase NO66 to embryonic stem cell genes during differentiation. Nat Struct Mol Biol *19*, 1273–1281. https://doi.org/10.1038/NSMB.2449.

Brier, A.S.B., Loft, A., Madsen, J.G.S., Rosengren, T., Nielsen, R., Schmidt, S.F., Liu, Z., Yan, Q., Gronemeyer, H., and Mandrup, S. (2017). The KDM5 family is required for activation of pro-proliferative cell cycle genes during adipocyte differentiation. Nucleic Acids Research *45*, 1743–1759. https://doi.org/10.1093/NAR/GKW1156.

Brown, D.A., di Cerbo, V., Feldmann, A., Ahn, J., Ito, S., Blackledge, N.P., Nakayama, M., McClellan, M., Dimitrova, E., Turberfield, A.H., et al. (2017). The SET1 Complex Selects Actively Transcribed Target Genes via Multivalent Interaction with CpG Island Chromatin. Cell Reports *20*, 2313–2327. https://doi.org/10.1016/j.celrep.2017.08.030.

Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. Nature Reviews Genetics *21*, 630–644. https://doi.org/10.1038/s41576-020-0258-4.

Busslinger, M., Hurst, J., and Flavell, R.A. (1983). DNA methylation and the regulation of globin gene expression. Cell *34*, 197–206. https://doi.org/10.1016/0092-8674(83)90150-2.

Cao, R., and Zhang, Y. (2004). SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. Molecular Cell *15*, 57–67. https://doi.org/10.1016/j.molcel.2004.06.020.

Cao, K., Collings, C.K., Morgan, M.A., Marshall, S.A., Rendleman, E.J., Ozark, P.A., Smith, E.R., and Shilatifard, A. (2018). An Mll4/COMPASS-Lsd1 epigenetic axis governs enhancer function and pluripotency transition in embryonic stem cells. Sci Adv *4*. https://doi.org/10.1126/SCIADV.AAP8747.

Cao, R., Wang, L., Wang, H., Xia, L., Erdjument-Bromage, H., Tempst, P., Jones, R.S., and Zhang, Y. (2002). Role of histone H3 lysine 27 methylation in polycomb-group silencing. Science (1979) *298*, 1039–1043. https://doi.org/10.1126/science.1076997.

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

Čermaková, K., Tesina, P., Demeulemeester, J., el Ashkar, S., Méreau, H., Schwaller, J., Řezáčová, P., Veverka, V., and de Rijck, J. (2014). Validation and Structural Characterization of the LEDGF/p75–MLL Interface as a New Target for the Treatment of MLL-Dependent Leukemia. Cancer Research *74*, 5139–5151. https://doi.org/10.1158/0008-5472.CAN-13-3602.

Chalkley, G.E., and Verrijzer, C.P. (1999). DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. EMBO J *18*, 4835–4845. https://doi.org/10.1093/EMBOJ/18.17.4835.

Cheah, J.S., and Yamada, S. (2017). A simple elution strategy for biotinylated proteins bound to streptavidin conjugated beads using excess biotin and heat. Biochem Biophys Res Commun *493*, 1522–1527. https://doi.org/10.1016/J.BBRC.2017.09.168.

Chen, H., and Boutros, P.C. (2011). VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics *12*, 1–7. https://doi.org/10.1186/1471-2105-12-35/TABLES/1.

Chen, Q., Chen, Y., Bian, C., Fujiki, R., and Yu, X. (2013). TET2 promotes histone O-GlcNAcylation during gene transcription. Nature *493*, 561–564. https://doi.org/10.1038/NATURE11742.

Chen, S., Yang, Z., Wilkinson, A.W., Deshpande, A.J., Sidoli, S., Krajewski, K., Strahl, B.D., Garcia, B.A., Armstrong, S.A., Patel, D.J., et al. (2015). The PZP Domain of AF10 Senses Unmodified H3K27 to Regulate DOT1L-Mediated Methylation of H3K79. Molecular Cell *60*, 319–327. https://doi.org/10.1016/J.MOLCEL.2015.08.019.

Cheng, X., and Blumenthal, R.M. (2008). Mammalian DNA Methyltransferases: A Structural Perspective. Structure *16*, 341–350. https://doi.org/10.1016/J.STR.2008.01.004.

Choi, J., Bachmann, A.L., Tauscher, K., Benda, C., Fierz, B., and Müller, J. (2017). DNA binding by PHF1 prolongs PRC2 residence time on chromatin and thereby promotes H3K27 methylation. Nature Structural and Molecular Biology *24*, 1039–1047. https://doi.org/10.1038/nsmb.3488.

Choi, W.Y., Hwang, J.H., Cho, A.N., Lee, A.J., Lee, J., Jung, I., Cho, S.W., Kim, L.K., and Kim, Y.J. (2020). DNA Methylation of Intragenic CpG Islands are Required for Differentiation from iPSC to NPC. Stem Cell Reviews and Reports *16*, 1316–1327. .

Chow, C.M., Georgiou, A., Szutorisz, H., Maia e Silva, A., Pombo, A., Barahona, I., Dargelos, E., Canzonetta, C., and Dillon, N. (2005). Variant histone H3.3 marks promoters of transcriptionally active genes during mammalian cell division. EMBO Reports *6*, 354–360. https://doi.org/10.1038/sj.embor.7400366.

Christofori, G., and Keller, W. (1988). 3' cleavage and polyadenylation of mRNA precursors in vitro requires a poly(A) polymerase, a cleavage factor, and a snRNP. Cell *54*, 875–889. https://doi.org/10.1016/S0092-8674(88)91263-9.

Christoforou, A., Mulvey, C.M., Breckels, L.M., Geladaki, A., Hurrell, T., Hayward, P.C., Naake, T., Gatto, L., Viner, R., Arias, A.M., et al. (2016). A draft map of the mouse pluripotent stem cell spatial proteome. Nature Communications *7*, 9992. https://doi.org/10.1038/ncomms9992.

Chronis, C., Fiziev, P., Papp, B., Butz, S., Bonora, G., Sabri, S., Ernst, J., and Plath, K. (2017). Cooperative Binding of Transcription Factors Orchestrates Reprogramming. Cell *168*, 442-459.e20. https://doi.org/10.1016/J.CELL.2016.12.016.

Chuang, L.S.H., Ian, H.I., Koh, T.W., Ng, H.H., Xu, G., and Li, B.F.L. (1997). Human DNA-(cytosine-5) methyltransferase-PCNA complex as a target for p21(WAF1). Science (1979) *277*, 1996–2000. .

Cianfrocco, M.A., Kassavetis, G.A., Grob, P., Fang, J., Juven-Gershon, T., Kadonaga, J.T., and Nogales, E. (2013). Human TFIID Binds to Core Promoter DNA in a Reorganized Structural State. Cell *152*, 120–131. https://doi.org/10.1016/J.CELL.2012.12.005.

Cierpicki, T., Risner, L.E., Grembecka, J., Lukasik, S.M., Popovic, R., Omonkowska, M., Shultis, D.D., Zeleznik-Le, N.J., and Bushweller, J.H. (2010). Structure of the MLL CXXC domain–DNA complex and its functional role in MLL-AF9 leukemia. Nature Structural & Molecular Biology *17*, 62–68. https://doi.org/10.1038/nsmb.1714.

Clapier, C.R., Iwasa, J., Cairns, B.R., and Peterson, C.L. (2017). Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. Nature Reviews Molecular Cell Biology *18*, 407–422. https://doi.org/10.1038/nrm.2017.26.

Cloos, P.A.C., Christensen, J., Agger, K., Maiolica, A., Rappsilber, J., Antal, T., Hansen, K.H., and Helin, K. (2006). The putative oncogene GASC1 demethylates tri- and dimethylated lysine 9 on histone H3. Nature *442*, 307–311. https://doi.org/10.1038/nature04837.

Clouaire, T., Webb, S., Skene, P., Illingworth, R., Kerr, A., Andrews, R., Lee, J.H., Skalnik, D., and Bird, A. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. Genes Dev *26*, 1714–1728. https://doi.org/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 Biol *15*, 451. https://doi.org/10.1186/S13059-014-0451-X.

Cooper, D.N., Taggart, M.H., and Bird, A.P. (1983). Unmethylated domains in vertebrate DNA. Nucleic Acids Research *11*, 647–658. https://doi.org/10.1093/nar/11.3.647.

Cooper, S., Dienstbier, M., Hassan, R., Schermelleh, L., Sharif, J., Blackledge, N.P., DeMarco, V., Elderkin, S., Koseki, H., Klose, R., et al. (2014). Targeting Polycomb to Pericentric Heterochromatin in Embryonic Stem Cells Reveals a Role for H2AK119u1 in PRC2 Recruitment. Cell Reports *7*, 1456–1470. https://doi.org/10.1016/J.CELREP.2014.04.012.

Core, L., and Adelman, K. (2019). Promoter-proximal pausing of RNA polymerase II: a nexus of gene regulation. Genes & Development *33*, 960–982. https://doi.org/10.1101/GAD.325142.119.

Corona, D.F.V., Längst, G., Clapier, C.R., Bonte, E.J., Ferrari, S., Tamkun, J.W., and Becker, P.B. (1999). ISWI is an ATP-dependent nucleosome remodeling factor. Mol Cell *3*, 239–245. https://doi.org/10.1016/S1097-2765(00)80314-7.

Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. Nature *274*, 775–780. https://doi.org/10.1038/274775A0.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology *26*, 1367–1372. https://doi.org/10.1038/nbt.1511.

Cramer, P. (2019). Organization and regulation of gene transcription. Nature *573*, 45–54. https://doi.org/10.1038/s41586-019-1517-4.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931–21936. https://doi.org/10.1073/pnas.1016071107.

Danino, Y.M., Even, D., Ideses, D., and Juven-Gershon, T. (2015). The core promoter: At the heart of gene expression. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms *1849*, 1116–1131. https://doi.org/10.1016/J.BBAGRM.2015.04.003.

Dawlaty, M.M., Breiling, A., Le, T., Raddatz, G., Barrasa, M.I., Cheng, A.W., Gao, Q., Powell, B.E., Li, Z., Xu, M., et al. (2013). Combined Deficiency of Tet1 and Tet2 Causes Epigenetic Abnormalities but Is Compatible with Postnatal Development. Developmental Cell *24*, 310–323. https://doi.org/10.1016/J.DEVCEL.2012.12.015.

Deaton, A., and Bird, A. (2011). CpG islands and the regulation of transcription. Genes Dev *25*, 1010–1022. https://doi.org/10.1101/gad.2037511.1010.

Deaton, M., Webb, S., Kerr, A.R.W., Illingworth, R.S., Guy, J., Andrews, R., and Bird, A. (2011). Cell type – specific DNA methylation at intragenic CpG islands in the immune system. Genome Researh *21*, 1074–1086. https://doi.org/10.1101/gr.118703.110.Freely.

Dellino, G.I., Schwartz, Y.B., Farkas, G., McCabe, D., Elgin, S.C.R., and Pirrotta, V. (2004). Polycomb silencing blocks transcription initiation. Mol Cell *13*, 887–893. https://doi.org/10.1016/S1097-2765(04)00128-5.

Demers, C., Chaturvedi, C.P., Ranish, J.A., Juban, G., Lai, P., Morle, F., Aebersold, R., Dilworth, F.J., Groudine, M., and Brand, M. (2007). Activator-Mediated Recruitment of the MLL2 Methyltransferase Complex to the β-Globin Locus. Molecular Cell *27*, 573–584. https://doi.org/10.1016/J.MOLCEL.2007.06.022.

Deng, W., and Roberts, S.G.E. (2005a). A core promoter element downstream of the TATA box that is recognized by TFIIB. Genes & Development *19*, 2418. https://doi.org/10.1101/GAD.342405.

Deng, W., and Roberts, S.G.E. (2005b). A core promoter element downstream of the TATA box that is recognized by TFIIB. Genes Dev *19*, 2418–2423. https://doi.org/10.1101/GAD.342405.

Denholtz, M., Bonora, G., Chronis, C., Splinter, E., de Laat, W., Ernst, J., Pellegrini, M., and Plath, K. (2013). Long-Range Chromatin Contacts in Embryonic Stem Cells Reveal a Role for Pluripotency Factors and Polycomb Proteins in Genome Organization. Cell Stem Cell *13*, 602–616. https://doi.org/10.1016/J.STEM.2013.08.013.

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

Deplus, R., Delatte, B., Schwinn, M.K., Defrance, M., Méndez, J., Murphy, N., Dawson, M.A., Volkmar, M., Putmans, P., Calonne, E., et al. (2013). TET2 and TET3 regulate GlcNAcylation and H3K4 methylation through OGT and SET1/COMPASS. EMBO J *32*, 645–655. https://doi.org/10.1038/EMBOJ.2012.357.

DiTacchio, L., Le, H.D., Vollmers, C., Hatori, M., Witcher, M., Secombe, J., and Panda, S. (2011). Histone lysine demethylase JARID1a activates CLOCK-BMAL1 and influences the circadian clock. Science (1979) *333*, 1881–1885. https://doi.org/10.1126/science.1206022.

Dobrinić, P., Szczurek, A.T., and Klose, R.J. (2021). PRC1 drives Polycomb-mediated gene repression by controlling transcription initiation and burst frequency (Springer US).

Dodge, J.E., Ramsahoye, B.H., Wo, Z.G., Okano, M., and Li, E. (2002). De novo methylation of MMLV provirus in embryonic stem cells: CpG versus non-CpG methylation. Gene *289*, 41–48. https://doi.org/10.1016/S0378-1119(02)00469-9.

Doncheva, N.T., Morris, J.H., Gorodkin, J., and Jensen, L.J. (2019). Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. Journal of Proteome Research *18*, 623–632. https://doi.org/10.1021/acs.jproteome.8b00702.

Dorighi, K.M., Swigut, T., Henriques, T., Bhanu, N. v., Scruggs, B.S., Nady, N., Still, C.D., Garcia, B.A., Adelman, K., and Wysocka, J. (2017). Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. Molecular Cell *66*, 568-576.e4. https://doi.org/10.1016/J.MOLCEL.2017.04.018.

Duncan, B.K., and Miller, J.H. (1980). Mutagenic deamination of cytosine residues in DNA. Nature *287*, 560–561. https://doi.org/10.1038/287560A0.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74. https://doi.org/10.1038/nature11247.

Earnshaw, W.C., and Rothfield, N. (1985). Identification of a family of human centromere proteins using autoimmune sera from patients with scleroderma. Chromosoma *91*, 313–321. https://doi.org/10.1007/BF00328227.

Eberl, H.C., Mann, M., and Vermeulen, M. (2011). Quantitative Proteomics for Epigenetics. ChemBioChem *12*, 224–234. https://doi.org/10.1002/cbic.201000429.

Eberl, H.C., Spruijt, C.G., Kelstrup, C.D., Vermeulen, M., and Mann, M. (2013). A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-free Interaction Proteomics. Molecular Cell *49*, 368–378. https://doi.org/10.1016/j.molcel.2012.10.026.

Edelheit, O., Hanukoglu, A., and Hanukoglu, I. (2009). Simple and efficient site-directed mutagenesis using two single-primer reactions in parallel to generate mutants for protein structure-function studies. BMC Biotechnology *9*, 61. https://doi.org/10.1186/1472-6750-9-61.

Edmunds, J.W., Mahadevan, L.C., and Clayton, A.L. (2008). Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. EMBO J *27*, 406–420. https://doi.org/10.1038/SJ.EMBOJ.7601967.

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D.J., Dash, C., Bazett-Jones, D.P., le Grice, S., McKay, R.D.G., Buetow, K.H., et al. (2008). Global Transcription in Pluripotent Embryonic Stem Cells. Cell Stem Cell *2*, 437–447. https://doi.org/10.1016/J.STEM.2008.03.021.

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., Mccune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Research *10*, 2709–2721. https://doi.org/10.1093/NAR/10.8.2709.

Eick, D., and Geyer, M. (2013). The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code. Chemical Reviews *113*, 8456–8490. https://doi.org/10.1021/CR400071F.

Endoh, M., Endo, T.A., Endoh, T., Fujimura, Y.I., Ohara, O., Toyoda, T., Otte, A.P., Okano, M., Brockdorff, N., Vidal, M., et al. (2008). Polycomb group proteins Ring1A/B are functionally linked to the core transcriptional regulatory circuitry to maintain ES cell identity. Development *135*, 1513–1524. https://doi.org/10.1242/dev.014340.

Endoh, M., Endo, T.A., Shinga, J., Hayashi, K., Farcas, A., Ma, K.W., Ito, S., Sharif, J., Endoh, T., Onaga, N., et al. (2017). PCGF6-PRC1 suppresses premature differentiation of mouse embryonic stem cells by regulating germ cell-related genes. Elife *6*. https://doi.org/10.7554/ELIFE.21064.

Eng, J.K., McCormack, A.L., and Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom *5*, 976–989. https://doi.org/10.1016/1044-0305(94)80016-2.

Engelen, E., Brandsma, J.H., Moen, M.J., Signorile, L., Dekkers, D.H.W., Demmers, J., Kockx, C.E.M., Ozgür, Z., van IJcken, W.F.J., van den Berg, D.L.C., et al. (2015). Proteins that bind regulatory regions identified by histone modification chromatin immunoprecipitations and mass spectrometry. Nature Communications *6*, 7155. https://doi.org/10.1038/ncomms8155.

Epsztejn-Litman, S., Feldman, N., Abu-Remaileh, M., Shufaro, Y., Gerson, A., Ueda, J., Deplus, R., Fuks, F., Shinkai, Y., Cedar, H., et al. (2008). De novo DNA methylation promoted by G9a prevents reprogramming of embryonically silenced genes. Nature Structural and Molecular Biology *15*, 1176–1183. https://doi.org/10.1038/nsmb.1476.

