Exploration into the MLL4/WRAD enzyme-substrate network:
Systematic identification of CFP1 as a non-histone substrate of
the MLL4 lysine methyltransferase

by

Ryan Collins

A thesis submitted to the Faculty of Graduate Studies and Research in partial
fulfillment of the requirements for the degree of

Master of Science

in

Biology

Carleton University

Ottawa, Ontario

© 2019

Ryan Collins

Exploration into the MLL4/WRAD enzyme-substrate network: Systematic identification of CFP1 as a non-histone substrate of the MLL4 lysine methyltransferase

Submitted by

Ryan Collins

In partial fulfillment of the requirements for the degree of Master of Science

Chair, Department of Biology _____

Thesis Supervisor _____

Carleton University

Abstract

Histone lysine methyltransferases (KMTs) are key actors in the regulation of the cell's most critical functions, including but by no means limited to gene expression, DNA damage repair, and cell differentiation. Histone KMTs impart control over these processes by remodelling chromatin through the modification of histone N-tail lysine residues. While historically studied exclusively within the context of histones, emerging evidence points to a broader role for histone KMTs, specifically through the lysine methylation of non-histone proteins. In this thesis, I investigate non-histone lysine methylation activity of the KMT2-class enzyme MLL4, characterized by its monomethylation activity on histone H3K4 at distal enhancer regions and devoid of other known substrates. Through tandem *shotgun* and *systematic* approaches, I identify the histone KMT-associated regulatory protein CXXC finger protein 1 as a novel substrate of MLL4, and conduct *in vitro* investigation of the potential functional consequences of this methylation event.

Acknowledgments

First and foremost, I would like to thank my advisor Dr. Kyle Biggar for providing me with constant support, guidance, and encouragement, without which I would not have been able to achieve this accomplishment. Thank you to the Couture lab at uOttawa for providing me with advice and plasmids used in these experiments. Thank you to Anand, Hemanta, Matt, and all of the other Biggar lab members. This work could not have been completed without their support in the conduct and troubleshooting of experiments, nor without the welcoming and fun working environment they helped create. I am grateful for the love, support and encouragement of my family and friends throughout my studies. I would like to thank my girlfriend Mireille for her constant support, and for her endless patience throughout the many challenges of grad school, be they long nights in the lab or long days spent writing. Finally, I am incredibly grateful for the opportunity to have worked in the Biggar lab. I can't imagine any other workplace that could facilitate the tremendous personal and professional growth I have experienced during my time here.

Table of Contents

	Page
Title Page	i
Acceptance Sheet	ii
Abstract	iii
Acknowledgements	iv
Table of Contents	v
List of Abbreviations	vi
List of Figures	ix
List of Tables	x
Chapter 1: General Introduction	1
Chapter 2: Divergent Approaches in Identifying Non-Histone Substrates of the Histone H3K4 Methyltransferase MLL4	22
Chapter 3: Histone H3K4 Methyltransferase MLL4 Methylates CXXC Finger Protein-1 Within a Lysine Rich Regulatory Domain	55
Chapter 4: MLL4-dependant Methylation of CXXC Finger Protein-1 Potentiates Interaction with Setd1A and May Control Interaction with MLL4 PHD Domains	88
Chapter 5: General Discussion	110

List of Abbreviations

ARHGAP32 – Rho GTPase activating protein 32
BSA – Bovine serum albumin
CFP1 – CXXC finger protein 1
CFP1me – Methylated CXXC finger protein 1
CGI – CpG island
COMPASS – Complex of proteins associated with Set1
DCM – dichloromethane
DDR – DNA-damage response
DMF – dimethylformide
DNMT1 – DNA methyltransferase 1
DSB – Double-strand break
DTT – Dithiothreitol
EDTA – Ethylenediaminetetraacetic acid
GC – Gene conversion
GST – Glutathione S-transferase
H2AK15ub – Histone H2A lysine 15 ubiquitinated
H3K4 – Histone H3 lysine 4
H3K4me1/2/3 – Histone H3 lysine 4 mono/di/trimethyl
H3K14ac – Histone H3 lysine 14 acetyl
H4K20me2 - Histone H4 lysine 20 dimethyl
HDR – Homology-directed repair
HKMT – Histone lysine methyltransferase
HRP – Horseradish peroxidase
HSPA8 – Heat Shock Protein Family A (Hsp70) Member 8
IPTG – Isopropyl β -D-1-thiogalactopyranoside
K328 – Lysine 328

KMT – Lysine methyltransferase
Kx – Lysine x, where x is the amino acid position
LB – Lysogeny broth
MAP3K2 – Mitogen-Activated Protein Kinase Kinase Kinase 2
Me0, me1, me2, me3 – Nil methyl, monomethyl, dimethyl, trimethyl
MLL – Mixed lineage leukemia
NHEJ – Non-homologous end joining
Ni-NTA – Nickle nitrilotriacetic acid
NUFIP2 – Nuclear FMR1 Interacting Protein 2
PAGE – Polyacrylamide gel electrophoresis
PBS – Phosphate-buffered saline
PHD – Plant homeodomain
PMSF – Phenylmethylsulfonyl fluoride
PPI – Protein-protein interaction
PTM – Post-translational modification
PVDF – polyvinylidene difluoride
SAM – S-adenosyl-L-methionine
SDS-PAGE – sodium dodecyl sulfate polyacrylamide gel electrophoresis
SET – Suppressor of variegation 3-9, Enhancer-of-zeste and Trithorax
SID – Setd1-interacting domain
SNW1 – SNW Domain Containing 1
SSA – Single-strand annealing
TBS – Tris-buffered saline
TBS-T – Tris-buffered saline, Tween 20
TFA – Trifluoroacetic acid
TIPS – Triisopropylsilane
TIRR – Tudor-interacting repair regulator
TP53BP1 or 53BP1 – Tumor suppressor p53 binding protein 1

TTD – Tandem Tudor domain

WRAD – WDR5, Ash2L, RbBP5, Dpy-30

List of Figures

#	Title	Page
1.1	The evolutionary expansion of H3K4 methyltransferases is accompanied by increasing domain architecture complexity.	16
2.1	“Shotgun” style approach to identifying new substrate interactions with MLL4	39
2.2	Ribbon cartoon depicting putative methyllysine site at K1563 within the Tudor 2 domain of the 53BP1 tandem Tudor domains.	40
2.3	Systematic identification of the MLL4 recognition motif using a solid-phase SPOT peptide array printed on aminated cellulose	41
2.4	2D and 3D representations of MLL3 ligand-binding pocket interaction	42
2.5	Quantification of known methyllysine proteins containing some or all of our putative MLL4 recognition motif	44
2.6	Domains and regions of CFP1	45
3.1	MLL4 WRAD methylates CFP1 (302-656) but not 53BP1 TTD in a bioluminescence-based end-point methyltransferase assay	73
3.2	Methylation window peptide array shows extensive MLL4 WRAD-dependent methylation of CFP but no methylation of 53BP1 TTD	74
3.3	The methylated lysines identified by the methylation window peptide array span the majority of the basic region and are directly adjacent to the Setd1 interaction domain (SID)	75
3.4	The MLL4 recognition motif poorly aligns with CFP1 methylation window sequences that underwent <i>in vitro</i> MLL4 WRAD-dependent methylation on a SPOT peptide array	76
3.5	Cladogram illustrating the evolutionary relatedness of CFP1 in <i>H. sapiens</i> (Human), <i>M. Musculus</i> (Mouse) <i>A. carolinensis</i> (Lizard) <i>D. rerio</i> (Fish) <i>D. melanogaster</i> (Fly) <i>C. elegans</i> (Worm), and <i>S. cerevisiae</i> (Yeast).	79
3.6	Sequence alignment of DNMT1 regions that facilitate interaction with CFP1	80
4.1	Setd1A pulldown assay	104
4.2	MLL4 PHD domain binding on CFP1 methyllysine peptide array	106
5.1	Figure 4A and B from Butler et al. 2008	125
5.2	Figure 5 from Butler et al. 2008	126
S.1	Purified proteins used in this study	127

List of Tables

#	Title	Page
2.1	MLL4 WRAD motif perfect match sequences in known methyllysine modified proteins.	46
2.2	MLL4 WRAD motif perfect match sequences in known methyllysine modified residues.	47

Chapter 1:

General Introduction

The central dogma of molecular biology states that genetic information is stored in DNA. As needed, this DNA is transcribed into RNA, which is in turn translated into proteins. Proteins – biological structures with emergent physical properties conferred by the arrangement of the amino acids with which they are constructed – are the main drivers of biological processes and comprise the majority of cell machinery. The large amount of data produced by many genome projects and decades of computational analysis of those genomic data provided us with a relatively complete set of genes and the proteins they encode. Despite our improved understanding in the genetic basis of cell function following the decoding of the human genome, we are still faced with a transitional gap in our knowledge of protein function (i.e., functional proteomics). The disparity between the enormous amount of readily available genomic data and our current understanding of how proteins ‘work’ in the cell calls for innovative strategies for proteomics research and the elucidation of protein function. While genomic sequencing can identify gene alterations with a predisposition towards a particular cellular phenotype, it is the function and connectivity of the encoded proteins that are the real drivers of cell function, and hence knowledge of how this function and connectivity occurs is of critical importance.

1.1 Regulation of protein function through dynamic post-translational modifications

Protein activity is complex. Beyond their discrete functionality, interactions between individual proteins synergistically enable various biological processes and pathways in a hierarchical fashion. The primary sequence of the protein, which is a linear sequence of combinatorial possibilities among 20 amino acids, dictates the folding of the protein into its 3-D structure. The resultant 3-D structure arranges groups of amino acids with particular physicochemical properties into functional regions called domains. The functional properties and

architectural arrangement of domains within a protein dictate with which small molecules and other proteins it is able to interact. This interaction network is further specified by the cellular environment in which a protein is localized, i.e. proteins that only occur in the nucleus can likewise only interact with small molecules and proteins also found in the nucleus. In some cases, this physical association facilitates the formation of stoichiometrically stable multimeric protein complexes that possess functionally unique to that of any of its constituents. Furthermore, these complexes may interact with additional individual proteins or with other complexes to form modules that likewise possess or otherwise enable distinct functionality. In this synergistic hierarchical manner, a limited number of proteins can dynamically interact to form myriad molecular species, enabling exponentially diverse modes of cellular function. These interactions are typically referred to as ‘protein-protein interactions’ (PPIs). The importance of PPIs cannot be understated – they are the arbiters of vast cellular communication networks and help control some of the most critical cellular processes, from gene expression to cell division.

A number of different mechanisms can act to regulate PPIs, such as (1) protein expression and (2) the availability of co-factors. However, these interactions can also be dynamically modulated through a reversible process called post-translational modification (PTM), providing the cell with a rapid means of mediating PPIs in response to environmental cues (i.e., cell stress, growth factor stimulation, etc.). As a result, PTMs can act to increase the complexity of the possible network of functional proteins. A typical PTM involves the addition or removal of small chemical moieties to specific amino acid residues within the structure of a protein. The presence (or absence) of a PTM can then act to alter the structure or physical characteristics of a protein, in turn directly altering the function of a protein or perhaps regulating its interaction within its PPI network [1].

There are several species of PTM that exist in the cell, most notably this includes phosphorylation, ubiquitination, acetylation, and methylation. Each type of modification can elicit different effects depending on the biological context. Protein lysine methylation was discovered in the mid-20th century [2]. Although this new discovery led to an initial surge in interest, focus on protein methylation waned for a number of decades in the shadow of a growing body of work on other PTMs, such as tyrosine phosphorylation. Indeed, while the discovery of lysine methylation predated that of tyrosine phosphorylation by two decades, the wealth of research regarding the former is dwarfed by that of the latter. However, developments in the field of protein lysine methylation in recent years have revealed a plurality of methyllysine-mediated biological processes [3, 4], in turn revealing a functional role for lysine methylation of similar importance to that of other PTMs. The dynamic regulation of protein function through lysine methylation will be discussed in further detail below.

1.2 Lysine methylation of histone proteins

Historically methylation has been viewed as a histone-specific modification. Histones are proteins which organize genomic DNA into a structure called a nucleosome – in a nucleosome DNA is coiled around histone proteins similarly to “beads on a string” [5]. A combination of multiple histone PTMs (including phosphorylation, acetylation, and ubiquitination in addition to methylation) are involved in the regulation of chromatin structure and gene expression, collectively representing what is known as the “histone code” [reviewed in 6]. These histone modifications ultimately modify surrounding chromatin structure and can either sequester or expose DNA to transcription factors depending on the specific modification. Through this mechanism PTMs can be described as key regulators of gene expression.

Protein methylation commonly occurs on lysine and arginine residues, but has also been observed to occur on histidine and cysteine residues. Lysine residues can be methylated up to three times at the ϵ -amino group, for a total of four possible methylation states (unmethylated, or mono-, di-, and trimethylated) [7]. The lysine-specific ‘histone code’ is known to be highly dynamic. On lysine residues, the methylation state can be either increased through the addition of methyl groups by lysine methyltransferase (KMT) enzymes or decreased through removal of these methyl groups by lysine demethylase (KDM) enzymes. As such, both KMT and KDM enzymes are sometimes referred to as “writer” and “eraser” proteins, respectively, that are able to act in conjunction with “reader” proteins that recognize and bind with specific methylation sites and modification states through the presence of a methyl-binding domain (MBD). This dynamic interplay has been referred to as the ‘writer-reader-eraser’ model [8].

Dynamic protein lysine methylation has been classically observed and studied on histone proteins, owing to its size, prevalence and importance in packaging eukaryotic DNA into chromatin. Histone lysine modification can contribute to both positive and negative regulation of gene expression, depending on the degree of methylation and the specific lysine residue modified. For example, methylation of H3K4 and H3K79 is generally associated with increased transcriptional activity, whereas H3K9 and H3K27 are often associated with gene inactivation [9, 10, 11]. Additionally, lysine modifications can be localized depending on genomic context. To illustrate, monomethylation of Lys 4 of histone H3 (i.e., H3K4me) is primarily associated with distal enhancer regions, whereas di- and trimethylation of H3K4 are primarily associated with sites of active transcription [12]. Histone methylation marks are also linked to double strand break (DSB) signalling and the DNA damage response. In one such case, the binding of tumor suppressor p53 binding protein 1 (TP53BP1) to di-methylated histone H4 at lysine 20 (H4K20me2) influences

DSB pathway repair choice while simultaneously recruiting other DSB-associated repair machinery to the break site [13]. Crucially, interaction between 53BP1 and the DSB-adjacent histone H4 is dependent on the di-methylation state of H4K20 and is facilitated by the tandem Tudor MBDS present within 53BP1. In this model, 53BP1 is a classic example of a methyllysine ‘reader’ protein.

1.3 Non-histone protein lysine methylation plays multiple regulatory roles

Recent studies have revealed that lysine methylation is a dynamic and important PTM with implications in biological processes well beyond that those resulting from histone modification [3]. While the methylation of lysine insignificantly affects charge, it reduces both hydration energy and hydrogen bonding potential of the modified residue [14]. Indeed, methyl group substitution of Nε hydrogen atoms slightly increases the size and hydrophobicity of the nitrogen moiety, while the loss of hydrogen atoms limits hydrogen bonding such that me3 modification is characterized by a total loss of hydrogen bond participation [15]. This modification of biophysical characteristics conveys a high degree of specificity towards MBD-driven methyllysine interactions and also provides a mechanism by which lysine methylation can disrupt PPIs. As such, non-histone protein lysine methylation is diverse in regulatory function. Regulatory function is achieved either directly through modification of the activity and function of the modified protein, or indirectly through enabling and inhibiting PPIs.

As a well-studied example, the tumor suppressor protein p53 undergoes a high degree of dynamic lysine methylation at several residues within its carboxy-terminal regulatory domain [16]. Di-methylation of p53 at lysine 370 (p53K370me2) by a yet unknown KMT facilitates an

interaction with 53BP1 that is driven by its tandem Tudor MBDs. This p53(K370me2)-53BP1 interaction enables p53-promoter interaction and ultimately aids control of anti-tumorigenic cell fate decisions [17]. Lysine-specific demethylase 1 (LSD1), was found to demethylate the dimethylation modification at Lys 370 (i.e., loss of p53K370me2) thereby disrupting the methyl-reader abilities of the Tudor domains of 53BP1. This ultimately results in the repression of p53 transcriptional activity. In contrast, the mono-methylation of p53K382 (p53K382me1) is associated with a repression of p53-dependant transcriptional activation in cancer [18]. This illustrates the possible combinatorial complexity that exists within the non-histone methyllysine proteome, where methylation events in proximate residues can have opposing regulatory outcomes on protein function.

Lysine methylation modifications can interact with and regulate other PTMs in a process referred to as ‘cross-talk’. For example, the receptor tyrosine kinase vascular endothelial growth factor receptor 1 (VEGFR1) is able to regulate signal transduction events that are associated with cell proliferation and angiogenesis [19]. Methylation of lysine 831 (VEGFR1-K831me) within its tyrosine kinase domain is thought to enhance kinase activity, indirectly linking lysine methylation with increases in tyrosine phosphorylation of VEGFR1 targets [20].

The lysine methylation of DNA-methyltransferase 1 (DNMT1) at lysine 142 is associated with the inactive form of the enzyme, whereas phosphorylation of serine 143 is associated with the active form. Interestingly, these modifications are mutually exclusive, demonstrating how different PTMs can function as a ‘switch’ to regulate protein activity [21]. A similar mechanism is seen in the MAP kinase signalling pathway member MAP3K2, where methylation at lysine 260 by the KMT SMYD3 (SET and MYND domain containing protein 3) potentiates pathway activity in cancer cells, promoting tumorigenesis [4]. Methylation at K260 of MAP3K2 inhibits interaction

with PP2A phosphatase complex, a negative regulator of MAP kinase signalling. Here, we observe the presence of a methyllysine PTM indirectly upregulating MAPK-dependant phosphorylation through modulating key protein-protein interactions.

1.4 Histone H3 lysine 4 methyltransferases

Lysine methylation is performed by enzymes known as KMTs. These KMTs are able to methylate their substrates through a reaction involving S-adenosyl L-methionine (SAM), which functions as the methyl-group donor [7]. Many KMTs were initially thought to be histone-specific, though it is now known that histone-associated KMTs are responsible for the methylation of some non-histone protein substrates [22].

With the sole exception of the DOT1L family of histone H3 lysine-79 (H3K79) methyltransferases, SET (Suppressor of variegation, Enhancer of zeste, Trithorax) domains confer the enzymatic activity that accomplishes all known lysine methylation events [14]. Notably, Suppressor of variegation, Enhancer of zeste and Trithorax are each genes in *Drosophila* possessing this conserved catalytic domain. To date, the list of SET domain-containing KMTs is still growing but is currently divided into eight classes of enzymes based on their histone-substrate specificity, KMT1 through KMT8 [23]. The KMT2 class contains the histone H3 lysine 4 (H3K4) methyltransferases.

Histone H3K4 methylation was first observed following identification of Set1 in *Saccharomyces cerevisiae* [23]. In this model, Set1 is capable of mono-, di- and tri-methylation of H3K4. Indeed, Set1 orthologs are present in all subsequent orders of life, all the way to humans (i.e., KMT2-class enzymes). *Drosophila* has four Set1 paralogs: dSet1, trithorax (trx), trithorax-

related (*trr*), and *Ash1*, each capable of H3K4 methylation [24]. These proteins make up part of an enzymatic family called trithorax group (TrxG). Early developmental biology research commonly utilized *Drosophila*, and it was here that role of TrxG proteins in epigenetic inheritance was first elucidated. TrxG proteins were shown to maintain an active transcriptional state in homeobox (HOX) genes during embryogenesis, after maternal and zygotic transcription factors are no longer able to control transcriptional memory in the developing organism [25]. Trithorax is named for the striking physiological effects of *trx* mutation in *Drosophila* embryos, where HOX dysregulation results in the transformation of posterior embryonic segments into anterior segments. Later studies revealed the functional diversity of TrxG proteins beyond HOX regulation, including roles in cell division, X chromosome inactivation, and tumorigenesis, ultimately revealing the ubiquity of H3K4 methylation-dependant cellular processes [reviewed in 26]. RNAi-mediated knockdown of *Drosophila* *dSet1*, *trx*, or *trr* is embryonic lethal, while *dSet1* depletion results in global H3K4 di- and trimethylation deficiency [27, 28]. This illustrates the requirement of H3K4 methylation activity in development, as well as the non-redundancy of *Set1* paralogs.

The functional diversity discovered in *Drosophila* TrxG proteins has continued to expand through evolutionary time. Indeed, each of the *Set1*-related TrxG proteins (*dSet1*, *trx*, *trr*) has two paralogs in mammals. The mammalian HKMTs *Setd1A* and *Setd1B* are most closely related to *Drosophila* *dSet1*, *MLL1* (KMT2A) and *MLL2* (KMT2B) to *trx*, and *MLL3* (KMT2C) and *MLL4* (KMT2D) to *trr*. Later discussed below, all six of these core catalytic KMTs assemble into protein complexes with distinct members, further modifying function and specificity.

The six mammalian *Set1* paralogs have both redundant and non-redundant functions. For example, both *Setd1A* and *Setd1B* together appear to be responsible for the majority of H3K4 di- and trimethylation, illustrated by the global reduction of H3K4me₃ observed in simultaneous

knockdowns, and knockdowns of complex members unique to the Setd1A/B COMPASS complex [29, 30, 31]. However, the Setd1 proteins also exhibit non-redundant function. Phenotypes of individual Setd1 knockouts differ – both are embryonic lethal, but Setd1A knockouts fail to gastrulate while Setd1B knockouts die at embryonic day (E)11.5 [32]. Additionally, Setd1A alone is critical for embryonic stem cell proliferation.

Redundant and non-redundant functions are also observed in the MLL family (MLL1-MLL4, or KMT2A-KMT2D) of H3K4 methyltransferases. This is discussed in detail below.

1.5 H3K4 methyltransferase associate with multimeric protein complexes

In yeast, Set1 is not active independently. It only gains activity upon associating with the multimeric protein complex COMPASS (COMplex of Proteins Associated with Set1). Each subunit of COMPASS plays a specific role in the biological function of Set1 as a H3K4 methyltransferase, from complex assembly to the direct regulation of KMT activity [23]. Though the requirement of a COMPASS-like protein complex for full methylation activity is conserved in all Set1 homologs, there are only four members which are totally conserved, constituting a ‘core complex’. These ‘core complex’ members are the proteins Ash2 (absent, small, or homeotic discs 2-like or Ash2L in metazoans); RbBP5 (retinoblastoma-binding protein 5) and WDR5 (WD repeat domain 5); and Dpy-30 (Dumpy-30). This complex is collectively referred to as WRAD.

