THE UNIVERSITY OF CHICAGO

CHARACTERIZATION OF OXIDIZED METHYLCYTOSINE BINDING PROTEIN ACTIVITIES IN THE MAMMALIAN BRAIN AND STEM CELLS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

BY

KATHRYN ELIZABETH MALECEK

CHICAGO, ILLINOIS

JUNE 2016

For Leo F. Slattery.

Live every day like it's brain day.

Table of Contents

List of Figures

List of Tables

ACKNOWLEDGMENTS

This work was deeply guided by Alex Ruthenburg's bold thinking and tireless commitment to the power of biochemistry. I am grateful to him for pushing me to pursue this project, especially when it was difficult and slow going. I am thankful for his countless helpful ideas, his patience, and his helping hands on brain collection days in Hobart, Indiana.

My colleagues in the lab provided a wonderful environment in which to pursue challenging research, and I thank them for their contributions of insight, encouragement, and good humor. I particularly enjoyed working with an undergraduate in the lab, Matthew Sullivan, who contributed to some of oligonucleotide purifications and candidate binding experiments presented in the appendices. I also acknowledge the support of many classmates, friends, and other mentors outside the lab for their enthusiasm for my work. I am very thankful for the support of my thesis committee members and graduate program administrators for their advice and guidance during my time in graduate school.

Finally, I am deeply grateful for the support of my family. I thank my brother, parents, and grandparents for rallying to support me in every endeavor. I would especially like to acknowledge those who taught me to love science: Leo Slattery, Nick Malecek, and Lee McCuller.

ABSTRACT

5-Methylcytosine embedded in mammalian DNA represses local transcription by recruiting modification-specific binding partners. Its active removal is initiated by sequential oxidation of the 5-methyl group by TET enzymes to produce three oxidized species, collectively referred to as [ox]mC. Although rare, the distribution of [ox]mC modifications is tissue-, gene-, and coding strand-specific and distinct from 5-methylcytosine, suggesting unique functions. To examine this possibility, I fractionated mammalian brain extracts to discover, isolate and characterize binding partners specific for [ox]mC. This purification reveals remarkably specific factors that are selective for each of the three oxidation states and sensitive to the 5-modification state on each strand. I demonstrate that one such factor, WDR76, is a highly 5-hydroxymethylcytosine-specific binding protein. I have begun to lay the foundation for further mechanistic studies of these specific binding proteins in mouse embryonic stem cells and leukemia. My results provide an essential bridge from studies of the distribution of [ox]mC and the effects of TET knockouts, to the possible functions of [ox]mC recognition in gene regulation or chromatin signaling.

Chapter 1

THE BIOLOGY OF OXIDIZED METHYLCYTOSINE

1.1 DNA methylation - classical mechanisms

The particular properties of a given cell are largely defined by its complement of protein and RNA expression, which is determined by its pattern of gene activity and the transcription factors that direct it. However, gene expression is not sufficiently constrained by the set of transcription factors within a given cell alone, as cells with similar transcription factor profiles can differ in their transcriptional behavior. Moreover, the gene expression potential of a given cell can be stably restricted such that is not necessarily reversible by exposure to a different set of transcription factors. The stable restriction of the output of the genome of different cell types within the same organism of has motivated the characterization of epigenetic mechanisms for modifying transcriptional activity. Paramount among such mechanisms are cytosine methylation within DNA, and histone modifications within chromatin. Both help explain substates of cell character and their stability, as well as illuminating the molecular basis of plasticity in cell identity at the chromatin level. My dissertation represents an expansion upon the classical mechanisms of DNA methylation to uncover a potential role for specific recognition of DNA demethylation intermediates as epigenetic modifications in their own right. Through their specific recognition by cognate binding partners identified herein, these molecules may also inform transcriptional output and cell identity, constituting an additional layer of chromatin information.

Across mammals, modification of DNA by methylation of cytosine regulates gene expression, cell identity, development, and disease processes. DNA methyl transferases (DNMTs) install methyl groups at the 5-position of cytosine within the bodies of developmentally regulated genes, at imprinted genes with monoallelic expression, and importantly, across much of the repetitive sequence within the genome. In each instance, the presence of methylcyto-

sine is critical for transcriptional regulation and genome stability. Methylcytosine (mC) is predominantly found within CpG dinucleotides, wherein, as a result of Watson-Crick base-pairing, the cytosine-phosphodiester-guanine sequence is found on both strands of the DNA helix. mC is therefore often found in a symmetric manner at both cytosines of the opposed CpG dinucleotide. mC can be found in other contexts such as CpA or CpH where H can be A, C or T. These contexts represent a relatively minor proportion of cytosine methylation, but are enriched in certain cell types and gene features, and may be recognized as biologically distinct from mCpG and its binding proteins. This may be of particular importance in the brain [Guo et al. 2014] and in stem cells [Ramsahoye et al. 2000], where methylation is most dynamic and where some mC-specific binding proteins may be sensitive to these different presentations [Gabel et al. 2015; Guo et al. 2014]. CpG methylation will be the focus of this introduction, but CpA methylation will be considered later with regard to demethylation intermediates and gene regulation in stem cells.