Ernst, P., and Vakoc, C.R. (2012). WRAD: enabler of the SET1-family of H3K4 methyltransferases. Brief Funct Genomics *11*, 217–226. https://doi.org/10.1093/BFGP/ELS017.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49. https://doi.org/10.1038/nature09906.

Ernst, P., Mabon, M., Davidson, A.J., Zon, L.I., and Korsmeyer, S.J. (2004). An Mll-dependent Hox program drives hematopoietic progenitor expansion. Current Biology *14*, 2063–2069. https://doi.org/10.1016/j.cub.2004.11.012.

Eskeland, R., Leeb, M., Grimes, G.R., Kress, C., Boyle, S., Sproul, D., Gilbert, N., Fan, Y., Skoultchi, A.I., Wutz, A., et al. (2010). Ring1B Compacts Chromatin Structure and Represses Gene Expression Independent of Histone Ubiquitination. Molecular Cell *38*, 452–464. https://doi.org/10.1016/J.MOLCEL.2010.02.032.

Esnault, C., Ghavi-Helm, Y., Brun, S., Soutourina, J., van Berkum, N., Boschiero, C., Holstege, F., and Werner, M. (2008). Mediator-dependent recruitment of TFIIH modules in preinitiation complex. Mol Cell *31*, 337–346. https://doi.org/10.1016/J.MOLCEL.2008.06.021.

Espejo, I., di Croce, L., and Aranda, S. (2020). The changing chromatome as a driver of disease: A panoramic view from different methodologies. BioEssays *42*, 2000203. https://doi.org/10.1002/bies.202000203.

Farcas, A.M., Blackledge, N.P., Sudbery, I., Long, H.K., McGouran, J.F., Rose, N.R., Lee, S., Sims, D., Cerase, A., Sheahan, T.W., et al. (2012). KDM2B links the polycomb repressive complex 1 (PRC1) to recognition of CpG islands. Elife *2012*, 1–26. https://doi.org/10.7554/eLife.00205.

Farthing, C.R., Ficz, G., Ng, R.K., Chan, C.F., Andrews, S., Dean, W., Hemberger, M., and Reik, W. (2008). Global Mapping of DNA Methylation in Mouse Promoters Reveals Epigenetic Reprogramming of Pluripotency Genes. PLOS Genetics *4*, e1000116. https://doi.org/10.1371/JOURNAL.PGEN.1000116.

Feaver, W.J., Gileadi, O., Li, Y., and Kornberg, R.D. (1991). CTD kinase associated with yeast RNA polymerase II initiation factor b. Cell *67*, 1223–1230. https://doi.org/10.1016/0092-8674(91)90298-D.

Fenn, J.B., Mann, M., Meng, C.K., Wong, S.F., and Whitehouse, C.M. (1989). Electrospray Ionization for Mass Spectrometry of Large Biomolecules. Science (1979) *246*, 64–71. https://doi.org/10.1126/SCIENCE.2675315.

Festuccia, N., Owens, N., Papadopoulou, T., Gonzalez, I., Tachtsidi, A., Vandoermel-Pournin, S., Gallego, E., Gutierrez, N., Dubois, A., Cohen-Tannoudji, M., et al. (2019). Transcription factor activity and nucleosome organization in mitosis. Genome Research *29*, 250–260. https://doi.org/10.1101/GR.243048.118/-/DC1.

Ficz, G., Branco, M.R., Seisenberger, S., Santos, F., Krueger, F., Hore, T.A., Marques, C.J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. Nature *473*, 398–404. https://doi.org/10.1038/nature10008.

Flanagan, J.F., Mi, L.Z., Chruszcz, M., Cymborowski, M., Clines, K.L., Kim, Y., Minor, W., Rastinejad, F., and Khorasanizadeh, S. (2005). Double chromodomains cooperate to recognize the methylated histone H3 tail. Nature *438*, 1181–1185. https://doi.org/10.1038/nature04290.

Fodor, B.D., Kubicek, S., Yonezawa, M., O'Sullivan, R.J., Sengupta, R., Perez-Burgos, L., Opravil, S., Mechtler, K., Schotta, G., and Jenuwein, T. (2006). Jmjd2b antagonizes H3K9 trimethylation at pericentric heterochromatin in mammalian cells. Genes & Development *20*, 1557–1562. https://doi.org/10.1101/GAD.388206.

Francis, N.J., Kingston, R.E., and Woodcock, C.L. (2004). Chromatin compaction by a polycomb group protein complex. Science (1979) *306*, 1574–1577. https://doi.org/10.1126/science.1100576.

Frauer, C., Rottach, A., Meilinger, D., Bultmann, S., Fellinger, K., Hasenöder, S., Wang, M., Qin, W., Söding, J., Spada, F., et al. (2011). Different Binding Properties and Function of CXXC Zinc Finger Domains in Dnmt1 and Tet1. PLOS ONE *6*, e16627. https://doi.org/10.1371/JOURNAL.PONE.0016627.

Fuchs, G., Hollander, D., Voichek, Y., Ast, G., and Oren, M. (2014). Cotranscriptional histone H2B monoubiquitylation is tightly coupled with RNA polymerase II elongation rate. Genome Research *24*, 1572–1583. https://doi.org/10.1101/GR.176487.114.

Fursova, N.A., Blackledge, N.P., Nakayama, M., Ito, S., Koseki, Y., Farcas, A.M., King, H.W., Koseki, H., and Klose, R.J. (2019). Synergy between Variant PRC1 Complexes Defines Polycomb-Mediated Gene Repression. Molecular Cell *74*, 1020-1036.e8. https://doi.org/10.1016/j.molcel.2019.03.024.

Fyodorov, D. v., and Kadonaga, J.T. (2002). Dynamics of ATP-dependent chromatin assembly by ACF. Nature *418*, 896–900. https://doi.org/10.1038/NATURE00929.

Gao, X., Wang, H., Yang, J.J., Liu, X., and Liu, Z.R. (2012a). Pyruvate Kinase M2 Regulates Gene Transcription by Acting as a Protein Kinase. Molecular Cell *45*, 598–609. https://doi.org/10.1016/J.MOLCEL.2012.01.001.

Gao, Z., Zhang, J., Bonasio, R., Strino, F., Sawai, A., Parisi, F., Kluger, Y., and Reinberg, D. (2012b). PCGF Homologs, CBX Proteins, and RYBP Define Functionally Distinct PRC1 Family Complexes. Molecular Cell *45*, 344–356. https://doi.org/10.1016/j.molcel.2012.01.002.

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

Gaspar-Maia, A., Alajem, A., Polesso, F., Sridharan, R., Mason, M.J., Heidersbach, A., Ramalho-Santos, J., McManus, M.T., Plath, K., Meshorer, E., et al. (2009). Chd1 regulates open chromatin and pluripotency of embryonic stem cells. Nature *460*, 863–868. https://doi.org/10.1038/nature08212.

Gatchalian, J., Malik, S., Ho, J., Lee, D.S., Kelso, T.W.R., Shokhirev, M.N., Dixon, J.R., and Hargreaves, D.C. (2018). A non-canonical BRD9-containing BAF chromatin remodeling complex regulates naive pluripotency in mouse embryonic stem cells. Nat Commun *9*. https://doi.org/10.1038/S41467-018-07528-9.

Gaubatz, J.W., and Cutler, R.G. (1978). Age-Related Differences in the Number of Ribosomal RNA Genes of Mouse Tissues. Gerontology *24*, 179–207. https://doi.org/10.1159/000212250.

Gauchier, M., van Mierlo, G., Vermeulen, M., and Déjardin, J. (2020). Purification and enrichment of specific chromatin loci. Nature Methods *17*, 380–389. https://doi.org/10.1038/s41592-020-0765-4.

Geiger, J.H., Hahn, S., Lee, S., and Sigler, P.B. (1996). Crystal structure of the yeast TFIIA/TBP/DNA complex. Science (1979) *272*, 830–836. https://doi.org/10.1126/science.272.5263.830.

Gibbons, B.J., Brignole, E.J., Azubel, M., Murakami, K., Voss, N.R., Bushnell, D.A., Asturias, F.J., and Kornberg, R.D. (2012). Subunit architecture of general transcription factor TFIIH. Proc Natl Acad Sci U S A *109*, 1949–1954. https://doi.org/10.1073/pnas.1105266109.

Gill, R.M., Gabor, T. v., Couzens, A.L., and Scheid, M.P. (2013). The MYC-associated protein CDCA7 is phosphorylated by AKT to regulate MYC-dependent apoptosis and transformation. Mol Cell Biol *33*, 498–513. https://doi.org/10.1128/MCB.00276-12.

Gillet, L.C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., and Aebersold, R. (2012). Targeted Data Extraction of the MS/MS Spectra Generated by Data-independent Acquisition: A New Concept for Consistent and Accurate Proteome Analysis. Molecular & Cellular Proteomics *11*, O111.016717. https://doi.org/10.1074/MCP.O111.016717.

Gilmartin, G.M., and Nevins, J.R. (1989). An ordered pathway of assembly of components required for polyadenylation site recognition and processing. Genes Dev *3*, 2180–2190. https://doi.org/10.1101/GAD.3.12B.2180.

Gingras, A.-C., Abe, K.T., and Raught, B. (2019). Getting to know the neighborhood: using proximity-dependent biotinylation to characterize protein complexes and map organelles. Current Opinion in Chemical Biology *48*, 44–54. https://doi.org/10.1016/J.CBPA.2018.10.017.

Ginno, P.A., Burger, L., Seebacher, J., Iesmantavicius, V., and Schübeler, D. (2018). Cell cycle-resolved chromatin proteomics reveals the extent of mitotic preservation of the genomic regulatory landscape. Nature Communications *9*, 4048. https://doi.org/10.1038/s41467-018-06007-5.

Glaser, S., Schaft, J., Lubitz, S., Vintersten, K., van der Hoeven, F., Tufftteland, K.R., Aasland, R., Anastassiadis, K., Ang, S.L., and Stewart, A.F. (2006). Multiple epigenetic maintenance factors implicated by the loss of Mll2 in mouse development. Development *133*, 1423–1432. https://doi.org/10.1242/DEV.02302.

Goldberg, M. (1979). Sequence analysis of Drosophila histone genes.

Goodwin, L.R., and Picketts, D.J. (2018). The role of ISWI chromatin remodeling complexes in brain development and neurodevelopmental disorders. Molecular and Cellular Neuroscience *87*, 55–64. https://doi.org/10.1016/J.MCN.2017.10.008.

Goto, Y., Hayashi, R., Muramatsu, T., Ogawa, H., Eguchi, I., Oshida, Y., Ohtani, K., and Yoshida, K. (2006). JPO1/CDCA7, a novel transcription factor E2F1-induced protein, possesses intrinsic transcriptional regulator activity. Biochimica et Biophysica Acta - Gene Structure and Expression *1759*, 60–68. https://doi.org/10.1016/j.bbaexp.2006.02.004.

Graham, J.M. (2001). Isolation of Nuclei and Nuclear Membranes From Animal Tissues. Current Protocols in Cell Biology *12*, 3.10.1-3.10.19. https://doi.org/10.1002/0471143030.cb0310s12.

Grand, R.S., Burger, L., Gräwe, C., Michael, A.K., Isbel, L., Hess, D., Hoerner, L., Iesmantavicius, V., Durdu, S., Pregnolato, M., et al. (2021). BANP opens chromatin and activates CpG-island-regulated genes. Nature *596*, 133–137. https://doi.org/10.1038/s41586-021-03689-8.

Grau, D.J., Chapman, B.A., Garlick, J.D., Borowsky, M., Francis, N.J., and Kingston, R.E. (2011). Compaction of chromatin by diverse Polycomb group proteins requires localized regions of high charge. Genes & Development *25*, 2210–2221. https://doi.org/10.1101/GAD.17288211.

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

Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. Nature *389*, 349–352. https://doi.org/10.1038/38664.

Guan, S., Taylor, P.P., Han, Z., Moran, M.F., and Ma, B. (2020). Data Dependent-Independent Acquisition (DDIA) Proteomics. Journal of Proteome Research *19*, 3230–3237. https://doi.org/10.1021/acs.jproteome.0c00186.

Guenther, M.G., Jenner, R.G., Chevalier, B., Nakamura, T., Croce, C.M., Canaani, E., and Young, R.A. (2005). Global and Hox-specific roles for the MLL1 methyltransferase. Proc Natl Acad Sci U S A *102*, 8603–8608. https://doi.org/10.1073/pnas.0503072102.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77–88. https://doi.org/10.1016/J.CELL.2007.05.042.

Guo, X., Wang, L., Li, J., Ding, Z., Xiao, J., Yin, X., He, S., Shi, P., Dong, L., Li, G., et al. (2015). Structural insight into autoinhibition and histone H3-induced activation of DNMT3A. Nature *517*, 640–644. https://doi.org/10.1038/nature13899.

Guzmán, E., and Lis, J.T. (1999). Transcription factor TFIIH is required for promoter melting in vivo. Mol Cell Biol *19*, 5652–5658. https://doi.org/10.1128/MCB.19.8.5652.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999). Correlation between Protein and mRNA Abundance in Yeast. Molecular and Cellular Biology *19*, 1720–1730. https://doi.org/10.1128/mcb.19.3.1720.

Ha, I., Roberts, S., Maldonado, E., Sun, X., Kim, L.U., Green, M., and Reinberg, D. (1993). Multiple functional domains of human transcription factor IIB: distinct interactions with two general transcription factors and RNA polymerase II. Genes Dev *7*, 1021–1032. https://doi.org/10.1101/GAD.7.6.1021.

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

Hajkova, P., Erhardt, S., Lane, N., Haaf, T., El-Maarri, O., Reik, W., Walter, J., and Surani, M.A. (2002). Epigenetic reprogramming in mouse primordial germ cells. Mechanisms of Development *117*, 15–23. https://doi.org/10.1016/S0925-4773(02)00181-8.

Hanson, R.D., Hess, J.L., Yu, B.D., Ernst, P., van Lohuizen, M., Berns, A., van der Lugt, N.M.T., Shashikant, C.S., Ruddle, F.H., Seto, M., et al. (1999). Mammalian Trithorax and Polycomb-group homologues are antagonistic regulators of homeotic development. Proc Natl Acad Sci U S A *96*, 14372–14377. https://doi.org/10.1073/pnas.96.25.14372.

Hata, K., Okano, M., Lei, H., and Li, E. (2002). Dnmt3L cooperates with the Dnmt3 family of de novo DNA methyltransferases to establish maternal imprints in mice. Development *129*, 1983–1993. https://doi.org/10.1242/dev.129.8.1983.

Hatzi, K., Jiang, Y., Huang, C., Garrett-Bakelman, F., Gearhart, M.D., Giannopoulou, E.G., Zumbo, P., Kirouac, K., Bhaskara, S., Polo, J.M., et al. (2013). A Hybrid Mechanism of Action for BCL6 in B Cells Defined by Formation of Functionally Distinct Complexes at Enhancers and Promoters. Cell Reports *4*, 578–588. https://doi.org/10.1016/J.CELREP.2013.06.016.

He, J., Kallin, E.M., Tsukada, Y.I., and Zhang, Y. (2008). The H3K36 demethylase Jhdm1b/Kdm2b regulates cell proliferation and senescence through p15(Ink4b). Nat Struct Mol Biol *15*, 1169–1175. https://doi.org/10.1038/NSMB.1499.

He, Y., Fang, J., Taatjes, D.J., and Nogales, E. (2013). Structural visualization of key steps in human transcription initiation. Nature *495*, 481–486. https://doi.org/10.1038/NATURE11991.

He, Y.F., Li, B.Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, N., et al. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. Science (1979) *333*, 1303–1307. https://doi.org/10.1126/science.1210944.

van Heeringen, S.J., and Veenstra, G.J.C. (2011). GimmeMotifs: A de novo motif prediction pipeline for ChIP-sequencing experiments. Bioinformatics *27*, 270–271. https://doi.org/10.1093/bioinformatics/btq636.

Hendrix, J., van Heertum, B., Vanstreels, E., Daelemans, D., and de Rijck, J. (2014). Dynamics of the Ternary Complex Formed by c-Myc Interactor JPO2, Transcriptional Co-activator LEDGF/p75, and Chromatin. Journal of Biological Chemistry *289*, 12494–12506. https://doi.org/10.1074/JBC.M113.525964.

Herz, H.M., Garruss, A., and Shilatifard, A. (2013). SET for life: biochemical activities and biological functions of SET domain-containing proteins. Trends Biochem Sci *38*, 621–639. https://doi.org/10.1016/J.TIBS.2013.09.004.

Herzing, L.B.K., Romer, J.T., Horn, J.M., and Ashworth, A. (1997). Xist has properties of the X-chromosome inactivation centre. Nature *386*, 272–275. https://doi.org/10.1038/386272A0.

Ho, L., and Crabtree, G.R. (2010). Chromatin remodelling during development. Nature *463*, 474–484. https://doi.org/10.1038/nature08911.

Hochheimer, A., and Tjian, R. (2003). Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. Genes & Development *17*, 1309–1320. https://doi.org/10.1101/GAD.1099903.

Hodges, H.C., Stanton, B.Z., Cermakova, K., Chang, C.Y., Miller, E.L., Kirkland, J.G., Ku, W.L., Veverka, V., Zhao, K., and Crabtree, G.R. (2018). Dominant-negative SMARCA4 mutants alter the accessibility landscape of tissue-unrestricted enhancers. Nature Structural and Molecular Biology *25*, 61–72. https://doi.org/10.1038/s41594-017-0007-3.

Hoffman, E.A., Frey, B.L., Smith, L.M., and Auble, D.T. (2015). Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes. Journal of Biological Chemistry *290*, 26404–26411. https://doi.org/10.1074/jbc.R115.651679.