In this model, RbBP5 and Ash2L must first form a heterodimer in order to interact with MLL and SET family KMTs. This interaction alone has been shown to substantially upregulate methylation activity in all six human Set1 homologs [33]. Evidence suggests that the RbBP5-Ash2L heterodimer interacts directly with MLL2-4, Setd1A and Setd1B, but requires the addition

of WDR5 to interact with MLL1 [33]. MLL1-RbBP5 fusion proteins exhibit robust methylation activity, suggesting WDR5 does not participate in catalytic activity, but rather functions as a promotor of complex assembly [33]. Interestingly, phosphorylation of RbBP5 on serine 350 has been shown to increase affinity with Ash2L and ultimately promotes WRAD formation [34]. This demonstrates a type of indirect PTM crosstalk, whereby PTMs of one type (i.e., phosphorylation) affect the formation of protein complexes that catalyse PTMs of another type (i.e., methylation). DPY-30 associates with Ash2L and assists in complex targeting and localization [35].

In addition to the core complex members, there are other complex members that are only found associated with particular Set1 homologs, including but by no means limited to: (1) the CpG-binding CXXC1 (also known as CFP1) that is only found in the Setd1A/Setd1B-associated complex, (2) menin tumor suppressor which is only found associated with MLL1 and MLL2 complexes, or (3) the histone H3K27 demethylase UTX which is only found associated with MLL3 and MLL4 complexes [23]. The modularity of COMPASS-like complexes contributes to specificity and localization of the associated methyltransferase.

1.6 The MLL family of methyltransferases

The MLL family of methyltransferases were discovered when physicians noticed a subset of acute lymphoblastic and myeloid leukemia patients fared much worse than others, particularly when those patients were children [36]. Research on these patients revealed chromosomal translocations at the locus 11q23, as well as leukemic blasts featuring markers from both myeloid and lymphoblastic leukemias (hence the name “mixed-lineage”). Sequencing of this locus revealed

a gene homologous to the trithorax (*Trx*) in *Drosophila*, and later studies revealed the presence of the highly conserved catalytic SET domain, responsible for H3K4 methylation activity [37].

MLL KMTs differ in their evolutionary lineage. MLL1 and MLL2 are paralogs descendant from *Drosophila* trithorax, whereas MLL3 and MLL4 are paralogs descendant from *Drosophila* trithorax-related. This reflects the different roles played by the two subgroups of MLL proteins. MLL1 and MLL2 mono- and di-methylate H3K4 and are found at both promoter and enhancer DNA regions, but MLL3 and MLL4 are predominantly mono-methyltransferases and are highly enriched at enhancer regions [38]. Owing to their homology with *Drosophila* trithorax and trithorax-related proteins, MLL-family methyltransferases are collectively regulators of homeotic gene expression. Interestingly, phylogenetic analysis of SET domain sequences among KMT2 proteins results in segregation that correlates with the lineage of each protein, i.e., MLL1 and MLL2 group with *Drosophila* *trx*, whereas MLL3 and MLL4 group with *Drosophila* *trr* [39]. The evolutionary divergence of SET domain sequences may correlate with the evolution of unique novel substrates among distinct MLL protein lineages. As such, there is considerable interest in probing individual KMTs for novel substrate activity.

1.7 MLL4 in human health and disease

Mixed-lineage leukemia 4 (MLL4, also known as KMT2D, MLL2 and ALR) is a KMT with only one currently known substrate. *In vivo* it is known to monomethylate H3K4 at distal enhancer regions primed for activation, a methyl mark both associated with an increase in transcriptional activity and required for regulation of cell-specific gene expression [40]. All MLL family members facilitate H3K4 monomethylation, while MLL4 exhibits partial redundancy

with its paralog MLL3 [41]. The MLL4 protein co-localizes with lineage determining transcription factors on transcriptional enhancers and is essential for cell differentiation and embryonic development. It also plays critical roles in regulating cell fate transition, metabolism and tumor suppression. As a reflection of its functional impact in cell biology, mutations in MLL4 have been associated with Kabuki Syndrome [42], congenital heart disease [43], and various forms of cancer [38].

Loss of function mutations in MLL4 have been identified in Kabuki syndrome [42], with mutational occurrence rates between 56% and 75%. Congenital heart disease has been associated with an excess of mutations in genes that regulate H3K4 methylation, including *MLL4* [43]. Frameshift and nonsense mutations in the SET and PHD domains affect 37% and 60%, respectively, of the total *MLL4* mutations in cancers [38]. Cancers with somatic mutations in *MLL4* occur most commonly in the brain, lymph nodes, blood, lungs, large intestine, and endometrium [38]. These cancers include medulloblastoma, pheochromocytoma, non-Hodgkin lymphomas, cutaneous T-cell lymphoma, bladder, lung, and endometrial carcinomas, esophageal squamous cell carcinoma, pancreatic cancer and prostate cancer. Given its clear impact in human health and disease, it has become of considerable interest to explore the function of MLL4 beyond H3K4 methylation. Remarkably, belying the crucial role of MLL4 in human health and animal development, H3K4 remains the only known enzymatic target. Indeed it is possible, and posited likely, that MLL4 is able to facilitate the methylation of more than only one methylation site within the methyllysine proteome. **It is therefore the goal of this thesis to broaden the scope of the MLL4 enzyme-substrate network to provide additional insights into the roles that MLL4 may be facilitating in the cell, and how its dysfunction may contribute to disease. I**

hypothesize that the methyltransferase MLL4 is able to facilitate methylation of proteins beyond its only known substrate H3K4.

The work presented in this thesis employs both *targeted* and *systematic* approaches to advance the study of MLL4 methylation targets. To explore my research hypothesis, I have outlined the following specific aims:

Aim 1 (Chapter 2) – *In vitro* exploration of the MLL4 lysine methylation nexus through peptide substrate screening approaches.

Aim 2 (Chapter 3) – Validation of *in vitro* methylation of newly identified targets of MLL4-regulated methylation.

Aim 3 (Chapter 4) – Functional analysis of the identified MLL4-regulated lysine methylation event. .

Figures

A.

B.

Figure 1.1. A. The evolutionary expansion of H3K4 methyltransferases is accompanied by increasing domain architecture complexity. **B.** Phylogenetic analysis of MLL protein SET results in grouping that mirrors grouping of full-length proteins. Adapted from [39]

References

- [1] Duan G., Walther D. (2015). The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. *PLoS Computational Biology* 11(2): e1004049.
- [2] Ambler, R. P. & Rees, M. W. (1959). Epsilon-N-methyl-lysine in bacterial flagellar protein. *Nature* 184: 56–57.
- [3] Biggar KK, Li SS-C. (2015). Non-histone protein methylation as a regulator of cellular signalling and function. *Nature Reviews Molecular Cell Biology* 16: 5-17.
- [4] Mazur PK et al (2014). SMYD3 links lysine methylation of MAP3K2 to Ras-driven cancer. *Nature* 510(7504): 283-287.
- [5] Luger K., Mader A.W., Richmond R.K., Sargent D.F., Richmond T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389: 251-260
- [6] Jenuwein T., & Allis C.D. (2001). Translating the Histone Code. *Science* 293(5532): 1074-1080.
- [7] Smith, B.C. & Denu J.M. (2008). Chemical mechanisms of histone lysine and arginine modifications. *Biochimica et Biophysica Acta* 1789: 45-57.
- [8] Chen, C., Nott, T. J., Jin, J. & Pawson, T. (2011). Deciphering arginine methylation: Tudor tells the tale. *Nature Reviews Molecular Cell Biology*. 12: 629–642.
- [9] Zhang Y., & Reginberg D. (2001). Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes and Development* 15: 2343-2360.

- [10] Farooq Z., Banday S., Pandita T.K., & Altar M. (2016). The many faces of histone H3K79 methylation. *Mutation Research/Reviews in Mutation Research* 768: 46-52.
- [11] Wiles E.T., & Selker E.U. (2016). H3K27 methylation: a promiscuous repressive chromatin mark. *Current Opinions in Genetics and Development* 43: 31-37.
- [12] Ernst J., Kheradpour P., Mikkelsen T.S., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473: 43-49.
- [13] Panier, S. & Boulton S.J. (2014). Double-strand break repair: 53BP1 comes into focus. *Nature Reviews Molecular Cell Biology* 15(1): 7-19.
- [14] Hamamoto, R., Saloura, V., & Nakamura, Y. (2015). Critical roles of non-histone protein lysine methylation in human tumorigenesis. *Nature Reviews Cancer* 15(2): 110–124.
- [15] Luo M. (2018). Chemical and Biochemical Perspectives of Protein Lysine Methylation. *Chemical Reviews* 118(14): 6656-6705.
- [16] West L.E. & Gozani O. (2011). Regulation of p53 function by lysine methylation. *Epigenomics* 3: 361–369.
- [17] Cuella-Martine R, Oliverira C, Lockstone HE, Snellenberg S, Grolmusova N. (2016). 53BP1 Integrates DNA Repair and p53-Dependant Cell Fate Decisions via Distinct Mechanisms. *Molecular Cell* 64: 51-64.
- [18] Shi X et al (2007). Modulation of p53 function by SET8-mediated methylation at lysine 382. *Molecular Cell* 27: 656-646.
- [19] Olsson A.K., Dimberg A., Kreuger J. & Claesson-Welsh L. (2006). VEGF receptor signalling – in control of vascular function. *Nature Reviews Molecular Cell Biology* 7: 359-371.

- [20] Kunizaki M et al. The lysine 831 of vascular endothelial growth factor receptor 1 is a novel target of methylation by SMYD3. *Cancer Research* 61: 10759-10765.
- [21] Esteve, P. O. et al. (2011). A methylation and phosphorylation switch between an adjacent lysine and serine determines human DNMT1 stability. *Nature Structural Molecular Biology* 18: 42–48.
- [22] Lee, T. Y. et al. (2014). Identification and characterization of lysine-methylated sites on histones and non histone proteins. *Computational Biological Chemistry* 50: 11–18.
- [23] Shilatifard, A. (2012). The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis. *Annual Review of Biochemistry* 81(1): 65–95.
- [24] Eissenberg J. C., Shilatifard A. 2010. Histone H3 lysine 4 (H3K4) methylation in development and differentiation. *Developmental Biology* 339: 240–249.
- [25] Schuettengruber B., Martinez A.-M., Iovino N., & Cavalli G. (2011). Trithorax group proteins: switching genes on and keeping them active. *Nature Reviews Molecular Cell Biology* 12: 799-814.
- [26] Schuettengruber B., Bourbon H.-M., Di Croce L., Cavalli G. (2017). Genome Regulation by Polycomb and Trithorax: 70 Years and Counting. *Cell* 171(1): 34-57.
- [27] Ardehali M.B., Mei A., Zobeck K.L., Caron M., Lis J.T., Kusch T. (2011) Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcriptional silencing. *PLOS Biology* 8: e1000453.

- [28] Mohan M., Herz H.M., Smith E.R., Zhang Y., Jackson J., et al. (2011). The COMPASS family of H3K4 methylases in *Drosophila*. *Molecular Cell Biology* 31:4310-4318.
- [29] Clouaire T., Webb S., Skene P., Illingworth R., Kerr A., et al. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes and Development* 26: 1714-1728.
- [30] Lee J.H., Skalnik D.G. (2008). Wdr82 is a C-terminal domain-binding protein that recruits the Setd1A Histone H3-Lys4 methyltransferase complex to transcription start sites of transcribed genes. *Molecular Cell Biology* 28: 609-618.
- [31] Wu M., Wang P.F., Lee J.S., Martin-Brown S., Florens L., et al. (2008). Molecular recognition of H3K4 trimethylation by Wdr82, a component of human Set1/COMPASS. *Molecular Cell Biology* 28: 7337-7344.
- [32] Bledau A.S., Schmidt K., Neumann K., Hill U., Ciotta G., et al. (2014). The H3K4 methyltransferase Setd1a is first required at the epiblast stage, whereas Setd1b becomes essential after gastrulation. *Development* 141: 1022-1035.
- [33] Li Y et al. (2016). Structural basis for activity regulation of MLL family methyltransferases. *Nature* 530(7591): 447–452.
- [34] Zhang et al (2015). A phosphorylation switch on RbBP5 regulates histone H3 Lys4 methylation. *Genes and Development* 29(2): 123-128.
- [35] Haddad et al (2018). Structural Analysis of the Ash2L/Dpy-30 Complex Reveals a Heterogeneity in H3K4 Methylation. *Structure* 26: 1-10.

- [36] Slany R.K. (2009). The molecular biology of mixed lineage leukemia. *Haematologica* 94(7): 984-993.
- [37] Milne T.A., Briggs S.D., Brock H.W., Martin M.E., Gibbs D., Allis C.D., Hess J.L. (2002). MLL targets SET domain methyltransferase activity to Hox gene promoters. *Molecular Cell* 10(5): 1107-1117.
- [38] Rao R.R., Dou Y. (2015). Hijacked in cancer: the MLL/KMT2 family of methyltransferases. *Nature Reviews Cancer* 15(6): 334-346.
- [39] Zhang Y., Mittal A., Reid J., Reich S., Gamblin S.J., & Wilson J.R. (2015). Evolving Catalytic Properties of the MLL Family SET Domain. *Structure* 23(10): 1921-1933.
- [40] Froimchuk E., Jang Y., Ge K. (2017). Histone H3 lysine 4 methyltransferase KMT2D. *Gene* 627: 337-342.
- [41] Lee J, Wang C, Xu S, Cho Y-W, Wang L, Feng X, Baldrige A, Sartorelli V, Zhaung L, Peng W, Ge K. (2013). H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *eLife* 2: e01503.
- [42] Ng S.B., Bigham A.W., Buckingham K.J., Hannibal M.C., McMillin M.J., Gildersleeve H.I., et al. (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature Genetics* 42(9): 790–793.
- [43] Zaidi S., Choi M., Wakimoto H., Ma L., Jiang J., Overton J.D., et al. (2013). De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 498(7453): 220–223.

Chapter 2

Divergent Approaches in Identifying Non-Histone Substrates of the
Histone H3 Lysine Methyltransferase MLL4

2.1 Introduction

Driven by the quickening pace of emerging technologies that have enabled researchers to look deeper into the methyllysine proteome, the list of proteins that are now known to undergo dynamic lysine methylation is rapidly expanding. Such technologies have included; (1) methyl-binding domain (MBD) affinity enrichment of methyllysine protein (such as the use of L3MBTL1 3xMBT domains) [1], (2) immune-affinity enrichment through the gradual development of pan-antibodies with specificity towards methyl-modified proteins/peptides [2], and (3) computational methods of predicting methyllysine occurrence, followed by targeted validation studies [3]. Despite that for decades lysine methylation was through to be predominantly a histone-specific modification, the development of these technologies has allowed lysine methylation to emerge as a ubiquitous and dynamic post-translational modification (PTM). As a result, we now know that lysine methylation can have dramatic implication in the regulation of protein activity, protein-protein interactions, and crosstalk interactions with other PTMs [4].

There are currently over 3,000 known lysine methylated proteins, collectively containing over 5,000 lysine methylation sites that consist of the methyllysine proteome (www.phosphositeplus.org; accessed July 2019). Although the growth of the methyllysine proteome has been expanding at an impressive pace, the lysine methyltransferases (KMTs) responsible for the vast majority of these modifications remain unknown. In fact, many KMTs – in particular histone methyltransferases (HKMTs) – have only one known substrate. Considering the sheer number of modification sites within the methyllysine proteome it is reasonable to suspect, and I posit likely, that many HKMTs participate in the methylation of substrates beyond histones and/or their currently known non-histone substrates. This idea is evidenced through a number of recent studies that have begun experimentally building, and validating, the substrate networks

associated with each HKMT. For example, recent studies have revealed methylation of several non-histone proteins by HKMT G9a, including tumor suppressor p53 at K373me2 [5]. This supports a broader implication for the currently annotated HKMT-substrate network beyond histones. As a result, all HKMTs will be referred to as KMTs throughout this thesis to reflect their broader roles within the human proteome.

As a direct result of the growing implications of lysine methylation in basic cell biology and its contributions to Human disease (i.e., cancer, intellectual disability, etc.), there is considerable incentive to uncover novel relationships between KMTs and known methyllysine substrates. This incentive is further highlighted in diseases where there is a known contribution attributable towards KMT dysfunction, where each new KMT-substrate relationship represents a potential therapeutic target or mechanism of disease progression. To date, a number of non-histone methylation sites have been implicated in cancer progression [6], human tumorigenesis [7], and neurogenerative disease [8]. As a result, a number of preclinical studies have identified KMTs as promising targets for anti-cancer drug development and there has been notable effort put in place to develop therapeutics against this family of enzymes [9]. As the known KMT enzyme-substrate network continues to expand, so too does the landscape of rational drug targets.

Given the importance in establishing causal relationships between methylation sites and their regulatory KMTs, there are several different approaches through which additional KMT substrates have been discovered. The simplest, and perhaps crudest, approach is through a *shotgun* or *target-driven exploration* of potential substrates. Typically, in this type of approach to KMT-substrate identification, known methyllysine substrates of interest (i.e., without a known associated KMT) are selected by a defined criterion (i.e., biological relevance, similarity to a known KMT-substrate, etc.) and are then screened against a specific KMT for methylation activity. This

approach was the initial impetus for the “shotgun” discovery of MAP3K2-K260 methylation by SMYD3, whereby SMYD3 methyltransferase activity towards the MAP3K2 protein was discovered through a protein microarray containing many different possible substrate proteins [10]. The benefit of this approach is that substrates that participate in a known KMT-regulated pathway (such as those implicated in a specific disease) can be pre-selected for screening. This ultimately facilitates the rapid generation of a deductive hypotheses. However, targeted substrate identification inherently suffers from researcher bias, and may fail to return any positive results. Furthermore, this approach to building the KMT-substrate network fails to uncover any fundamental information on how KMTs select for specific substrates and methylation sites (i.e., modifiable lysine residues). A thorough and systematic understanding of KMT substrate selection preference would have the benefit of allowing researchers to continually build KMT-substrate networks in real-time, as new lysine methylation sites are uncovered.

In contrast to a targeted or shotgun approach to substrate identification, the guiding principles of enzyme substrate selection can be realized through a **systematic analysis of substrate preference**. This process begins through the initial elucidation of a candidate recognition motif for a KMT of interest. For the purpose of this thesis, a recognition motif will be defined as the specific sequence of amino acids surrounding (and including) a methylatable lysine residue that facilitates the methyltransferase activity of a specific KMT. Once the KMT recognition motif is resolved, it can then be queried against a database of known lysine methylation sites together with their flanking amino acids (i.e., the “methylation window”). Methylation windows containing all, or a sub-section, of the identified recognition motif are considered “substrate candidates”, and each computationally identified candidate can be ranked in accordance with how well their methylation window matches the queried KMT recognition motif (i.e., 0

residue mismatches, 1 residue mismatch, etc.). While the relevance of the hits with regard to disease implications is inherently not guaranteed, the systematic nature of this approach eschews researcher bias and may return a number of interesting results. Additionally, solving the recognition motif provides valuable insight regarding the specific biochemical interactions that facilitate enzyme-substrate binding, which can be exploited for the development of inhibitor drugs. A limitation of this approach is that it requires at least one substrate to be known in order create the single residue permutations that are necessary for the experiment (described in detail below). However, it should also be mentioned that this type of systematic approach can also suffer from inherent bias in experimental design. These experiments are typically designed around a sequence of a single known KMT substrate, as a result, all single residue permutations of that substrate will always retain significant similarity to the original substrate. This limitation may obscure unknown substrates that are recognized through a motif dissimilar to that of the known substrate. Therefore, a number of different substrates (if known) should also be used in an attempt to reduce single-substrate bias.

Recent studies have employed cellulose-based SPOT peptide arrays in the creation of these substrate permutations [11]. These SPOT peptide arrays are a robust, low-cost, and medium-throughput means (1-1,200 peptides) of characterizing protein-peptide interactions *in vitro* [12]. In this approach, the methylation window of a known substrate is systematically permuted (or mutated) to generate a matrix of peptides where every wild-type residue is substituted with each of the 19 other amino acids. Probing the final SPOT array with a purified KMT of interest and a radiolabelled methyl group donor (i.e., adenosyl-L-methionine, S-[methyl-3H]) indicates which of the mutated peptides retain significant enzyme activity at a specific threshold (i.e., >50% 3H incorporation vs. WT substrate). This method is able to highlight residue positions where particular

amino acids may be required for substrate binding/activity (e.g., a hydrophobic pocket interaction) and also positions where there is a greater range of residue tolerability (e.g., an externally oriented residue side chain). The recognition motif is subsequently generated from the sequence of amino acid substitutions which retain or improve upon wild-type activity.

For this study, I have employed both shotgun-driven and systematic analytical approaches in identifying novel substrates for the KMT mixed lineage leukemia 4 (MLL4) WRAD complex. As previously discussed in Chapter 1, MLL4 is a histone H3K4 methyltransferase with no other known substrates.

2.2 Materials & Methods

2.2.1 Protein expression and purification

p3E-MLL4 (5382–5537) SET/Post-SET domain construct was expressed as a glutathione S-transferase (GST) fusion protein. *Escherichia coli* BL21 cells containing either construct were grown at 37 °C to an OD₆₀₀ of ~0.6 in lysogeny broth (LB) containing 1 mM MgCl₂ and 1 μM ZnSO₄. Expression was induced using 0.1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG), and proceeded for 18 hours at 16 °C. After centrifugation at 2,500xg for 15 minutes, the bacterial pellet was lysed in 50 mM Tris pH 7.5 buffer containing 300 mM NaCl, 3 mM DTT, 1 μM ZnSO₄, 0.01% Triton X-100, 1 mg/mL chicken egg lysozyme (Sigma), 10% glycerol; and a protease inhibitor cocktail consisting of 100 μM PMSF, 1 μM E-64, 1 μM pepstatin, and 5 μM bestatin. The lysate was clarified through centrifugation at 18,000xg for 45 minutes. The fusion protein was isolated from clarified lysate by affinity chromatography with glutathione sepharose 4B beads (GE Healthcare) and eluted using 50 mM Tris pH 8 containing 10 mM glutathione.