1.1.1 The DNA methyltransferases

DNA methylation is mediated by two enzymatic activities: *de novo* and maintenance methylation. DNMT3A and DNMT3B introduce *de novo* DNA methylation at unmodified CpG sites, while DNMT1 recognizes asymmetrically methylated CpG dinucleotides and deposits a second methyl group on the unmodified DNA strand. The activity of DNMT1 during DNA replication allows DNA methylation patterns to be permuted to subsequent generations, allowing stable maintenance of the methylated state through repeated semi-conservative DNA replications and cell division cycles [Klose and Bird 2006; Bird and Wolffe 1999]. The precise mechanistic details of how the domains of the DNMTs contribute to this activity are well reviewed elsewhere [Bestor 2000].

The activity of DNMT3a and DNMT3b across the genome is modulated by the co-factor DNMT3L, and guided by various means. First, the DNMTs themselves or proteins with

which they associate can recognize particular chromatin domains by their histone modifications either directly, such as by the PWWP domain of the DNMT3a and DNMT3b which recognize sites of H3K36 methylation [Qin and Min 2014], or indirectly by DNMT3L [Ooi et al. 2007] excluding H3K4me3 modified chromatin and other DNMT-associated factors that localize to regions of H3K27, H3K36, and H3K9 methylation [Viré et al. 2006; Cedar and Bergman 2009; Morselli et al. 2015]. Through coupling with histone modifications, DNA methylation therefore seems to act at genes that have already been partially silenced by transcriptionally repressive histone modifications [Brinkman et al. 2012; Statham et al. 2012]. Second, DNA methylation can be directed by cell-type specific DNA binding proteins that recruit DNMTs to their cognate sequences, a mechanism that can promote the transcriptional repression of tumor suppressors, [Di Croce et al. 2002], such as recruitment of DNMT activity to Myc E-box binding sites [Brenner et al. 2005]. Third, *de novo* DNA methylation activity can be directed by promoter or centromere derived RNAi that recruits DNMTs to sites of homology [Kawasaki and Taira 2004; Kanellopoulou et al. 2005]. The generality of this mechanism remains controversial as gene silencing via RNAi is also observed at some of the examined promoters in the absence of DNA methylation. However, in plants, short RNAs homologous to DNA sequences can initiate *de novo* DNA methylation [Chan et al. 2004; Mette et al. 2000], and this has been demonstrated to work in human cells using shRNAs to target the promoter of tumor suppressor genes [Castanotto et al. 2005]. At present, there remains some expectation that this is a minor mechanism.

All three DNMTs are essential for proper embryonic development [Okano et al. 1999] and viability of somatic cells, though they contribute somewhat differently in each context. DNMT3b and DNMT1 are most essential for maintenance of early embryonic DNA methylation patterns whereas DNMT3a contributes more so to *de novo* methylation in adult somatic tissues and CpA methylation in general.

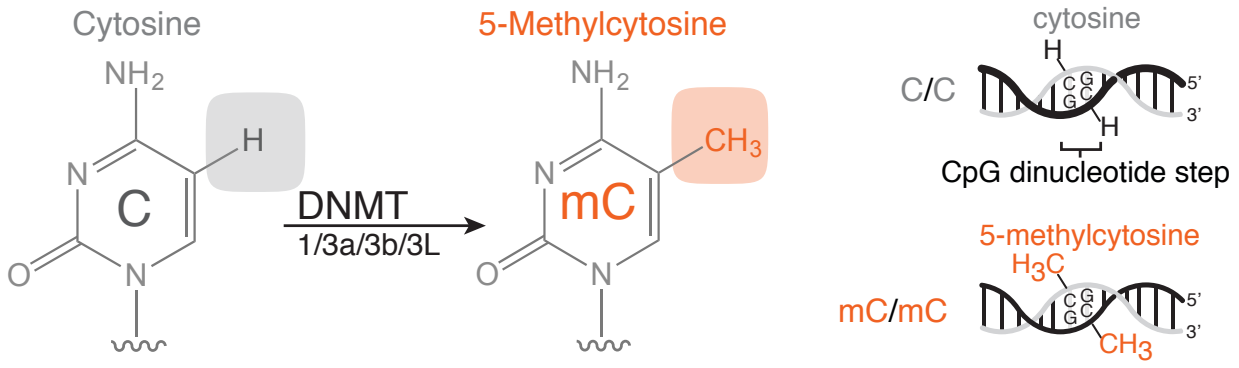


Figure 1.1: Methylation of cytosine by DNA Methyl Transferases (DNMTs) occurs at CpG dinucleotides. The symmetry of CpG dinucleotide step allows for symmetric modification of these sites on both strands of the DNA helix.