Hoffmeister, H., Fuchs, A., Erdel, F., Pinz, S., Gröbner-Ferreira, R., Bruckmann, A., Deutzmann, R., Schwartz, U., Maldonado, R., Huber, C., et al. (2017). CHD3 and CHD4 form distinct NuRD complexes with different yet overlapping functionality. Nucleic Acids Res *45*, 10534–10554. https://doi.org/10.1093/NAR/GKX711.

Holdeman, R., Nehrt, S., and Strome, S. (1998). MES-2, a maternal protein essential for viability of the germline in Caenorhabditis elegans, is homologous to a Drosophila Polycomb group protein. Development *125*, 2457–2467. https://doi.org/10.1242/DEV.125.13.2457.

Holliday, R., and Pugh, J.E. (1975). DNA Modification Mechanisms and Gene Activity During Development. Science (1979) *187*, 226–232. https://doi.org/10.1126/SCIENCE.187.4173.226.

Holmberg, A., Blomstergren, A., Nord, O., Lukacs, M., Lundeberg, J., and Uhlén, M. (2005). The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. ELECTROPHORESIS *26*, 501–510. https://doi.org/10.1002/elps.200410070.

Holstege, F.C.P., van der Vliet, P.C., and Timmers, H.T.M. (1996). Opening of an RNA polymerase II promoter occurs in two distinct steps and requires the basal transcription factors IIE and IIH. The EMBO Journal *15*, 1666–1677. https://doi.org/10.1002/J.1460-2075.1996.TB00512.X.

Hong, S.H., Cho, Y.W., Yu, L.R., Yu, H., Veenstra, T.D., and Ge, K. (2007). Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. Proceedings of the National Academy of Sciences *104*, 18439–18444. https://doi.org/10.1073/PNAS.0707292104.

Horii, T., and Hatada, I. (2016). Regulation of CpG methylation by Dnmt and Tet in pluripotent stem cells. The Journal of Reproduction and Development *62*, 331. https://doi.org/10.1262/JRD.2016-046.

Horton, J.R., Upadhyay, A.K., Qi, H.H., Zhang, X., Shi, Y., and Cheng, X. (2010). Enzymatic and structural insights for substrate specificity of a family of jumonji histone lysine demethylases. Nature Structural and Molecular Biology *17*, 38–44. https://doi.org/10.1038/nsmb.1753.

Hosogane, M., Funayama, R., Nishida, Y., Nagashima, T., and Nakayama, K. (2013). Ras-induced changes in H3K27me3 occur after those in transcriptional activity. PLoS Genet *9*. https://doi.org/10.1371/JOURNAL.PGEN.1003698.

Hotchkiss, R.D. (1948). The quantitative separation of purines, pyrimidines, and nucleosides by paper chromatography. Journal of Biological Chemistry *175*, 315–332. https://doi.org/10.1016/S0021-9258(18)57261-6.

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

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

Hu, D., Gao, X., Cao, K., Morgan, M.A., Mas, G., Smith, E.R., Volk, A.G., Bartom, E.T., Crispino, J.D., di Croce, L., et al. (2017). Not All H3K4 Methylations Are Created Equal: Mll2/COMPASS Dependency in Primordial Germ Cell Specification. Mol Cell *65*, 460-475.e6. https://doi.org/10.1016/J.MOLCEL.2017.01.013.

Hu, J., Lutz, C.S., Wilusz, J., and Tian, B. (2005). Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. RNA *11*, 1485–1493. https://doi.org/10.1261/rna.2107305.

Hu, L., Li, Z., Cheng, J., Rao, Q., Gong, W., Liu, M., Shi, Y.G., Zhu, J., Wang, P., and Xu, Y. (2013c). Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation. Cell *155*, 1545–1555. https://doi.org/10.1016/J.CELL.2013.11.020.

Hu, L., Lu, J., Cheng, J., Rao, Q., Li, Z., Hou, H., Lou, Z., Zhang, L., Li, W., Gong, W., et al. (2015). Structural insight into substrate preference for TET-mediated oxidation. Nature *527*, 118–122. https://doi.org/10.1038/nature15713.

Huang, A., Ho, C.S.W., Ponzielli, R., Barsyte-Lovejoy, D., Bouffet, E., Picard, D., Hawkins, C.E., and Penn, L.Z. (2005). Identification of a novel c-Myc protein interactor, JPO2, with transforming activity in medulloblastoma cells. Cancer Res *65*, 5607–5619. https://doi.org/10.1158/0008-5472.CAN-05-0500.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res *37*, 1–13. https://doi.org/10.1093/NAR/GKN923.

Huang, H., Sabari, B.R., Garcia, B.A., David Allis, C., and Zhao, Y. (2014). SnapShot: Histone modifications. Cell *159*, 458-458.e1. https://doi.org/10.1016/j.cell.2014.09.037.

Huang, J., Gurung, B., Wan, B., Matkar, S., Veniaminova, N.A., Wan, K., Merchant, J.L., Hua, X., and Lei, M. (2012). The same pocket in menin binds both MLL and JUND but has opposite effects on transcription. Nature 2012 482:7386 *482*, 542–546. https://doi.org/10.1038/nature10806.

Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., et al. (2008). A novel CpG island set identifies tissue-specific methylation at developmental gene loci. PLoS Biology *6*, 0037–0051. https://doi.org/10.1371/journal.pbio.0060022.

Illingworth, R.S., Gruenewald-Schneider, U., Webb, S., Kerr, A.R.W., James, K.D., Turner, D.J., Smith, C., Harrison, D.J., Andrews, R., and Bird, A.P. (2010). Orphan CpG Islands Identify Numerous Conserved Promoters in the Mammalian Genome. PLoS Genetics *6*, e1001134. https://doi.org/10.1371/journal.pgen.1001134.

Ingham, P.W. (1983). Differential expression of bithorax complex genes in the absence of the extra sex combs and trithorax genes. Nature *306*, 591–593. https://doi.org/10.1038/306591a0.

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*, 607–614. https://doi.org/10.1007/BF00271751.

Issaeva, I., Zonis, Y., Rozovskaia, T., Orlovsky, K., Croce, C.M., Nakamura, T., Mazo, A., Eisenbach, L., and Canaani, E. (2007). Knockdown of ALR (MLL2) Reveals ALR Target Genes and Leads to Alterations in Cell Adhesion and Growth. Molecular and Cellular Biology *27*, 1889–1903. https://doi.org/10.1128/mcb.01506-06.

Ito, S., Dalessio, A.C., Taranova, O. v., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133. https://doi.org/10.1038/NATURE09303.

Ito, S., Shen, L., Dai, Q., Wu, S.C., Collins, L.B., Swenberg, J.A., He, C., and Zhang, Y. (2011). Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science (1979) *333*, 1300–1303. https://doi.org/10.1126/science.1210597.

Izaguirre-Carbonell, J., Christiansen, L., Burns, R., Schmitz, J., Li, C., Mokry, R.L., Bluemn, T., Zheng, Y., Shen, J., Carlson, K.S., et al. (2019). Critical role of Jumonji domain of JMJD1C in MLL-rearranged leukemia. Blood Advances *3*, 1499–1511. https://doi.org/10.1182/BLOODADVANCES.2018026054.

Jacobson, R.H., Ladurner, A.G., King, D.S., and Tjian, R. (2000). Structure and function of a human TAF(II)250 double bromodomain module. Science (1979) *288*, 1422–1425. https://doi.org/10.1126/science.288.5470.1422.

Jain, B.P., and Pandey, S. (2018). WD40 Repeat Proteins: Signalling Scaffold with Diverse Functions. The Protein Journal *37*, 391–406. https://doi.org/10.1007/S10930-018-9785-7.

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. Nucleic Acids Res *48*, D498–D503. https://doi.org/10.1093/NAR/GKZ1031.

Jelinic, P., Pellegrino, J., and David, G. (2011). A Novel Mammalian Complex Containing Sin3B Mitigates Histone Acetylation and RNA Polymerase II Progression within Transcribed Loci. Molecular and Cellular Biology *31*, 54–62. https://doi.org/10.1128/mcb.00840-10.

Jenness, C., Giunta, S., Müller, M.M., Kimura, H., Muir, T.W., and Funabiki, H. (2018). HELLS and CDCA7 comprise a bipartite nucleosome remodeling complex defective in ICF syndrome. Proc Natl Acad Sci U S A *115*, E876–E885. https://doi.org/10.1073/PNAS.1717509115.

Jermann, P., Hoerner, L., Burger, L., and Schübeler, D. (2014). Short sequences can efficiently recruit histone H3 lysine 27 trimethylation in the absence of enhancer activity and DNA methylation. Proc Natl Acad Sci U S A *111*. https://doi.org/10.1073/PNAS.1400672111.

Jeziorska, D.M., Murray, R.J.S., de Gobbi, M., Gaentzsch, R., Garrick, D., Ayyub, H., Chen, T., Li, E., Telenius, J., Lynch, M., et al. (2017). DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. Proceedings of the National Academy of Sciences *114*, E7526–E7535. https://doi.org/10.1073/pnas.1703087114.

Ji, X., Dadon, D.B., Abraham, B.J., Lee, T.I., Jaenisch, R., Bradner, J.E., and Young, R.A. (2015). Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions. Proceedings of the National Academy of Sciences *112*, 3841–3846. https://doi.org/10.1073/pnas.1502971112.

Jia, D., Jurkowska, R.Z., Zhang, X., Jeltsch, A., and Cheng, X. (2007). Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. Nature *449*, 248–251. https://doi.org/10.1038/nature06146.

Jiao, L., and Liu, X. (2015). Structural basis of histone H3K27 trimethylation by an active polycomb repressive complex 2. Science (1979) *350*. https://doi.org/10.1126/science.aac4383.

Jin, C., and Felsenfeld, G. (2007). Nucleosome stability mediated by histone variants H3.3 and H2A.Z. Genes and Development *21*, 1519–1529. https://doi.org/10.1101/gad.1547707.

Jin, B., Ernst, J., Tiedemann, R.L., Xu, H., Sureshchandra, S., Kellis, M., Dalton, S., Liu, C., Choi, J.H., and Robertson, K.D. (2012). Linking DNA Methyltransferases to Epigenetic Marks and Nucleosome Structure Genome-wide in Human Tumor Cells. Cell Reports *2*, 1411–1424. https://doi.org/10.1016/J.CELREP.2012.10.017.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., and Felsenfeld, G. (2009). H3.3/H2A.Z double variant-containing nucleosomes mark “nucleosome-free regions” of active promoters and other regulatory regions. Nature Genetics *41*, 941–945. https://doi.org/10.1038/ng.409.

Josling, G.A., Selvarajah, S.A., Petter, M., and Duffy, M.F. (2012). The role of bromodomain proteins in regulating gene expression. Genes (Basel) *3*, 320–343. https://doi.org/10.3390/GENES3020320.

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

Jungblut, A., Hopfner, K.P., and Eustermann, S. (2020). Megadalton chromatin remodelers: common principles for versatile functions. Current Opinion in Structural Biology *64*, 134–144. https://doi.org/10.1016/J.SBI.2020.06.024.

Kadoch, C., Williams, R.T., Calarco, J.P., Miller, E.L., Weber, C.M., Braun, S.M.G., Pulice, J.L., Chory, E.J., and Crabtree, G.R. (2017). Dynamics of BAF-Polycomb complex opposition on heterochromatin in normal and oncogenic states. Nat Genet *49*, 213–222. https://doi.org/10.1038/NG.3734.

Kalb, R., Latwiel, S., Baymaz, H.I., Jansen, P.W.T.C., Müller, C.W., Vermeulen, M., and Müller, J. (2014). Histone H2A monoubiquitination promotes histone H3 methylation in Polycomb repression. Nat Struct Mol Biol *21*, 569–571. https://doi.org/10.1038/NSMB.2833.

Kaneko, S., Son, J., Shen, S.S., Reinberg, D., and Bonasio, R. (2013). PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. Nat Struct Mol Biol *20*, 1258–1264. https://doi.org/10.1038/NSMB.2700.

Kasinath, V., Beck, C., Sauer, P., Poepsel, S., Kosmatka, J., Faini, M., Toso, D., Aebersold, R., and Nogales, E. (2021). JARID2 and AEBP2 regulate PRC2 in the presence of H2AK119ub1 and other histone modifications. Science (1979) *371*. https://doi.org/10.1126/science.abc3393.

Kennedy-Darling, J., and Smith, L.M. (2014). Measuring the Formaldehyde Protein–DNA Cross-Link Reversal Rate. Analytical Chemistry *86*, 5678–5681. https://doi.org/10.1021/ac501354y.

Kim, D.I., and Roux, K.J. (2016). Filling the Void: Proximity-Based Labeling of Proteins in Living Cells. Trends in Cell Biology *26*, 804–817. https://doi.org/10.1016/j.tcb.2016.09.004.

Kim, Y., Geiger, J.H., Hahn, S., and Sigler, P.B. (1993). Crystal structure of a yeast TBP/TATA-box complex. Nature *365*, 512–520. https://doi.org/10.1038/365512a0.

King, A.D., Huang, K., Rubbi, L., Liu, S., Wang, C.-Y., Wang, Y., Pellegrini, M., and Fan, G. (2016). Reversible Regulation of Promoter and Enhancer Histone Landscape by DNA Methylation in Mouse Embryonic Stem Cells. Cell Reports *17*, 289–302. https://doi.org/10.1016/j.celrep.2016.08.083.

King, H.W., Fursova, N.A., Blackledge, N.P., and Klose, R.J. (2018). Polycomb repressive complex 1 shapes the nucleosome landscape but not accessibility at target genes. Genome Research *28*, 1494–1507. https://doi.org/10.1101/gr.237180.118.

Kireeva, M.L., Hancock, B., Cremona, G.H., Walter, W., Studitsky, V.M., and Kashlev, M. (2005). Nature of the Nucleosomal Barrier to RNA Polymerase II. Molecular Cell *18*, 97–108. https://doi.org/10.1016/J.MOLCEL.2005.02.027.

Kirtana, R., Manna, S., and Patra, S.K. (2020). Molecular mechanisms of KDM5A in cellular functions: Facets during development and disease. Experimental Cell Research *396*, 112314. https://doi.org/10.1016/J.YEXCR.2020.112314.

Kishikawa, S., Murata, T., Ugai, H., Yamazaki, T., and Yokoyama, K.K. (2003). Control elements of Dnmt1 gene are regulated in cell-cycle dependent manner. Nucleic Acids Symposium Series *3*, 307–308. https://doi.org/10.1093/NASS/3.1.307.

Kleinjan, D.A., Seawright, A., Childs, A.J., and van Heyningen, V. (2004). Conserved elements in Pax6 intron 7 involved in (auto)regulation and alternative transcription. Developmental Biology *265*, 462–477. https://doi.org/10.1016/J.YDBIO.2003.09.011.

Kliszczak, A.E., Rainey, M.D., Harhen, B., Boisvert, F.M., and Santocanale, C. (2011). DNA mediated chromatin pull-down for the study of chromatin replication. Scientific Reports *1*, 1–7. https://doi.org/10.1038/srep00095.

Kloet, S.L., Makowski, M.M., Baymaz, H.I., van Voorthuijsen, L., Karemaker, I.D., Santanach, A., Jansen, P.W.T.C., di Croce, L., and Vermeulen, M. (2016). The dynamic interactome and genomic targets of Polycomb complexes during stem-cell differentiation. Nature Structural and Molecular Biology *23*, 682–690. https://doi.org/10.1038/nsmb.3248.

Klose, R.J., Yamane, K., Bae, Y., Zhang, D., Erdjument-Bromage, H., Tempst, P., Wong, J., and Zhang, Y. (2006). The transcriptional repressor JHDM3A demethylates trimethyl histone H3 lysine 9 and lysine 36. Nature *442*, 312–316. https://doi.org/10.1038/nature04853.

Klose, R.J., Yan, Q., Tothova, Z., Yamane, K., Erdjument-Bromage, H., Tempst, P., Gilliland, D.G., Zhang, Y., and Kaelin, W.G. (2007). The Retinoblastoma Binding Protein RBP2 Is an H3K4 Demethylase. Cell *128*, 889–900. https://doi.org/10.1016/J.CELL.2007.02.013.

Konev, A.Y., Tribus, M., Sung, Y.P., Podhraski, V., Chin, Y.L., Emelyanov, A. v., Vershilova, E., Pirrotta, V., Kadonaga, J.T., Lusser, A., et al. (2007). CHD1 motor protein is required for deposition of histone variant H3.3 into chromatin in vivo. Science *317*, 1087–1090. https://doi.org/10.1126/SCIENCE.1145339.

Kostrewa, D., Zeller, M.E., Armache, K.J., Seizl, M., Leike, K., Thomm, M., and Cramer, P. (2009). RNA polymerase II-TFIIB structure and mechanism of transcription initiation. Nature *462*, 323–330. https://doi.org/10.1038/NATURE08548.

Krasny, L., and Huang, P.H. (2021). Data-independent acquisition mass spectrometry (DIA-MS) for proteomic applications in oncology. Molecular Omics *17*, 29–42. https://doi.org/10.1039/D0MO00072H.

Kriaucionis, S., and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in purkinje neurons and the brain. Science (1979) *324*, 929–930. https://doi.org/10.1126/science.1169786.

Krietenstein, N., Wal, M., Watanabe, S., Park, B., Peterson, C.L., Pugh, B.F., and Korber, P. (2016). Genomic Nucleosome Organization Reconstituted with Pure Proteins. Cell *167*, 709-721.e12. https://doi.org/10.1016/J.CELL.2016.09.045.