Full length pOPINJ-WDR5, p2E-RbBP5, pOPINS-Ash2L, and pGst-Dpy-30 are each expressed as 6xHistadine fusion proteins. Each construct was grown in *E. coli* BL21 cells at 37 °C to an OD₆₀₀ of ~0.6 in lysogeny broth (LB) containing 1 mM MgCl₂. Expression was induced using 0.1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG), and proceeded for 18 hours at 16 °C. After centrifugation at 2,500xg for 15 minutes, the bacterial pellet was lysed in P5 (50 mM NaHPO₄ pH 7.0, 500 mM NaCl, 5 mM imidazole, 10% glycerol, 0.05% Triton X-100) containing; protease inhibitor cocktail (100 μM PMSF, 1 μM E-64, 1 μM pepstatin, and 5 μM bestatin). The lysate was clarified through centrifugation at 18,000xg for 45 minutes. The fusion proteins were then each isolated from clarified lysate by affinity chromatography with Ni-NTA beads (Qiagen) and eluted using P500 (P5 with 500 mM imidazole).

2.2.2 Array design, SPOT synthesis, and array processing

For the ‘shotgun-approach’ array, known methyllysine sites of interest (criterion for selection was based on biological interest of the protein and used for screening substrates of multiple KMTs) were selected for synthesis. Each site of interest was printed in duplicate, as a non-, mono-, and dimethyl moiety (Figure 2.1A). For the systematic array, the wild-type sequence for histone H3 (amino acids 1-11) was systematically mutated from position -3 to +5 relative to the K4 position, such that each wild-type residue was mutated to all other 19 amino acids. All peptides containing each systematic residue mutant were synthesized. Solid-phase peptides were all synthesized directly on a sheet of aminated cellulose filter paper with an Intavis Multiprep instrument using Fmoc chemistry at an approximate concentration of 0.2 μmol/SPOT [3]. At the end of the synthesis, the arrays were deprotected by three sequential washes of 20% piperidine dissolved in dimethylformamide (DMF), followed by three washes in DMF, and three final washes

with anhydrous ethanol. Arrays were then dried by air. After drying, side-chain protecting groups were cleaved from the peptides with a trifluoroacetic acid solution (95% TFA, 3% TIPS and 2% H₂O). The arrays were again washed three times with dichloromethane (DCM), followed by three washes with DMF, and three washes of anhydrous ethanol, and allowed to air dry overnight in a fume hood. Immediately before use, arrays were briefly washed with anhydrous ethanol, then equilibrated in a solution of 50% ethanol in methylation buffer (50 mM Tris pH 8.5, 3 mM MgCl₂, 1mM EDTA, 1 mM dithiothreitol (DTT)), and then finally only the methylation buffer.

2.2.3 *In vitro* array-based MLL4 WRAD methylation

The MLL4 WRAD complex was assembled by combining equal molar ratios of MLL4, WDR5, RbBP5, Ash2L and Dpy-30. The assembled MLL4 WRAD complex was then aliquoted and snap-frozen for storage at -80 °C in the presence of 20% glycerol. For array-based methylation experiments, a total of 2 µM of MLL4 WRAD in methylation buffer was incubated at 4 °C overnight with 0.466 µM adenosyl-L-methionine S-[methyl-3H] (specific activity: 80Ci/mmol) (Perkin Elmer) on a rocking platform. After incubation with MLL4 WRAD, arrays were washed with 10 mL of wash buffer (100 mM NH₄HCO₃, 1% sodium dodecyl sulfate (SDS)) and the array was incubated on a rocker for 3 minutes at room temperature. Washes were repeated 5 x 3 minutes. Following washing of the array, the arrays were transferred to a new containing enough anhydrous ethanol to completely submerge the array. Arrays were then incubated in ethanol on a rocking platform for 2 minutes. Following incubation of the array in ethanol, the ethanol was removed and enough 7% DPO solution (7 g of 2,5-Diphenyloxazole in 100 mL ethanol) was added to complete submerge the array. Arrays were then gently agitated for 2 minutes to ensure uniform distribution

of the DPO solution. The arrays were then removed from the DPO solution and allowed to dry completely on filter paper overnight prior to array fluorography.

For fluorography of the methylated arrays, the dried arrays were wrapped with a single layer of plastic food wrap and placed into an autoradiography cassette containing intensifying screens (Carestream BioMax TransScreen LE). In complete darkness, a sheet of detection film (Amersham Hyperfilm MP) was added to the cassette above the array. The loaded cassette was closed, and the film was then exposed to array fluorography at -80°C for 1 week. After exposure the film developed using Ilford ilfosol 3 film developer in complete darkness and then fixed with Ilford rapid fixer according to manufacturer's instructions under a safe light. Densitometric analysis of the array spots was performed using ImageJ with 'Protein Array Analyzer' macro (<https://imagej.nih.gov/ij/macros/toolsets/>). SPOTs producing signal greater than 50% of the wild-type H3K4 peptide substrate signal were determined to be a tolerable substrate.

2.2.4 MLL4 WRAD candidate substrate motif analysis and substrate prediction

A candidate MLL4 WRAD substrate recognition motif was generated base on the results of the H3K4-based permutation substrate array. For this analysis, SPOTs producing signal greater than 50% of the wild-type H3K4 peptide substrate signal were determined to be a tolerable mutation and were included in the candidate recognition motif. The raw densitometry data was normalized to the array background signal and was then analysed using a custom software created in the Biggar lab, PeSA (v.1.5) (<http://www.github.com/EmineTopcu/PeSA>). This software works by creating a weight-based motif using the residues at every specific position along a peptide substrate that has a quantified value greater than a user-defined threshold (i.e., 50% of H3K4 for this thesis).

Candidate MLL4 WRAD substrates were identified from a list containing all lysine residues from proteins classified to be located in the nucleus by Uniprot. This list contained 218,254 lysine residues to be explored as possible MLL4 WRAD substrates. Candidate substrates were identified based on a binary approach of motif sequence matches. In this approach, all sites were then classified as either containing a perfect match (i.e., H3K4 sequence) to the candidate recognition motif, or any possibility of 1 to 8 residue mismatches.

2.3 Results & Discussion

2.3.1 A “shotgun-style” approach to novel substrate identification reveals TP53BP1 as a possible MLL4 WRAD substrate

MLL4 WRAD-dependant methylation events were observed on several peptides, suggesting potential *in vitro* enzyme-substrate relationships (Figure 2.1B). Although no peptide substrate met our criteria of 50% SPOT intensity relative to the H3K4 substrate, several potential substrates with qualitatively lower levels of relative methylation did emerge from the array. Peptides of interest shown to be methylated (below our defined threshold) on the array include K1563 of Tumor Suppressor p53 Binding Protein 1 (TP53BP1, also 53BP1), K338 and K341 of Rho GTPase activating protein 32 (ARHGAP32), K569 of Nuclear FMR1 Interacting Protein 2 (NUFIP2), K146 of SNW Domain Containing 1 (SNW1), K108 of Heat Shock Protein Family A (Hsp70) Member 8 (HSPA8), and K260 of Mitogen-Activated Protein Kinase Kinase Kinase 2 (MAP3K2). A high degree of methylation was also observed on the positive control H3K4 and H3K9 peptides. While H3K9 isn't itself a substrate of MLL4, the peptide did contain the K4 residue. This illustrates the potential for error in identifying which specific lysine was methylated

and the importance of conducting follow-up experiments designed for greater resolution before a specific methyllysine modification can be identified with complete certainty.

While each potential substrate identified on this array merits further validation and further analysis, in-depth investigation of every hit is beyond the scope of this thesis. As such, a substrate was selected for validation based on two criteria, (1) an implication of MLL4 in an interesting biological pathway, and (2) a methyllysine residue within a domain of predictable functional importance. In accordance with these criteria, we selected 53BP1 as a candidate substrate for additional investigation.

First reported in 1994 as a p53-interacting protein of unknown function and importance [13], subsequent decades of research have unveiled 53BP1 as a crucial actor within the DNA-damage response (DDR) pathway. Though devoid of enzymatic activity, 53BP1 participates in a plurality of PTM-dependant and -independent protein-protein interactions (PPIs) that ultimately facilitate cellular control of DNA double-strand break (DSB) repair decisions, [reviewed in 14, 15], recruit additional DSB-responsive proteins, and stimulate p53-dependant gene expression [16].

DNA DSBs are highly deleterious DNA lesions. When left unrepaired or otherwise erroneously repaired, DSBs can result in a permanent arrest of growth leading to cell death [17], and in some cases the respective activation and abatement of oncogenes and tumor suppressors by mutations introduced through chromosomal rearrangement [18]. There are two pathways through which cells can respond to and repair DSBs: non-homologous end joining (NHEJ) and homology-directed repair (HDR). In NHEJ, minimally or unprocessed DNA ends are ligated at the break locus. Conversely, HDR requires DNA ends to undergo resection, which ultimately facilitates repair using homologous DNA (e.g. from a sister chromatid) as a template [19]. DSB repair

pathway choice is multifactorial and complex, but 53BP1 plays an important role. Indeed, 53BP1 has been established as a key promoter of NHEJ [reviewed in 14]. Though thought for many years to be a suppressor of HDR, recent studies suggest that 53BP1 instead suppresses only the error-prone mode of HDR known as single-strand annealing (SSA) and is required for the execution of gene conversion (GC), the ‘error-free’ mode [15]. It is noteworthy that a functional MLL4-dependant methylation event in 53BP1 would implicate MLL4 in DSB pathway selection, or more broadly, the DDR. This implication influenced our decision to select 53BP1 for further investigation.

Interestingly, the putative K1563 methyllysine is located within the tandem Tudor domains (TTDs) of 53BP1 (Figure 2.2). A Tudor domain is a methyl-binding domain (MBD) which folds in an antiparallel five-stranded β -barrel [reviewed in 20]. The methyllysine residue is recognized through cation- π interactions between the residue methylammonium group and the Tudor domain’s aromatic binding cage. 53BP1 contains two Tudor domains in a tandem arrangement, though studies suggest only the first β -barrel (“Tudor 1”, Figure 2.2) is functional [21], leaving the role of Tudor 2 unclear. 53BP1 is recruited to DSB regions through the binding of di-methylated histone H4 lysine 20 (i.e., H4K20me2) by the TTD [reviewed in 13], and through an interaction with ubiquitinated histone H2A lysine 15 (i.e., H2AK15ub) facilitated by 53BP1 oligomerization. 53BP1 binding activity is regulated by a protein known as Tudor-interacting repair regulator (TIRR), which abrogates 53BP1 access to H4K20me2 by binding to the TTD and blocking the H4K20me2 interaction surface [22].

While the bulk of 53BP1-p53 interaction is mediated through 53BP1’s tandem BRCT domains [15], TTD-p53 methyllysine binding is positively associated with antitumorigenic

activity [23, 24]. This interaction occurs between the TTD and methyllysine residues located within p53's C-terminal regulatory domain, specifically p53K370me2 and p53K382me2.

Structural analysis of the 53BP1 TTD indicated the putative methylation site is in fact located on Tudor 2 (Figure 2), the second β -barrel whose function is unclear. However, 53BP1 K1563 has also been noted as both a ubiquitination and SUMOylation site [25, 26]. It is possible that 53BP1 K1563me is part of an unelucidated PTM cross-talk system which confers an unknown regulatory function to Tudor 2. Lysine methylation has been shown to inhibit modification of adjacent residues [27]. As such, K1563 methylation could affect the PTM of adjacent residues, such as ubiquitination at K1559 [25], or phosphorylation at K1569 [28]. Further investigation into the function of Tudor 2 and its PTMs could reveal a role for K1563 methylation.

Furthermore, TIRR-53BP1 binding is mediated through hydrogen bonds within close proximity to the putative K1563 methylation site, specifically Y1552 and M1584 [29]. It is also intriguing to hypothesize that 53BP1 K1563 methylation attenuates TIRR binding strength through increased steric bulk, disrupting hydrogen bond formation.

Although not located in Tudor 1 of 53BP1, the fact remains that the putative MLL4-dependant methylation site at K1563 resides within a domain of functional importance and implicates MLL4 in an interesting biological pathway. As such, our aim in further studying this link is to first validate the methylation event, and if valid elucidate potential methylation-dependant mechanism *in vitro*. The experimental approach and results of these further studies are described in Chapter 3.

2.3.2 Elucidation of an MLL4 recognition motif through systematic mutation of H3K4

Analysis of the methylation pattern generated through systematic mutation of H3K4 on a peptide array (Figure 2.3B) provided us with a motif which may represent the (or one of the) substrate sequences recognized by the active site of MLL4 (Figure 2.3C). As such, the set of proteins containing this sequence are likely candidates for MLL4-dependent methylation. This recognition motif is characterized by the presence of residues with similar chemical properties in distinct positions (Figure 2.3D). The -3 position (relative to the central K) primarily tolerates residues with hydrophobic side chains, but also amino acids with unique properties such as proline, rigid and containing a secondary amine, and glycine, which is highly compact. The -2 position has similar tolerances, with the inclusion of positively charged K and R. The -1-position primarily tolerates hydrophobics but also the polar uncharged residues serine and the wild-type residue threonine. There appears to be no preference for any particular amino acid or physical property in positions +1, +2, and +3. The +4 position may contain proline, glycine or the positive charged residues, the latter of which is required in position +5.

Preference for particular amino acids or physical properties within distinct positions of a recognition sequence is often characteristic of the emergent physical topology of the associated active site. With this in mind, we qualitatively compared this recognition sequence with the binding pocket of MLL3, an MLL4 paralog (Figure 2.4). We chose to analyse MLL3 SET, as the solved crystal structure [30] of MLL4 SET was reported to be potentially inaccurate due to the presence of a tag required for crystallization [31 supplementary information]. Because the substrate binding residues in the MLL family are highly conserved [30], insights gleaned from the analysis of MLL3 SET are likely applicable to MLL4 SET. We note that the positions in the recognition motif that tolerate multiple substitutions (i.e. positions +1, +2, +3) (Figure 2.3) tend to

be residues that make fewer or no contact with the binding pocket, whereas residues that contact the binding pocket tend to be less tolerable of substitutions (Figure 2.4A)

2.3.4 Bioinformatic analysis reveals several hundred known methyllysine proteins containing the MLL4 recognition motif, including Setd1A/B COMPASS member CXXC Finger Protein-1

We conducted a bioinformatic search of the recognition motif to identify proteins containing some or all of the sequence. The motif was queried against 218,254 lysine residues in known nuclear proteins, of which 452 lysines and their surrounding residues were a perfect match (i.e., 9 out of 9 motif positions) (Figure 2.5). 192 of those perfect matches were found to be located to exist within proteins that are known to undergo methylation (Table 2.1). Interestingly, 5 of those lysine residues are actual known methylation sites (Table 2.2), which lends credence to the validity of our approach.

Next, we sought to identify which of the proteins containing a perfect match of the MLL4 recognition motif could be of interest for further investigation. We noted that the Setd1A/B complex member CXXC finger protein 1 (CFP1) was included in our results. CFP1 functions to regulate Setd1-mediated H3K4 by targeting the complex to actively transcribed promoters through a multivalent interaction with non-methylation CpG and H3K4me3 [32]. Specifically, our analysis revealed K328 as the potential methylation site, which has been previously identified (but not characterized) as a putative methylation site using a broad-range chemical proteomics screen [33]. It is intriguing that this lysine is adjacent to the Setd1 interaction domain (SID) (Figure 2.6). The SID is a key facilitator of interaction between CFP1 and Setd1: point mutations within this domain have been shown to completely abolish CFP1-Setd1 binding and result in aberrant Setd1

methylation activity [34]. Though reported to be comprised by amino acids 361 to 422, the rationale for this exact definition is unclear, and it is not unlikely that modifications to adjacent residues could affect SID function. Furthermore, CFP1 has been found to interact with the DNA CpG methyltransferase DNMT1 [33], and in *C. elegans* the Sin3 histone deacetylase complex [35], and this association appears to function independently from Setd1 association. This makes CFP1 a particularly interesting candidate for MLL4-dependant methylation. It is interesting to hypothesize that methylation of K328 with the SID acts as a methyl-switch that contributes to the regulating of CFP1 partner binding.

2.3.5 Conclusions

Through two distinct approaches we have identified two proteins that are promising candidates as new MLL4 substrates. 53BP1 K1563 methylation could implicate MLL4 in the DDR through non-histone protein methylation activity. CFP1 K328 methylation may represent a novel methyl-switch mechanism that regulates partner association between distinct epigenetic regulatory mechanisms. To investigate these hypotheses, we first had to confirm *in vitro* MLL4-dependant of 53BP1 and CFP1 in solution. The results of these experiments are presented in Chapter 3.

Figures

Figure 2.1. “Shotgun” style approach to identifying new substrate interactions with MLL4. **A.** Cartoon representation of array layout. Each peptide is printed as a “triplet” me0, me1, and me2 moiety to assay mono-, di-, or trimethylase activity, respectively. Each triplet is printed in duplicate. **B.** A solid-phase peptide array featuring known methyllysine substrates of interested was printed in triplicate on aminated cellulose using standard Fmoc chemistry. The array was incubated with 1 μ M MLL4 WRAD and 0.8 μ M adenosyl-L-methionine, S-[methyl-3H], and imaged by exposure to Kodak x-ray film for 7 days. Dark spots indicate the presence of radiolabelled SAM, signifying enzymatic activity.

Figure 2.2. Ribbon cartoon depicting putative methyllysine site at K1563 within the Tudor 2 domain of the 53BP1 tandem Tudor domains. Image generated using PyMOL.

A.

○ polar	→ sidechain acceptor	○ solvent residue	⊗ arene-arene
○ acidic	→ sidechain donor	○ metal complex	⊗+ arene-cation
○ basic	→ backbone acceptor	○ solvent contact	
○ greasy	→ backbone donor	○ metal contact	
○ proximity	○ ligand exposure	○ receptor contact	
○ contour			

B.

Figure 2.4. A. 2D representation of MLL3 ligand-binding pocket interaction B. 3D representation of MLL3 ligand-binding pocket interaction. Images generated using MOE software from crystal structure described in [30] (PDBe 5f59).

A.

B.

Matches	Hits	Percentage
9	451	0.2066
8	4458	2.0425
7	19138	8.7686
6	46383	21.251
5	63369	29.034
4	50484	23.130
3	24801	11.363
2	8559	3.9215
1	611	0.2799

Figure 2.5. Quantification of known methyllysine proteins containing some or all of our putative MLL4 recognition motif. “Hits” refers to the number of proteins containing some or all of the recognition motif. “Matches” refers to the number of residues within a hit that are identical to the recognition motif, where 9 out of 9 is considered a perfect match.

Figure 2.6. Domains and regions of CFP1. Figure originally from Butler et al., 2008 [34]

Table 2.1. MLL4 WRAD motif perfect match sequences in known methyllysine modified proteins

Uniprot ID	Gene name	Residue	Sequence
O00567	NOP56_HUMAN	448	EKKRLKKEKKR
O14646	CHD1_HUMAN	1498	LHKLYKHAIKK
O15391	TYY2_HUMAN	143	SKKPSKKPSGK
O43290	SNUT1_HUMAN	403	TFKKTkRRVKK
O43395	PRPF3_HUMAN	110	SSGVKKRRIPR
O60673	REV3L_HUMAN	1189	PSIVTKKRNRK
O60885	BRD4_HUMAN	285	QPVKTKKGVKR
O75151	PHF2_HUMAN	696	FPIRRKKNAPK
O75151	PHF2_HUMAN	953	KSKKKKSAKRK
O75367	H2AY_HUMAN	147	KKPVSKKAGGK
O75376	NCOR1_HUMAN	2329	PKLISKSNSRK
O75528	TADA3_HUMAN	124	GRPKSKNLQPK
O75592	MYCB2_HUMAN	2823	APVKTkLDPPR
O75592	MYCB2_HUMAN	3168	ENKKSKEKKK
O76021	RL1D1_HUMAN	347	EHGKKKRGRGK
O95260	ATE1_HUMAN	247	FPPKAKSNQPK
O95359	TACC2_HUMAN	2287	NPPPTKKIGKK
O96028	NSD2_HUMAN	1080	WGLVAKRDIRK
P00533	EGFR_HUMAN	676	RHIVRKRTLRR
P07305	H10_HUMAN	111	AFKKTKEIKK
P07900	HS90A_HUMAN	414	LKVIRKNLVKK
P08238	HS90B_HUMAN	406	LKVIRKNIVKK
P10242	MYB_HUMAN	120	WSVIAKHLKGR
P11388	TOP2A_HUMAN	1461	DPAKTKNRRKR
P11388	TOP2A_HUMAN	1286	AFKPIKKGKKR
P16402	H13_HUMAN	179	SAKKVKTPQPK
P16403	H12_HUMAN	22	KAPVKKKAACK
P25490	TYY1_HUMAN	173	GGGRVKKGGGK
P27695	APEX1_HUMAN	27	EAKKSKTAAKK
P31629	ZEP2_HUMAN	2280	PYVLSKQHEKR
P38432	COIL_HUMAN	183	EEAKRKSPKKK
P43243	MATR3_HUMAN	478	KYKRIKKPEGK
P46013	KI67_HUMAN	2647	GIKVLKQRAKK
P46087	NOP2_HUMAN	699	VTGKLKQRSPK
P48382	RFX5_HUMAN	465	SDAKRKRGRPR
P48634	PRC2A_HUMAN	1190	WSPPAKSLAPK
P49711	CTCF_HUMAN	259	PTKIKKKGVKK
P49711	CTCF_HUMAN	595	ETKKSkrGRKR
P49711	CTCF_HUMAN	649	APPPAKRRRGR
P49756	RBM25_HUMAN	236	SHPRKKKKEKK
P49792	RBP2_HUMAN	614	VLPLLKIIKKK
P51532	SMCA4_HUMAN	583	KEKKKKKKKKK
P51532	SMCA4_HUMAN	1596	RSVKVKIKLGR
P55769	NH2L1_HUMAN	86	VFVRSKQALGR
P68431	H31_HUMAN	5	AMARTKQTARK
P78527	PRKDC_HUMAN	2908	AELPAKRVRGK
P78545	ELF3_HUMAN	246	KHGKRKRGRPR