1.1.2 Methylcytosine binding proteins

DNA methylation constitutes a heritable mechanism by which the genome can be annotated without altering its coding potential. Moreover, this annotation informs local gene regulation through its recognition by mC-specific binding proteins. It is this connection between chemical modification and functional readout in the form of altered transcriptional activity on chromatin that has come to define mC as an epigenetic regulator and a paradigm among epigenetic mechanisms.

mC is specifically bound by families of proteins containing methylcytosine binding domains (MBDs) or SET and RING-finger associated (SRA) domains. These proteins act to recruit chromatin-modifying enzymes, namely histone deacetylase complexes, that through downstream chromatin remodeling alter the transcriptional activity of the underlying genes. With the exception of MBD3, members of the MBD family bind tightly to symmetrically modified mCpG dinucleotides with mid to low nanomolar binding affinities. [Jones et al. 1998; Hendrich and Bird 1998]

The SRA domains of UHRF1 and UHRF2 recognize asymmetric, hemi-methylated DNA sequences and bind them in an extra-helical fashion, flipping the methylated base out of the double helix to interrogate the mC/C modification state [Arita et al. 2008; Avvakumov et al.

2008]. This activity makes UHRF1 an ideal partner to promote the activity of DNMT, with which it associates during DNA replication [Bostick et al. 2007; Hashimoto et al. 2009]. These SRA domain-containing proteins also contain tandem Tudor domains that recognize H3K9 methylation and exclude H3K4 methylation [Nady et al. 2011], and therefore are important for promoting DNA methylation of chromatin bearing this modification state [Pichler et al. 2011].

Finally, certain Cys2-His2 zinc finger proteins (C_2H_2) bind DNA in a methylation-dependent manner over longer stretches of specific sequences when internal CpG dinucleotides embedded in these cognate sequences are methylated. These most notably include the transcription factors Kaiso, Klf4, and Zfp57, whose roles range from DNA repair to pluripotency transcription factors.

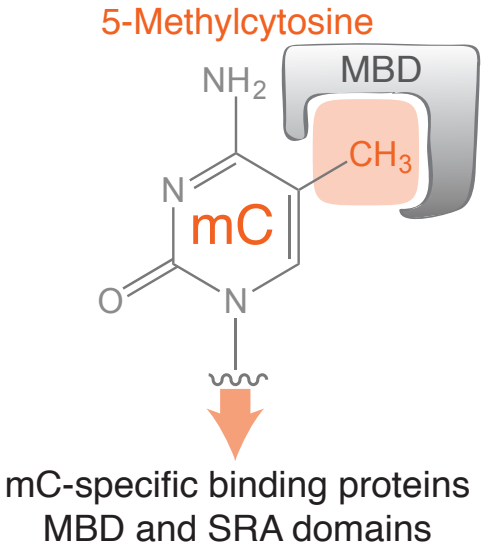


Figure 1.2: Methylcytosine is specifically bound by proteins containing MBD and SRA domains

1.1.3 Mapping of methylcytosine, functions and dynamics

Sites of modified methylcytosine can be resolved from sites of unmodified cytosine by bisulfite sequencing methods [Clark et al. 1994]. (These methods will be reviewed in section 1.3.)

mC can now be mapped with single nucleotide resolution and with increasing depth on the complementary strands of individual DNA fragments, allowing methylation frequency to be estimated for a particular cytosine within a given cell type or lineage [Zhao et al. 2014]. From extensive application of bisulfite sequencing over the last 20 years, it is apparent that mC is often installed at CpG dinucleotides within gene bodies. The acquisition of mC is often coupled to other forms of stable chromatin silencing, such as H3K9 and H3K27 methylation, and chromatin compaction [Jones et al. 1998; Bird et al. 1998]. These mechanisms and their misregulation in disease are well reviewed elsewhere [Klose and Bird 2006; Robertson 2005; Bergman and Cedar 2013; Cedar and Bergman 2009].

Bisulfite sequencing studies have revealed that cytosine methylation is deployed during early mammalian development in order to silence pluripotency genes and cell type specific genes of other lineages. However, it is also clear that *de novo* cytosine methylation occurs in the adult, particularly in the neuronal and immune systems as cells are further specified in their cell identity and activity.