Krogan, N.J., Kim, M., Ahn, S.H., Zhong, G., Kobor, M.S., Cagney, G., Emili, A., Shilatifard, A., Buratowski, S., and Greenblatt, J.F. (2002). RNA Polymerase II Elongation Factors of Saccharomyces cerevisiae: a Targeted Proteomics Approach. Molecular and Cellular Biology *22*, 6979–6992. https://doi.org/10.1128/mcb.22.20.6979-6992.2002.

Krogan, N.J., Kim, M., Tong, A., Golshani, A., Cagney, G., Canadien, V., Richards, D.P., Beattie, B.K., Emili, A., Boone, C., et al. (2003). Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II. Mol Cell Biol *23*, 4207–4218. https://doi.org/10.1128/MCB.23.12.4207-4218.2003.

Kruczek, I., and Doerfler, W. (1982). The unmethylated state of the promoter/leader and 5′-regions of integrated adenovirus genes correlates with gene expression. The EMBO Journal *1*, 409–414. https://doi.org/10.1002/J.1460-2075.1982.TB01183.X.

Ku, M., Koche, R.P., Rheinbay, E., Mendenhall, E.M., Endoh, M., Mikkelsen, T.S., Presser, A., Nusbaum, C., Xie, X., Chi, A.S., et al. (2008). Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains. PLoS Genetics *4*, e1000242. https://doi.org/10.1371/journal.pgen.1000242.

Kukimoto, I., Elderkin, S., Grimaldi, M., Oelgeschläger, T., and Varga-Weisz, P.D. (2004). The Histone-Fold Protein Complex CHRAC-15/17 Enhances Nucleosome Sliding and Assembly Mediated by ACF. Molecular Cell *13*, 265–277. https://doi.org/10.1016/S1097-2765(03)00523-9.

Kundu, S., Ji, F., Sunwoo, H., Jain, G., Lee, J.T., Sadreyev, R.I., Dekker, J., and Kingston, R.E. (2017). Polycomb Repressive Complex 1 Generates Discrete Compacted Domains that Change during Differentiation. Mol Cell *65*, 432-446.e5. https://doi.org/10.1016/J.MOLCEL.2017.01.009.

Kurdistani, S.K., and Grunstein, M. (2003). In vivo protein–protein and protein–DNA crosslinking for genomewide binding microarray. Methods *31*, 90–95. https://doi.org/10.1016/S1046-2023(03)00092-6.

Kuroda, M.I., Kang, H., De, S., and Kassis, J.A. (2020). Dynamic Competition of Polycomb and Trithorax in Transcriptional Programming. Annual Review of Biochemistry *89*, 235–253. https://doi.org/10.1146/annurev-biochem-120219-103641.

Kustatscher, G., Wills, K.L.H., Furlan, C., and Rappsilber, J. (2014a). Chromatin enrichment for proteomics. Nature Protocols *9*, 2090–2099. https://doi.org/10.1038/nprot.2014.142.

Kustatscher, G., Hegarat, N., Wills, K.L.H., Furlan, C., Bukowski-Wills, J.-C., Hochegger, H., and Rappsilber, J. (2014b). Proteomics of a fuzzy organelle: interphase chromatin. The EMBO Journal *33*, 648–664. https://doi.org/10.1002/embj.201387614.

Kuzmichev, A., Nishioka, K., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2002). Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. Genes Dev *16*, 2893–2905. https://doi.org/10.1101/GAD.1035902.

Lafon, A., Taranum, S., Pietrocola, F., Dingli, F., Loew, D., Brahma, S., Bartholomew, B., and Papamichos-Chronakis, M. (2015). INO80 Chromatin Remodeler Facilitates Release of RNA Polymerase II from Chromatin for Ubiquitin-Mediated Proteasomal Degradation. Mol Cell *60*, 784–796. https://doi.org/10.1016/J.MOLCEL.2015.10.028.

Lagha, M., Bothma, J.P., Esposito, E., Ng, S., Stefanik, L., Tsui, C., Johnston, J., Chen, K., Gilmour, D.S., Zeitlinger, J., et al. (2013). Paused Pol II coordinates tissue morphogenesis in the Drosophila embryo. Cell *153*, 976. https://doi.org/10.1016/J.CELL.2013.04.045.

Lagrange, T., Kapanidis, A.N., Tang, H., Reinberg, D., and Ebright, R.H. (1998a). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev *12*, 34–44. https://doi.org/10.1101/GAD.12.1.34.

Lagrange, T., Kapanidis, A.N., Tang, H., Reinberg, D., and Ebright, R.H. (1998b). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev *12*, 34–44. https://doi.org/10.1101/GAD.12.1.34.

Lan, F., Bayliss, P.E., Rinn, J.L., Whetstine, J.R., Wang, J.K., Chen, S., Iwase, S., Alpatov, R., Issaeva, I., Canaani, E., et al. (2007). A histone H3 lysine 27 demethylase regulates animal posterior development. Nature *449*, 689–694. https://doi.org/10.1038/nature06192.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.

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

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095–1107. https://doi.org/10.1016/0888-7543(92)90024-M.

Lau, M.S., Schwartz, M.G., Kundu, S., Savol, A.J., Wang, P.I., Marr, S.K., Grau, D.J., Schorderet, P., Sadreyev, R.I., Tabin, C.J., et al. (2017). Mutation of a nucleosome compaction region disrupts Polycomb-mediated axial patterning. Science (1979) *355*, 1081–1084. https://doi.org/10.1126/science.aah5403.

Laue, K., Daujat, S., Crump, J.G., Plaster, N., Roehl, H.H., van Bebber, F., Busch-Nentwich, E., Dahm, R., Frohnhöfer, H.G., Geiger, H., et al. (2008). The multidomain protein Brpf1 binds histones and is required for Hox gene expression and segmental identity. Development *135*, 1935–1946. https://doi.org/10.1242/DEV.017160.

Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. Trends in Genetics *32*, 42–56. https://doi.org/10.1016/j.tig.2015.10.007.

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*, 41725–41731. https://doi.org/10.1074/JBC.M508312200.

Lee, D.-H., Gershenzon, N., Gupta, M., Ioshikhes, I.P., Reinberg, D., and Lewis, B.A. (2005). Functional Characterization of Core Promoter Elements: the Downstream Core Element Is Recognized by TAF1. Molecular and Cellular Biology *25*, 9674–9686. https://doi.org/10.1128/mcb.25.21.9674-9686.2005.

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

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*, 44669–44676. https://doi.org/10.1074/JBC.M107179200.

Lee, J.T., Davidow, L.S., and Warshawsky, D. (1999). Tsix, a gene antisense to Xist at the X-inactivation centre. Nat Genet *21*, 400–404. https://doi.org/10.1038/7734.

Lee, S.-M., Lee, J., Noh, K.-M., Choi, W.-Y., Jeon, S., Oh, G.T., Kim-Ha, J., Jin, Y., Cho, S.-W., and Kim, Y.-J. (2017). Intragenic CpG islands play important roles in bivalent chromatin assembly of developmental genes. Proceedings of the National Academy of Sciences *114*, E1885–E1894. https://doi.org/10.1073/pnas.1613300114.

Lees, A.M., Deconinck, A.E., Campbell, B.D., and Lees, R.S. (2005). Atherin: a newly identified, lesion-specific, LDL-binding protein in human atherosclerosis. Atherosclerosis *182*, 219–230. https://doi.org/10.1016/J.ATHEROSCLEROSIS.2005.01.041.

Lehmann, L., Ferrari, R., Vashisht, A.A., Wohlschlegel, J.A., Kurdistani, S.K., and Careys, M. (2012). Polycomb repressive complex 1 (PRC1) disassembles RNA polymerase II preinitiation complexes. J Biol Chem *287*, 35784–35794. https://doi.org/10.1074/JBC.M112.397430.

Lehnertz, B., Ueda, Y., Derijck, A.A.H.A., Braunschweig, U., Perez-Burgos, L., Kubicek, S., Chen, T., Li, E., Jenuwein, T., and Peters, A.H.F.M. (2003). Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Current Biology *13*, 1192–1200. https://doi.org/10.1016/S0960-9822(03)00432-9.

Lewis, E.B. (1978). A Gene Complex Controlling Segmentation in Drosophila. In Genes, Development and Cancer, (Springer, Boston, MA), pp. 205–217.

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

Li, H., Ilin, S., Wang, W., Duncan, E.M., Wysocka, J., Allis, C.D., and Patel, D.J. (2006). Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. Nature *442*, 91–95. https://doi.org/10.1038/NATURE04802.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Li, H., Liefke, R., Jiang, J., Kurland, J.V., Tian, W., Deng, P., Zhang, W., He, Q., Patel, D.J., Bulyk, M.L., et al. (2017). Polycomb-like proteins link the PRC2 complex to CpG islands. Nature *549*, 287–291. https://doi.org/10.1038/nature23881.

Li, X., Egervari, G., Wang, Y., Berger, S.L., and Lu, Z. (2018). Regulation of chromatin and gene expression by metabolic enzymes and metabolites. Nature Reviews Molecular Cell Biology *19*, 563–578. https://doi.org/10.1038/s41580-018-0029-7.

Li, Y., Wen, H., Xi, Y., Tanaka, K., Wang, H., Peng, D., Ren, Y., Jin, Q., Dent, S.Y.R., Li, W., et al. (2014). AF9 YEATS Domain Links Histone Acetylation to DOT1L-Mediated H3K79 Methylation. Cell *159*, 558–571. https://doi.org/10.1016/j.cell.2014.09.049.

Li, Y., Han, J., Zhang, Y., Cao, F., Liu, Z., Li, S., Wu, J., Hu, C., Wang, Y., Shuai, J., et al. (2016). Structural basis for activity regulation of MLL family methyltransferases. Nature *530*, 447–452. https://doi.org/10.1038/nature16952.

Liebler, D.C. (2004). Shotgun mass spec goes independent. Nature Methods *1*, 16–17. https://doi.org/10.1038/nmeth1004-16.

Liefke, R., Karwacki-Neisius, V., and Shi, Y. (2016). EPOP Interacts with Elongin BC and USP7 to Modulate the Chromatin Landscape. Molecular Cell *64*, 659–672. https://doi.org/10.1016/J.MOLCEL.2016.10.019.

Lim, M.Y., Paulo, J.A., and Gygi, S.P. (2019). Evaluating False Transfer Rates from the Match-between-Runs Algorithm with a Two-Proteome Model. Journal of Proteome Research *18*, 4020–4026. https://doi.org/10.1021/acs.jproteome.9b00492.

Lin, J.J., Lehmann, L.W., Bonora, G., Sridharan, R., Vashisht, A.A., Tran, N., Plath, K., Wohlschlegel, J.A., and Carey, M. (2011). Mediator coordinates PIC assembly with recruitment of CHD1. Genes & Development *25*, 2198–2209. https://doi.org/10.1101/GAD.17554711.

Lin, K., Yan, Q., Mitchell, A., Funk, N., Lu, C., and Xiao, H. (2019). A simple method for non-denaturing purification of biotin-tagged proteins through competitive elution with free biotin. Biotechniques *68*, 41–44. https://doi.org/10.2144/btn-2019-0088.

Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., Garvik, B.M., and Yates, J.R. (1999). Direct analysis of protein complexes using mass spectrometry. Nature Biotechnology *17*, 676–682. https://doi.org/10.1038/10890.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322. https://doi.org/10.1038/nature08514.

Littlefield, O., Korkhin, Y., and Sigler, P.B. (1999). The structural basis for the oriented assembly of a TBP/TFB/promoter complex. Proc Natl Acad Sci U S A *96*, 13668–13673. https://doi.org/10.1073/PNAS.96.24.13668.

Liu, H., Sadygov, R.G., and Yates, J.R. (2004). A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Analytical Chemistry *76*, 4193–4201. https://doi.org/10.1021/ac0498563.

Liu, W., Ma, Q., Wong, K., Li, W., Ohgi, K., Zhang, J., Aggarwal, A.K., and Rosenfeld, M.G. (2013). Brd4 and JMJD6-Associated Anti-Pause Enhancers in Regulation of Transcriptional Pause Release. Cell *155*, 1581–1595. https://doi.org/10.1016/J.CELL.2013.10.056.

Liu, W.L., Coleman, R.A., Ma, E., Grob, P., Yang, J.L., Zhang, Y., Dailey, G., Nogales, E., and Tjian, R. (2009). Structures of three distinct activator-TFIID complexes. Genes Dev *23*, 1510–1521. https://doi.org/10.1101/GAD.1790709.

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. Cell *165*, 535–550. https://doi.org/10.1016/J.CELL.2016.03.014.

Ljungman, M., Parks, L., Hulbatte, R., and Bedi, K. (2019). The role of H3K79 methylation in transcription and the DNA damage response. Mutation Research/Reviews in Mutation Research *780*, 48–54. https://doi.org/10.1016/j.mrrev.2017.11.001.

Loh, C.H., van Genesen, S., Perino, M., Bark, M.R., and Veenstra, G.J.C. (2021). Loss of PRC2 subunits primes lineage choice during exit of pluripotency. Nature Communications *12*, 1–14. https://doi.org/10.1038/s41467-021-27314-4.

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

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

Long, Y., Hwang, T., Gooding, A.R., Goodrich, K.J., Rinn, J.L., and Cech, T.R. (2020). RNA is essential for PRC2 chromatin occupancy and function in human pluripotent stem cells. Nat Genet *52*, 931–938. https://doi.org/10.1038/S41588-020-0662-X.

Lorch, Y., Maier-Davis, B., and Kornberg, R.D. (2006). Chromatin remodeling by nucleosome disassembly in vitro. Proc Natl Acad Sci U S A *103*, 3090–3093. https://doi.org/10.1073/PNAS.0511050103.

Low, J.K.K., Webb, S.R., Silva, A.P.G., Saathoff, H., Ryan, D.P., Torrado, M., Brofelth, M., Parker, B.L., Shepherd, N.E., and Mackay, J.P. (2016). CHD4 Is a Peripheral Component of the Nucleosome Remodeling and Deacetylase Complex. Journal of Biological Chemistry *291*, 15853–15866. https://doi.org/10.1074/JBC.M115.707018.

Lu, K., Ye, W.J., Zhou, L., Collins, L.B., Chen, X., Gold, A., Ball, L.M., and Swenberg, J.A. (2010). Structural Characterization of Formaldehyde-Induced Cross-Links Between Amino Acids and Deoxynucleosides and Their Oligomers. J Am Chem Soc *132*, 3388–3399. https://doi.org/10.1021/ja908282t.

Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408. https://doi.org/10.1038/s41586-020-2188-x.

Luco, R.F., Pan, Q., Tominaga, K., Blencowe, B.J., Pereira-Smith, O.M., and Misteli, T. (2010). Regulation of alternative splicing by histone modifications. Science (1979) *327*, 996–1000. https://doi.org/10.1126/science.1184208.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature *389*, 251–260. https://doi.org/10.1038/38444.

Luo, Z., Lin, C., and Shilatifard, A. (2012). The super elongation complex (SEC) family in transcriptional control. Nature Reviews Molecular Cell Biology *13*, 543–547. https://doi.org/10.1038/nrm3417.

Lusser, A., Urwin, D.L., and Kadonaga, J.T. (2005). Distinct activities of CHD1 and ACF in ATP-dependent chromatin assembly. Nat Struct Mol Biol *12*, 160–166. https://doi.org/10.1038/NSMB884.

Mahmoudi, T., Boj, S.F., Hatzis, P., Li, V.S.W., Taouatas, N., Vries, R.G.J., Teunissen, H., Begthel, H., Korving, J., Mohammed, S., et al. (2010). The Leukemia-Associated Mllt10/Af10-Dot1l Are Tcf4/β-Catenin Coactivators Essential for Intestinal Homeostasis. PLOS Biology *8*, e1000539. https://doi.org/10.1371/JOURNAL.PBIO.1000539.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Letters *583*, 3966–3973. https://doi.org/10.1016/J.FEBSLET.2009.10.036.

Maiti, A., and Drohat, A.C. (2011). Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine: POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES *. Journal of Biological Chemistry *286*, 35334–35338. https://doi.org/10.1074/JBC.C111.284620.

Malovannaya, A., Li, Y., Bulynko, Y., Jung, S.Y., Wang, Y., Lanz, R.B., O'malley, B.W., Qin, J., Lindqvist, C., Schuster, S.C., et al. (2010). Streamlined analysis schema for high-throughput identification of endogenous protein complexes. Proceedings of the National Academy of Sciences *107*, 2431–2436. https://doi.org/10.1073/PNAS.0912599106.

Margaritis, T., Oreal, V., Brabers, N., Maestroni, L., Vitaliano-Prunier, A., Benschop, J.J., van Hooff, S., van Leenen, D., Dargemont, C., Géli, V., et al. (2012). Two distinct repressive mechanisms for histone 3 lysine 4 methylation through promoting 3'-end antisense transcription. PLoS Genet *8*. https://doi.org/10.1371/JOURNAL.PGEN.1002952.

Marshall, N.F., and Price, D.H. (1995). Purification of P-TEFb, a transcription factor required for the transition into productive elongation. J Biol Chem *270*, 12335–12338. https://doi.org/10.1074/JBC.270.21.12335.

Mas, G., Blanco, E., Ballaré, C., Sansó, M., Spill, Y.G., Hu, D., Aoi, Y., le Dily, F., Shilatifard, A., Marti-Renom, M.A., et al. (2018). Promoter bivalency favors an open chromatin architecture in embryonic stem cells. Nature Genetics *50*, 1452–1462. https://doi.org/10.1038/s41588-018-0218-5.

Maston, G.A., Zhu, L.J., Chamberlain, L., Lin, L., Fang, M., and Green, M.R. (2012). Non-canonical TAF complexes regulate active promoters in human embryonic stem cells. Elife *1*. https://doi.org/10.7554/ELIFE.00068.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., et al. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature *466*, 253–257. https://doi.org/10.1038/nature09165.