Q03164	KMT2A_HUMAN	234	PGVKKIKITHGK
Q03164	KMT2A_HUMAN	272	ATKIKKLRAGK
Q03164	KMT2A_HUMAN	959	TAVKTKILIKK
Q05639	EF1A2_HUMAN	439	AVGVKNVEKK
Q09666	AHNK_HUMAN	5711	KLKKSIIKMPK
Q12873	CHD3_HUMAN	283	PDGRKKLRGKK
Q12873	CHD3_HUMAN	51	GVKKRKRGPCK
Q12906	ILF3_HUMAN	388	PSKKKKKIQKK
Q13023	AKAP6_HUMAN	600	VPLLSKHKSKK
Q13111	CAF1A_HUMAN	795	AAIPSKSRLKR
Q13123	RED_HUMAN	300	NKKLKKKDKGK
Q13330	MTA1_HUMAN	527	SPLVLKQAVRK
Q13415	ORC1_HUMAN	460	TKVPKSLKPR
Q13428	TCOF_HUMAN	1439	SNPKSKKEKKK
Q13435	SF3B2_HUMAN	451	APKLSKKKLRR
Q13523	PRP4B_HUMAN	224	ERKKS KSPSKR
Q13620	CUL4B_HUMAN	795	ARVLAKNPKGK
Q13823	NOG2_HUMAN	654	DRAPSKKGKKR
Q13895	BYST_HUMAN	194	REVL SKYRSGK
Q14498	RBM39_HUMAN	51	ERKRSKSKERK
Q14498	RBM39_HUMAN	67	ERKKS KSRERK
Q14562	DHX8_HUMAN	703	FGLLKKT VQKR
Q14669	TRIPC_HUMAN	1003	DVLKRKR LPKR
Q14781	CBX2_HUMAN	162	KDPIRKKRGRK
Q14865	ARI5B_HUMAN	445	GTKRIKHEIPK
Q14978	NOLC1_HUMAN	76	NGPVAKKAKKK
Q14978	NOLC1_HUMAN	615	TFPKRKKGEKR
Q15046	SYK_HUMAN	20	EPKLSKNELKR
Q15059	BRD3_HUMAN	239	PPVVKKKG VKR
Q15059	BRD3_HUMAN	490	VNKP KKKKEKK
Q15361	TTF1_HUMAN	276	HKKKS KKKKKK
Q15398	DLGP5_HUMAN	22	EMIRTKIAHRK
Q15545	TAF7_HUMAN	141	ITLPLKNVRKR
Q1ED39	KNOP1_HUMAN	65	EMPLVKKKKKK
Q1ED39	KNOP1_HUMAN	66	MPLVKKKKKKK
Q1ED39	KNOP1_HUMAN	67	PLVKKKKKKKK
Q1ED39	KNOP1_HUMAN	289	PALKRKKKKKR
Q1ED39	KNOP1_HUMAN	290	ALKRKKKKKRK
Q32MZ4	LRRF1_HUMAN	570	SQKKT KNKKKK
Q32MZ4	LRRF1_HUMAN	575	KNKKKKKNKKK
Q3L8U1	CHD9_HUMAN	2429	NQPIVKKRRGR
Q3L8U1	CHD9_HUMAN	2430	QPIVKKRRGRR
Q562F6	SGO2_HUMAN	614	INKLRKKVNRK
Q56NI9	ESCO2_HUMAN	547	RLKRRKRRIARR
Q5T8A7	PPR26_HUMAN	704	DLLRSKRKLKK
Q66PJ3	AR6P4_HUMAN	297	KDKRRKKKKKR
Q66PJ3	AR6P4_HUMAN	303	KKKKRKKLKKK
Q6PD62	CTR9_HUMAN	933	DLPISKKKKRR
Q6PD62	CTR9_HUMAN	983	GPKPKRRRPPK
Q6UB98	ANR12_HUMAN	1845	LHIRKKIEEKR
Q6UB99	ANR11_HUMAN	2447	LQIRKKIEEKR
Q7KZF4	SND1_HUMAN	409	EFLRKKLIGKK
Q7Z699	SPRE1_HUMAN	322	SLKIKKSKRRK

Q7Z6E9	RBBP6_HUMAN	1731	SKKKKKKKEKK
Q7Z6E9	RBBP6_HUMAN	1732	KKKKKKKKEKKK
Q86UE4	LYRIC_HUMAN	443	PTGKSKKKKKK
Q86UE4	LYRIC_HUMAN	446	KSKKKKKKKKK
Q86UE4	LYRIC_HUMAN	575	KQIKKKKKKARR
Q86X95	CIR1_HUMAN	237	EKKKKKKDRKK
Q86XP3	DDX42_HUMAN	298	MIGIAKTGSGK
Q8IVF2	AHNK2_HUMAN	3614	QAPKAKLDAGR
Q8IVL0	NAV3_HUMAN	2076	EYVITKSGRKK
Q8IWS0	PHF6_HUMAN	159	SSPKSKKKSRK
Q8IY81	SPB1_HUMAN	804	TYVVAKKGVGR
Q8IY92	SLX4_HUMAN	1661	PKGPAKTKGPR
Q8IZA3	H1FOO_HUMAN	174	NVGKVKKAAKR
Q8IZT6	ASPM_HUMAN	674	FIKPLKTDIPR
Q8N8L2	ZN491_HUMAN	76	LEKPYKHKQRR
Q8NEZ4	KMT2C_HUMAN	3605	EKGKKKRTRKK
Q8NF91	SYNE1_HUMAN	6675	ASAVLKRAHKR
Q8NI36	WDR36_HUMAN	773	LDVIKKKNKPK
Q8TDD1	DDX54_HUMAN	752	DKKKIKTESGR
Q8TDI0	CHD5_HUMAN	50	VSLPKKKKPKK
Q8TDI0	CHD5_HUMAN	324	LGKSKRRRKK
Q8TEK3	DOT1L_HUMAN	397	SKARKKKLNKK
Q8TEQ6	GEMI5_HUMAN	828	KVILLKKEPPK
Q8WUI4	HDAC7_HUMAN	167	RYKPKKSLERR
Q8WYP5	ELYS_HUMAN	1973	SAIPRKRGRR
Q8WZ42	TITIN_HUMAN	2516	DEGPYKLIVGR
Q8WZ42	TITIN_HUMAN	10606	EEKVLKLKPKR
Q8WZ42	TITIN_HUMAN	11877	EEVVLKSVLRK
Q8WZ42	TITIN_HUMAN	17902	TCKVSKLLEGK
Q8WZ42	TITIN_HUMAN	31951	VEVPAKIHLPK
Q8WZ42	TITIN_HUMAN	33594	SPPRVKSPEPR
Q92541	RTF1_HUMAN	229	IKKKLKTAKKK
Q92831	KAT2B_HUMAN	672	KEIIKKLIERK
Q96BK5	PINX1_HUMAN	295	TLKPKKRRGKK
Q96BR9	ZBT8A_HUMAN	199	QQPLAKHEPRK
Q96JP5	ZFP91_HUMAN	286	EEPPRKRGRRR
Q96KM6	Z512B_HUMAN	859	EEPVALPPRR
Q96KQ7	EHMT2_HUMAN	361	SRKRRKREPPR
Q96L73	NSD1_HUMAN	1959	WGLRTKTDIKK
Q96L91	EP400_HUMAN	1432	AELLSKKKIPR
Q96Q15	SMG1_HUMAN	2134	TILPTKTKPKK
Q96QE3	ATAD5_HUMAN	479	KNKKLKKKNKK
Q9BQ70	TCF25_HUMAN	131	RKKKKKQKNKK
Q9BQG0	MBB1A_HUMAN	1167	QSPISKRRKKK
Q9BQI6	SLF1_HUMAN	313	YTLRRKRKKGK
Q9BVI0	PHF20_HUMAN	543	RVKPKKKKKKK
Q9BVI0	PHF20_HUMAN	544	VKPKKKKKKKK
Q9BVI0	PHF20_HUMAN	545	KPKKKKKKKKK
Q9BVP2	GNL3_HUMAN	7	KRPKLKKASKR
Q9BW19	KIFC1_HUMAN	89	AQKVSKKTGPR
Q9BYG3	MK67I_HUMAN	174	ERLLRKKLAKK
Q9BZ95	NSD3_HUMAN	1306	KNAKLKQKRRK
Q9C0D4	Z518B_HUMAN	227	VKAVAKLEPKR

Q9H0H5	RGAP1_HUMAN	179	TFKLKKREKRR
Q9H2G4	TSYL2_HUMAN	407	RIKRKKQEMKK
Q9H2P0	ADNP_HUMAN	616	AAVPYKKDVGK
Q9H2Y7	ZN106_HUMAN	1040	KNKRRRIKGKK
Q9H2Y7	ZN106_HUMAN	1353	GSKKKKKLRKK
Q9H2Y7	ZN106_HUMAN	1354	SKKKKKLRKKK
Q9H6F5	CCD86_HUMAN	329	NPAKLKRAKKK
Q9HCG8	CWC22_HUMAN	822	GDPKKKRGERR
Q9NPF5	DMAP1_HUMAN	292	KAPKKKLQKK
Q9NPF5	DMAP1_HUMAN	86	RTVKAKLGSKK
Q9NPG3	UBN1_HUMAN	186	KEKKKKSPKKR
Q9NQZ2	SAS10_HUMAN	342	LIPKAKSTKPK
Q9NR30	DDX21_HUMAN	75	NSPKSKKAKKK
Q9NR30	DDX21_HUMAN	56	VFPKAKQVKKK
Q9NR30	DDX21_HUMAN	93	ISPKTSLRKK
Q9NR48	ASH1L_HUMAN	711	SKPLKKRKGRK
Q9NR48	ASH1L_HUMAN	710	QSKPLKKRKGR
Q9NRL2	BAZ1A_HUMAN	1357	LSPRRKRRGRK
Q9NUQ6	SPS2L_HUMAN	66	MTGKKKNNKRR
Q9NVU7	SDA1_HUMAN		MKVLKKQKKKK
Q9NVU7	SDA1_HUMAN	272	MKVLKKQKKKK
Q9NZI8	IF2B1_HUMAN	77	HSVPPKKQRSRK
Q9P0U4	CXXC1_HUMAN	328	RAVKVKHVKRR
Q9P275	UBP36_HUMAN	845	PHGKRKRKKKK
Q9P2D0	IBTK_HUMAN	979	MFKKAKTKAKK
Q9UIF9	BAZ2A_HUMAN	670	EVPKVKRGRGR
Q9UIG0	BAZ1B_HUMAN	895	GIAKAKLVMRR
Q9UK61	TASOR_HUMAN	105	FQIPRKSREKK
Q9UPP1	PHF8_HUMAN	523	LSLPSKNGSKK
Q9UPV0	CE164_HUMAN	114	IKKKKKKKEKK
Q9UPV0	CE164_HUMAN	111	SGAIKKKKKKK
Q9UQ35	SRRM2_HUMAN	193	KKKKKKKDRGR
Q9UQ35	SRRM2_HUMAN	1754	SSPRTKTTSTR
Q9UQR1	ZN148_HUMAN	394	KLVLKKINSKR
Q9Y2L1	RRP44_HUMAN	618	LNKLAKILKKR
Q9Y4C8	RBM19_HUMAN	823	TLARKKQVPRK
Q9Y620	RA54B_HUMAN	135	WCKPSKKKHKK
Q9Y6X0	SETBP_HUMAN	586	TPVKKKRGPRK
Q9Y6X0	SETBP_HUMAN	1016	SDLKSKKKRGR

Table 2.2. MLL4 WRAD motif perfect match sequences in known methyllysine modified residues.

Uniprot ID	Gene name	Residue	Sequence
P11388	TOP2A_HUMAN	1286	AFKPIKKGKKR
P68431	H31_HUMAN	5	AMARTKQTARK
Q92831	KAT2B_HUMAN	672	KEIIKKLIERK
Q9H2G4	TSYL2_HUMAN	407	RIKRKKQEMKK
Q9P0U4	CXXC1_HUMAN	328	RAVKVKHVKRR

References

- [1] Moore K.E., Carlson S.M., Camp N.D., Cheung P., James R.G., Katrin F.C., Wolf-Yadlin A., & Gozani O. (2013). A general molecular affinity strategy for global detection and proteomic analysis of lysine methylation. *Molecular Cell* 50(3): 444-456.
- [2] Cao X.-J., Arnaudo A.M., & Garcia B.A. (2013). Large-scale identification of protein lysine methylation in vivo. *Epigenetics* 8(5): 477-485.
- [3] Chan W.C. & White P.D. (2000). *Fmoc Solid Phase Peptide Synthesis*. Oxford University Press.
- [4] Wu Z, Connolly J, Biggar KK. (2017). Beyond histones – the expanding roles of protein lysine methylation. *The FEBS Journal* 284(17).
- [5] Shinkai Y, Tachibana M. (2011). H3K9 methyltransferase G9a and the related molecule GLP. *Genes and Development* 25: 781-788.
- [6] Carlson SM, Gozani O. (2016). Nonhistone Lysine Methylation in the Regulation of Cancer Pathways. *Cold Spring Harbor Perspectives in Medicine* 6(11): a026435.
- [7] Hamamoto R, Saloura V, Nakamura Y. (2015). Critical roles of non-histone protein lysine methylation in human tumorigenesis. *Nature Reviews Cancer* 15(2): 110-124.
- [8] Rowe EM, Xing V, Biggar KK. (2019). Lysine methylation: Implications in neurodegenerative disease. *Brain Research* 1707:164-171.
- [9] Wagner T, Jung M. (2012). New lysine methyltransferase drug targets in cancer. *Nature Biotechnology* 30(7): 622-623.
- [10] Mazur PK et al (2014). SMYD3 links lysine methylation of MAP3K2 to Ras-driven cancer. *Nature* 510(7504): 283-287.

- [11] Lanouette S et al. (2015). Discovery of Substrates for a SET Domain Lysine Methyltransferase Predicted by Multistate Computational Protein Design. *Structure* 23(1): 206-215.
- [12] Hilpert K, Winkler DFH, Hancock REW. (2007). Peptide arrays on cellulose support: SPOT synthesis, a time and cost efficient method for synthesis of large numbers of peptides in a parallel and addressable fashion. *Nature Protocols* 2: 1333-1349.
- [13] Iwabuchi K., Bartel P. L., Li B., Marraccino R., & Fields S. (1994) Two cellular proteins that bind to wild-type but not mutant p53. *Proceedings of the National Academy of Sciences of the USA* 91: 6098–6102.
- [14] Panier S., Boulton S.J. (2014) Double-strand break repair: 53BP1 comes into focus. *Nature Reviews Molecular Cell Biology* 15: 7-18.
- [15] Ochs F., Somyajit K., Altmeyer M., Rask M.-B. Lukas J., & Lukas C. (2016). 53BP1 fosters fidelity of homology-directed DNA repair. *Nature Structural & Molecular Biology* 23(8): 713-721.
- [16] Cuella-Martin R., Oliveira C., Lockstone H.E., Snellenberg S., Grolmusova N., & Chapman J.R. (2016). 53BP1 Integrates DNA Repair and p53-Dependant Cell Fate Decisions via Distinct Mechanisms. *Molecular Cell* 64, 51-64.
- [17] Bennett C. B., Lewis A. L., Baldwin K. K. & Resnick M. A. (1993). Lethality induced by a single site-specific double-strand break in a dispensable yeast plasmid. *Proceedings of the National Academy of Sciences of the USA* 90: 5613–5617.
- [18] Jackson S. P. & Bartek J. (2009). The DNA-damage response in human biology and disease. *Nature* 461: 1071–1078.

- [19] Mehta A., & Haber J.E. (2014). Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair. *Cold Spring Harbor Perspectives in Biology* 6(9): a016428.
- [20] Botuyan M.V., Mer G. (2016). Tudor Domains as Methyl-Lysine and Methyl-Arginine Readers. In Binda O., & Fernandez-Zapico M.E. (Eds.) *Chromatin Signalling and Diseases* (pp. 149-165). Academic Press.
- [21] Tong Q., Cui G., Botuyan M.V., et al. (2015). Structural Plasticity of Methyllysine Recognition by the Tandem Tudor Domain of 53BP1. *Structure* 23(2): 312-321.
- [22] Drané, P. et al. (2017). TIRR regulates 53BP1 by masking its histone methyl-lysine binding function. *Nature* 543: 211–216.
- [23] J. Huang, R. Sengupta, A.B. Espejo, M.G. Lee, J.A. Dorsey, M. Richter, S. Opravil, R. Shiekhattar, M.T. Bedford, T. Jenuwein, S.L. Berger. (2007). p53 is regulated by the lysine demethylase LSD1. *Nature* 449: 105-108
- [24] I. Kachirskaja, X. Shi, H. Yamaguchi, K. Tanoue, H. Wen, E.W. Wang, E. Appella, O. Gozani. (2008). Role for 53BP1 Tudor domain recognition of p53 dimethylated at lysine 382 in DNA damage signaling. *Journal of Biological Chemistry* 283: 34660-34666.
- [25] Akimov V, et al. (2018). UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites. *Nature Structural Molecular Biology* 25: 631-640.
- [26] Hendriks IA, et al. (2014). Uncovering global SUMOylation signaling networks in a site-specific manner. *Nature Structural Molecular Biology* 21: 927-936
- [27] Esteve, P. O. et al. (2011). A methylation and phosphorylation switch between an adjacent lysine and serine determines human DNMT1 stability. *Nature Structural Molecular Biology* 18: 42–48.

- [28] Tsai CF, et al. (2015). Large-scale determination of absolute phosphorylation stoichiometries in human cells by motif-targeting quantitative proteomics. *Nature Communications* 6: 6622.
- [29] Dai Y., Zhang A., Shan S., Gong Z., & Zhou Z. (2018). Structural basis for recognition of 53BP1 tandem Tudor domain by TIRR. *Nature Communications* 9: 2123.
- [30] Li Y., Han J., Zhang Y., Cao F., et al. Structural basis for activity regulation of MLL family methyltransferases. *Nature* 530: 447-452.
- [31] Zhang Y., Mittal A., Reid J., Reich S., Gamblin S.J., & Wilson J.R. (2015). Evolving Catalytic Properties of the MLL Family SET Domain. *Structure* 23(10): 1921-1933.
- [32] Brown D.A., Di Cerbo V., Feldmann A., et al. The SET1 Complex Selects Actively Transcribed Target Genes via Multivalent Interaction with CpG Island Chromatin. *Cell Reports* 20(10): 2313-2327.
- [33] Wu Z., Cheng Z., Sun M., Wan X., Liu P., He T., Tan M., & Zhao Y. (2015). A chemical proteomics approach for global analysis of lysine monomethylome profiling. *Molecular Cell Proteomics* 14(2): 329-339.
- [34] Butler J.S., Lee J.-H., & Skalnik D.G. (2008). CFP1 Interacts with DNMT1 Independently of Association with the Setd1 Histone H3K4 Methyltransferase Complexes. *DNA and Cell Biology* 27(10): 533-543.
- [35] Beurton F., Stempor P., & Maron M. (2018). Physical and functional interaction between SET1/COMPASS complex component CFP-1 and a Sin3 HDAC complex. bioRxiv 436147

Chapter 3

Histone Methyltransferase MLL4 Methylates CXXC Finger

Protein-1 Within a Lysine-Rich Regulatory Domain

3.1 Introduction

In Chapter 2, we employed two different approaches in the identification of novel MLL4 substrates; (1) shotgun and (2) systematic exploration. From these approaches we selected two candidate substrates: K1563 of the DNA damage response-associated 53BP1 (shotgun identified), and K382 of the Setd1A/B complex member CFP1 (systematic identified).

In the case of 53BP1, it has been well-established that this protein is a key regulator of DNA double strand break (DSB) repair [reviewed in 1]. In this role, 53BP1 integrates several molecular pathways to control DSB repair pathway choice. 53BP1 promotes non-homologous end joining (NHEJ) DNA repair, the dominant G₁ DSB repair pathway where non- or minimally processed DNA ends are re-ligated [1]. MLL4-dependant methylation of 53BP1 is intriguing as a possible enzyme-substrate relationship to explore, as it implicates MLL4 in DSB repair decisions and more broadly in the DNA damage response.

We identified K328 of CXXC finger protein 1 (CFP1) as a potential MLL4 substrate through our systematic exploration of MLL4 substrate specificity. CFP1 plays an integral role in histone modification. Histone modifications must be both temporally and spatially coordinated in order to properly regulate modification-dependant chromatin accessibility and gene expression. The COMPASS (COMplex of Proteins ASSociated with Set1) and COMPASS-like protein complexes that associate with KMT2-class KMTs and enhance methyltransferase activity towards histone H3 proteins are thought to contribute to this special and temporal coordination through complex interactions with chromatin-associated molecular structures. The COMPASS complex that associates with the mammalian KMTs Set1A and Set1B contains CFP1 [2], where it acts to guide the KMT complex to appropriate targets. This occurs through bivalent interaction CFP1, H3K4me₃, and genomic DNA CGIs – regions of around one kilobase in length with high

CpG content lacking DNA methylation, often found in regions containing H3K4me3 [3]. This bivalent interaction with H3K4me3 and CGIs occurs through the methyl-binding plant homeodomain (PHD) and CXXC domain of CFP1, respectively [4, 5]. This interaction is the primary means through which Set1A/B KMTs are recruited to chromatin, and their activity restricted to appropriate targets [4, 6].

H3K4me3 is a modification found at the promoters of actively expressed genes in yeast [7] and is associated with increased gene activity in mammals [8], although the exact function remains unclear at this time. It has been suggested that H3K4me3 may play a role in transcriptional responsiveness (i.e., faster transcriptional response at key genes under stressful environmental conditions) and transcriptional consistency [9]. CFP1 is required for the deposition and maintenance of H3K4me3 marks at many gene promoters [10, 11]. CFP1-deficiency also leads to reductions in the relative levels of H3K9 and K14 acetylation [11], demonstrating a possible CFP1-dependant crosstalk between H3 lysine methylation and acetylation modifications. Interestingly, lack of CFP1-dependant H3K4me3 does not appear to drastically affect steady-state transcription nor transcriptional induction output [10, 11], further highlighting the unclear contribution of H3K4me3. Furthermore, mouse embryos lacking CFP1 fail to gastrulate and are non-viable at 4.5-6.5 days post-coitus [12], and while CFP1-deficient mouse embryonic stem cells are viable, they fail to differentiate and exhibit a relative three-fold increase in apoptosis [13]. Collectively, these results highlight the importance of CFP1-dependant H3K4 trimethylation despite the apparent minor effect on transcription.

Structurally, CXXC domains are cysteine-rich small zinc-finger domains that functionally recognize and bind to patterns of cytosine and guanidine nucleotides in DNA [14]. There are four classes of CXXC domains in humans, each class defined by the specific DNA motif recognized

by the domain. The CXXC domain of CFP1 is the sole member of Class 1, preferentially binding the CcgG motif [15]. Interestingly, the motif recognized by CXXC domains can themselves be considered regulatory structures. Artificial CpG clusters are able to recruit CFP1-bound Set1 to induce local H3K4me3, positively influencing gene expression [16].