Cytosine methylation is also observed by bisulfite sequencing to dramatically change during the course of development and during cell differentiation processes in the adult. Cytosine methylation is rapidly lost across the genomes of primordial germ cells and in paternal pronuclei of fertilized zygotes. In these cell types, the loss of mC is nearly global [Lister et al. 2011; Lister et al. 2013; Meissner et al. 2008]. This loss is particularly striking during development, when global and site specific demethylation is observed in order to reestablish a pluripotent state in primordial germ cells [Oswald et al. 2000], and in developing embryos, when cell type specific genes are demethylated as cells commit to particular fates [Meissner et al. 2008]. Cytosine methylation is lost at specific loci in several biological contexts. It is critical in somatic tissues for diversification of antibody production in B cells and in activated T lymphocytes [Scharer et al. 2013; Bruniquel and Schwartz 2003], in the adult brain during learning [Day and Sweatt 2010], and in response to hormonal signaling

[Métivier et al. 2008]. The observation of both passive and rapid, site-specific and global loss of mC suggests that a dynamically regulated enzymatic activity functions to remove methylcytosine. One particularly striking example is the loss of methylcytosine across the CpG dinucleotides of the brain derived neurotrophic growth factor promoter when cultured mouse neurons are stimulated to depolarize with salt solutions [Martinowich et al. 2003]. As shown in fig. 1.3, the rapid reduction in the relative methylation level of many bisulfite sequenced clones indicates that mC is lost instantaneously upon neuronal stimulation. The basis for this activity was not known at the time, though many were speculated. However, given its rapid onset in stimulated neurons, this loss of mC clearly could not be attributed to passive dilution - a targeted enzymatic process must be at work.

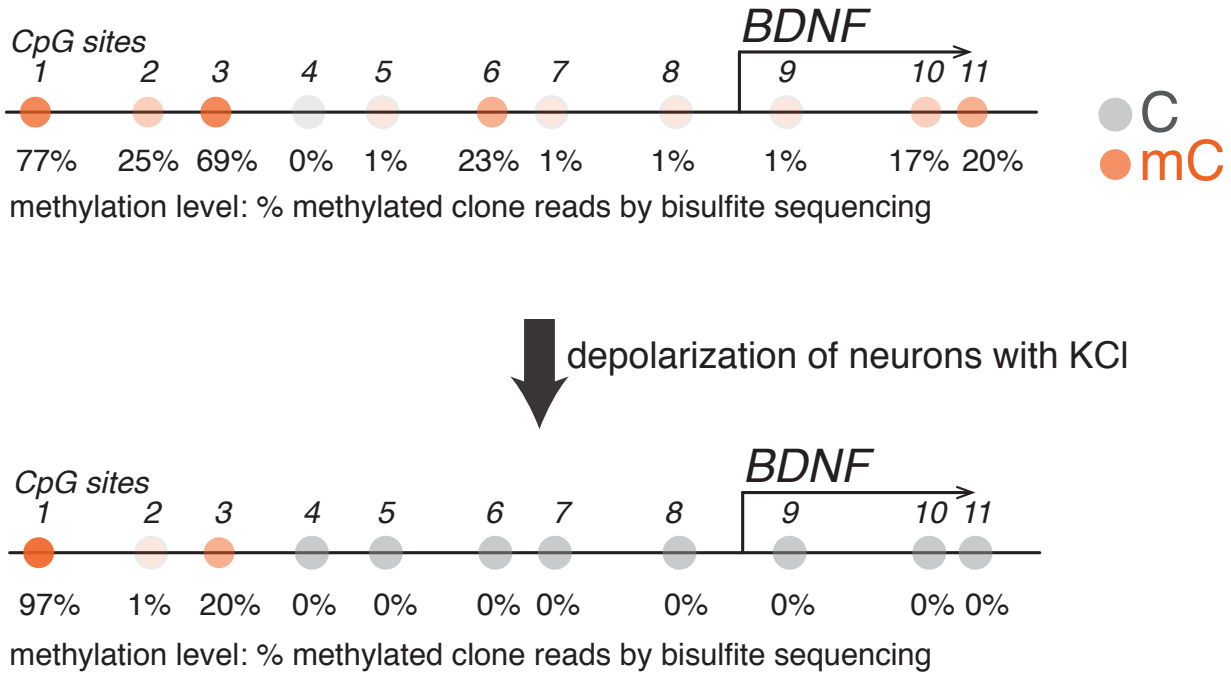


Figure 1.3: CpG methylation of the *BDNF* (brain-derived neurotrophic growth factor) promoter before and after KCl stimulation of mouse E14 cortical neurons. Individual CpG dinucleotides are represented by circles shaded according to their degree of methylation estimated from bisulfite sequencing of 35 clones. The promoter is observed to lose methylation upon depolarization of the neurons, in the absence of cell division. Adapted from [Martinowich et al. 2003]