Maxon, M.E., Goodrich, J.A., and Tjian, R. (1994). Transcription factor IIE binds preferentially to RNA polymerase IIa and recruits TFIIH: a model for promoter clearance. Genes Dev *8*, 515–524. https://doi.org/10.1101/GAD.8.5.515.

Mayer, W., Niveleau, A., Walter, J., Fundele, R., and Haaf, T. (2000). Demethylation of the zygotic paternal genome. Nature *403*, 501–502. https://doi.org/10.1038/35000656.

McStay, B., and Grummt, I. (2008). The epigenetics of rRNA genes: From molecular to chromosome biology. Annual Review of Cell and Developmental Biology *24*, 131–157. https://doi.org/10.1146/annurev.cellbio.24.110707.175259.

Meisenheimer, K.M., and Koch, T.H. (1997). Photocross-Linking of Nucleic Acids to Associated Proteins. Critical Reviews in Biochemistry and Molecular Biology *32*, 10–11. https://doi.org/10.3109/10409239709108550.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766–770. https://doi.org/10.1038/nature07107.

Mellacheruvu, D., Wright, Z., Couzens, A.L., Lambert, J.P., St-Denis, N.A., Li, T., Miteva, Y. v., Hauri, S., Sardiu, M.E., Low, T.Y., et al. (2013). The CRAPome: A contaminant repository for affinity purification-mass spectrometry data. Nature Methods *10*, 730–736. https://doi.org/10.1038/nmeth.2557.

Meshorer, E., Yellajoshula, D., George, E., Scambler, P.J., Brown, D.T., and Misteli, T. (2006). Hyperdynamic Plasticity of Chromatin Proteins in Pluripotent Embryonic Stem Cells. Developmental Cell *10*, 105–116. https://doi.org/10.1016/j.devcel.2005.10.017.

Michael Green, N. (1990). [5] Avidin and streptavidin. Methods in Enzymology *184*, 51–67. https://doi.org/10.1016/0076-6879(90)84259-J.

Michel, B.C., D'Avino, A.R., Cassel, S.H., Mashtalir, N., McKenzie, Z.M., McBride, M.J., Valencia, A.M., Zhou, Q., Bocker, M., Soares, L.M.M., et al. (2018). A non-canonical SWI/SNF complex is a synthetic lethal target in cancers driven by BAF complex perturbation. Nature Cell Biology *20*, 1410–1420. https://doi.org/10.1038/s41556-018-0221-1.

van Mierlo, G., and Vermeulen, M. (2021). Chromatin Proteomics to Study Epigenetics — Challenges and Opportunities. Molecular & Cellular Proteomics *20*, 100056. https://doi.org/10.1074/mcp.R120.002208.

van Mierlo, G., Veenstra, G.J.C., Vermeulen, M., and Marks, H. (2019). The Complexity of PRC2 Subcomplexes. Trends in Cell Biology *29*, 660–671. https://doi.org/10.1016/j.tcb.2019.05.004.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553–560. https://doi.org/10.1038/nature06008.

Mills, E.W., and Green, R. (2017). Ribosomopathies: There's strength in numbers. Science (1979) *358*. https://doi.org/10.1126/SCIENCE.AAN2755/ASSET/67B9B601-3F89-4685-9ADD-870A7F2BC6F9/ASSETS/GRAPHIC/358_AAN2755_F4.JPEG.

Milne, T.A., Dou, Y., Martin, M.E., Brock, H.W., Roeder, R.G., and Hess, J.L. (2005). MLL associates specifically with a subset of transcriptionally active target genes. Proc Natl Acad Sci U S A *102*, 14765–14770. https://doi.org/10.1073/PNAS.0503630102.

Milne, T.A., Kim, J., Wang, G.G., Stadler, S.C., Basrur, V., Whitcomb, S.J., Wang, Z., Ruthenburg, A.J., Elenitoba-Johnson, K.S.J., Roeder, R.G., et al. (2010). Multiple Interactions Recruit MLL1 and MLL1 Fusion Proteins to the HOXA9 Locus in Leukemogenesis. Molecular Cell *38*, 853–863. https://doi.org/10.1016/J.MOLCEL.2010.05.011.

Min, G.L., Villa, R., Trojer, P., Norman, J., Yan, K.P., Reinberg, D., di Croce, L., and Shiekhattar, R. (2007). Demethylation of H3K27 regulates polycomb recruitment and H2A ubiquitination. Science (1979) *318*, 447–450. https://doi.org/10.1126/science.1149042.

Miyamoto, R., Okuda, H., Kanai, A., Takahashi, S., Kawamura, T., Matsui, H., Kitamura, T., Kitabayashi, I., Inaba, T., and Yokoyama, A. (2020). Activation of CpG-Rich Promoters Mediated by MLL Drives MOZ-Rearranged Leukemia. Cell Reports *32*, 108200. https://doi.org/10.1016/j.celrep.2020.108200.

Mizuguchi, G., Shen, X., Landry, J., Wu, W.H., Sen, S., and Wu, C. (2004). ATP-Driven Exchange of Histone H2AZ Variant Catalyzed by SWR1 Chromatin Remodeling Complex. Science (1979) *303*, 343–348. https://doi.org/10.1126/science.1090701.

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 (1979) *211*, 393–396. https://doi.org/10.1126/SCIENCE.6164095.

Mohn, F., Weber, M., Rebhan, M., Roloff, T.C., Richter, J., Stadler, M.B., Bibel, M., and Schübeler, D. (2008). Lineage-Specific Polycomb Targets and De Novo DNA Methylation Define Restriction and Potential of Neuronal Progenitors. Molecular Cell *30*, 755–766. https://doi.org/10.1016/J.MOLCEL.2008.05.007.

Moon, K.J., Mochizuki, K., Zhou, M., Jeong, H.S., Brady, J.N., and Ozato, K. (2005). The Bromodomain Protein Brd4 Is a Positive Regulatory Component of P-TEFb and Stimulates RNA Polymerase II-Dependent Transcription. Molecular Cell *19*, 523–534. https://doi.org/10.1016/J.MOLCEL.2005.06.027.

Moore, L.D., Le, T., and Fan, G. (2013). DNA methylation and its basic function. Neuropsychopharmacology *38*, 23–38. https://doi.org/10.1038/npp.2012.112.

Morey, L., Pascual, G., Cozzuto, L., Roma, G., Wutz, A., Benitah, S.A., and di Croce, L. (2012). Nonoverlapping functions of the polycomb group Cbx family of proteins in embryonic stem cells. Cell Stem Cell *10*, 47–62. https://doi.org/10.1016/j.stem.2011.12.006.

Morey, L., Aloia, L., Cozzuto, L., Benitah, S.A., and di Croce, L. (2013). RYBP and Cbx7 Define Specific Biological Functions of Polycomb Complexes in Mouse Embryonic Stem Cells. Cell Reports *3*, 60–69. https://doi.org/10.1016/j.celrep.2012.11.026.

Morselli, M., Pastor, W.A., Montanini, B., Nee, K., Ferrari, R., Fu, K., Bonora, G., Rubbi, L., Clark, A.T., Ottonello, S., et al. (2015). In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse. Elife *2015*. https://doi.org/10.7554/ELIFE.06205.

Mortusewicz, O., Schermelleh, L., Walter, J., Cardoso, M.C., and Leonhardt, H. (2005). Recruitment of DNA methyltransferase I to DNA repair sites. Proc Natl Acad Sci U S A *102*, 8905–8909. https://doi.org/10.1073/PNAS.0501034102.

Motea, E.A., and Berdis, A.J. (2010). Terminal deoxynucleotidyl transferase: The story of a misguided DNA polymerase. Biochimica et Biophysica Acta - Proteins and Proteomics *1804*, 1151–1166. https://doi.org/10.1016/j.bbapap.2009.06.030.

Mühlbacher, W., Sainsbury, S., Hemann, M., Hantsche, M., Neyer, S., Herzog, F., and Cramer, P. (2014). Conserved architecture of the core RNA polymerase II initiation complex. Nature Communications *5*, 1–6. https://doi.org/10.1038/ncomms5310.

Müller, J., Hart, C.M., Francis, N.J., Vargas, M.L., Sengupta, A., Wild, B., Miller, E.L., O'Connor, M.B., Kingston, R.E., and Simon, J.A. (2002). Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. Cell *111*, 197–208. https://doi.org/10.1016/S0092-8674(02)00976-5.

Murai, M.J., Pollock, J., He, S., Miao, H., Purohit, T., Yokom, A., Hess, J.L., Muntean, A.G., Grembecka, J., and Cierpicki, T. (2014). The same site on the integrase-binding domain of lens epithelium–derived growth factor is a therapeutic target for MLL leukemia and HIV. Blood *124*, 3730–3737. https://doi.org/10.1182/BLOOD-2014-01-550079.

Murzina, N. v., Pei, X.Y., Zhang, W., Sparkes, M., Vicente-Garcia, J., Pratap, J.V., McLaughlin, S.H., Ben-Shahar, T.R., Verreault, A., Luisi, B.F., et al. (2008). Structural Basis for the Recognition of Histone H4 by the Histone-Chaperone RbAp46. Structure *16*, 1077–1085. https://doi.org/10.1016/j.str.2008.05.006.

Musselman, C.A., Mansfield, R.E., Garske, A.L., Davrazou, F., Kwan, A.H., Oliver, S.S., O'Leary, H., Denu, J.M., Mackay, J.P., and Kutateladze, T.G. (2009). Binding of the CHD4 PHD2 finger to histone H3 is modulated by covalent modifications. Biochemical Journal *423*, 179–187. https://doi.org/10.1042/BJ20090870.

Myant, K., and Stancheva, I. (2008). LSH cooperates with DNA methyltransferases to repress transcription. Mol Cell Biol *28*, 215–226. https://doi.org/10.1128/MCB.01073-07.

Myant, K., Termanis, A., Sundaram, A.Y.M., Boe, T., Li, C., Merusi, C., Burrage, J., de Las Heras, J.I., and Stancheva, I. (2011). LSH and G9a/GLP complex are required for developmentally programmed DNA methylation. Genome Res *21*, 83–94. https://doi.org/10.1101/GR.108498.110.

Nakagawa, T., Kajitani, T., Togo, S., Masuko, N., Ohdan, H., Hishikawa, Y., Koji, T., Matsuyama, T., Ikura, T., Muramatsu, M., et al. (2008). Deubiquitylation of histone H2A activates transcriptional initiation via trans-histone cross-talk with H3K4 di- and trimethylation. Genes Dev *22*, 37–49. https://doi.org/10.1101/GAD.1609708.

de Napoles, M., Mermoud, J.E., Wakao, R., Tang, Y.A., Endoh, M., Appanah, R., Nesterova, T.B., Silva, J., Otte, A.P., Vidal, M., et al. (2004). Polycomb Group Proteins Ring1A/B Link Ubiquitylation of Histone H2A to Heritable Gene Silencing and X Inactivation. Developmental Cell *7*, 663–676. https://doi.org/10.1016/J.DEVCEL.2004.10.005.

Naveh-Many, T., and Cedar, H. (1981). Active gene sequences are undermethylated. Proc Natl Acad Sci U S A *78*, 4246–4250. https://doi.org/10.1073/PNAS.78.7.4246.

Nekrasov, M., Wild, B., and Müller, J. (2005). Nucleosome binding and histone methyltransferase activity of Drosophila PRC2. EMBO Rep *6*, 348–353. https://doi.org/10.1038/SJ.EMBOR.7400376.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. Nature *543*, 72–77. https://doi.org/10.1038/nature21373.

Ng, H.H., Xu, R.M., Zhang, Y., and Struhl, K. (2002). Ubiquitination of Histone H2B by Rad6 Is Required for Efficient Dot1-mediated Methylation of Histone H3 Lysine 79 *. Journal of Biological Chemistry *277*, 34655–34657. https://doi.org/10.1074/JBC.C200433200.

Ng, H.H., Dole, S., and Struhl, K. (2003). The Rtf1 component of the Paf1 transcriptional elongation complex is required for ubiquitination of histone H2B. J Biol Chem *278*, 33625–33628. https://doi.org/10.1074/JBC.C300270200.

Nikolov, D.B., Chen, H., Halay, E.D., Usheva, A.A., Hisatake, K., Lee, D.K., Roeder, R.G., and Burley, S.K. (1995). Crystal structure of a TFIIB-TBP-TATA-element ternary complex. Nature *377*, 119–128. https://doi.org/10.1038/377119A0.

Nowak, D.E., Tian, B., and Brasier, A.R. (2005). Two-step cross-linking method for identification of NF-κB gene network by chromatin immunoprecipitation. Biotechniques *39*, 715–724. https://doi.org/10.2144/000112014.

Okamoto, I., and Heard, E. (2009). Lessons from comparative analysis of X-chromosome inactivation in mammals. Chromosome Research *17*, 659–669. https://doi.org/10.1007/s10577-009-9057-7.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell *99*, 247–257. https://doi.org/10.1016/S0092-8674(00)81656-6.

Oksuz, O., Narendra, V., Lee, C.H., Descostes, N., LeRoy, G., Raviram, R., Blumenberg, L., Karch, K., Rocha, P.P., Garcia, B.A., et al. (2018). Capturing the Onset of PRC2-Mediated Repressive Domain Formation. Mol Cell *70*, 1149-1162.e5. https://doi.org/10.1016/J.MOLCEL.2018.05.023.

Ooi, S.K.T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.P., Allis, C.D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. Nature *448*, 714–717. https://doi.org/10.1038/nature05987.

Orlando, V., Strutt, H., and Paro, R. (1997). Analysis of Chromatin Structure by in Vivo Formaldehyde Cross-Linking. Methods *11*, 205–214. https://doi.org/10.1006/meth.1996.0407.

Orphanides, G., LeRoy, G., Chang, C.H., Luse, D.S., and Reinberg, D. (1998). FACT, a factor that facilitates transcript elongation through nucleosomes. Cell *92*, 105–116. https://doi.org/10.1016/S0092-8674(00)80903-4.

Orphanides, G., Wu, W.H., Lane, W.S., Hampsey, M., and Reinberg, D. (1999). The chromatin-specific transcription elongation factor FACT comprises human SPT16 and SSRP1 proteins. Nature *400*, 284–288. https://doi.org/10.1038/22350.

Osman, S., and Cramer, P. (2020). Structural Biology of RNA Polymerase II Transcription: 20 Years on. Annual Review of Cell and Developmental Biology *36*, 1–34. https://doi.org/10.1146/annurev-cellbio-042020-021954.

van Oss, S.B., Cucinotta, C.E., and Arndt, K.M. (2017). Emerging Insights into the Roles of the Paf1 Complex in Gene Regulation. Trends in Biochemical Sciences *42*, 788–798. https://doi.org/10.1016/j.tibs.2017.08.003.

Oswald, J., Engemann, S., Lane, N., Mayer, W., Olek, A., Fundele, R., Dean, W., Reik, W., and Walter, J. (2000). Active demethylation of the paternal genome in the mouse zygote. Current Biology *10*, 475–478. https://doi.org/10.1016/S0960-9822(00)00448-6.

Otani, J., Nankumo, T., Arita, K., Inamoto, S., Ariyoshi, M., and Shirakawa, M. (2009). Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX–DNMT3–DNMT3L domain. EMBO Rep *10*, 1235–1241. https://doi.org/10.1038/EMBOR.2009.218.

Pachano, T., Sánchez-Gaya, V., Ealo, T., Mariner-Faulí, M., Bleckwehl, T., Asenjo, H.G., Respuela, P., Cruz-Molina, S., Muñoz-San Martín, M., Haro, E., et al. (2021). Orphan CpG islands amplify poised enhancer regulatory activity and determine target gene responsiveness. Nat Genet *53*, 1036–1049. https://doi.org/10.1038/S41588-021-00888-X.

Paggetti, J., Largeot, A., Aucagne, R., Jacquel, A., Lagrange, B., Yang, X.J., Solary, E., Bastie, J.N., and Delva, L. (2010). Crosstalk between leukemia-associated proteins MOZ and MLL regulates HOX gene expression in human cord blood CD34 cells. Oncogene *29*, 5019–5031. https://doi.org/10.1038/onc.2010.254.

Panning, B., and Jaenisch, R. (1996). DNA hypomethylation can activate Xist expression and silence X-linked genes. Genes Dev *10*, 1991–2002. https://doi.org/10.1101/GAD.10.16.1991.

Papamichos-Chronakis, M., Watanabe, S., Rando, O.J., and Peterson, C.L. (2011). Global regulation of H2A.Z localization by the INO80 chromatin-remodeling enzyme is essential for genome integrity. Cell *144*, 200–213. https://doi.org/10.1016/J.CELL.2010.12.021.

Papin, C., le Gras, S., Ibrahim, A., Salem, H., Karimi, M.M., Stoll, I., Ugrinova, I., Schröder, M., Fontaine-Pelletier, E., Omran, Z., et al. (2021). CpG Islands Shape the Epigenome Landscape. Journal of Molecular Biology *433*. https://doi.org/10.1016/j.jmb.2020.09.018.

Parry, T.J., Theisen, J.W.M., Hsu, J.Y., Wang, Y.L., Corcoran, D.L., Eustice, M., Ohler, U., and Kadonaga, J.T. (2010). The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. Genes and Development *24*, 2013–2018. https://doi.org/10.1101/gad.1951110.

Pasini, D., Bracken, A.P., Jensen, M.R., Denchi, E.L., and Helin, K. (2004a). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. EMBO Journal *23*, 4061–4071. https://doi.org/10.1038/sj.emboj.7600402.

Pasini, D., Bracken, A.P., Jensen, M.R., Denchi, E.L., and Helin, K. (2004b). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. EMBO Journal *23*, 4061–4071. https://doi.org/10.1038/sj.emboj.7600402.