The carboxy-terminus of CFP1 contains the Setd1-interacting domain (SID), and it is reportedly located between conserved basic and coiled-coil domains at amino acids 361-422 (Figure 2.6) [17]. It is noteworthy that the reasoning for its exact definition was not made clear in the publication that first reported the SID. Association of CFP1 with the Setd1 complexes occurs through a multivalent interaction between the SID, C-terminal PHD2 domain and the most distal C-terminal amino acids. Truncation or mutation within any of these sites abolishes Setd1 interaction [17].

Recent reports show that CFP1 associates with a Sin3 histone deacetylase (HDAC) complex in *C. elegans* [18]. This interaction is mediated through a cysteine-rich region in the C-terminus of CFP1, and recruits Sin3 with the associated HDAC to H3K4me3 enriched promoters. Interestingly, mutations in CFP1, Sin3 and the *C. elegans* SET1 homolog result in similar phenotypes, suggesting a key role for CFP1 in histone acetyl- and methyllysine crosstalk.

Further complicating the function of CFP1, this protein has also been implicated in cytosine methylation. CFP1-deficient murine embryonic stem cells exhibited a 70% reduction of global cytosine methylation affecting single-copy genes, imprinted loci, and repetitive elements [13]; however, this deficiency is a consequence of impaired maintenance cytosine methylation as is typical of DNMT1 activity, rather than a disruption of *de novo* methylation [19]. Defects in CFP1-deficient embryonic stem cells are reversed upon introduction of a CFP1 expression vector, underlining the direct causal relationship between the perturbed epigenetic phenotype and

CFP1 deficiently [19]. Reductions of global cytosine methylation are also observed in CFP1-depleted zebrafish embryos [20].

The DNMT1 interaction has been shown to occur through the CFP1 via N-terminal (amino acids 1-123), central (103-367) and C-terminal (361-656) regions [17]. Interestingly, this appears to be independent of Setd1, as mutations within the SID that abolish Setd1 binding do not affect the binding of DNMT1 [17]. Together, this suggests that CFP1 can independently associate with cytosine and histone methylation machinery. Direct interaction between bacterially expressed CFP1 and DNMT1 has not been demonstrated *in vitro*, potentially introducing the possibility of post-translational modification or chromatin-associated protein interaction in regulating the association [17]. In the results presented in Chapter 2, we identified a region within CFP1 which is highly methylated by the Setd1 paralog mixed lineage leukemia 4 (MLL4) *in vitro*. It is intriguing to hypothesize that dynamic MLL4-regulated methylation events within the lysine-rich basic region of the SID may be a mechanism through which CFP1 can ‘switch’ between association with Setd1 and DNMT.

For this study, I have employed *in vitro* methylation assays to validate candidate MLL4 WRAD substrates, CFP1 and 53BP1, previously identified in Chapter 2. This will be accomplished by expanding potential peptide-methylation activity to the methylation of recombinant protein.

3.2 Materials & Methods

3.2.1 Protein expression and purification

p3E-MLL4 (5382–5537) SET/Post-SET domain construct was expressed as a glutathione S-transferase (GST) fusion protein. *Escherichia coli* BL21 cells containing either construct were grown at 37 °C to an OD₆₀₀ of ~0.6 in lysogeny broth (LB) containing 1 mM MgCl₂ and 1 µM ZnSO₄. Expression was induced using 0.1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG), and proceeded for 18 hours at 16 °C. After centrifugation at 2,500xg for 15 minutes, the bacterial pellet was lysed in 50 mM Tris pH 7.5 buffer containing 300 mM NaCl, 3 mM DTT, 1 µM ZnSO₄, 0.01% Triton X-100, 1 mg/mL chicken egg lysozyme (Sigma), 10% glycerol; and a protease inhibitor cocktail consisting of 100 µM PMSF, 1 µM E-64, 1 µM pepstatin, and 5 µM bestatin. The lysate was clarified through centrifugation at 18,000xg for 45 minutes. The fusion protein was isolated from clarified lysate by affinity chromatography with glutathione sepharose 4B beads (GE Healthcare) and eluted using 50 mM Tris pH 8 containing 10 mM glutathione.

53BP1 TTD (1484-1603) was expressed as a glutathione S-transferase (GST) fusion protein. *Escherichia coli* BL21 cells containing the construct were grown at 37 degrees C to an OD₆₀₀ of ~0.6 in lysogeny broth (LB) containing 1 mM MgCl₂. Expression was induced using 0.1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG), and proceeded for 18 hours at 16 °C. After centrifugation at 2,500xg for 15 minutes, the bacterial pellet was lysed in 50 mM Tris pH 7.5 buffer containing 200 mM NaCl, 1 mM DTT, 0.01% Triton X-100, 1 mg/mL chicken egg lysozyme (Sigma), 10% glycerol; and a protease inhibitor cocktail consisting of 100 µM PMSF, 1 µM E-64, 1 µM pepstatin, and 5 µM bestatin. The lysate was clarified through centrifugation at 18,000xg for 45 minutes. The fusion protein was isolated from clarified lysate by affinity chromatography with

glutathione sepharose 4B beads (GE Healthcare) and eluted using 50 mM Tris pH 8 containing 10 mM glutathione.

Full length WDR5, RbBP5, Ash2L, Dpy-30, and CFP1 (302-656) were each expressed as 6xHistidine fusion proteins. Each construct was grown in *E. coli* BL21 cells at 37 °C to an OD₆₀₀ of ~0.6 in lysogeny broth (LB) containing 1 mM MgCl₂. Expression was induced using 0.1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG), and proceeded for 18 hours at 16 °C. After centrifugation at 2,500xg for 15 minutes, the bacterial pellet was lysed in P5 (50 mM NaHPO₄ pH 7.0, 500 mM NaCl, 5 mM imidazole, 10% glycerol, 0.05% Triton X-100) containing; protease inhibitor cocktail (100 μM PMSF, 1 μM E-64, 1 μM pepstatin, and 5 μM bestatin). The lysate was clarified through centrifugation at 18,000xg for 45 minutes. The fusion proteins were then each isolated from clarified lysate by affinity chromatography with Ni-NTA beads (QIAGEN) and eluted using P500 (P5 with 500mM imidazole).

3.2.2 Bioluminescence-based end-point methyltransferase activity assays

Methyltransferase activity was assayed using the Promega MTase-Glo™ Methyltransferase assay kit (Cat# V7601). MLL4 (5382–5537), WDR5, Ash2L, RbBP5, Dpy30, 53BP1 (1484-1603), CFP1 (302-656), and H3K4 fractions were dialysed in reaction buffer (20mM Tris pH 8.0, 50mM NaCl, 1mM EDTA, 3mM MgCl₂, 1mM DTT) for 2 hours at 4 °C. Substrate solutions were prepared in 4X concentrated reaction buffer with 0.4 mg/mL bovine serum albumin (BSA). After addition of enzyme solution, the final concentrations of substrate buffer components were 15 μM of substrate (either 53BP1 (1484-1603), CFP1 (302-656), or H3K4), 10 μM S-adenyl-L-methionine (SAM), and 0.1 mg/mL BSA in 1X concentrated reaction

buffer. Equimolar mixtures of the MLL4 COMPASS-like complex members were combined such that the final MLL4 complex concentration was 0.5 μ M after addition to substrate buffer.

Each individual reaction contained 4 μ L each of substrate buffer and enzyme solution combined in a LumiTrax 200 opaque white 384-well plate (Greiner). Reactions were plated in triplicate. The plate was gently and briefly agitated on a benchtop plate shaker and centrifuged for 1 minute at 500xg. The reaction proceeded for 16 hours at room temperature in a dark location. Following incubation, 2 μ L of 5X MTase-Glo™ reagent (400 mM Tris buffer pH 8.0, 40 mM MgCl₂, 75 mM NaCl, 125 μ M dGTP, 200 μ M sodium metaphosphate, 80 μ g/ml polyphosphate - AMP phosphotransferase, 200 μ g/ml SAH-Hydrolase, 50 μ g/ml adenosine kinase, 5 mM DTT [21]) was added to each well. The plate was then gently agitated and centrifuged for 1 minute at 500xg, after which the reaction was incubated at room temperature for 60 minutes in a dark location. Next, 10 μ L of detection solution (1 mM potassium phosphoenolpyruvate, 8 U/ml pyruvate kinase [21]) was added to each well, and the plate was once again reacted for 60 minutes at room temperature in the dark following agitation and centrifugation as described above. At the end of this incubation period, End-point relative luminescence readings were taken using a BioTek Cytation 5 plate reader. Statistical significance was calculated with iterative one-way ANOVAs using GraphPad Prism 7.

3.2.3 Methylation window array design, SPOT synthesis, and array processing

To identify the specific lysine methylated within each substrate, every possible methylation window within 53BP1 (1484-1603) and CFP1 (302-656) was selected for screening. The methylation window was defined at +/- 10 residues flanking a central lysine (K), resulting in peptides of 21 residues in length. All peptides were separated from the array surface through the

addition of a N-terminal flexible 6-aminohexanoic acid group. This +/- 10 residue length was chosen as, (1) it is near the maximum possible length reliably allowed by our synthesis procedure, and (2) using larger windows lowers the risk of excluding key distal residues. Three control peptides were printed for each methylation window to confirm that the central lysine is in fact the methylated residue. Each control SPOT contained lysine to arginine (K-to-R) substitutions at the following loci: (1) the central K residue, (2) all flanking K residues (excluding the central K), and (3) at all K residues within the window (i.e., flanking and central K residues). H3K4 was included as a positive control.

As described in Chapter 2 Section 2.2.2, solid-phase peptides were synthesized directly on a sheet of aminated cellulose filter paper with an Intavis Multiprep instrument using Fmoc chemistry at an approximate concentration of 0.2 μmol / SPOT. Following synthesis, the arrays were deprotected by three sequential washes of 20% piperidine dissolved in dimethylformamide (DMF), followed by three washes in DMF, and three final washes with anhydrous ethanol. Arrays were then dried by air. After drying, sidechain protecting groups were cleaved from the peptides with a trifluoroacetic acid solution (95% TFA, 3% TIPS and 2% H_2O). The arrays were again washed three times with dichloromethane (DCM), followed by three washes with DMF, and three washes of anhydrous ethanol, and allowed to air dry overnight in a fume hood. Immediately before use, arrays were briefly washed with anhydrous ethanol, then equilibrated in a solution of 50% ethanol in methylation buffer (20 mM Tris pH 8.0, 50 mM NaCl, 3 mM MgCl_2 , 1 mM EDTA, 1 mM dithiothreitol (DTT)), and then finally only the methylation buffer.

3.2.4 Array-based systematic probing of 53BP1 (1484-1603) and CFP1 (302-656) methylation windows

The MLL4 WRAD complex was assembled by combining equal molar ratios of MLL4, WDR5, RbBP5, Ash2L and Dpy-30. The assembled MLL4 WRAD complex was then aliquoted and snap-frozen for storage at -80 °C in the presence of 20% glycerol. For array-based methylation experiments, a total of 2 µM of MLL4 WRAD in methylation buffer was incubated at 4 °C overnight with 0.466 µM adenosyl-L-methionine S-[methyl-3H] (specific activity: 80Ci/mmol) (Perkin Elmer) on a rocking platform. After incubation with MLL4 WRAD, arrays were washed with 10 mL of wash buffer (100 mM NH₄HCO₃, 1% sodium dodecyl sulfate (SDS)) and the array was incubated on a rocker for 3 minutes at room temperature. Washes were repeated 5 x 3 minutes. Following washing, the array was transferred to a new containing enough anhydrous ethanol to completely submerge the array. The array was then incubated in ethanol on a rocking platform for 2 min. Following incubation of the array in ethanol, the ethanol was removed and enough 7% DPO solution (7 g of 2,5-Diphenyloxazole in 100 mL ethanol) was added to complete submerge the array. Arrays were then gently agitated for 2 min to ensure uniform distribution of the DPO solution. The arrays were then removed from the DPO solution and allowed to dry completely on filter paper overnight prior to array fluorography. For fluorography of the methylated arrays, the dried arrays were wrapped with a single layer of plastic food wrap and placed into an autoradiography cassette containing intensifying screens (Carestream BioMax TransScreen LE). In complete darkness, a sheet of detection film (Amersham Hyperfilm MP) was added to the cassette above the array. The loaded cassette was closed, and the film was then exposed to array fluorography at -80 °C for 1 week. After exposure the film developed using Ilford ilfosol 3 film

developer in complete darkness and then fixed with Ilford rapid fixer according to manufacturer's instructions under a safe light.

3.3 Results & Discussion

3.3.1 *In vitro* methyltransferase activity assays reveal MLL4 WRAD-dependant methylation of CFP1 (302-656), but not 53BP1 TTD (1484-1603)

Surprisingly, luminescence values for H3K4, 53BP1 (1484-1603), and a no substrate negative control appeared to be similar and were found not to be statistically different (Figure 3.1). The no substrate control contained only MLL4 WRAD and reaction buffer in place of the substrate. This treatment represents the background luminescence of the MLL4 WRAD solution and possible auto-methylation of one or more members of the complex itself. Indeed, it is known that MLL1 undergoes a robust auto-methylation reaction in the absence of histone H3, and that MLL4 may do the same in addition to methylating complex member Ash2L [22, 23]. Because the no substrate control luminescence is significantly greater than the no enzyme control, which contains H3K4, but not MLL4 WRAD ($p < 0.0001$), we can assume this value indeed represents the background auto-methylation activity of MLL4 WRAD in absence of a valid substrate. This indicates that there was likely no true *in vitro* methylation of 53BP1. However, the statistically similar luminescence between the no substrate control and the H3K4 control likely indicates that H3K4 methylation and the auto-methylation produced a similar amount of luminescence, rather than the enzyme complex being inactive. This is exemplified when comparing to the MLL4 WRAD methylation of CFP1 (302-656) (Figure 3.1), which would not be possible in the absence of an inactive complex, and because the complex was able to methylate H3K4 and numerous CFP1 sites on the substrate peptide array (Figure 3.2).

Indeed, the luminescence generated by the CFP1 substrate reaction was over ten-fold greater than that generated by H3K4 ($p \leq 0.0001$). As the concentration of both the H3K4 and CFP1 substrates were synthesized as the same concentration (and therefore, should be similar), this increase could be due to CFP1 (302-656) undergoing a greater degree of relative methylation by MLL4 WRAD, suggesting the construct is more readily methylated *in vitro* than H3K4. This can be achieved through one of two mechanisms, (1) a lower K_m for a specific site, or (2) the existence of multiple methylatable residues beyond K328. To test for the later possibility, we next employed a peptide array featuring each possible methylation window within our constructs to identify the specific lysine (or multiple lysine residues) subject to MLL4 methylation.

3.3.2 MLL4-COMPASS methylates several lysine residues within the CFP1 basic domain, a lysine-rich regulatory region

Unsurprisingly, little to no 53BP1 methylation was observed on the methylation window array (Figure 3.2). A low-strength signal was detected at the SPOT corresponding with the K1514 residue, but otherwise there was no indication of methylation activity on the 53BP1 TTD peptides. Ultimately, these results cast doubt on the existence of an enzyme-substrate relationship between MLL4 and 53BP1, and thus presents 53BP1 as a lower priority for future exploration in this thesis.

In contrast, a number of CFP1 peptides from the window substrate array were found to be methylated. This result agrees with observations from Section 3.3.1 that suggest the possibility of multiple methylation events within the CFP1 (302-656) construct (Figure 3.2). Methylation was found to occur on residues K331, K335, K336, K339, K340, K341, K346, K351, K353, K355, K357, and K359 (Figure 3.2). Surprisingly, the predicted K328 substrate which initially led us to investigate CFP1, did not appear to undergo *in vitro* methylation on the array. In fact, sequence

alignment between our recognition motif and the methylated CFP1 peptides shows little overlap except in peptides where K328 is a flanking residue, in which case the motif perfectly aligns (Figure 3.3). The proximity of these methylation events combined with the length of the window array peptides resulted in the methylation of the central K-to-R control peptides. Arginine substitution of the central K residue didn't appear to impede methylation of the flanking K residues. Methylation was also abolished in the 'flanking K-to-R' control SPOTs, suggesting central K methylation is dependent on the presence of at least one flanking K residue(s), whereas arginine substitution is intolerable. The negative control 'all K-to-R' peptides displayed no methylation signal. Overall, these results suggest that MLL4 methylates a plurality of lysine residues spanning amino acids 321-369 in CFP1.

3.3.3 MLL4-dependant basic domain methylation could regulate CFP1 activity through an undiscovered methyl-switch mechanism

Interestingly, the high degree of methylation within CFP1 peptides almost directly coincides with the defined basic region of CFP1 that spans amino acids 321-360, with K328 being the only currently known putative methylation site within the whole protein (Figure 3.4) [24]. This basic region of CFP1 was first described upon discovery of CFP1 nearly two decades ago, where it was speculated that the region may facilitate CFP1 homodimerization, driven by results showing that a fragment of CFP1 containing the basic region formed a second lower-mobility complex in EMSA assays [25]. Intriguingly, this observation pre-dated discovery of the SID, and in fact the fragment described (106 to 345) actually contains the basic domain as well as a significant portion of the SID. Though poorly studied in the interim, recent studies have revealed that heterologously expressed CFP1 forms homodimers in yeast [18]. Furthermore, stoichiometric analysis of

Setd1A/B complex pull-downs revealed that CFP1 is 1.6 times more abundant than the core KMT subunit [26]. Finally, it is of note that the basic domain has been shown to contribute to Setd1 binding through an unknown mechanism. Though not discussed in their publication, Butler et al. showed that a CFP1 fragment lacking the basic domain (361-656) reduced Setd1 association by approximately 50% compared to a similar fragment containing the basic domain (302-656) [17]. It is interesting to speculate that MLL4-dependant methylation of the basic domain may be a contributor to Setd1 binding. Furthermore, investigation of the link between the basic domain and possible homodimerization appears merited. Considering the implication of the basic domain in this proposed interaction, it is interesting to ponder a possible role for MLL4-dependant methylation in its mediation.

To date, the biochemical nature of the interactions between CFP1 SID and Setd1 proteins has not been studied. This interaction occurs between the CFP1 SID and a region within Setd1 (amino acids 1415-1450 in Setd1A) [17, 30]. It is known that mutations within the SID abolish association [17], but the specific physiochemical cause of this abolishment is unresolved. Furthermore, the reasoning for defining the SID as occurring between amino acids 361 and 422 is unclear. The role of the SID, and more broadly the nature of CFP1-Setd1 association is further confounded by studies showing a remarkable redundancy among CFP1 domains in rescuing appropriate cytosine methylation, histone methylation, plating efficiency, and *in vitro* differentiation in *CXXCI* knockout embryonic stem (ES) cells [6]. CFP1 is required for restriction of the Setd1 complex to euchromatin, [6], and as such *CXXCI* knockout ES cells display inappropriately high levels of H3K4me3 owing to an increase in Setd1 promiscuity [27]. Surprisingly, full-length CFP1 with a C375A mutation within the SID was able to rescue appropriate H3K4me3 levels, despite previous studies showing that this mutation abolished Setd1

binding [17]. Similarly, truncated CFP1 lacking the PHD2 domain was shown to fail to interact with Setd1, and yet similar (surprisingly, shorter) fragments are able to rescue H3K4me3 activity [27]. This strongly suggests that the H3K4me3 rescue enabled by these fragments is due to a mechanism independent of Setd1 association. Perhaps rescue activity displayed by CFP1 fragments unable to associate with Setd1 is a product of an undescribed CFP1 protein complex.

It is interesting to speculate mutations that abolish Setd1 interaction mimic the action of an MLL4-dependant methyl switch. MLL4-dependant methylation of the basic domain could modulate CFP1 association with Setd1 and other associated proteins, described or otherwise. The methylation state associated with the abrogation of Setd1 association may conversely promote association with these other binding partners. To test this hypothesis, we designed a pull-down experiment to compared Setd1 affinity between methylated and unmethylated CFP1. The results of this experiment are presented in Chapter 4.

3.3.3 Basic domain methylation could promote CFP1 association with DNMT1 or MLL4 PHD domains

It is worth noting that while association while CFP1 association with DNMT1 is well established, the specific function of the interaction is not. CFP1-deficient murine ES cells exhibited a ~70% reduction of global cytosine methylation, a 45% reduction in total DNA methyltransferase activity, and a similar reduction in DNMT1 expression [19]. These deficits are reversed by introduction of a CFP1 expression vector, directly implicating CFP1-DNMT1 in the regulation of these processes. It is claimed that CFP1 binding increases the stability and the translation of DNMT1, but this conclusion is drawn from unpublished data with no further justification or explanation [17]. This is indeed plausible, but in the absence of supporting evidence it remains an

unproven claim, leaving the function of CFP1-DNMT1 interaction unknown. Intriguingly, CFP1 amino acids 1-123, 103-367, and 361-656 are each sufficient for interaction with DNMT [17]. Additionally, at least one of these regions was present in every CFP1 fragment used to probe rescue function [27]. In line with our hypothesis, Setd1-binding-deficient CFP1 may in turn preferentially interact with DNMT1. It is interesting to speculate that an unelucidated function of CFP1-DNMT1 could be directly or indirectly responsible for the rescued maintenance of appropriate H3K4me3 levels in *CXXC1* knockout ES cells. It is also possible that MLL4-dependant methylation of CFP1 could facilitate interaction with the MLL4 PHD domains, which could subsequently induce a novel function in the MLL4 complex. As such, we tested the binding of MLL4 PHD domains with CFP1 of various methylation states. The results of those experiments are presented in Chapter 4.

3.3.4 Three SID-adjacent lysines shown to be methylated by the mammalian KMT MLL4 exist only in mammals

Because MLL4 is a KMT that is only found in vertebrates, it is likely that the existence of an MLL4-dependant CFP1 regulatory mechanism would appear around the same evolutionary period. Excitingly, that seems to be the case. Each methylated lysine observed in Figure 3.2 appears to have first emerged in the fish *D. rerio*, which also possesses MLL4 (Figure 3.5). The appearance of a string of methylated lysines within a domain implicated in dimerization coinciding with the appearance of a KMT able to methylate them is conspicuous. Intriguingly, K355, K357, and K359 are each only present in mammals (Figure 3.5). It is noteworthy that these residues are also the ones closest to the SID. It is possible that their addition further refined the function of the mechanism.