1.1.4 Previously proposed mechanisms for the removal of methylcytosine

An active demethylation activity has long been sought to explain how cells remodel their gene expression programs at the chromatin level through mC, and is well summarized in the following review [Wu and Zhang 2010]. For many years, the field invoked passive dilution as a somehow site-specific mechanism for removal of of methylcytosine [Lin et al. 2000], with some evidence in mammalian primordial germ cells that passive loss being the predominant mode [Kagiwada et al. 2013; Seisenberger et al. 2012] . Several activities have been proposed to enzymatically remove mC by direct nucleotide- or base-excision repair pathways, perhaps initiated by Gadd45 [Niehrs and Schäfer 2012; Barreto et al. 2007], a cytosine deaminase activity provided by either AID or APOBEC [Popp et al. 2010; Guo et al. 2011; Bhutani et al. 2010], or radical oxygen mechanisms to break the C-C bond directly [Okada et al. 2010]. Each of these mechanisms has received mixed support in the literature. The association of Gadd45a with demethylation is not reproducible, and it is not expressed in key cell types that undergo dramatic active demethylation, namely oocytes and zygotes [Jin, Guo, and Pfeifer 2008]. The activity of AID and APOBEC is preferential for unmodified cytosine, and not mC (or hydroxymethylcytosine, the deamination product of which, hydroxymethyuracil, is not detected in cells with abundant hmC) [Nabel et al. 2012]. But perhaps most problematic for resolving the enzymatic basis of any demethylation pathway, deletion of these proposed active demethylase genes does not disrupt the loss of methylation observed in embryonic stem cells [Engel et al. 2009; Revy et al. 2000; Muramatsu et al. 2000], indicating that each of these are not the *major* pathway by which mC is actively removed from the genome in this relevant cell context.

While clear biochemical and genetic evidence in plants supports the ability of a direct DNA glycosylase, Demeter, to remove mC, there is no compelling mammalian orthologue. However, consideration of the another demethylase solution from bacterial systems, AlkB, ultimately provided an instructive lead in the search for an active demethylation activity.

AlkB functions to remove alkylating damage from DNA using iron, 2-oxoglutarate, and oxygen as enzymatic cofactors [Falnes, Johansen, and Seeberg 2002; Trewick et al. 2002]. Similar biochemical activities have been described in fungi and trypanosomes for the base-J binding proteins [Cliffe et al. 2009]. Base J is a modified form of thymine produced by base J binding proteins (JBPs) via hydroxylation and glycosylation of the methyl group at the 5-position of thymine. Similar thymine hydroxylase activities are used in thymine salvage pathways to yield uracil through sequential oxidation of a hydroxylated thymine intermediate. Searches for similar dioxygenase domains led to the identification of the ten-eleven translocation (TET) family of proteins in mammals [Tahiliani et al. 2009; Ito et al. 2010].

1.2 The discovery of TET enzymes and oxidized methylcytosine

The TET enzymes, of which there are three in humans and related othrologs across metazoans, possess dioxgyenase domains similar to that of JBPs for coordinating iron and 2-oxoglutarate cofactors. TET1 and TET3 also contain CXXC cysteine rich zinc chelating motifs that are a common protein domain for binding CpG dinucleotides, and may specify modification states therein. The TET enzymes were found to modify cytosine alone or within CpG dinucleotides to a previously reported but somewhat forgotten base: 5-hydroxymethylcytosine, which was found at low levels in mouse embryonic stem cells [Tahiliani et al. 2009]. The initial report of the TET enzymes coincided with a report that this modified nucleobase is also present at very low abundance in the mammalian brain, particularly within the Purkinje neurons of the cerebellum [Kriaucionis and Heintz 2009]. It was further shown that the TET enzymes are each capable of catalyzing the sequential oxidation of mC to not only 5-hydroxymethylcytosine (hmC) but also to two further oxidized species: 5-formylcytosine (fC), and 5- carboxylcytosine (carC) [Ito et al. 2011; He et al. 2011]. These modified nucleobases will be referred to herein by their abbreviations and collectively as

'oxidized methylcytosine'or '[ox]mC.'

Whereas hmC represents less than 1% of total cytosine even in the most enriched tissue, the mammalian brain cortex, the abundance of fC and carC is even less than that of hmC, but can be detected in stem cells at approximately 20 and parts per million respectively in initial mass spectrometry quantification experiments [Lu et al. 2013]. The potential of [ox]mC to contribute to active demethylation, perhaps by disrupting the affinity of mC-specific binding proteins, or the altering activity of DNMTs, particularly DNMT1, at hmC-modified CpG dinucleotides was very open at this point. These new findings certainly encouraged a reevaluation of active demethylation, but raised the question of what these modified nucleobases could possibly be doing at such low levels in two very distinct tissues - highly differentiated somatic cells such as neurons, that seldom divide, and pluripotent stem cells, that have yet to specify their methylation patterns in the course of many cell divisions.

The first phenotypic descriptions of the TET enzymes in mouse embryonic stem cells indicated that lack of TET1 correlated with aberrant methylation of the *Nanog* promoter, suggesting that TET1 functions to prevent hypermethylation of this promoter [Ito et al. 2010]. Overall, diminished TET1 levels were associated with misregulation of genes associated with inner cell mass specification, and biased differentiation toward the trophoectoderm lineage in developing embryos [Ito et al. 2010].