Pasini, D., Bracken, A.P., Jensen, M.R., Denchi, E.L., and Helin, K. (2004c). Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. EMBO Journal *23*, 4061–4071. https://doi.org/10.1038/sj.emboj.7600402.

Pasini, D., Bracken, A.P., Hansen, J.B., Capillo, M., and Helin, K. (2007). The Polycomb Group Protein Suz12 Is Required for Embryonic Stem Cell Differentiation. Molecular and Cellular Biology *27*, 3769–3779. https://doi.org/10.1128/MCB.01432-06.

Pasini, D., Malatesta, M., Jung, H.R., Walfridsson, J., Willer, A., Olsson, L., Skotte, J., Wutz, A., Porse, B., Jensen, O.N., et al. (2010a). Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. Nucleic Acids Res *38*, 4958–4969. https://doi.org/10.1093/NAR/GKQ244.

Pasini, D., Cloos, P.A.C., Walfridsson, J., Olsson, L., Bukowski, J.P., Johansen, J. v., Bak, M., Tommerup, N., Rappsilber, J., and Helin, K. (2010b). JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. Nature *464*, 306–310. https://doi.org/10.1038/nature08788.

Pasini, D., Malatesta, M., Jung, H.R., Walfridsson, J., Willer, A., Olsson, L., Skotte, J., Wutz, A., Porse, B., Jensen, O.N., et al. (2010c). Characterization of an antagonistic switch between histone H3 lysine 27 methylation and acetylation in the transcriptional regulation of Polycomb group target genes. Nucleic Acids Research *38*, 4958–4969. https://doi.org/10.1093/NAR/GKQ244.

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*, 17–24. https://doi.org/10.1016/J.SBI.2019.10.001.

Patikoglou, G.A., Kim, J.L., Sun, L., Yang, S.H., Kodadek, T., and Burley, S.K. (1999). TATA element recognition by the TATA box-binding protein has been conserved throughout evolution. Genes Dev *13*, 3217–3230. https://doi.org/10.1101/GAD.13.24.3217.

Pavri, R., Zhu, B., Li, G., Trojer, P., Mandal, S., Shilatifard, A., and Reinberg, D. (2006). Histone H2B Monoubiquitination Functions Cooperatively with FACT to Regulate Elongation by RNA Polymerase II. Cell *125*, 703–717. https://doi.org/10.1016/J.CELL.2006.04.029.

Payer, B., and Lee, J.T. (2008). X chromosome dosage compensation: How mammals keep the balance. Annual Review of Genetics *42*, 733–772. https://doi.org/10.1146/annurev.genet.42.110807.091711.

Pearson, J.C., Lemons, D., and McGinnis, W. (2005). Modulating Hox gene functions during animal body patterning. Nature Reviews Genetics *6*, 893–904. https://doi.org/10.1038/nrg1726.

Pedersen, M.T., Kooistra, S.M., Radzisheuskaya, A., Laugesen, A., Johansen, J.V., Hayward, D.G., Nilsson, J., Agger, K., and Helin, K. (2016). Continual removal of H3K9 promoter methylation by Jmjd2 demethylases is vital for ESC self-renewal and early development. The EMBO Journal *35*, 1550–1564. https://doi.org/10.15252/EMBJ.201593317.

Peng, J.C., Valouev, A., Swigut, T., Zhang, J., Zhao, Y., Sidow, A., and Wysocka, J. (2009). Jarid2/Jumonji Coordinates Control of PRC2 Enzymatic Activity and Target Gene Occupancy in Pluripotent Cells. Cell *139*, 1290–1302. https://doi.org/10.1016/j.cell.2009.12.002.

Peng, L., Li, Y., Xi, Y., Li, W., Li, J., Lv, R., Zhang, L., Zou, Q., Dong, S., Luo, H., et al. (2016). MBD3L2 promotes Tet2 enzymatic activity for mediating 5-methylcytosine oxidation. J Cell Sci *129*, 1059–1071. https://doi.org/10.1242/JCS.179044.

Perino, M., van Mierlo, G., Karemaker, I.D., van Genesen, S., Vermeulen, M., Marks, H., van Heeringen, S.J., and Veenstra, G.J.C. (2018). MTF2 recruits Polycomb Repressive Complex 2 by helical-shape-selective DNA binding. Nature Genetics *50*, 1002–1010. https://doi.org/10.1038/s41588-018-0134-8.

Petryk, N., Bultmann, S., Bartke, T., and Defossez, P.A. (2021). Staying true to yourself: mechanisms of DNA methylation maintenance in mammals. Nucleic Acids Research *49*, 3020–3032. https://doi.org/10.1093/NAR/GKAA1154.

Piunti, A., and Shilatifard, A. (2021). The roles of Polycomb repressive complexes in mammalian development and cancer. Nature Reviews Molecular Cell Biology *22*, 326–345. https://doi.org/10.1038/s41580-021-00341-1.

Plys, A.J., Davis, C.P., Kim, J., Rizki, G., Keenen, M.M., Marr, S.K., and Kingston, R.E. (2019a). Phase separation of Polycomb-repressive complex 1 is governed by a charged disordered region of CBX2. Genes Dev *33*, 799–813. https://doi.org/10.1101/GAD.326488.119.

Plys, A.J., Davis, C.P., Kim, J., Rizki, G., Keenen, M.M., Marr, S.K., and Kingston, R.E. (2019b). Phase separation of Polycomb-repressive complex 1 is governed by a charged disordered region of CBX2. Genes Dev *33*, 799–813. https://doi.org/10.1101/GAD.326488.119.

Poli, J., Gerhold, C.B., Tosi, A., Hustedt, N., Seeber, A., Sack, R., Herzog, F., Pasero, P., Shimada, K., Hopfner, K.P., et al. (2016). Mec1, INO80, and the PAF1 complex cooperate to limit transcription replication conflicts through RNAPII removal during replication stress. Genes Dev *30*, 337–354. https://doi.org/10.1101/GAD.273813.115.

Poli, J., Gasser, S.M., and Papamichos-Chronakis, M. (2017). The INO80 remodeller in transcription, replication and repair. Philosophical Transactions of the Royal Society B: Biological Sciences *372*. https://doi.org/10.1098/RSTB.2016.0290.

Poot, R.A., Dellaire, G., Hülsmann, B.B., Grimaldi, M.A., Corona, D.F.V., Becker, P.B., Bickmore, W.A., and Varga-Weisz, P.D. (2000). HuCHRAC, a human ISWI chromatin remodelling complex contains hACF1 and two novel histone-fold proteins. The EMBO Journal *19*, 3377–3387. https://doi.org/10.1093/EMBOJ/19.13.3377.

Porrua, O., and Libri, D. (2015). Transcription termination and the control of the transcriptome: Why, where and how to stop. Nature Reviews Molecular Cell Biology *16*, 190–202. https://doi.org/10.1038/nrm3943.

Poynter, S.T., and Kadoch, C. (2016). Polycomb and trithorax opposition in development and disease. Wiley Interdisciplinary Reviews: Developmental Biology *5*, 659–688. https://doi.org/10.1002/wdev.244.

Pray-Grant, M.G., Daniel, J.A., Schieltz, D., Yates, J.R., and Grant, P.A. (2005). Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. Nature *433*, 434–438. https://doi.org/10.1038/nature03242.

Pugh, B.F., and Tjian, R. (1991). Transcription from a TATA-less promoter requires a multisubunit TFIID complex. Genes & Development *5*, 1935–1945. https://doi.org/10.1101/GAD.5.11.1935.

Qiu, H., Hu, C., Gaur, N.A., and Hinnebusch, A.G. (2012). Pol II CTD kinases Bur1 and Kin28 promote Spt5 CTR-independent recruitment of Paf1 complex. The EMBO Journal *31*, 3494–3505. https://doi.org/10.1038/EMBOJ.2012.188.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Rafiee, M.R., Girardot, C., Sigismondo, G., and Krijgsveld, J. (2016). Expanding the Circuitry of Pluripotency by Selective Isolation of Chromatin-Associated Proteins. Molecular Cell *64*, 624–635. https://doi.org/10.1016/j.molcel.2016.09.019.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research *44*, W160–W165. https://doi.org/10.1093/NAR/GKW257.

Rani, P.G., Ranish, J.A., and Hahn, S. (2004). RNA polymerase II (Pol II)-TFIIF and Pol II-mediator complexes: the major stable Pol II complexes and their activity in transcription initiation and reinitiation. Mol Cell Biol *24*, 1709–1720. https://doi.org/10.1128/MCB.24.4.1709-1720.2004.

Ranjan, A., Mizuguchi, G., Fitzgerald, P.C., Wei, D., Wang, F., Huang, Y., Luk, E., Woodcock, C.L., and Wu, C. (2013). Nucleosome-free region dominates histone acetylation in targeting SWR1 to promoters for H2A.Z replacement. Cell *154*, 1232. https://doi.org/10.1016/J.CELL.2013.08.005.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop And Go Extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Analytical Chemistry *75*, 663–670. https://doi.org/10.1021/ac026117i.

Rasmussen, K.D., Berest, I., Keβler, S., Nishimura, K., Simón-Carrasco, L., Vassiliou, G.S., Pedersen, M.T., Christensen, J., Zaugg, J.B., and Helin, K. (2019). TET2 binding to enhancers facilitates transcription factor recruitment in hematopoietic cells. Genome Research *29*, 564–575. https://doi.org/10.1101/GR.239277.118/-/DC1.

Rauch, T.A., Wu, X., Zhong, X., Riggs, A.D., and Pfeifer, G.P. (2009). A human B cell methylome at 100-base pair resolution. Proc Natl Acad Sci U S A *106*, 671–678. https://doi.org/10.1073/pnas.0812399106.

Razin, A., and Cedar, H. (1977). Distribution of 5-methylcytosine in chromatin. Proc Natl Acad Sci U S A *74*, 2725–2728. https://doi.org/10.1073/PNAS.74.7.2725.

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtier, K., Ponting, C.P., Allis, C.D., et al. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature *406*, 593–599. https://doi.org/10.1038/35020506.

Reiter, F., Wienerroither, S., and Stark, A. (2017). Combinatorial function of transcription factors and cofactors. Current Opinion in Genetics & Development *43*, 73–81. https://doi.org/10.1016/J.GDE.2016.12.007.

Ren, X., Vincenz, C., and Kerppola, T.K. (2008). Changes in the distributions and dynamics of polycomb repressive complexes during embryonic stem cell differentiation. Mol Cell Biol *28*, 2884–2895. https://doi.org/10.1128/MCB.00949-07.

Reyes, A.A., Marcum, R.D., and He, Y. (2021). Structure and Function of Chromatin Remodelers. Journal of Molecular Biology *433*, 166929. https://doi.org/10.1016/j.jmb.2021.166929.

Reynolds, N., Salmon-Divon, M., Dvinge, H., Hynes-Allen, A., Balasooriya, G., Leaford, D., Behrens, A., Bertone, P., and Hendrich, B. (2012). NuRD-mediated deacetylation of H3K27 facilitates recruitment of Polycomb Repressive Complex 2 to direct gene repression. The EMBO Journal *31*, 593–605. https://doi.org/10.1038/EMBOJ.2011.431.

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

Riising, E.M., Comet, I., Leblanc, B., Wu, X., Johansen, J.V., and Helin, K. (2014). Gene silencing triggers polycomb repressive complex 2 recruitment to CpG Islands genome wide. Molecular Cell *55*, 347–360. https://doi.org/10.1016/j.molcel.2014.06.005.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Brugmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., et al. (2007). Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. Cell *129*, 1311–1323. https://doi.org/10.1016/J.CELL.2007.05.022.

Robert, O., Sures, I., Ciossek, T., Fuchs, M., and Ullrich, A. (1996). Isolation and developmental expression analysis of Enx-1, a novel mouse Polycomb group gene. Mechanisms of Development *55*, 171–184. https://doi.org/10.1016/0925-4773(96)00499-6.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nature Biotechnology *29*, 24–26. https://doi.org/10.1038/nbt.1754.

Roeder, R.G. (1996). The role of general initiation factors in transcription by RNA polymerase II. Trends in Biochemical Sciences *21*, 327–335. https://doi.org/10.1016/s0968-0004(96)10050-5.

Roeder, R.G., and Rutter, W.J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. Nature *224*, 234–237. https://doi.org/10.1038/224234a0.

Roeder, R.G., and Rutter, W.J. (1970). Specific Nucleolar and Nucleoplasmic RNA Polymerases. Proceedings of the National Academy of Sciences *65*, 675–682. https://doi.org/10.1073/PNAS.65.3.675.

Rose, N.R., King, H.W., Blackledge, N.P., Fursova, N.A., Ember, K.J., Fischer, R., Kessler, B.M., and Klose, R.J. (2016). RYBP stimulates PRC1 to shape chromatin-based communication between Polycomb repressive complexes. Elife *5*, 1–29. https://doi.org/10.7554/eLife.18591.

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*, 795–804. https://doi.org/10.1016/S0092-8674(88)91087-2.

Ruhl, D.D., Jin, J., Cai, Y., Swanson, S., Florens, L., Washburn, M.P., Conaway, R.C., Conaway, J.W., and Chrivia, J.C. (2006). Purification of a human SRCAP complex that remodels chromatin by incorporating the histone variant H2A.Z into nucleosomes. Biochemistry *45*, 5671–5677. https://doi.org/10.1021/BI060043D.

Ruthenburg, A.J., Li, H., Milne, T.A., Dewell, S., McGinty, R.K., Yuen, M., Ueberheide, B., Dou, Y., Muir, T.W., Patel, D.J., et al. (2011). Recognition of a Mononucleosomal Histone Modification Pattern by BPTF via Multivalent Interactions. Cell *145*, 692–706. https://doi.org/10.1016/J.CELL.2011.03.053.

Sainsbury, S., Niesser, J., and Cramer, P. (2013). Structure and function of the initially transcribing RNA polymerase II-TFIIB complex. Nature *493*, 437–440. https://doi.org/10.1038/NATURE11715.

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

Sakai, D., and Trainor, P.A. (2009). Treacher Collins syndrome: unmasking the role of Tcof1/treacle. Int J Biochem Cell Biol *41*, 1229–1232. https://doi.org/10.1016/j.biocel.2008.10.026.

Saksouk, N., Simboeck, E., and Déjardin, J. (2015). Constitutive heterochromatin formation and transcription in mammals. Epigenetics and Chromatin *8*, 1–17. https://doi.org/10.1186/1756-8935-8-3.

de Santa, F., Totaro, M.G., Prosperini, E., Notarbartolo, S., Testa, G., and Natoli, G. (2007). The Histone H3 Lysine-27 Demethylase Jmjd3 Links Inflammation to Inhibition of Polycomb-Mediated Gene Silencing. Cell *130*, 1083–1094. https://doi.org/10.1016/J.CELL.2007.08.019.

Santoro, R., and Grummt, I. (2001). Molecular Mechanisms Mediating Methylation-Dependent Silencing of Ribosomal Gene Transcription. Molecular Cell *8*, 719–725. https://doi.org/10.1016/S1097-2765(01)00317-3.

Saurin, A.J., Shiels, C., Williamson, J., Satijn, D.P.E., Otte, A.P., Sheer, D., and Freemont, P.S. (1998). The human polycomb group complex associates with pericentromeric heterochromatin to form a novel nuclear domain. J Cell Biol *142*, 887–898. https://doi.org/10.1083/JCB.142.4.887.

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 *103*, 1412–1417. https://doi.org/10.1073/pnas.0510310103.

Scelfo, A., Fernández-Pérez, D., Tamburri, S., Zanotti, M., Lavarone, E., Soldi, M., Bonaldi, T., Ferrari, K.J., and Pasini, D. (2019). Functional Landscape of PCGF Proteins Reveals Both RING1A/B-Dependent-and RING1A/B-Independent-Specific Activities. Molecular Cell *74*, 1037-1052.e7. https://doi.org/10.1016/j.molcel.2019.04.002.

Schmitges, F.W., Prusty, A.B., Faty, M., Stützer, A., Lingaraju, G.M., Aiwazian, J., Sack, R., Hess, D., Li, L., Zhou, S., et al. (2011). Histone Methylation by PRC2 Is Inhibited by Active Chromatin Marks. Molecular Cell *42*, 330–341. https://doi.org/10.1016/J.MOLCEL.2011.03.025.

Schoenfelder, S., Sugar, R., Dimond, A., Javierre, B.M., Armstrong, H., Mifsud, B., Dimitrova, E., Matheson, L., Tavares-Cadete, F., Furlan-Magaril, M., et al. (2015). Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nat Genet *47*, 1179–1186. https://doi.org/10.1038/NG.3393.

Schübeler, D. (2015). Function and information content of DNA methylation. Nature *517*, 321–326. https://doi.org/10.1038/nature14192.

Schuettengruber, B., and Cavalli, G. (2009). Recruitment of Polycomb group complexes and their role in the dynamic regulation of cell fate choice. Development *136*, 3531–3542. https://doi.org/10.1242/dev.033902.

Schuettengruber, B., Bourbon, H.M., di Croce, L., and Cavalli, G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. Cell *171*, 34–57. https://doi.org/10.1016/j.cell.2017.08.002.

Seif, E., Kang, J.J., Sasseville, C., Senkovich, O., Kaltashov, A., Boulier, E.L., Kapur, I., Kim, C.A., and Francis, N.J. (2020). Phase separation by the polyhomeotic sterile alpha motif compartmentalizes Polycomb Group proteins and enhances their activity. Nat Commun *11*. https://doi.org/10.1038/S41467-020-19435-Z.

Serizawa, H., Mäkelä, T.P., Conaway, J.W., Conaway, R.C., Weinberg, R.A., and Young, R.A. (1995). Association of Cdk-activating kinase subunits with transcription factor TFIIH. Nature *374*, 280–282. https://doi.org/10.1038/374280A0.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498–2504. https://doi.org/10.1101/GR.1239303.