Following these results, we wondered if a similar pattern would be seen in DNMT1. Though the specific interactions that facilitate CFP1 binding are unknown, we know that the regions within DNMT1 amino acids 167-493 and 970-1616 are sufficient for association [17]. Additionally, an inhibitor peptide that bound DNMT1 at amino acids 1081-1097 abolished CFP1 binding [28]. With this in mind, we performed a multiple sequence alignment using Clustal Omega to identify any emergent substitutions within these regions exclusive to vertebrates (Figure 3.6). Unfortunately, there were no obvious candidates for newly evolved interaction regions, and the majority of both 167-493 and 970-1616 sequences is highly conserved, even in *T. castaneum*, which possesses DNMT1 despite the absence of DNA CpG methylation in its genome [31]. Further investigation into the means through which DNMT1 and CFP1 interact could provide additional insight regarding its evolutionary conservation.

Figures

Figure 3.1. MLL4 WRAD methylates CFP1 (302-656) but not 53BP1 TTD in a bioluminescence-based end-point methyltransferase assay. 0.5 μ M MLL4 WRAD supplemented with 10 μ M SAM was reacted with 15 μ M of either H3K4, 53BP1 TTD, CFP1 (302-656) using a Promega MTase-Glo™ kit (Cat# V7601). The reaction buffer contained 20mM Tris pH 8.0, 50mM NaCl, 1mM EDTA, 3mM MgCl₂, 1mM DTT and 0.1 mg/mL BSA. 'No enzyme' contained H3K4 and an equivalent volume of reaction buffer in lieu of MLL4 WRAD. 'No substrate' contained MLL4 WRAD and reaction buffer in lieu of substrate solution. 'H3K4 (MLL4 only)' contained both MLL4 and H3K4, but WRAD was replaced with reaction buffer. Reactions ran for 16 hours at room temperature in a dark location and were processed according to the manufacturers instructions. Luminescence readings were taken using a BioTek Cytation 5 plate reader. **A.** Graph showing full results. **B.** Graph with CFP1 removed to better illustrate smaller magnitude results.

Figure 3.2. Methylation window peptide array shows extensive MLL4 WRAD-dependent methylation of CFP but no methylation of 53BP1 TTD. Every possible CFP1 (302-656) and 53BP1 TTD methylation window (central K +/- 10 residues) was printed on an aminated cellulose array using SPOT synthesis. For each peptide control peptides were printed either substituting the central K with R, all Ks flanking the central K with R, or all Ks substituted with R. Array was incubated with 2 μ M of MLL4 WRAD in methylation buffer (50 mM Tris pH 8.5, 3 mM MgCl₂, 1mM EDTA, 1 mM DTT) and incubated at 4 °C overnight with 0.466 μ M SAM S-[methyl-3H]. Amersham Hyperfilm MP film was exposed to array fluorography at -80 °C for 1 week and developed using ilfisol 3 developer and fixer.

Figure 3.3. The methylated lysines identified by the methylation window peptide array span the majority of the basic region and are directly adjacent to the Setd1 interaction domain (SID). Figure adapted from [17].

A.

K331

B.

K335

C.

K335

A.

K339

B.

K340

C.

K359

Figure 3.4. The MLL4 recognition motif poorly aligns with CFP1 methylation window sequences that underwent *in vitro* MLL4 WRAD-dependent methylation on a SPOT peptide array. Alignment was performed using Tomtom Motif Comparison Tool [29, (<http://meme-suite.org/doc/tomtom.html>)]. **A.** K331. **B.** K335. **C.** K336. **D.** K339 **E.** K340. **F.** K359. Methylated peptides which did not align with the recognition motif are omitted.

A.

B.

Figure 3.5. A. Cladogram illustrating the evolutionary relatedness of CFP1 in *H. sapiens* (Human), *M. Musculus* (Mouse) *A. carolinensis* (Lizard) *D. rerio* (Fish) *D. melanogaster* (Fly) *C. elegans* (Worm), and *S. cerevisiae* (Yeast). **B.** Methylated lysines in the basic region of CFP1 are conserved

Figure 3.6. Sequence alignment of DNMT1 regions that facilitate interaction with CFP1. Alignments performed by Clustal Omega Multiple Sequence Alignment (EMBL-EBI) using the default parameters and colored with the Clustalx scheme. Alignment visualized using Jalview. **A.** N-terminal CFP1 interaction region. **B.** C-terminal CFP1 interaction region. Box indicates proposed interaction region as described in [29]

References

- [1] Mehta A., & Haber J.E. (2014). Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair. *Cold Spring Harbor Perspectives in Biology* 6(9): a016428
- [2] Lee J-H, Skalnik DG. (2005). CpG-binding protein (CXXC Finger Protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of yeast Set1/COMPASS complex. *Journal of Biological Chemistry* 280: 41725-41731.
- [3] Illingworth RS, Bird AP. 2009. CpG islands–‘a rough guide.’ *FEBS Letters* 583: 1713–1720.
- [4] Lee JH, Voo KS, Skalnik DG. 2001. Identification and characterization of the DNA binding domain of CpG-binding protein. *Journal of Biological Chemistry* 276: 44669–44676.
- [5] Brown et al. (2017). The SET1 Complex Selects Actively Transcribed Target Genes via Multivalent Interaction with CpG Island Chromatin. *Cell Reports* 20, 2313-2327.
- [6] Tate C.M., Lee J.-H., & Skalnik D.G. (2010). CXXC Finger Protein 1 Restricts the Setd1A Histone H3-Lys4 Methyltransferase Complex to Euchromatin. *FEBS Journal* 277(1): 210-223.
- [7] Santos-Rosa H., Schneider R., Bannister A.J., Sherriff J, Bernstein BE, Emre NC, Schreiber SL, Mellor J, Kouzarides T. (2002) Active genes are tri-methylated at K4 of histone H3. *Nature*. 419: 407-411.
- [8] Kusch T. (2012). Histone H3 lysine 4 methylation revisited. *Transcription* 3: 310–314.
- [9] Howe FS, Fischl H, Murray SC, Mellor J. (2016). Is H3K4me3 instructive for transcription activation? *Bioessays* 39(1): 1600095.
- [10] Clouaire T, Webb S, Skene P, Illingworth R, Kerr A, Andrews R, et al. (2012). Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes and Development* 26(15): 1714–1728.

- [11] Clouaire T, Webb S, Bird A. (2014). Cfp1 is required for gene expression-dependent H3K4 trimethylation and H3K9 acetylation in embryonic stem cells. *Genome Biology* 15, 451.
- [12] Denissov S et al. (2014). Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant. *Development* 141: 526-537.
- [13] Carlone DL, Skalnik DG. (2001). CpG binding protein is crucial for early development. *Molecular Cell Biology* 21: 7601-7606.
- [14] K.S. Voo, D.L. Carlone, B.M. Jacobsen, A. Flodin, D.G. Skalnik. Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Molecular Cell Biology* 20(6): 2108-2121.
- [15] Chao X., Lie K, Lei M, Yang A, Li Y, Hughes TR, Min J. (2018). DNA Sequence Recognition of Human CXXC Domains and Their Structural Determinants. *Structure* 26(1): 85-95.
- [16] Thomson JP et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* 464, 1082-1086.
- [17] Butler JS, Lee JH, Skalnik DG. (2008). CFP1 interacts with DNMT1 independently of association with the Setd1 Histone H3K4 methyltransferase complexes. *DNA Cell Biology* 27(10): 533-543.
- [18] Beurton F., Stempor P., Caron M., Appert A., Dong R., Chen D., Cluet D., Coute Y., Herbette M., Huang N., Polveche H., Spichty M., Bedet C., Ahringer J., & Palladino F. (2018). Physical and functional interaction between SET1/COMPASS complex component CFP-1 and a Sin3 HDAC complex. bioRxiv 436147.

- [19] Carlone D.L. et al. (2005). Reduced genomic cytosine methylation and defective cellular differentiation in embryonic stem cells lacking CpG binding protein. *Molecular Cell Biology* 21: 7601-7606.
- [20] Young SRL, Mumaw C, Marrs JA, Skalnik DG. (2006). Antisense targeting of CXXC finger protein 1 inhibits genomic cytosine methylation and primitive hematopoiesis in zebrafish. *Journal of Biological Chemistry* 281: 37034-37044.
- [21] Hsiao K., Zegzouti H., & Goueli S.A. (2016). Methyltransferase-Glo: a universal, bioluminescent and homogenous assay for monitoring all classes of methyltransferases. *Epigenomics* 8(3): 321-339.
- [22] Patel A., Vought V.E., Swatkoski S., et al. (2014). Automethylation Activities within the Mixed Lineage Leukemia-1 (MLL1) Core Complex Reveal Evidence Supporting a “Two-active Site” Model for Multiple Histone H3 Lysine 4 Methylation. *Journal of Biological Chemistry* 289(2): 868-884.
- [23] Shinsky S.A., Monteith K.E., Viggiano S., & Cosgrove M.S. (2015). Biochemical Reconstitution and Phylogenetic Comparison of Human SET1 Family Core Complexes Involved in Histone Methylation. *Journal of Biological Chemistry* 290: 6361-6375.
- [24] Wu Z., Cheng Z., Sun M., Wan X., Liu P., He T., Tan M., & Zhao Y. (2015). A chemical proteomics approach for global analysis of lysine monomethylome profiling. *Molecular Cell Proteomics* 14(2): 329-339.
- [25] Voo K.S., Carlone D.L., Jacobsen B.M., Flodin A., & Skalnik D.G. (2000). Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA Methyltransferase, Human Trithorax, and Methyl-CpG Binding Domain Protein 1. *Molecular Cell Biology* 20(6): 2108-2121.

- [26] van Nuland R., Smits A.H., Pallaki P., Jansen P.W., Vermeulen M., Timmers H.T. (2013). Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes. *Molecular Cell Biology* 33(10): 2067-2077.
- [27] Tate CM, Lee J-H, Skalnik DG. (2009). CXXC Finger Protein 1 Contains Redundant Functional Domains That Support Embryonic Stem Cell Cytosine Methylation, Histone Methylation and Differentiation. *Molecular and Cell Biology* 29, 3817-3831.
- [28] Cheray M., Nadaradjane A., Bonnet P., Routier S., Vallette F.M., & Cartron P.F. (2014). Specific inhibition of DNMT1/CFP1 reduces cancer phenotypes and enhances chemotherapy effectiveness. *Epigenomics* 6(3): 267-275
- [29] Gupta S., Stamatoyannopolous J.A., Bailey T., & Noble W.S. (2007). Quantifying similarity between motifs. *Genome Biology* 8(2): R24.
- [30] Lee J.-H. & Skalnik D.G. (2008). Wdr82 Is a C-Terminal Domain-Binding Protein That Recruits the Setd1A Histone H3-Lys4 Methyltransferase Complex to Transcription Start Sites of Transcribed Human Genes. *Molecular and Cellular Biology* 28(2): 609-618.
- [31] Schulz N.K.E., Wagner C.I., Ebeling J., et al. (2018). Dnmt1 has an essential function despite the absence of CpG DNA methylation in the red flour beetle *Tribolium castaneum*. *Scientific Reports* 8: 16462.

Chapter 4

MLL4-dependant Methylation of CXXC Finger Protein-1 Potentiates
Interaction with Setd1A and May Influence Interaction with MLL4 PHD
Domains

4.1 Introduction

4.1.1 MLL4-dependant methylation of CFP1 could modulate association with Setd1 methyltransferases

In Chapter 3, we observed MLL4-dependant methylation of CFP1 within a lysine rich region that may be a key contributor to Setd1A/B association and therefore chromatin remodelling. Synthesis of research examining CFP1-Setd1 association revealed the possibility that dimerization could potentiate this interaction, and that dimerization could occur through the SID and neighboring basic domain. We consequently hypothesized that the methylation of key lysine residues within the basic domain, namely K331, K335, K336, K339, K340, K341, K346, K351, K353, K355, K357, and K359 (Figure 3.2), could modulate dimerization, either promoting or discouraging Setd1 binding.

To explore this hypothesis, we designed a pulldown experiment using Setd1A (1362-1563) as an affinity capture trap, evaluating the binding ability of both methylated and unmethylated CFP1 (302-656).

4.1.2 MLL family PHD fingers regulate H3K4-specific activity, but other functions are unknown

Plant homeodomain (PHD) zinc fingers are regulatory structures usually found in chromatin modifying and transcription regulating proteins [reviewed in 1]. They are typified by a conserved tertiary structure known as a PHD fold stabilized by zinc atoms within a cysteine-histidine-cysteine motif. PHD fingers primarily bind histones and associated histone modifications, including but not limited to H3R2, H3K14ac, and H3K4me3. PHD fingers interact with trimethylated H3K4 in a means similar to that seen in the Tudor domain: an aromatic cage

surrounds the trimethyl ammonium group and is stabilized by van der Waals and cation- π interactions [1, 2]. PHD fingers are interesting in that they allow for combinatorial histone mark reading, either through a single domain or through tandem PHD fingers. This leads us to speculate on the role of PHD fingers in MLL proteins, and specifically the role of the triple PHD finger cassettes found in MLL4.

Indeed, all MLL proteins contain PHD fingers, though the number and location of these domains are variable (Figure 1.1). MLL PHD fingers are important enablers of enzyme function, though the specific interactions responsible are often unique among proteins [3]. For example, MLL1-dependant gene expression is regulated in part through its PHD₃ domain. Binding of H3K4me₃ by MLL1 PHD₃ activates target gene transcription [4], whereas PHD₃-dependant association with prolyl isomerase Cyp₃ has a negative regulatory role [4]. While this positive regulatory H3K4me₃ interaction is conserved in MLL2, the Cyp₃ interaction is not, despite a high degree of sequence similarity shared between the MLL1 and MLL2 PHD₃ domains [6]. The loss of this association is credited to a single substitution of methionine for aspartate in a key region. This demonstrates how even minor differences in regulatory domain structure can modify enzyme function and highlights the likely significance of drastic changes in KMT2 domain architecture through evolutionary time.

MLL4 contains 7 PHD fingers organized into two tandem PHD cassettes (PHD₁₋₃, PHD₄₋₆), and a standalone domain PHD₇. Though the exact function of many of these domains is not well studied, a growing body of evidence implicates MLL4 PHD fingers in key regulatory interactions. MLL4 PHD₂ exhibits E3 ubiquitin ligase activity in the presence of the E2 conjugating enzyme CDC34, a function which is shared with MLL1 PHD₂ but not MLL2 or MLL3 PHD₂ [7]. This activity appears to negative regulate MLL1, and it is reasonable to speculate this is

also true for MLL4. PHD₄₋₆ cassette of MLL4 acts in tandem to associate with either unmodified histone H4 or histone H4 asymmetrically methylated at R3 [8]. This association is disrupted by symmetric methylation, which subsequently inhibits methyltransferase activity. Mutations thought to impede the PHD fold in MLL4 PHD₅ are associated with Kabuki syndrome, a developmental disorder strongly linked with MLL4 dysregulation. Disruption of PHD₅ leads to an overall reduction of MLL4 activity, and likewise MLL4-dependant differentiation. Recent studies indicate MLL4 PHD₆ acts as a reader of acetylated H4K16 and present a link between MLL4 chromatin occupancy and the activity of H4K16-specific acetyltransferase MOF [9]. An extended PHD₆ domain, defined as PHD₆ plus an adjacent N-terminal zinc finger, specifically recognizes histone H4H18 [10]. This interaction is required for MLL4 methylation activity, and specifically impedes neuronal differentiation of NT2/D1 cells.

Interestingly, some PHD fingers have been reported to simultaneously bind with histone and non-histone proteins [11]. This is the case for the above described MLL1 PHD₃-Cyp3 interaction, which forms a ternary regulatory complex. It is interesting to speculate that MLL4-dependant CFP1 methylation could promote CFP1me association with the PHD domains and induce a novel regulatory state. As such, we designed a peptide array containing peptides representing each of the methylated CFP1 lysines found in Chapter 3 and incubated it with PHD₁₋₃, PHD₄₋₆, and PHD₇ constructs to assay binding activity.

4.2 Materials & Methods

4.2.1 Setd1A (1362-1563) cloning

The Setd1A region coding for amino acids 1362-1563 was amplified by PCR using custom primers (F: 5'-TAAGCAAAGCTTGCGAGGAGGGCGAAGAGGAG-3', R: 5'-TAAGCACTCGAGGCCCATGATGGCGGAGGTACC-3'). The PCR product was subjected to gel electrophoresis on at 1% agarose gel and purified by gel extraction of the appropriate band using a BioBasic PCR Products Purification Kit according to the manufacturer's instructions. The purified PCR product and the vector pET28a were each subjected to double digest with HindIII and XhoI restriction enzymes in 2.1 buffer (New England Biolabs). Digested pET28a and PCR product were ligated at a molar ratio of 1:10 with DNA T4 ligase (New England Biolabs) overnight at 16 °C. The ligation was transformed into DH5 α and BL21 cells.

We subjected several DH5 α transformants to colony PCR using the Setd1A primers (Table S1). Colonies containing what appeared to be our insert were sequenced for confirmation.

4.2.2 Protein expression and purification

p3E-MLL4 (5382–5537) SET/Post-SET domain construct was expressed as a glutathione S-transferase (GST) fusion protein. *Escherichia coli* BL21 cells containing either construct were grown at 37 °C to an OD₆₀₀ of ~0.6 in lysogeny broth (LB) containing 1 mM MgCl₂ and 1 μ M ZnSO₄. Expression was induced using 0.1 mM Isopropyl β -D-1-thiogalactopyranoside (IPTG), and proceeded for 18 hours at 16 °C. After centrifugation at 2,500xg for 15 minutes, the bacterial pellet was lysed in 50 mM Tris pH 7.5 buffer containing 300 mM NaCl, 3 mM DTT, 1 μ M ZnSO₄, 0.01% Triton X-100, 1 mg/mL chicken egg lysozyme (Sigma), 10% glycerol; and a protease inhibitor cocktail consisting of 100 μ M PMSF, 1 μ M E-64, 1 μ M pepstatin, and 5 μ M bestatin.

The lysate was clarified through centrifugation at 18,000xg for 45 minutes. The fusion protein was isolated from clarified lysate by affinity chromatography with glutathione sepharose 4B beads (GE Healthcare) and eluted using 50 mM Tris pH 8 containing 10 mM glutathione.

Full length pOPINJ-WDR5, p2E-RbBP5, pOPINS-Ash2L, pGst-Dpy-30, pET28a-CFP1 (302-656), and pET28a-Setd1(1362-1563) were each expressed as 6xHistadine fusion proteins. Each construct was grown in *E. coli* BL21 cells at 37 °C to an OD₆₀₀ of ~0.6 in lysogeny broth (LB) containing 1 mM MgCl₂. Expression was induced using 0.1 mM Isopropyl β-D-1-thiogalactopyranoside (IPTG), and proceeded for 18 hours at 16 °C. After centrifugation at 2,500xg for 15 minutes, the bacterial pellet was lysed in P5 (50 mM NaHPO₄ pH 7.0, 500 mM NaCl, 5 mM imidazole, 10% glycerol, 0.05% Triton X-100) containing; protease inhibitor cocktail (100 μM PMSF, 1 μM E-64, 1 μM pepstatin, and 5 μM bestatin). The lysate was clarified through centrifugation at 18,000xg for 45 minutes. The fusion proteins were then each isolated from clarified lysate by affinity chromatography with Ni-NTA beads (Qiagen) and eluted using P500 (P5 with 500 mM imidazole).

GST-tagged MLL4 PHD domains of were expressed in *Escherichia coli* BL21 (DE3) (Stratagene), which were cultured in LB medium (50 μM IPTG) for 14–16 h at 16°C. Bacterial cells were pelleted, resuspended in the lysis buffer (20 mM Tris-HCl at pH 8.0, 137 mM NaCl, 1.5 mM MgCl₂, 1 mM EDTA, 0.2 mM phenylmethylsulfonyl fluoride [PMSF], 10% [v/v] glycerol) and lysed with lysozyme of a concentration of 0.5 mg/mL for 1 h at 4°C. Cells were further lysed by sonication, and the supernatants were collected after centrifugation. The whole-cell lysates were incubated with glutathione beads for 14–16 h at 4°C (Amersham), followed by extensive washing with lysis buffer supplemented with 0.1% Triton X-100. The GST fusion proteins were eluted with the glutathione-containing elution buffer (100 mM Tris-HCl at pH 8.0).

4.2.3 Thrombin cleavage and methylation of CFP1 (302-656)

Because both the CFP1 (302-656) and Setd1A (1362-1563) contained 6xHis fusion tags, the CFP1 (302-656) tag needed to be cleaved before a pulldown experiment was possible. This was accomplished using the Thrombin CleanCleave kit (Sigma-Aldrich Cat# RECOMT). The thrombin bead slurry was thoroughly washed in cleavage buffer (50 mM Tris-HCl pH 8.0 with 10 mM CaCl₂ and the supernatant removed. We found that the CaCl₂ component of the cleavage buffer consistently precipitated the CFP1 construct, so the beads were next equilibrated in 50 mM Tris pH 8.0. CFP1(302-656) in 50 mM Tris pH 8.0 was added to the beads, which were rotated at 4 °C overnight for 16 hours. The beads were centrifuged at 500 xg and the supernatant collected.

To ensure cleavage we ran both uncleaved and cleaved CFP1 (302-656) on a 10% agarose SDS-PAGE gel. The gel was transferred to a 0.2 µm PVDF membrane at 90V for 120 minutes. The membrane was blocked with 0.5% non-fat skim milk for 1 hour at room temperature on a rocker. The membrane was briefly washed in Tris buffered saline with 0.05% Triton X-100 (TBS-T) (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.05% Triton X-100) to remove excess milk, and subsequently probed with Anti-6xHis-HRP in TBS-T at a concentration of 1:5000 for 1 hour at room temperature on a rocker. The membrane was washed 5x 5 minutes with TBS-T and imaged using Clarity Western ECL Blotting Substrate (BioRad). The membrane was imaged using a BioRad ChemiDoc XRS+. To estimate recovery, we subsequently stained the western blot in Coomassie brilliant blue and imaged it using a BioRad ChemiDoc XRS+.

4.2.4 Methylation of CFP1(302-656) by MLL4

Samples of cleaved CFP1(302-656) and the MLL4-WRAD complex were dialysed in methylation buffer (20 mM Tris pH 8.0, 50 mM NaCl, 1 mM EDTA, 3 mM MgCl₂, and 1 mM

DTT. Enzyme and substrate were combined for a final concentration of 5 μ M cleaved CFP1(302-656) and 0.2 μ M MLL4-WRAD complex with 10 μ M S-adenosyl L-methionine (SAM). The methylation reaction was shielded from light and rotated overnight at room temperature. Glutathione Sepharose 4B beads (GE Healthcare) were equilibrated in methylation buffer and rotated with the methylation reaction for 2 hours at 4 °C. The beads were centrifuged at 800 xg and the supernatant collected as a “flow-through” sample.