Other studies suggest unique roles for each TET in oncogenesis pathways, particularly hematopoeitic malignancies [Moran-Crusio et al. 2011; Ko et al. 2011; Huang et al. 2013] from which the name of the TETs - *ten-eleven translocation* is derived. In various leukemias, TETs were first noted as oncogenic fusion genes with the histone methyltransferase *MLL* in patients with acute myelogenous leukemia [Lorsbach et al. 2003; Ono et al. 2002].

Some division of labor among the TET enzymes is apparent as the phenotypes of each knockout and expression levels of each TET are explored. Generally, stem cell and neural phenotypes are observed for TET1 and TET3 knockouts [Li et al. 2014], while primordial

germ cell and gamete phenotypes are observed for TET3 knockouts [Gu et al. 2011]. The particular enzymology, including any distinctions in their preferred substrates for binding or oxidation, in addition to any cell type specific cofactors for the TET enzymes, remains to be clarified. Recent work has started to reveal some distinctions among the TETs and their behavior, namely that TET1 and TET2 favor oxidation of mC to hmC over subsequent steps. The structure of the active site of TET1 and TET2 suggests that this is due to increasingly steric hindrance of hydrogen abstraction when the enzyme is bound to hmC and fC as compared to mC [Hu et al. 2015]. While the binding affinities of TET2 for different modified cytosine species are not significantly different, the activity of TET2 is greater for mC substrates than hmC substrates. Similar studies of TET3 have not been reported. The reduced reactivity of hmC and these TET enzymes seems to confer some kinetic stability to hmC within the genome, perhaps partially explaining why the first oxidation state, hmC, is far more abundant than the latter two, fC and carC. Whereas many other chromatin modifying enzymes show a great deal of allosteric regulation of their activity by association with cell-type specific protein factors, these early studies suggest that absent other protein interactions, the TET enzymes are intrinsically biased towards the first oxidation of methylcytosine only.

The latter two [ox]mC species, fC and carC, may also be less abundant because they can be excised and replaced with unmodified cytosine via base excision repair pathways by the action of thymine deglycosylase (TDG) [He et al. 2011]. A decarboxylase activity of the DNMTs or other unidentified putative enzymes has also been proposed [Liutkeviciute et al. 2014; Schiesser et al. 2012] but remains controversial as compelling *in vivo* evidence for such a specific activity is lacking, and the conditions required to observe the activity *in vitro* are quite biochemically forcing in their requirements for reducing power [Schiesser et al. 2012].

The generation of [ox]mC species and their replacement with unmodified cytosine via base excision repair together constitute an active demethylation pathway that can operate in the

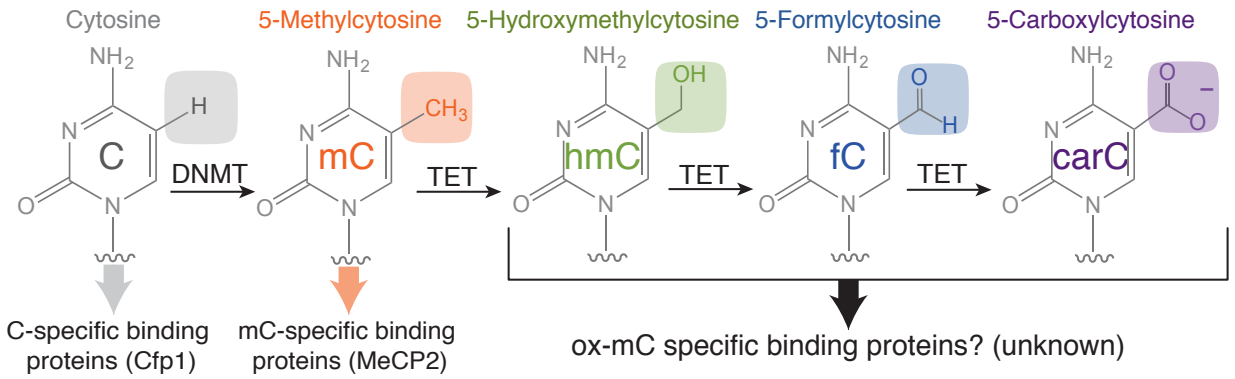


Figure 1.4: DNMTs and TETs generate methylcytosine and oxidized methylcytosine. Methylcytosine (mC) is generated from unmodified cytosine (C) by the DNA Methyltransferase enzymes (DNMTs). It was discovered in 2009 that the TET enzymes, of which there are three family members, can oxidize mC to hydroxymethylcytosine (hmC) [Tahiliani et al. 2009]. The same enzymes can perform two more iterative oxidations of hmC to form formylcytosine (fC) and carboxylcytosine (carC) [Ito et al. 2011; He et al. 2011]. Specific binding proteins for C and mC have been described, but it is not known if distinct specific proteins exist for the oxidized methylcytosine states.

absence of cell division, which can therefore account for the rapid erasure of mC observed in the early developing embryo and in non-mitotic adult neurons and immune cells. While there may be some cell types and contexts, namely primordial germ cells, where passive demethylation via dilution of mC through cell division with limited DNMT maintenance are important mechanisms [Seisenberger, Peat, and Reik 2013; Kagiwada et al. 2013], the accumulation of mC in TET and TDG knockouts indicates that the predominant means by which mC is normally lost is via the generation and excision of [ox]mC species [Wu et al. 2014; Shen et al. 2013; Hu et al. 2014]. This pathway therefore provides a long-sought mechanistic basis for dynamic cytosine methylation levels that is both more biochemically reasonable and genetically supported than previously proposed mechanisms.