Shao, W., and Zeitlinger, J. (2017). Paused RNA polymerase II inhibits new transcriptional initiation. Nat Genet *49*, 1045–1051. https://doi.org/10.1038/NG.3867.

Shao, H., Revach, M., Moshonov, S., Tzuman, Y., Gazit, K., Albeck, S., Unger, T., and Dikstein, R. (2005). Core Promoter Binding by Histone-Like TAF Complexes. Molecular and Cellular Biology *25*, 206–219. https://doi.org/10.1128/mcb.25.1.206-219.2005.

Sharif, J., Muto, M., Takebayashi, S.I., Suetake, I., Iwamatsu, A., Endo, T.A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., et al. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. Nature *450*, 908–912. https://doi.org/10.1038/NATURE06397.

Shechter, D., Dormann, H.L., Allis, C.D., and Hake, S.B. (2007). Extraction, purification and analysis of histones. Nature Protocols *2*, 1445–1457. https://doi.org/10.1038/nprot.2007.202.

Sheikh, B.N., Downer, N.L., Phipson, B., Vanyai, H.K., Kueh, A.J., McCarthy, D.J., Smyth, G.K., Thomas, T., and Voss, A.K. (2015). MOZ and BMI1 play opposing roles during Hox gene activation in ES cells and in body segment identity specification in vivo. Proc Natl Acad Sci U S A *112*, 5437–5442. https://doi.org/10.1073/pnas.1422872112.

Shemer, R., Hershko, A.Y., Perk, J., Mostoslavsky, R., Tsuberi, B.Z., Cedar, H., Buiting, K., and Razin, A. (2000). The imprinting box of the Prader-Willi/Angelman syndrome domain. Nature Genetics *26*, 440–443. https://doi.org/10.1038/82571.

Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R.A., and Issa, J.P.J. (2007). Genome-Wide Profiling of DNA Methylation Reveals a Class of Normally Methylated CpG Island Promoters. PLOS Genetics *3*, e181. https://doi.org/10.1371/JOURNAL.PGEN.0030181.

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

Shin, Y., and Brangwynne, C.P. (2017). Liquid phase condensation in cell physiology and disease. Science (1979) *357*. https://doi.org/10.1126/science.aaf4382.

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

Sibler, A.P., Kempf, E., Glacet, A., Orfanoudakis, G., Bourel, D., and Weiss, E. (1999). In vivo biotinylated recombinant antibodies: high efficiency of labelling and application to the cloning of active anti-human IgG1 Fab fragments. Journal of Immunological Methods *224*, 129–140. https://doi.org/10.1016/S0022-1759(99)00016-2.

Sidoli, S., Cheng, L., and Jensen, O.N. (2012). Proteomics in chromatin biology and epigenetics: Elucidation of post-translational modifications of histone proteins by mass spectrometry. Journal of Proteomics *75*, 3419–3433. https://doi.org/10.1016/j.jprot.2011.12.029.

Simic, R., Lindstrom, D.L., Tran, H.G., Roinick, K.L., Costa, P.J., Johnson, A.D., Hartzog, G.A., and Arndt, K.M. (2003). Chromatin remodeling protein Chd1 interacts with transcription elongation factors and localizes to transcribed genes. The EMBO Journal *22*, 1846–1856. https://doi.org/10.1093/EMBOJ/CDG179.

Sims, R.J., Chen, C.F., Santos-Rosa, H., Kouzarides, T., Patel, S.S., and Reinberg, D. (2005). Human but Not Yeast CHD1 Binds Directly and Selectively to Histone H3 Methylated at Lysine 4 via Its Tandem Chromodomains *. Journal of Biological Chemistry *280*, 41789–41792. https://doi.org/10.1074/JBC.C500395200.

Sirbu, B.M., Couch, F.B., Feigerle, J.T., Bhaskara, S., Hiebert, S.W., and Cortez, D. (2011). Analysis of protein dynamics at active, stalled, and collapsed replication forks. Genes and Development *25*, 1320–1327. https://doi.org/10.1101/gad.2053211.

Smith, Z.D., and Meissner, A. (2013). DNA methylation: roles in mammalian development. Nature Reviews Genetics *14*, 204–220. https://doi.org/10.1038/nrg3354.

Smith, P.A., Tripp, B.C., DiBlasio-Smith, E.A., Lu, Z., LaVallie, E.R., and McCoy, J.M. (1998). A plasmid expression system for quantitative in vivo biotinylation of thioredoxin fusion proteins in Escherichia coli. Nucleic Acids Research *26*, 1414–1420. https://doi.org/10.1093/NAR/26.6.1414.

Solage, A., and Cedar, H. (1978). Organization of 5-methylcytosine in chromosomal DNA. Biochemistry *17*, 2934–2938. https://doi.org/10.1021/BI00607A036.

Soldi, M., and Bonaldi, T. (2013). The proteomic investigation of chromatin functional domains reveals novel synergisms among distinct heterochromatin components. Molecular and Cellular Proteomics *12*, 764–780. https://doi.org/10.1074/mcp.M112.024307.

Song, J., Rechkoblit, O., Bestor, T.H., and Patel, D.J. (2011). Structure of DNMT1-DNA Complex Reveals a Role for Autoinhibition in Maintenance DNA Methylation. Science (1979) *331*.

de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., and Vogel, C. (2009). Global signatures of protein and mRNA expression levels. Molecular BioSystems *5*, 1512–1526. https://doi.org/10.1039/B908315D.

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

Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W.T.C., Bauer, C., Münzel, M., Wagner, M., Müller, M., Khan, F., et al. (2013a). Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. Cell *152*, 1146–1159. https://doi.org/10.1016/J.CELL.2013.02.004.

Spruijt, C.G., Gnerlich, F., Smits, A.H., Pfaffeneder, T., Jansen, P.W.T.C., Bauer, C., Münzel, M., Wagner, M., Müller, M., Khan, F., et al. (2013b). Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. Cell *152*, 1146–1159. https://doi.org/10.1016/j.cell.2013.02.004.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., Tiwari, V.K., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495. https://doi.org/10.1038/nature10716.

Stanton, B.Z., Hodges, C., Calarco, J.P., Braun, S.M.G., Ku, W.L., Kadoch, C., Zhao, K., and Crabtree, G.R. (2017). Smarca4 ATPase mutations disrupt direct eviction of PRC1 from chromatin. Nat Genet *49*, 282–288. https://doi.org/10.1038/NG.3735.

Steen, H., and Jensen, O.N. (2002). Analysis of protein-nucleic acid interactions by photochemical cross-linking and mass spectrometry. Mass Spectrometry Reviews *21*, 163–182. https://doi.org/10.1002/mas.10024.

Stein, R., Sciaky-Gallili, N., Razin, A., and Cedar, H. (1983). Pattern of methylation of two genes coding for housekeeping functions. Proc Natl Acad Sci U S A *80*, 2422–2426. https://doi.org/10.1073/PNAS.80.9.2422.

Stielow, B., Finkernagel, F., Stiewe, T., Nist, A., and Suske, G. (2018). MGA, L3MBTL2 and E2F6 determine genomic binding of the non-canonical Polycomb repressive complex PRC1.6. PLOS Genetics *14*, e1007193. https://doi.org/10.1371/journal.pgen.1007193.

Stielow, B., Zhou, Y., Cao, Y., Simon, C., Pogoda, H.-M., Jiang, J., Ren, Y., Phanor, S.K., Rohner, I., Nist, A., et al. (2021). The SAM domain-containing protein 1 (SAMD1) acts as a repressive chromatin regulator at unmethylated CpG islands. Science Advances *7*, eabf2229. https://doi.org/10.1126/sciadv.abf2229.

Suetake, I., Shinozaki, F., Miyagawa, J., Takeshima, H., and Tajima, S. (2004). DNMT3L stimulates the DNA methylation activity of Dnmt3a and Dnmt3b through a direct interaction. J Biol Chem *279*, 27816–27823. https://doi.org/10.1074/JBC.M400181200.

Sze, C.C., and Shilatifard, A. (2016). MLL3/MLL4/COMPASS Family on Epigenetic Regulation of Enhancer Function and Cancer. Cold Spring Harb Perspect Med *6*. https://doi.org/10.1101/CSHPERSPECT.A026427.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research *47*, D607–D613. https://doi.org/10.1093/nar/gky1131.

Szwagierczak, A., Bultmann, S., Schmidt, C.S., Spada, F., and Leonhardt, H. (2010). Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. Nucleic Acids Research *38*, e181–e181. https://doi.org/10.1093/nar/gkq684.

Taherbhoy, A.M., Huang, O.W., and Cochran, A.G. (2015). BMI1-RING1B is an autoinhibited RING E3 ubiquitin ligase. Nat Commun *6*. https://doi.org/10.1038/NCOMMS8621.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., et al. (2009). Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. Science (1979) *324*, 930–935. https://doi.org/10.1126/SCIENCE.1170116.

Takagaki, Y., Ryner, L.C., and Manley, J.L. (1989). Four factors are required for 3'-end cleavage of pre-mRNAs. Genes Dev *3*, 1711–1724. https://doi.org/10.1101/GAD.3.11.1711.

Takai, D., and Jones, P.A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A *99*, 3740–3745. https://doi.org/10.1073/pnas.052410099.

Talbert, P.B., and Henikoff, S. (2010). Histone variants — ancient wrap artists of the epigenome. Nature Reviews Molecular Cell Biology *11*, 264–275. https://doi.org/10.1038/nrm2861.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics *31*, 2032–2034. https://doi.org/10.1093/BIOINFORMATICS/BTV098.

Tatavosian, R., Kent, S., Brown, K., Yao, T., Duc, H.N., Huynh, T.N., Zhen, C.Y., Ma, B., Wang, H., and Ren, X. (2019b). Nuclear condensates of the Polycomb protein chromobox 2 (CBX2) assemble through phase separation. J Biol Chem *294*, 1451–1463. https://doi.org/10.1074/JBC.RA118.006620.

Tavares, L., Dimitrova, E., Oxley, D., Webster, J., Poot, R., Demmers, J., Bezstarosti, K., Taylor, S., Ura, H., Koide, H., et al. (2012). RYBP-PRC1 complexes mediate H2A ubiquitylation at polycomb target sites independently of PRC2 and H3K27me3. Cell *148*, 664–678. https://doi.org/10.1016/j.cell.2011.12.029.

Taylor, G.C.A., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M.M., and Bickmore, W.A. (2013). H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. Genome Research *23*, 2053–2065. https://doi.org/10.1101/gr.155028.113.

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

Teo, G., Liu, G., Zhang, J., Nesvizhskii, A.I., Gingras, A.-C., and Choi, H. (2014). SAINTexpress: Improvements and additional features in Significance Analysis of INTeractome software. Journal of Proteomics *100*, 37–43. https://doi.org/10.1016/j.jprot.2013.10.023.

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

Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics *14*, 178–192. https://doi.org/10.1093/BIB/BBS017.

Tie, F., Banerjee, R., Stratton, C.A., Prasad-Sinha, J., Stepanik, V., Zlobin, A., Diaz, M.O., Scacheri, P.C., and Harte, P.J. (2009). CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing. Development *136*, 3131–3141. https://doi.org/10.1242/dev.037127.

Torigoe, S.E., Urwin, D.L., Ishii, H., Smith, D.E., and Kadonaga, J.T. (2011). Identification of a rapidly formed nonnucleosomal histone-DNA intermediate that is converted into chromatin by ACF. Mol Cell *43*, 638–648. https://doi.org/10.1016/J.MOLCEL.2011.07.017.

Tremblay, K.D., Duran, K.L., and Bartolomei, M.S. (1997). A 5' 2-kilobase-pair region of the imprinted mouse H19 gene exhibits exclusive paternal methylation throughout development. Molecular and Cellular Biology *17*, 4322–4329. https://doi.org/10.1128/MCB.17.8.4322.

Tsumura, A., Hayakawa, T., Kumaki, Y., Takebayashi, S.I., Sakaue, M., Matsuoka, C., Shimotohno, K., Ishikawa, F., Li, E., Ueda, H.R., et al. (2006). Maintenance of self-renewal ability of mouse embryonic stem cells in the absence of DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Genes to Cells *11*, 805–814. https://doi.org/10.1111/J.1365-2443.2006.00984.X.

Tu, S., Teng, Y.C., Yuan, C., Wu, Y.T., Chan, M.Y., Cheng, A.N., Lin, P.H., Juan, L.J., and Tsai, M.D. (2008). The ARID domain of the H3K4 demethylase RBP2 binds to a DNA CCGCCC motif. Nature Structural and Molecular Biology *15*, 419–421. https://doi.org/10.1038/nsmb.1400.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M.Y., Geiger, T., Mann, M., and Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature Methods *13*, 731–740. https://doi.org/10.1038/nmeth.3901.

Tykocinski, M.L., and Max, E.E. (1984). CG dinucleotide clusters in MHC genes and in 5' demethylated genes. Nucleic Acids Res *12*, 4385–4396. .

Unoki, M., Funabiki, H., Velasco, G., Francastel, C., and Sasaki, H. (2019). CDCA7 and HELLS mutations undermine nonhomologous end joining in centromeric instability syndrome. J Clin Invest *129*. https://doi.org/10.1172/JCI99751.

Varga, J., Kube, M., Luck, K., and Schick, S. (2021). The BAF chromatin remodeling complexes: structure, function, and synthetic lethalities. Biochemical Society Transactions *49*, 1489–1503. https://doi.org/10.1042/BST20190960.

Varga-Weisz, P.D., Wilm, M., Bonte, E., Dumas, K., Mann, M., and Becker, P.B. (1997). Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. Nature *388*, 598–602. https://doi.org/10.1038/41587.

Varnaitė, R., and MacNeill, S.A. (2016). Meet the neighbors: Mapping local protein interactomes by proximity-dependent labeling with BioID. PROTEOMICS *16*, 2503–2518. https://doi.org/10.1002/pmic.201600123.

Venable, J.D., Dong, M.Q., Wohlschlegel, J., Dillin, A., and Yates, J.R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. Nature Methods *1*, 39–45. https://doi.org/10.1038/nmeth705.

Venolia, L., Gartler, S.M., Wassman, E.R., Yen, P., Mohandas, T., and Shapiro, L.J. (1982). Transformation with DNA from 5-azacytidine-reactivated X chromosomes. Proceedings of the National Academy of Sciences *79*, 2352–2354. https://doi.org/10.1073/PNAS.79.7.2352.

Vermeulen, M., and Déjardin, J. (2020). Locus-specific chromatin isolation. Nature Reviews Molecular Cell Biology *21*, 249–250. https://doi.org/10.1038/s41580-020-0217-0.

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

Villaseñor, R., Pfaendler, R., Ambrosi, C., Butz, S., Giuliani, S., Bryan, E., Sheahan, T.W., Gable, A.L., Schmolka, N., Manzo, M., et al. (2020). ChromID identifies the protein interactome at chromatin marks. Nature Biotechnology 1–9. https://doi.org/10.1038/s41587-020-0434-2.

Viturawong, T., Meissner, F., Butter, F., and Mann, M. (2013). A DNA-Centric Protein Interaction Map of Ultraconserved Elements Reveals Contribution of Transcription Factor Binding Hubs to Conservation. Cell Reports *5*, 531–545. https://doi.org/10.1016/j.celrep.2013.09.022.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nature Reviews Genetics *13*, 227–232. https://doi.org/10.1038/nrg3185.

Voigt, P., LeRoy, G., Drury, W.J., Zee, B.M., Son, J., Beck, D.B., Young, N.L., Garcia, B.A., and Reinberg, D. (2012). Asymmetrically Modified Nucleosomes. Cell *151*, 181–193. https://doi.org/10.1016/J.CELL.2012.09.002.

Voo, K.S., Carlone, D.L., Jacobsen, B.M., Flodin, A., and Skalnik, D.G. (2000a). 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. Mol Cell Biol *20*, 2108–2121. https://doi.org/10.1128/MCB.20.6.2108-2121.2000.

Wang, C., Lee, J.E., Lai, B., Macfarlan, T.S., Xu, S., Zhuang, L., Liu, C., Peng, W., and Ge, K. (2016). Enhancer priming by H3K4 methyltransferase MLL4 controls cell fate transition. Proc Natl Acad Sci U S A *113*, 11871–11876. https://doi.org/10.1073/pnas.1606857113.

Wang, H., Wang, L., Erdjument-Bromage, H., Vidal, M., Tempst, P., Jones, R.S., and Zhang, Y. (2004). Role of histone H2A ubiquitination in Polycomb silencing. Nature *431*, 873–878. https://doi.org/10.1038/NATURE02985.

Wang, H., Chang-Wong, T., Tang, H.Y., and Speicher, D.W. (2010). Comparison of extensive protein fractionation and repetitive LC-MS/MS analyses on depth of analysis for complex proteomes. Journal of Proteome Research *9*, 1032–1040. https://doi.org/10.1021/PR900927Y/.

Wang, L.H., Aberin, M.A.E., Wu, S., and Wang, S.P. (2021a). The MLL3/4 H3K4 methyltransferase complex in establishing an active enhancer landscape. Biochem Soc Trans *49*, 1041–1054. https://doi.org/10.1042/BST20191164.

Wang, P., Lin, C., Smith, E.R., Guo, H., Sanderson, B.W., Wu, M., Gogol, M., Alexander, T., Seidel, C., Wiedemann, L.M., 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*, 6074–6085. https://doi.org/10.1128/MCB.00924-09/.

Wang, S.P., Tang, Z., Chen, C.W., Shimada, M., Koche, R.P., Wang, L.H., Nakadai, T., Chramiec, A., Krivtsov, A. v., Armstrong, S.A., et al. (2017a). A UTX-MLL4-p300 Transcriptional Regulatory Network Coordinately Shapes Active Enhancer Landscapes for Eliciting Transcription. Molecular Cell *67*, 308-321.e6. https://doi.org/10.1016/J.MOLCEL.2017.06.028.