We identified the possibility that CFP1 fragments could associate with the MLL4 complex or non-specifically with the glutathione beads. As such, we washed the beads in methylation buffer with successively increasing NaCl concentrations, i.e. 50 mM (base concentration), 75 mM, 150 mM, 200 mM, 300 mM, 375 mM, and 450 mM. A sample of the flow-through and each wash was subjected to SDS-PAGE on a 10% agarose gel, followed by Coomassie brilliant blue staining and imaging with the BioRad ChemiDoc XRS+. To increase the concentration of recovered CFP1me (302-656) we subjected the containing fraction to a speedvac apparatus until it was approximately 2X concentrated.

4.2.5 Bioluminescence-based end-point methyltransferase activity assays

Before conducting the pulldown experiment, we wanted to ensure that the CFP1 used was methylated. To do this, we used the Promega MTase-Glo™ Methyltransferase assay kit (Cat# V7601) as described in Chapter 3.2.2. MLL4(5382–5537), WDR5, Ash2L, RbBP5, Dpy30, and, CFP1(302-656), and H3K4 fractions were all dialysed in reaction buffer (20 mM Tris pH 8.0, 50 mM NaCl, 1 mM EDTA, 3 mM MgCl₂, 1 mM DTT) for 2 hours at 4 °C. Substrate solutions were prepared in 4X concentrated reaction buffer with 0.4 mg/mL bovine serum albumin (BSA). After addition of enzyme solution, the final concentrations of substrate

buffer components were 15 μ M of substrate (either CFP1(302-656) or H3K4), 10 μ M S-adenyl-L-methionine (SAM), and 0.1 mg/mL BSA in 1X concentrated reaction buffer. Equimolar mixtures of the MLL4 COMPASS-like complex members were combined such that the final MLL4 complex concentration was 0.5 μ M after addition to substrate buffer.

Each individual reaction contained 4 μ L each of substrate buffer and enzyme solution combined in a LumiTrax 200 opaque white 384-well plate (Greiner). Reactions were plated in triplicate. The plate was gently and briefly agitated on a benchtop plate shaker and centrifuged for 1 minute at 500 xg. The reaction proceeded for 16 hours at room temperature in a dark location. Following incubation, 2 μ L of 5X MTase-Glo™ reagent (400 mM Tris buffer pH 8.0, 40 mM MgCl_2 , 75 mM NaCl, 125 μ M dGTP, 200 μ M sodium metaphosphate, 80 μ g/ml polyphosphate - AMP phosphotransferase, 200 μ g/ml SAH-Hydrolase, 50 μ g/ml adenosine kinase, 5 mM DTT) was added to each well. The plate was then gently agitated and centrifuged for 1 minute at 500xg, after which the reaction was incubated at room temperature for 60 minutes in a dark location. Next, 10 μ L of detection solution (1 mM potassium phosphoenolpyruvate, 8 U/ml pyruvate kinase) was added to each well, and the plate was once again reacted for 60 minutes at room temperature in the dark following agitation and centrifugation as described above. At the end of this incubation period, End-point relative luminescence readings were taken using a BioTek Cytation 5 plate reader.

4.2.6 CFP1(302-656) Pulldown via Setd1(1362-1563) affinity

CFP1(302-656), CFP1me(302-656) and Setd1A (1362-1563) were each dialysed into phosphate buffered saline (PBS) (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4 , 1.8 mM KH_2PO_4). 0.25 μ M of either CFP1 or CFP1me was reacted with 0.75 μ M Setd1A(1362-1563) at

rotating overnight at 4 °C. 25 μ L of Ni-NTA beads (Qiagen) were washed and equilibrated in PBS and added to the reaction solution. The solution was rotated for 2 hours at 4 °C, centrifuged at 800 xg and the supernatant was collected as the “flow-through” sample.

As described in 4.2.3, the Ni-NTA beads and bound Setd1A(1362-1563) were subjected to 50 μ L salt washes in an attempt to dissociate potentially bound CFP1/CFP1me(302-656) from Setd1A(1362-1563). Base buffer NaCl concentration (137 mM) was deemed “ Δ 0” and increased with each successive wash in the following manner: Δ 50 mM, Δ 100 mM, Δ 200 mM, Δ 300 mM, Δ 500 mM. 20 μ L of each sample and sample containing Ni-NTA beads in PBS were melted in an appropriate volume of 5X Laemeli buffer and subjected to SDS-PAGE on a 10% agarose gel. The gel was stained in Coomassie brilliant blue and imaged using a BioRad ChemiDoc XRS+.

4.2.7 CFP1me array design, SPOT synthesis, and array processing

To investigate the ability of MLL4 PHD₁₋₃, PHD₄₋₆, and PHD₇ to interact with CFP1 in a methylation-dependent fashion, we designed and printed peptide arrays containing each of the methylated lysine residues we identified within CFP1 (K331, K335, K336, K339, K340, K341, K346, K351, K353, K355, K357, and K359), +/- 10 flanking residues. Each lysine was printed as a nil, mono-, di-, and tri-methylated peptides. All peptides were separated from the array surface through the addition of a N-terminal flexible 6-aminohexanoic acid group. H4K20, H4K20me, H4K20me₂, and H4K20me₃ served as control SPOTs, as the former two of these peptides have been shown to interact with MLL4 PHD₄₋₆ [8].

As described in Chapter 2 Section 2.2.2, solid-phase peptides were synthesized directly on a sheet of aminated cellulose filter paper with an Intavis Multiprep instrument using Fmoc chemistry at an approximate concentration of 0.2 μ mol/ SPOT. Following synthesis, the arrays

were deprotected by three sequential washes of 20% piperidine dissolved in dimethylformamide (DMF), followed by three washes in DMF, and three final washes with anhydrous ethanol. Arrays were then dried by air. After drying, sidechain protecting groups were cleaved from the peptides with a trifluoroacetic acid solution (95% TFA, 3% TIPS and 2% H₂O). The arrays were again washed three times with dichloromethane (DCM), followed by three washes with DMF, and three washes of anhydrous ethanol, and allowed to air dry overnight in a fume hood. Immediately before use, arrays were briefly washed with anhydrous ethanol, then equilibrated in a solution of 50% ethanol in water, and then finally washed in only water.

4.2.8 Probing CFP1me arrays with MLL4 PHD domains

The array was equilibrated with several short washes of TBS-T. 1 μ M of either MLL4 PHD₁₋₃(156–341), PHD₄₋₆(1358–1572), or PHD₇(5079–5152) was incubated with the peptide array overnight at 4 °C in TBS-T. On the next day, the protein solution was poured off and the array quickly rinsed again with TBS-T. The array surface was blocked using 0.5% non-fat skim milk in TBS-T and allowed to incubate for 1 hour on a rocking platform at room temperature. Following a brief rinse in TBS-T to remove any traces of milk solution, the array was washed 3x 5minutes with TBS-T. Next, the arrays were probed with Anti-6xHis-HRP antibody at a 1:5000 concentration for 2 hours on a rocking platform at room temperature. The array was reacted with Clarity western blot ECL substrate (BioRad) and imaged using BioRad ChemiDoc XRS+. Densitometric analysis of the array spots was performed using ImageJ with ‘Protein Array Analyzer’ macro (<https://imagej.nih.gov/ij/macros/toolsets/>). SPOTs producing signal greater than 50% of the wild-type H4K20 peptide substrate signal were identified as potentially of interest.

4.3 Results & Discussion

4.3.1 CFP1me appears to increase/facilitate Setd1 binding

Interestingly, initial analysis of the pulldown assay suggests that MLL4-mediated methylation of CFP1(302-656) does influence binding with Setd1A(1362-1563) (Figure 4.1C). Methylated CFP1me(302-656) was not eluted in the salt washes, and only appeared in the lane containing beads boiled in 2X Laemmli buffer. This suggests that CFP1me(302-656) was bound to the Ni-NTA bound Setd1A(1362-1563). Conversely, unmethylated CFP1(302-656) appeared in the flow-through – the supernatant collected after centrifugation of the solution containing CFP1(302-656), Setd1A(1362-1563) and the Ni-NTA beads. This suggests that the unmethylated CFP1 fragment was unable to associate with Ni-NTA bound Setd1A(1362-1563). Taken together, these results support the hypothesis of an MLL4-driven methyl-switch controlling Setd1A/B-CFP1 interaction. This hypothesis could be further supported by conducting a native PAGE experiment, wherein unmethylated and methylated CFP1 are interacted with Setd1A and observed for differences in electromobility, indicating the presence or absence of protein-protein interaction. This experiment could also provide insight regarding a role for methylation in the formation of CFP1 homodimers.

This mechanism suggests a novel role for MLL4 – primarily thought to be a mono-methyltransferase – in the indirect regulation of H3K4 di- and tri-methylation through controlling Setd1A/B complex formation. It should be noted that MLL4, Setd1A, and CFP1 localize at different regions within the nucleus. MLL4 is highly enriched among distal enhancer regions in the genome, whereas Setd1A is found at sites of active transcription [12, 13]. Unbound CFP1 is localized within nuclear speckles, dynamic nucleoplasmic structures that contain various cellular machinery associated with gene expression [14]. MLL4 could interact with and methylated CFP1

from nuclear speckles, which can then move to form complexes with Setd1A. It is also possible that the distal enhancer regions containing MLL4 are proximate to Setd1A-associated active transcription sites *in vivo*, and this permits interaction between the three proteins.

4.3.2 CFP1me maybe bound to MLL4 through PHD

Because emerging evidence suggests PHD domains may participate in interactions beyond their canonical role in histone recognition [11], we were interested in the idea that interaction with MLL4 could promote the association of CFP1 and the MLL4 PHD domains. Indeed, histone lysine methyltransferases were long thought to be specific only to histones, but it is now clear that histone lysine methyltransferases are involved in a plurality of enzyme-substrate reactions [15]. To investigate a possible mechanism involving PHD₁₋₃, PHD₄₋₆, or PHD₇ of MLL4 with CFP1 of various methyl states, we probed a printed peptide array containing each of the CFP1 basic domain methyllysine windows discovered in Chapter 3 with constructs containing each of the MLL4 PHD cassettes and the lone PHD finger.

We found that the methylation state of particular lysine residues in the CFP1 basic domain affected *in vitro* binding of PHD₁₋₃ and to a lesser degree PHD₄₋₆ constructs (Figure 4.3B). For PHD₁₋₃, CFP1 K331me1 and K339me1 preferentially interacted compared to me0, me2, or me3. All methyllysine states of K335 were bound with an affinity greater than 50% of the wild-type H4K20me0 event, but me3 appeared to potentiate this effect. K353 was bound in the me0 and me1 state, but the me2 and me3 moieties abrogated interaction. Interestingly, one of the most conclusive results in this experiment was seen in the three most SID-proximal lysines K355, K357, and to a lesser degree K359. In these residues, the significant binding seen in me0 and me1 SPOTs was completely abolished by me2 and me3 modifications. Though K359 wasn't bound above 50% of

the WT signal in any residues, this pattern was conserved. These residues were identified as of potential interest in Chapter 3, as they were observed to be only conserved in mammals. The abrogation of binding with increasing methyl state at these residues could be indicative of hydrogen bonding between the CFP1 peptides and the PHD₄₋₆ cassette. Methyl group substitution of Nε hydrogen atoms increases the size and hydrophobicity of the nitrogen moiety, while limiting hydrogen bonding such that me₃ modification is characterized by a total loss of hydrogen bond participation [16].

The relationship between lysine position and methyl state similarly impacted PHD₄₋₆ binding (Figure 4.3C). K335 exhibited binding at all states, but me₂ and me₃ had the greatest effect. In contrast with PHD₁₋₃, trimethylation seemed to be key important for interaction with residues K335, and K340. Dimethylation of K351 increased interaction with PHD₄₋₆. Finally, the most SID-proximal lysines K353, K355, K357 and K359 once again best facilitated binding when in a me₀ or me₁ state. It is intriguing that this result mirrors that seen with PHD₁₋₃.

PHD₇ interacted with the array to some degree, but few interactions appeared meaningful (Figure 4.3D). Of those that did, K335 exhibited greater binding while unmethylated, and K345 binding activity was significantly diminished by trimethylation.

The biological relevance of this finding is unclear but could merit additional investigation should the field of non-histone PHD finger binding develop further. A shortfall to this approach is that PHD fingers, especially PHD finger cassettes, are canonically combinatorial readers that have been shown to integrate a several modifications into one reader state [1]. Assessing the individual contribution of each methyllysine contributes to the understanding of known interactions but isn't necessarily meaningful in the pursuit of novel substrate investigation. It would be interesting to

interact MLL4 PHD domain constructs with CFP1 of various methyl states in solution. If there is indeed an association, these results could provide insight regarding its biochemical nature.

4.4 Conclusions

We observed additional evidence depicting MLL4-mediation of CFP1 and Setd1A association through a pulldown experiment that suggested methylation promotes CFP1 binding. Overall these suggest the possible existence of a novel methyl switch effecting control of chromatin modification, but further studies should be performed to confirm this mechanism. Furthermore, we laid the foundation for future investigation into novel PHD finger-non-histone substrate interaction between the PHD₁₋₃ and PHD₄₋₆ MLL4 PHD cassettes and CFP1 basic region lysine residues at distinct methyl states. The synthesis of these results with work done in the previous chapters is presented in Chapter 5.

Figures

Figure 4.1. **A.** Anti-6xHis western blot of CFP1 fragment pre- and post-thrombin cleavage (left). Coomassie stain of western membrane (right). **B.** CFP1 fragment was recovered from the MLL4 WRAD-dependant methylation reaction in the flow-through. **C.** SDS-PAGE of methylated (above) and unmethylated (below) CFP1 (302-656) pulldowns using Setd1A for affinity capture. Methylated CFP1 fragments appear to bind Setd1A(fragment) with a greater affinity than unmethylated fragments. CFP1 fragments were incubated with a molar excess of Setd1A (1362-1563) PBS rotating overnight at 4 °C, and then incubated with Ni-NTA beads rotating for 2 hours at 4 °C. The beads were centrifuged, and the flow-through collected was labeled “FT”, representing the base salt concentration. The beads were subsequently washed in buffer solution with increasing concentrations of NaCl. The increasing molarity of NaCl is represented as ΔX , where X represents the increase in NaCl concentration compared to the reaction buffer. Finally, a portion of the beads was collected and melted in 2X Laemmli buffer (LB). 25uL of each sample was run on an SDS-PAGE gel and stained with Coomassie brilliant blue.

A.

D.

E.

Figure 4.3 MLL4 PHD domain binding on CFP1 methyllysine peptide array. **A.** Black and white image shows raw data from one binding array. Binding activity for each of the arrays was quantified by densitometry analysis. Densitometry measurements were normalized to H4K20 control SPOT. **B.** Relative binding on array probed with MLL4 PHD1-3. **C.** Relative binding on array probed with probed with MLL4 PHD4-6. **D.** Relative binding on array probed with MLL4 PHD 7. Dashed line indicates 50% relative binding compared to H4K20me0.

References

- [1] Sanchez R. & Zhou M.-M. (2012). The PHD Finger: A Versatile Epigenome Reader. *Trends in Biochemical Science* 36(7): 364-372.
- [2] Botuyan M.V., Mer G. (2016). Tudor Domains as Methyl-Lysine and Methyl-Arginine Readers. In Binda O., & Fernandez-Zapico M.E. (Eds.) *Chromatin Signaling and Diseases* (pp. 149-165). Academic Press.
- [3] Ali M., Hom R.A., Blakeslee W., Ikenouye L., & Kutateladze T.G. (2013). Diverse functions of PHD fingers of the MLL/KMT2 subfamily. *Biochimica et Biophysica Acta* 1843(2): 366-371.
- [4] Chang P.-Y., Hom R.A., Musselman C.A., Zhu L., Kuo A., Gozani O., Kutateladze T.G., & Cleary M.L. (2010). Binding of the MLL PHD3 finger to histone H3K4me3 is required for MLL-dependent gene transcription. *Journal of Molecular Biology* 400(2): 137-144.
- [5] Wang Z., Song J., Milne T.A., et al. (2010). Pro Isomerization in MLL1 PHD3-Bromo Cassette Connects H3K4me Readout to Cyp33 and HDAC-Mediated Repression. *Cell* 141(7): 1183-1194.
- [6] Hom R.A., Chang P.-Y., Roy S., et al. (2010). Molecular mechanism of MLL PHD3 and RNA recognition by the Cyp33 RRM domain. *Journal of Molecular Biology* 400(2): 145-154.
- [7] Wang, J., Muntean, A. G., Wu, L., & Hess, J. L. (2012). A Subset of Mixed Lineage Leukemia Proteins Has Plant Homeodomain (PHD)-mediated E3 Ligase Activity. *Journal of Biological Chemistry* 287(52): 43410–43416.
- [8] Dhar S.S., Lee S.-H., Kan P.-Y., et al. (2012). Trans-tail regulation of MLL4-catalyzed H3K4 methylation by H4R3 symmetric dimethylation is mediated by a tandem PHD of MLL4. *Genes and Development* 26: 2749-2762.

- [9] Zhang Y., Jang Y., Lee J.-E., et al. (2019). Selective binding of the PHD6 finger of MLL4 to histone H4K16ac links MLL4 and MOF. *Nature Communications* 10: 2314.
- [10] Liu Y., Qin S., Chen T.-Y., et al. (2019). Structural insights into trans-histone regulation of H3K4 methylation by unique histone H4 binding of MLL3/4. *Nature Communications* 10: 36.
- [11] Li Y., & Li H. (2012). Many keys to push: diversifying the ‘readership’ of plant homeodomain fingers. *Acta Biochimica et Biophysica Sinica* 44(1): 28-39.
- [12] Lee J.-E., Wang C., Xu S., et al. (2013). H3K4 mono- and di-methyltransferase MLL4 is required for enhancer activation during cell differentiation. *eLife* 2: e01503.
- [13] Lee J.H., & Skalnik D.G. (2008). Wdr82 is a C-terminal domain-binding protein that recruits the Setd1A Histone H3-Lys4 methyltransferase complex to transcription start sites of transcribed human genes. *Molecular Cell Biology* 28(2): 609-618.
- [14] Lee, J.-H., and Skalnik, D. G. (2002). CpG-binding Protein Is a Nuclear Matrix- and Euchromatin-associated Protein Localized to Nuclear Speckles Containing Human Trithorax. *Journal of Biological Chemistry* 27: 42259-42267.
- [15] Biggar K.K., Li S.S. (2015). Non-histone protein methylation as a regulator of cellular signaling and function. *Nature Reviews Molecular Cell Biology* 16(1):5-17.
- [16] Luo M. (2018). Chemical and Biochemical Perspectives of Protein Lysine Methylation. *Chemical Reviews* 118(14): 6656-6705.

Chapter 5

General Discussion

5.1 Non-histone lysine methylation is an important field of study that continues to grow

It is unlikely that Ambler could have ever predicted the functional importance of the small chromatographic peak between lysine and histidine first observed in 1959 [1]. Growing interest in epigenetic regulation and the availability of improved instrumentation has brought lysine methylation to forefront of dynamic chromatin remodeling research [reviewed in 2]. The complexity of histone lysine methylation is a function of the possibility of four distinct dynamically regulated methylation states per site (i.e., me0, me1, me2, me3) and the ability of reader proteins to specifically recognize each distinct state. It is thus unsurprising that this information-rich modification is a critical actor in protein regulation beyond chromatin remodeling, and that the rapidly growing interest in non-histone lysine methylation has resulted in the generation of complex and functionally diverse protein-protein interaction (PPI) networks [2].

For this thesis we selected the KMT2 class H3K4 methyltransferase (KMT) mixed lineage leukemia 4 (MLL4, KMT2D) as a candidate non-histone protein lysine methyltransferase. MLL4 is a descendant of yeast Set1 and like its paralog MLL3, is homologous to the *Drosophila* homeotic regulator Trithorax-related (trr). MLL4 KMT activity is conferred by the catalytic SET (Suppressor of variegation, Enhancer of zeste, Trithorax) domain [reviewed in 3]. The SET domain is highly conserved, yet phylogenetic analysis of KMT2 SET domains reveals subtle distinctions that may indicate the evolution of unique substrate recognition [4] (Figure 1.1). Additionally, MLL4 exhibits complex domain architecture compared to its descendants (Figure 1.1). This includes 7 PHD finger domains, arranged into two cassettes (PHD₁₋₃, PHD₄₋₆) and a single domain PHD₇. Though the exact function of each of these domains is poorly understood, it is thought that they in part contribute to combinatorial

histone mark reading required for MLL4 activity [5, 6, 7]. In spite of an evolutionary history that indicates expansion of functional diversity, H3K4 is the only known MLL4 substrate.

As in other KMT2 enzymes, MLL4 associates with a multimeric protein complex sometimes referred to as COMPASS-like (COMplex Associated with Set1-like) [3]. This complex consists of the core units WDR5, RbBP5, Ash2L, and Dpy30 present in all KMT2 enzymes, but also members unique trr-descendant KMTs including UTX, PAXIP1, PAGR1, and NCOA6 [3]. Ultimately, the COMPASS-like complex potentiates KMT activity and confers unique properties to each KMT2 sub-family. MLL4 and the associated COMPASS-like complex localizes to distal enhancer regions in the genome where it exhibits mono-methylation activity on H3K4 at inter- and intra-geneic regions [reviewed in 8].

MLL4 has been implicated in cancer and congenital heart disease [9, 10]. MLL4 is unique among KMT2 proteins in that it is a causative agent of the developmental disorder Kabuki syndrome [11].

Overall, the structurally diverse domain architecture and unique associated protein complex of MLL4 combined with its role in several disease states make it an interesting target for substrate exploration. The elucidation of novel enzyme substrates for MLL4 could further our understanding of dynamic non-histone lysine methylation and the regulatory interactions that depend upon this PTM. Identification of new regulatory interactions expands the landscape of rational drug targets for the treatment of myriad diseases.

5.2 There are several means of substrate discovery, each with distinct pros and cons

In our efforts to identify novel MLL4 substrates we employed two distinct medium-throughput approaches, described in detail in Chapter 2. Specifically, we used printed peptide

arrays to facilitate both shotgun-driven and systematic analytical approaches. Results from both arrays were ultimately generated by incubating each array with active MLL4-WRAD complex and a radioactive methyl donor i.e. adenosyl L-methionine S-[methyl-³H], and imaging radiation from methylated peptides using x-ray film. Further information on this procedure, the processing of those data and the generation of the arrays is detailed in 2.2.