1.2.1 The potential biological functions of [ox]mC species

The precise role of [ox]mC species in such a pathway, however, is not immediately clear. Ahead of further characterization, several possibilities were discussed in the field:

1. *Strictly intermediates*

It could be that [ox]mC species are merely transient intermediates generated rapidly by iterative TET oxidation events and removed rapidly by base excision repair machinery, with no meaningful lifetime on their own. It may be that the binding of conventional mC binding proteins is not significantly perturbed by [ox]mC. In this case, the fleeting period of [ox]mC modification does not have a distinct regulatory effect since any binding effect would be as shortlived as the brief lifetime of the [ox]mC species which are being generated to ultimately yield unmodified cytosine, a substrate the mC binding proteins already exclude. Given this, [ox]mC species and TET enzymes would only function as an enzymatically accessible means to revert mC to unmodified C, but the particular chemical identity of the [ox]mC species has no biological implications and no meaningful lifetime within DNA, and TETs function only as a foil to DNMTs, not to generate [ox]mC to exist for their own inherent purposes. One expectation of this model is that there is no reason for [ox]mC modifications to arise anywhere that mC does not already arise and act. [ox]mC species should only be found at sites where mC is already observed, and observed to be lost. A second expectation is that the absence of the TET enzymes would only be problematic for genes that need to lose methylation rapidly or in the absence of cell division. All other genes may be able to lose methylation passively or by another mechanism.

2. *Passive diversions from methylcytosine*

It could be that the function of [ox]mC species, either uniquely or collectively for all three forms, is to toggle or disrupt the affinity of mC or C specific binding proteins for these mod-

ifications. For example, a given MBD-containing protein might tolerate binding hmC but have lower affinity for hmC as compared to mC, and therefore the presence of hmC serves to tune down the probability of binding occupancy by this protein. Similarly, the presence of hmC opposite mC at an asymmetrically modified CpG dinucleotide could disrupt the binding of an SRA domain protein and the activity of the maintenance methyltransferase DNMT1 in the next generation of cells. Through cell division, this affinity disruption could lead to slow loss of mC across from [ox]mC species at CpG dinucleotides in subsequent generations not through active removal of any particular mC within a CpG nucleotide, but as CpGs bearing [ox]mC species preclude DNMT1 from maintaining the methylation pattern on new daughter strands of DNA. This would be a semi-passive role for [ox]mC species in that they 'distract'conventional mC binding proteins and maintenance DNMT enzymes, but do not exert other protein partner recruitment or gene regulation effects on their own. Again, an expectation of this model is that there is no reason for [ox]mC modifications to arise anywhere that mC does not already arise and act.

3. *An active and orthogonal signaling pathway*

Finally, it could be that [ox]mC species function to recruit new specific binding proteins to the sites within the genome where they arise and have some stable lifetime. In order to be recruited with a high probability of binding occupancy, these proteins would have to recognize the very rare [ox]mC modifications with tremendous specificity. However, recognition of these modifications uniquely would serve to recruit new protein activities to sites undergoing changes in DNA modification state. Such proteins could be useful effectors of gene regulation at a remodeling locus. It could also be that [ox]mC modifications have a biologically meaningful lifetime in DNA, such that they arise and persist without being removed and replaced with unmodified cytosine. In this case, it is possible that [ox]mC species would be found in places where mC is not otherwise stable or long-lived, but that perhaps have been

overlooked or undetected until now. If so, some distinct signaling effect orthogonal to that of mC could be initiated by [ox]mC-specific binding proteins at sites where the TETs act to generate these species for their own stable lifetime.

These potential models of [ox]mC function are certainly not mutually exclusive, and it could be that at different sites in the genome or times during development, different mechanisms could be at play. Essentially, each model makes predictions about two basic questions regarding [ox]mC biology:

(1) Where are [ox]mC modifications and for how long?
(2) Is [ox]mC modified DNA uniquely bound by an uncharacterized set of binding proteins, or by known mC binding proteins?

In an effort to distinguish between these models but ahead of undertaking a biochemical search for [ox]mC specific protein activities, one can first look to the recently developed sequencing strategies for mapping sites of [ox]mC modifications genome wide. The distribution and abundance of [ox]mC species can help indicate if their localization and lifetime is distinct from mC, and therefore hint whether it is biologically reasonable to expect that these modifications could have functions separate from that of their precursor, mC.