Wang, X., Paucek, R.D., Gooding, A.R., Brown, Z.Z., Ge, E.J., Muir, T.W., and Cech, T.R. (2017b). Molecular analysis of PRC2 recruitment to DNA in chromatin and its inhibition by RNA. Nat Struct Mol Biol *24*, 1028–1038. https://doi.org/10.1038/NSMB.3487.

Wang, X., Wang, H.Y., Hu, G.S., Tang, W.S., Weng, L., Zhang, Y., Guo, H., Yao, S.S., Liu, S.Y., Zhang, G.L., et al. (2021b). DDB1 binds histone reader BRWD3 to activate the transcriptional cascade in adipogenesis and promote onset of obesity. Cell Reports *35*, 109281. https://doi.org/10.1016/J.CELREP.2021.109281.

Wang, Y.L., Duttke, S.H.C., Chen, K., Johnston, J., Kassavetis, G.A., Zeitlinger, J., and Kadonaga, J.T. (2014). TRF2, but not TBP, mediates the transcription of ribosomal protein genes. Genes Dev *28*, 1550–1555. https://doi.org/10.1101/GAD.245662.114.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Peng, W., Zhang, M.Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nature Genetics *40*, 897–903. https://doi.org/10.1038/ng.154.

Wani, A.H., Boettiger, A.N., Schorderet, P., Ergun, A., Munger, C., Sadreyev, R.I., Zhuang, X., Kingston, R.E., and Francis, N.J. (2016). Chromatin topology is coupled to Polycomb group protein subnuclear organization. Nature Communications *7*, 1–13. https://doi.org/10.1038/ncomms10291.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520–562. https://doi.org/10.1038/nature01262.

Weaver, T.M., Liu, J., Connelly, K.E., Coble, C., Varzavand, K., Dykhuizen, E.C., and Musselman, C.A. (2019). The EZH2 SANT1 domain is a histone reader providing sensitivity to the modification state of the H4 tail. Sci Rep *9*. https://doi.org/10.1038/S41598-018-37699-W.

Wei, C., Xiao, R., Chen, L., Cui, H., Zhou, Y., Xue, Y., Hu, J., Zhou, B., Tsutsui, T., Qiu, J., et al. (2016). RBFox2 Binds Nascent RNA to Globally Regulate Polycomb Complex 2 Targeting in Mammalian Genomes. Mol Cell *62*, 875–889. https://doi.org/10.1016/J.MOLCEL.2016.04.013.

Whetstine, J.R., Nottke, A., Lan, F., Huarte, M., Smolikov, S., Chen, Z., Spooner, E., Li, E., Zhang, G., Colaiacovo, M., et al. (2006). Reversal of Histone Lysine Trimethylation by the JMJD2 Family of Histone Demethylases. Cell *125*, 467–481. https://doi.org/10.1016/J.CELL.2006.03.028.

Wickham, H. (2009). ggplot2 - Elegant Graphics for Data Analysis (New York: Springer).

Willforss, J., Chawade, A., and Levander, F. (2019). NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis. Journal of Proteome Research *18*, 732–740. https://doi.org/10.1021/acs.jproteome.8b00523.

Williams, K., Christensen, J., Pedersen, M.T., Johansen, J. v., Cloos, P.A.C., Rappsilber, J., and Helin, K. (2011). TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. Nature *473*, 343–349. https://doi.org/10.1038/nature10066.

Williams, L.H., Fromm, G., Gokey, N.G., Henriques, T., Muse, G.W., Burkholder, A., Fargo, D.C., Hu, G., and Adelman, K. (2015). Pausing of RNA Polymerase II Regulates Mammalian Developmental Potential through Control of Signaling Networks. Molecular Cell *58*, 311–322. https://doi.org/10.1016/J.MOLCEL.2015.02.003.

Wilson, B.C., Boehme, L., Annibali, A., Hodgkinson, A., Carroll, T.S., Oakey, R.J., and Seitan, V.C. (2020). Intellectual disability-associated factor Zbtb11 cooperates with NRF-2/GABP to control mitochondrial function. Nature Communications *11*, 1–20. https://doi.org/10.1038/s41467-020-19205-x.

Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nature Methods *6*, 359–362. https://doi.org/10.1038/nmeth.1322.

Wong, K.H., Jin, Y., and Struhl, K. (2014). TFIIH phosphorylation of the Pol II CTD stimulates mediator dissociation from the preinitiation complex and promoter escape. Mol Cell *54*, 601–612. https://doi.org/10.1016/J.MOLCEL.2014.03.024.

Wu, H., and Zhang, Y. (2014). Reversing DNA methylation: Mechanisms, genomics, and biological functions. Cell *156*, 45–68. https://doi.org/10.1016/j.cell.2013.12.019.

Wu, H., D'Alessio, A.C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Eve Sun, Y., and Zhang, Y. (2011a). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. Nature *473*, 389–394. https://doi.org/10.1038/nature09934.

Wu, J., Chen, Y., Lu, L.Y., Wu, Y., Paulsen, M.T., Ljungman, M., Ferguson, D.O., and Yu, X. (2011b). Chfr and RNF8 synergistically regulate ATM activation. Nature Structural and Molecular Biology *18*, 761–768. https://doi.org/10.1038/nsmb.2078.

Wu, M., Wang, P.F., Lee, J.S., Martin-Brown, S., Florens, L., Washburn, M., and Shilatifard, A. (2008). Molecular regulation of H3K4 trimethylation by Wdr82, a component of human Set1/COMPASS. Mol Cell Biol *28*, 7337–7344. https://doi.org/10.1128/MCB.00976-08.

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

Xiao, S., Xie, D., Cao, X., Yu, P., Xing, X., Chen, C.C., Musselman, M., Xie, M., West, F.D., Lewin, H.A., et al. (2012). Comparative Epigenomic Annotation of Regulatory DNA. Cell *149*, 1381–1392. https://doi.org/10.1016/J.CELL.2012.04.029.

Xie, S., Wang, Z., Okano, M., Nogami, M., Li, Y., He, W.W., Okumura, K., and Li, E. (1999). Cloning, expression and chromosome locations of the human DNMT3 gene family. Gene *236*, 87–95. https://doi.org/10.1016/S0378-1119(99)00252-8.

Xie, W., Schultz, M.D., Lister, R., Hou, Z., Rajagopal, N., Ray, P., Whitaker, J.W., Tian, S., Hawkins, R.D., Leung, D., et al. (2013). Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. Cell *153*, 1134–1148. https://doi.org/10.1016/J.CELL.2013.04.022.

Xiong, J., Zhang, Z., Chen, J., Huang, H., Xu, Y., Ding, X., Zheng, Y., Nishinakamura, R., Xu, G.-L., Wang, H., et al. (2016). Cooperative Action between SALL4A and TET Proteins in Stepwise Oxidation of 5-Methylcytosine. Molecular Cell *64*, 913–925. https://doi.org/10.1016/j.molcel.2016.10.013.

Xu, C., Bian, C., Lam, R., Dong, A., and Min, J. (2011). The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. Nature Communications *2*, 227–228. https://doi.org/10.1038/ncomms1237.

Xu, C., Liu, K., Lei, M., Yang, A., Li, Y., Hughes, T.R., and Min, J. (2018). DNA Sequence Recognition of Human CXXC Domains and Their Structural Determinants. Structure *26*, 85–95. https://doi.org/10.1016/j.str.2017.11.022.

Xu, T.H., Liu, M., Zhou, X.E., Liang, G., Zhao, G., Xu, H.E., Melcher, K., and Jones, P.A. (2020). Structure of nucleosome-bound DNA methyltransferases DNMT3A and DNMT3B. Nature *586*, 151–155. https://doi.org/10.1038/s41586-020-2747-1.

Xu, Y., Xu, C., Kato, A., Tempel, W., Abreu, J.G., Bian, C., Hu, Y., Hu, D., Zhao, B., Cerovina, T., et al. (2012). Tet3 CXXC Domain and Dioxygenase Activity Cooperatively Regulate Key Genes for Xenopus Eye and Neural Development. Cell *151*, 1200–1213. https://doi.org/10.1016/J.CELL.2012.11.014.

Yagi, H., Deguchi, K., Aono, A., Tani, Y., Kishimoto, T., and Komori, T. (1998). Growth Disturbance in Fetal Liver Hematopoiesis of Mll-Mutant Mice. Blood *92*, 108–117. https://doi.org/10.1182/BLOOD.V92.1.108.413K11_108_117.

Yamaguchi, S., Hong, K., Liu, R., Shen, L., Inoue, A., Diep, D., Zhang, K., and Zhang, Y. (2012). Tet1 controls meiosis by regulating meiotic gene expression. Nature *492*, 443–447. https://doi.org/10.1038/nature11709.

Yamaguchi, Y., Shibata, H., and Handa, H. (2013). Transcription elongation factors DSIF and NELF: promoter-proximal pausing and beyond. Biochim Biophys Acta *1829*, 98–104. https://doi.org/10.1016/J.BBAGRM.2012.11.007.

Yan, P.S., Rodriguez, F.J., Laux, D.E., Perry, M.R., Standiford, S.B., and Huang, T.H.M. (2000). Hypermethylation of ribosomal DNA in human breast carcinoma. British Journal of Cancer 2000 82:3 *82*, 514–517. https://doi.org/10.1054/bjoc.1999.0955.

Yang, X.J., and Seto, E. (2008). The Rpd3/Hda1 family of lysine deacetylases: From bacteria and yeast to mice and men. Nature Reviews Molecular Cell Biology *9*, 206–218. https://doi.org/10.1038/nrm2346.

Yang, C., Bolotin, E., Jiang, T., Sladek, F.M., and Martinez, E. (2007a). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. Gene *389*, 52–65. https://doi.org/10.1016/J.GENE.2006.09.029.

Yang, J.G., Madrid, T.S., Sevastopoulos, E., and Narlikar, G.J. (2006). The chromatin-remodeling enzyme ACF is an ATP-dependent DNA length sensor that regulates nucleosome spacing. Nat Struct Mol Biol *13*, 1078–1083. https://doi.org/10.1038/NSMB1170.

Yang, W., Xia, Y., Hawke, D., Li, X., Liang, J., Xing, D., Aldape, K., Hunter, T., Alfred Yung, W.K., and Lu, Z. (2012). PKM2 Phosphorylates Histone H3 and Promotes Gene Transcription and Tumorigenesis. Cell *150*, 685–696. https://doi.org/10.1016/J.CELL.2012.07.018.

Yang, X., Zaurin, R., Beato, M., and Peterson, C.L. (2007b). Swi3p controls SWI/SNF assembly and ATP-dependent H2A-H2B displacement. Nat Struct Mol Biol *14*, 540–547. https://doi.org/10.1038/NSMB1238.

Yang, Z., Yik, J.H.N., Chen, R., He, N., Moon, K.J., Ozato, K., and Zhou, Q. (2005). Recruitment of P-TEFb for Stimulation of Transcriptional Elongation by the Bromodomain Protein Brd4. Molecular Cell *19*, 535–545. https://doi.org/10.1016/J.MOLCEL.2005.06.029.

Yen, K., Vinayachandran, V., and Pugh, B.F. (2013). SWR-C and INO80 chromatin remodelers recognize nucleosome-free regions near +1 nucleosomes. Cell *154*, 1246. https://doi.org/10.1016/J.CELL.2013.08.043.

Yen, R. whay C., Vertino, P.M., Nelkin, B.D., Yu, J.J., El-deiry, W., Cumaraswamy, A., Lennon, G.G., Trask, B.J., Celano, P., and Baylin, S.B. (1992). Isolation and characterization of the cDNA encoding human DNA methyltransferase. Nucleic Acids Res *20*, 2287–2291. https://doi.org/10.1093/NAR/20.9.2287.

Yu, B.D., Hess, J.L., Horning, S.E., Brown, G.A.J., and Korsmeyer, S.J. (1995). Altered Hox expression and segmental identity in Mll-mutant mice. Nature *378*, 505–508. https://doi.org/10.1038/378505A0.

Yu, B.D., Hanson, R.D., Hess, J.L., Horning, S.E., and Korsmeyer, S.J. (1998). MLL, a mammalian trithorax-group gene, functions as a transcriptional maintenance factor in morphogenesis. Proc Natl Acad Sci U S A *95*, 10632–10636. https://doi.org/10.1073/pnas.95.18.10632.

Yu, C., Fan, X., Sha, Q.Q., Wang, H.H., Li, B.T., Dai, X.X., Shen, L., Liu, J., Wang, L., Liu, K., et al. (2017). CFP1 Regulates Histone H3K4 Trimethylation and Developmental Potential in Mouse Oocytes. Cell Reports *20*, 1161–1172. https://doi.org/10.1016/J.CELREP.2017.07.011.

Yuan, W., Wu, T., Fu, H., Dai, C., Wu, H., Liu, N., Li, X., Xu, M., Zhang, Z., Niu, T., et al. (2012). Dense chromatin activates polycomb repressive complex 2 to regulate H3 lysine 27 methylation. Science (1979) *337*, 971–975. https://doi.org/10.1126/science.1225237.

Zegerman, P., Canas, B., Pappin, D., and Kouzarides, T. (2002). Histone H3 Lysine 4 Methylation Disrupts Binding of Nucleosome Remodeling and Deacetylase (NuRD) Repressor Complex *. Journal of Biological Chemistry *277*, 11621–11624. https://doi.org/10.1074/JBC.C200045200.

Zeng, P.-Y., Vakoc, C.R., Chen, Z.-C., Blobel, G.A., and Berger, S.L. (2006). In vivo dual cross-linking for identification of indirect DNA-associated proteins by chromatin immunoprecipitation. Biotechniques *41*, 694–698. https://doi.org/10.2144/000112297.

Zhang, J., Bonasio, R., Strino, F., Kluger, Y., Holloway, J.K., Modzelewski, A.J., Cohen, P.E., and Reinberg, D. (2013). SFMBT1 functions with LSD1 to regulate expression of canonical histone genes and chromatin-related factors. Genes and Development *27*, 749–766. https://doi.org/10.1101/gad.210963.112.

Zhang, L., Zhang, K., Prändl, R., and Schöffl, F. (2004). Detecting DNA-binding of proteins in vivo by UV-crosslinking and immunoprecipitation. Biochemical and Biophysical Research Communications *322*, 705–711. https://doi.org/10.1016/J.BBRC.2004.07.202.

Zhang, Q., McKenzie, N.J., Warneford-Thomson, R., Gail, E.H., Flanigan, S.F., Owen, B.M., Lauman, R., Levina, V., Garcia, B.A., Schittenhelm, R.B., et al. (2019a). RNA exploits an exposed regulatory site to inhibit the enzymatic activity of PRC2. Nat Struct Mol Biol *26*, 237–247. https://doi.org/10.1038/S41594-019-0197-Y.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biology *9*, 1–9. https://doi.org/10.1186/gb-2008-9-9-r137.

Zhang, Z., Wu, S., Stenoien, D.L., and Paša-Tolić, L. (2014). High-throughput proteomics. Annual Review of Analytical Chemistry *7*, 427–454. https://doi.org/10.1146/annurev-anchem-071213-020216.

Zhang, Z., Jones, A.E., Wu, W., Kim, J., Kang, Y., Bi, X., Gu, Y., Popov, I.K., Renfrow, M.B., Vassylyeva, M.N., et al. (2017). Role of remodeling and spacing factor 1 in histone H2A ubiquitination-mediated gene silencing. Proc Natl Acad Sci U S A *114*, E7949–E7958. https://doi.org/10.1073/pnas.1711158114.

Zhang, Z., Deng, X., Liu, Y., Liu, Y., Sun, L., and Chen, F. (2019b). PKM2, function and expression and regulation. Cell and Bioscience *9*, 1–25. https://doi.org/10.1186/s13578-019-0317-8.

Zhao, J., Wang, M., Chang, L., Yu, J., Song, A., Liu, C., Huang, W., Zhang, T., Wu, X., Shen, X., et al. (2020). RYBP/YAF2-PRC1 complexes and histone H1-dependent chromatin compaction mediate propagation of H2AK119ub1 during cell division. Nat Cell Biol *22*, 439–452. https://doi.org/10.1038/S41556-020-0484-1.

Zhou, J.C., Blackledge, N.P., Farcas, A.M., and Klose, R.J. (2012). Recognition of CpG island chromatin by KDM2A requires direct and specific interaction with linker DNA. Mol Cell Biol *32*, 479–489. https://doi.org/10.1128/MCB.06332-11.

Zhou, K., Gaullier, G., and Luger, K. (2019). Nucleosome structure and dynamics are coming of age. Nature Structural and Molecular Biology *26*, 3–13. https://doi.org/10.1038/s41594-018-0166-x.

Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nature Reviews Genetics *12*, 7–18. https://doi.org/10.1038/nrg2905.

Zhu, H., Geiman, T.M., Xi, S., Jiang, Q., Schmidtmann, A., Chen, T., Li, E., and Muegge, K. (2006). Lsh is involved in de novo methylation of DNA. EMBO J *25*, 335–345. https://doi.org/10.1038/SJ.EMBOJ.7600925.

Zhu, L.J., Gazin, C., Lawson, N.D., Pagès, H., Lin, S.M., Lapointe, D.S., and Green, M.R. (2010). ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics *11*, 1–10. https://doi.org/10.1186/1471-2105-11-237.

Zylber, E.A., and Penman, S. (1971). Products of RNA Polymerases in HeLa Cell Nuclei. Proceedings of the National Academy of Sciences *68*, 2861–2865. https://doi.org/10.1073/PNAS.68.11.2861.