The shotgun or target-driven exploration simplest, and perhaps crudest approach. In this method known methyllysine substrates of interest (i.e., without a known associated KMT) are selected by a defined criterion (i.e., biological relevance, similarity to a known KMT-substrate, etc.) and are then screened against a specific KMT for methylation activity. The benefit of this approach is that substrates which participate in a known KMT-regulated pathway – such as those implicated in a specific disease – can be pre-selected for targeted screening. This ultimately facilitates the rapid generation of a deductive hypothesis. However, targeted substrate identification can inherently suffer from researcher bias, while shotgun exploration of KMT substrates and may fail to return any positive results. Furthermore, these approaches to building the KMT-substrate network fails to uncover any fundamental information on how KMTs select for specific substrates and methylation sites (i.e., modifiable lysine residues).

In contrast to a targeted or shotgun approach to substrate identification, the guiding principles of enzyme substrate selection can be realized through a systematic analysis of substrate preference. This process involves the elucidation of a candidate recognition motif for a KMT of interest. For the purpose of this thesis we defined a recognition motif as the specific sequence of amino acids surrounding a methylatable lysine residue that facilitates methyltransferase activity of a specific KMT. Once the KMT recognition motif is resolved, the resulting recognition motif can then be queried against a database of known lysine methylation sites together with their flanking

amino acids (i.e., the “methylation window”). While the relevance of the hits with regard to disease implications is inherently not guaranteed, the systematic nature of this approach eschews researcher bias and may return a number of interesting results. Additionally, solving the recognition motif provides valuable insight regarding the specific biochemical interactions that facilitate enzyme-substrate binding, which can be exploited for the development of inhibitor drugs. A limitation of this approach is that it requires at least one substrate to be known in order create the single residue permutations that are necessary for the experiment. However, it should also be mentioned that this type of systematic approach can also suffer from inherent bias in experimental design. These experiments are typically designed around a sequence of a single known KMT substrate, as a result, all single residue permutations of that substrate will always retain significant similarity to the original substrate. Because MLL4 has only one known substrate, the results gleaned from this screen were indeed subject to this mode of bias.

From these divergent approaches we narrowed our search to two substrate candidates: the DNA double-strand break repair pathway modulator p53 binding protein 1 (53BP1) [12] (Figure 2.1), and CXXC finger protein 1 (Figure 2.3), a key member of the H3K4 KMTs Setd1A and Setd1B COMPASS complexes [13], as well as binding partner of the DNA CpG methyltransferase DNMT1 [14]. Specifically, data from the shotgun array indicated MLL4-dependent methylation of lysine 1563 of 53BP1. This potential substrate was selected from the other hits both from a signal-dependant and functional standpoint. 53BP1 K1563 is of interest because it occurs within the tandem Tudor domains (Figure 2.2) of 53BP1, domains crucial for proper 53BP1 function because they allow the protein to bind H4K20me2 and initiate a signal cascade in regions of DNA damage [12]. Interestingly, if validated this methylation event would implicate MLL4 in the DNA damage response, and more specifically double-strand break repair

pathway decision making. Thus, we hypothesized the MLL4 may play a regulatory role in these mechanisms through 53BP1 methylation.

Using the recognition motif generated from the systematic array (Figure 2.3), we queried a large database containing all lysine residues within known nuclear proteins and sequentially filtered these results to obtain a list of proteins containing a perfect match of the recognition motif that have known methylation sites but no associated KMT (Figure 2.5, Table 2.1, Table 2.2). From this curated list we identified CFP1 K328 as a potential MLL4 substrate. This protein was chosen in part because it's interesting to speculate it could represent a regulatory link between enhancer localized MLL4 and Setd1A/B COMPASS localized to active transcription sites. Additionally, the proposed methylation site K328 occurs within, or at least adjacent to the domain responsible for CFP1 interaction with Setd1 proteins – the Setd1 interaction domain (SID) (Figure 2.6). Experiments involving selective mutation of key residues within the SID abolished interaction between CFP1 and Setd1 proteins [14]. It is noteworthy that there doesn't appear to be a rationale for the boundaries by which the SID was initially defined, i.e. amino acids 361-422. As such, a PTM that is highly proximal to this important functional domain i.e. K328 may actually be located within it. Finally, CFP1 has also been shown to associate with the DNA CpG methyltransferase DNMT1, through a mechanism that appears distinct from that of Setd1A/B association [14]. We hypothesized that a PTM methyl-switch located either within or proximal to the SID could be the means by which differential CFP1 association is regulated.

5.3 Validation of predicted MLL4 substrates 53BP1 and CFP1

As discussed above, there are scenarios for both approaches in which the results gleaned aren't particularly meaningful, though it's difficult to predict at which point in the experimental

pipeline that fact will become apparent. In Chapter 3, we aimed to validate the methylation of 53BP1 K1563 and CFP K328 *in vitro*, first through determining enzyme activity using a bioluminescence-based end-point methyltransferase activity assays, specifically the Promega MTase-Glo™ kit (Cat# V7601). In essence, the kit works by enzymatically converting the by-products of lysine methylation reactions to substrates for a luciferase indicator. A KMT (in this case MLL4 and the core complex WRAD) and substrate are reacted with S-adenyl L-methionine (SAM) for an appropriate amount of time, after which the enzymes required for generation of luminescence are added to the solution (detailed in 3.2.2). The greater the degree and/or frequency of methylation, the greater the amount of light produced, which can be quantified using a plate reader. The results of this assay indicated a surprising amount of methylation activity on our CFP1 fragment (302-656), while our 53BP1 construct (1484-1603) did not appear to undergo methylation at all (Figure 3.1, interpreted in 3.3.1).

5.4. Identifying a potential functional consequence of MLL4-dependant CFP1 methylation

Moving forward, we designed a peptide substrate array containing every lysine present within our CFP1 and 53BP1 constructs, plus the 10 flanking amino acids N- and C-terminal to the central K (i.e., a 21 amino acid methylation window). The aim of this experiment was to identify the specific residue or residues that undergo *in vitro* MLL4-dependant methylation. Interestingly and in agreement with our results in Figure 3.1, we identified several methylated lysines (specifically, K331, K335, K336, K339, K340, K341, K346, K351, K353, K355, K357, and K359) that occur with high frequency in adjacent positions within what has been described as the CFP1 basic domain, or basic region (Figure 3.2, Figure 3.3). Interestingly, K331 first appeared in vertebrates, and K355, K357 and K359 are only found in mammals (Figure 3.5). As

MLL4 is a vertebrate KMT, it is interesting to speculate that these mutations arose as a consequence of MLL4 activity.

Surprisingly, the substrate K328 predicted by our recognition motif was not among the methylated lysines. In our attempt to qualitatively score the methylation window of each methylated lysine against our MLL4 recognition motif, and subsequently determine which of the windows were more alike or dissimilar, we observed a surprising finding. K328 – i.e. the genesis of our rationale for selecting CFP1 a substrate candidate – did not appear to be methylated on the array (Figure 3.2). In a somewhat amusing outcome, our recognition motif failed to properly align with any of the MLL4-reactive methylation windows (Figure 3.4), but instead in every case aligned exclusively with K328. In fact, and the only valid alignments as returned by the software were those in which K328 was present as a flanking lysine. As such, it seems that we identified a plurality of potentially functionally significant methylation events by the pure chance that our recognition motif aligned perfectly with an adjacent but ultimately insignificant residue. In some ways this is arguably a more interesting result than if one or both approaches had worked, as it demonstrates the reality of exploratory scientific research. A hypothesis generated through rational analysis won't always produce the expected result, but the process of following that thread can in some rare instances result in a serendipitous discovery.

As an example of the former, we observed no significant methylation on any of the 53BP1 methylation window peptides. As such, the investigative effort was shifted to focus entirely on CFP1.

5.5 A proposed MLL4-dependent methyl-switch mechanism for controlling CFP1-Setd1 association

The proximity of the lysines that undergo apparent MLL4-dependent methylation in CFP1 is considerable, and the fact that this highly methylated lysine rich region occurs between amino acids 331 to 359) is very intriguing. As stated above, the rationale for defining the CFP1 SID as occurring strictly within amino acids 361-422 is unclear and does not appear to be stated in the literature [14]. This combined with the impressive density of apparent methyllysine modifications located within a region immediately proximal to what is defined as the SID led us to speculate that MLL4-dependant methylation of this region could serve a regulatory purpose. As such, we hypothesized that MLL4-dependant methylation of a lysine-rich domain proximal the SID could be involved in regulating the association of CFP1 with Setd1A/B.

To test this theory, we designed a pulldown experiment utilizing 6xHis-tagged Ni-NTA-bound Setd1A (1362-1563) as an affinity trap for either methylated or unmethylated un-tagged CFP1 (302-656). We methylated un-tagged CFP1 (302-656) samples in solution with MLL4-WRAD (detailed in 4.2.4) and confirmed methylation had taken place by running a concurrent bioluminescent assay with the same samples. The methylated and unmethylated CFP1 (302-656) fragments were separately combined with a molar excess of Setd1A (1362-1563) and allowed to co-incubate overnight. Each sample was incubated with Ni-NTA beads for two hours to facilitate interaction with the His tag of the Setd1A construct. The solution of beads and protein was centrifuged, and the supernatant was collected. This supernatant sample represents the “flow-through”, i.e. proteins in the solution that were unable to interact with any of the other constituents. The beads were washed in buffer of increasing salt concentration, with the aim of disrupting hydrostatic interaction between the CFP1/CFP1me (302-656) and Setd1A (1362-1563) constructs. The final sample was a fraction of beads melted in Laemmli buffer, which

should contain the Setd1A construct and any CFP1/CFP1me (302-656) not affected by the salt washes. This procedure is discussed in greater detail in 4.2.6.

Interestingly, we observed that MLL4-methylated CFP1 (302-656) appeared to bind with the Setd1A construct with enough strength that it was only present in the Laemmli sample (Figure 4.1C). Conversely, we observed that unmethylated CFP1 (302-656) was present only in the flow-through solution, suggesting it was mostly unable to associate with the Ni-NTA bound Setd1A construct. Taken together, these findings support our hypothesis that MLL4-dependant methylation of CFP1 within the lysine-rich basic domain forms the basis of a methyl-switch mechanism controls CFP1-Setd1A association, and more broadly indirectly links MLL4 activity with the regulation of Setd1A/B COMPASS activity through modulation of COMPASS formation. This hypothesis could be further supported by conducting a native PAGE experiment, wherein unmethylated and methylated CFP1 are interacted with Setd1A and observed for differences in electromobility, indicating the presence or absence of protein-protein interaction. This experiment could also provide insight regarding a role for methylation in the formation of CFP1 homodimers. It would also be interesting to examine the biomolecular implications of CFP1 basic domain methylation. The SID interacts with a CFP1-interacting region on Setd1 proteins (1415-1450 in Setd1A) [15]. A crystal structure of CFP1 and Setd1A interacting through these regions could provide insight regarding the effect of methylation on this interaction, though none exist at the time of publication.

5.6 The question of CFP1 homodimerization

In Chapter 3.3.3, we examined evidence from previous publications that hinted the basic domain of CFP1 may facilitate dimerization [16, 17], and that CFP1 has been observed to associate

with Setd1 COMPASS at a stoichiometric ratio of 1.6:1 [18]. In Butler et al. 2008 Figure 4B (Figure 5.1), it is clear that the truncated CFP1 fragment lacking the basic domain (361-656) abrogated Setd1A association by approximately 50% compared to the same fragment (302-656) containing with a basic domain [14]. Surprisingly, this detail wasn't discussed in the publication, and may represent otherwise unreported evidence that the basic domain in somehow involved in Setd1A association. It is interesting to speculate that these findings are related. What if homodimerization is required for Setd1 association? Could MLL4-dependant basic domain methylation enhance Setd1A association through inducing homodimer formation?

With further speculation, we wonder if the fact that CFP1 is known to associate with DNMT1 through a means distinct from that of Setd1 involves the basic domain methyl-switch [14]. If so, this methyl-switch could be the yet undiscovered mechanism controlling the CFP1 binding partner decision. Surprisingly, undiscussed results in Butler et al. 2008 also lends credence to this hypothesis. Figure 5 in this publication (Figure 5.2) clearly shows that CFP1 (361-656) containing a C375A mutation which abolishes Setd1A binding actually associates with DNMT1 approximately twice as much as CFP1 lacking the mutation [14]. While these fragments don't contain the basic domain, this result does demonstrate that mutations abrogating Setd1 binding appear to promote association with DNMT1. It should be noted that we observed no obvious homological evidence for the evolution of this mechanism in tandem with MLL4 (Figure 3.6), but it is possible that it is in fact conserved, as little research has been done regarding the specific interaction site between CFP1 and DNMT1.

In summary, we observe there is merit to investigate the idea that MLL4-dependant methylation of the C-terminal CFP1 basic domain induces CFP1 homodimerization. As a homodimer, CFP1 could more readily associate with Setd1 protein complexes. Conversely,

monomeric CFP1 preferentially associates with DNMT1. Mutations in the SID disrupt the ability of CFP1 to dimerize, and thus mimic the unmethylated state. Thus, CFP1 partner choice may be indirectly controlled through MLL4-mediated methylation. While admittedly speculative, the synthesis of the above described data suggests the existence of a “bigger picture” mechanism, and we posit investigation of this hypothesis would be a good place to start. The above-described native PAGE experiment is a logical first step in this investigation.

5.7 Preliminary investigation of non-histone PHD finger binding between MLL4 PHD₁₋₃, PHD₄₋₆, and PHD₇

Though PHD domains are typified by their ability to read histones and histone modifications, we thought it could be enlightening to screen our PHD₁₋₃, PHD₄₋₆, and PHD₇ constructs against a peptide array containing each of the putative CFP1 methylation sites at methylation states of me0, me1, me2, and me3. It is noteworthy that emerging evidence suggests PHD domains may participate in interactions beyond their canonical role in histone recognition. [19]. After all, histone lysine methyltransferases were long thought to be specific only to histones, but it is now clear that histone lysine methyltransferases are involved in a plurality of enzyme-substrate reactions [2].

We found that the methylation state of particular lysine residues in the CFP1 basic domain affected *in vitro* binding of PHD₁₋₃ and to a lesser degree PHD₄₋₆ constructs (Figure 4.3B). For PHD₁₋₃, CFP1 K331me1 and K339me1 preferentially interacted compared to me0, me2, or me3. All methyllysine states of K335 were bound with an affinity greater than 50% of the wild-type H4K20me0 event, but me3 appeared to potentiate this effect. K353 was bound in the me0 and me1 state, but the me2 and me3 moieties abrogated interaction. Interestingly, one of the most conclusive

results in this experiment was seen in the three most SID-proximal lysines K355, K357, and to a lesser degree K359. In these residues, the significant binding seen in me0 and me1 SPOTs was completely abolished by me2 and me3 modifications. Though K359 wasn't bound above 50% of the WT signal in any residues, this pattern was conserved. These residues were identified as of potential interest in Chapter 3, as they were observed to be only conserved in mammals. The abrogation of binding with increasing methyl state at these residues could be indicative of hydrogen bonding between the CFP1 peptides and the PHD₄₋₆ cassette. Methyl group substitution of Nε hydrogen atoms increases the size and hydrophobicity of the nitrogen moiety, while limiting hydrogen bonding such that me3 modification is characterized by a total loss of hydrogen bond participation [20].

The relationship between lysine position and methyl state similarly impacted PHD₄₋₆ binding (Figure 4.3C). K335 exhibited binding at all states, but me2 and me3 had the greatest effect. In contrast with PHD₁₋₃, trimethylation seemed to be key important for interaction with residues K335, and K340. Di-methylation of K351 increased interaction with PHD₄₋₆. Finally, the most SID-proximal lysines K353, K355, K357 and K359 once again best facilitated binding when in a me0 or me1 state. It is intriguing that this result mirrors that seen with PHD₁₋₃.

PHD₇ interacted with the array to some degree, but few interactions appeared meaningful (Figure 4.3D). Of those that did, K335 exhibited greater binding while unmethylated, and K345 binding activity was significantly diminished by trimethylation.

As discussed in 4.3.3, the biological relevance of this finding is unclear. There does appear to be a pattern among certain residues, and this may merit additional investigation should the field of non-histone PHD finger binding develop further.

5.8 Final Conclusions

For this thesis, we aimed to define and characterize new methyllysine substrates for the H3K4 methyltransferase MLL4. We undertook both shotgun and systematic analytical approaches which led us to 53BP1 K1563 and CFP1 K328 as potential MLL substrates. While we failed to validate 53BP1 K1563, CFP1 appeared to be a promising result. Investigation into which CFP1 residues were being methylated revealed a methyllysine rich region within the basic domain, N-terminal and proximal to the SID, which is required for association with Setd1A/B. Homological analysis of the methylated region revealed that most sites were conserved except for the three residues most proximal to the SID. We compared interaction between a Setd1A construct and either methylated or unmethylated CFP fragments, and showed that methylation appears to promote binding. From these data we proposed the existence of methyl-switch mechanism that controls CFP1 association with the Setd1 COMPASS, and further speculated that this mechanism may be involved in CFP1 binding partner decisions between Setd1A and DNMT1. Finally, assayed the ability of MLL PHD domains to interact with CFP1 peptides of various methyl states, and observed that particular methyl states in some positions appear to influence binding.

Figures

Figure 5.1 Figure 4A and B from Butler et al. 2008 [15]

Figure 5.2 Figure 5 from Butler et al. 2008 [15]

References

- [1] Ambler R.P. & Rees M.W. (1959). ϵ -N-Methyl-lysine in Bacterial Flagellar Protein. *Nature* 184: 56-57.
- [2] Biggar K.K., & Li S.S. (2015). Non-histone protein methylation as a regulator of cellular signaling and function. *Nature Reviews Molecular Cell Biology* 16(1):5-17.
- [3] Shilatifard, A. (2012). The COMPASS Family of Histone H3K4 Methylases: Mechanisms of Regulation in Development and Disease Pathogenesis. *Annual Review of Biochemistry*, 81(1), 65–95. doi:10.1146/annurev-biochem-051710-134100
- [4] Zhang Y., Mittal A., Reid J., Reich S., Gamblin S.J. & Wilson J.R. (2015). Evolving Catalytic Properties of the MLL Family SET Domain. *Structure* 23(10): 1921-1933.
- [5] Liu Y., Qin S., Chen T.-Y., et al. (2019). Structural insights into trans-histone regulation of H3K4 methylation by unique histone H4 binding of MLL3/4. *Nature Communications* 10: 36.
- [6] Dhar S.S., Lee S.-H., Kan P.-Y., et al. (2012). Trans-tail regulation of MLL4-catalyzed H3K4 methylation by H4R3 symmetric dimethylation is mediated by a tandem PHD of MLL4. *Genes and Development* 26: 2749-2762.
- [7] Zhang Y., Jang Y., Lee J.-E., et al. (2019). Selective binding of the PHD6 finger of MLL4 to histone H4K16ac links MLL4 and MOF. *Nature Communications* 10: 2314.
- [8] Sze C.C., & Shilatifard A. (2016). MLL3/MLL4/COMPASS Family on Epigenetic Regulation of Enhancer Function and Cancer. *Cold Spring Harbor Perspectives in Medicine* 6(11): a026427.
- [9] Rao R.C., Dou Y. (2015). Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. *Nature Rev. Cancer*. 15(6): 334–46.

- [10] Zaidi S., Choi M., Wakimoto H., Ma L., Jiang J., Overton J.D., et al. (2013). De novo mutations in histone-modifying genes in congenital heart disease. *Nature* 498(7453): 220–3. [39]
- Rao R.C., Dou Y. (2015). Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. *Nature Rev. Cancer*. 15(6): 334–46.
- [11] Ng S.B., Bigham A.W., Buckingham K.J., Hannibal M.C., McMillin M.J., Gildersleeve H.I., et al. (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genetics* 42(9): 790–3.
- [12] Panier S., & Boulton S.J. (2014). Double-strand break repair: 53BP1 comes into focus. *Nature Reviews Molecular Cell Biology* 15: 7-18.
- [13] Tate C.M., Lee J.-H., & Skalnik D.G. (2011). CXXC Finger Protein 1 Restricts the Setd1A Histone H3-Lys4 Methyltransferase Complex to Euchromatin. *FEBS Journal* 277(1): 210-223.
- [14] Butler J.S., Lee J.H., & Skalnik D.G. (2008). CFP1 interacts with DNMT1 independently of association with the Setd1 Histone H3K4 methyltransferase complexes. *DNA and Cellular Biology* 27(10): 533-543.
- [15] Lee J.-H. & Skalnik D.G. (2008). Wdr82 Is a C-Terminal Domain-Binding Protein That Recruits the Setd1A Histone H3-Lys4 Methyltransferase Complex to Transcription Start Sites of Transcribed Human Genes. *Molecular and Cellular Biology* 28(2): 609-618.
- [16] Voo K.S., Carlone D.L., Jacobsen B.M., Flodin A., & Skalnik D.G. (2000). Cloning of a Mammalian Transcriptional Activator That Binds Unmethylated CpG Motifs and Shares a CXXC Domain with DNA Methyltransferase, Human Trithorax, and Methyl-CpG Binding Domain Protein 1. *Molecular Cell Biology* 20(6): 2108-2121.
- [17] Beurton F., Stempor P., Caron M., Appert A., Dong R., Chen D., Cluet D., Coute Y., Herbette M., Huang N., Polveche H., Spichty M., Bedet C., Ahringer J., & Palladino F. (2018).

Physical and functional interaction between SET1/COMPASS complex component CFP-1 and a Sin3 HDAC complex. bioRxiv 436147.

[18] van Nuland R., Smits A.H., Pallaki P., Jansen P.W., Vermeulen M., Timmers H.T. (2013). Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes. *Molecular Cell Biology* 33(10): 2067-2077.

[19] Li Y., & Li H. (2012). Many keys to push: diversifying the ‘readership’ of plant homeodomain fingers. *Acta Biochimica et Biophysica Sinica* 44(1): 28-39.

[20] Luo M. (2018). Chemical and Biochemical Perspectives of Protein Lysine Methylation. *Chemical Reviews* 118(14): 6656-6705.

Supplementary Figures

MLL4 SET (5382-5537) – 44 kDa

WDR5 – 37 kDa

RbBP5 – 60.6 kDa

Ash2L – 69.4 kDa

Dpy30 – 37.3 kDa

53BP1 (1484-1603) – 18.5 kDa

CFP1 (302-656) – 42.5 kDa

Setd1A (1362-1563) – 28.5 kDa

Figure S1. Purified proteins used in this study