1.3 Overview of sequencing method development to map sites of mC and [ox]mC.

Bilsulfite sequencing strategies were originally developed to modify sites of C and not mC in order to allow the two to be distinguished at the level of sequence. During bisulfite treatment, cytosine undergoes sulfurization at the six position to yield cytosine sulfonate. In the subsequent hydrolytic deamination, the cytosine sulfonate is converted to uracil sulphonate

and ultimately uracil, as shown in fig. 1.5. Through subsequent PCR amplification prior to sequencing, the cytosine will ultimately be read out as thymine. Methylcytosine is not susceptible to bisulfite treatment and only converts very slowly to methylcytosine sulfonate. For this reason, methylcytosine is not converted to uracil by bisulfite deamination, and is still read out as C in sequencing experiments. When two sequencing runs, one with bisulfite treatment and one without, are performed in parallel on a sample, the differences between the two can be used to determine sites of methylcytosine in the input. Sites that undergo C to T conversions between the untreated and the bisulfite treated sample represent unmodified cytosines, whereas sites of C that persist in the bisulfite treated sample represent sites of methylcytosine.

Or so it was thought, until the low abundance of [ox]mC species in the genome was finally appreciated. Bisulfite treatment also modifies [ox]mC species to a great extent, and therefore they are lost and somewhat construed in bisulfite sequencing results. hmC reacts with bisulfite to form cytosine 5-methylsylfonate (CMS). This species is not readily deaminated to uracil, and the bulky adduct is problematic to PCR amplification in the usual bisulfite sequencing workflow [Huang et al. 2010; Jin, Kadam, and Pfeifer 2010], presumably by stalling DNA polymerase at CMS sites, particularly CMS clusters. This means that conventional bisulfite sequencing either conflates mC and hmC in its readout, or perhaps more likely, hydroxymethylated regions of DNA (though rare) are underrepresented in bisulfite sequencing experiments.

Information about sites of fC and carC is also convoluted by conventional bisulfite sequencing. Several groups report a bis-adduct formed from bisulfite treatment of formylcytosine. This species is observed to deformylate (decarbonylate) back to cytosine and then proceed to be deaminated to uracil, and so it is ultimately read as T in sequencing experiments, conflating it with unmodified C. carC is also read as T, thereby conflating it with unmodified C, though the exact sulfonate intermediate has not been reported in this case.

Early efforts to sequence for [ox]mC bases specifically sought to map hmC by its CMS intermediate by using a CMS-specific antibody to immunoprecipitate hmC-enriched DNA after conventional bisulfite sequencing [Pastor et al. 2011]. Other antibody-based methods were also used to recover hmC enriched DNA fragments directly [Tan et al. 2013; Stroud et al. 2011; Williams et al. 2011] and gave early hints at hmC's genomic distribution. The results were very limited in their ability to detect rare sites of hmC or hmC near mC. Given these limitations, site-resolved sequencing strategies were needed.

The first efforts map hmC at single base resolution adapted single molecule sequencing methods to take advantage of the chemical availability of the hydroxyl group to react and form bulky adducts [Song et al. 2012] and labeled hmC specifically with azide-substituted glucose using β-glucosyltransferase from T4 bacteriophage and click chemistry to add a biotin tag at the 5 position [Song et al. 2011]. This labeling provides a handle for pull down experiments to enrich for hmC-modified DNA with slightly better specificity than afforded by antibodies, but the same density-dependent detection concerns as before. At the same time, this bulky product (β-6-azide-glucosyl- 5-hydroxymethyl-cytosine (N3-5-gmC)) reliably stalls DNA polymerase such that it provides useful kinetic signal in single molecule real time DNA sequencing studies, wherein a paused DNA polymerase is strongly correlated with sites of hmC [Song et al. 2012].

While this approach is not easily amenable to a genome-wide sequencing, the ability to selectively modify hmC at the hydroxyl group when embedded in DNA is fundamental to broader strategies to map hmC specifically on a larger scale. Treatment of genomic DNA fragments with T4 β-glucosyltransferase and UDP-glucose results in specific glucosylation of hmC nucleobases at their hydroxyl group. When the DNA is then treated with recombinant TET enzymes, the glucosylation protects hmC nucleobases from oxidation, but oxidizes mC and fC to carC. After classic bisulfite treatment and deamination, carC is read as T, but the original hmC will be read as C. This procedure, termed TET-assisted bisulfite sequencing

(TAB-seq), can be used in parallel with traditional bisulfite sequencing to map hmC and mC uniquely at base resolution, and relative quantification of the abundance of each [Yu et al. 2012a; Yu et al. 2012b].

At the same time, other groups reported similar ways to sequence for hmC uniquely by manipulating the oxidation status of mC and hmC chemically using potassium perruthenate ($KRuO_4$). This reagent selectively oxidizes hmC to fC but does not react with mC or other bases. When followed by traditional bisulfite treatment to deaminate fC to uracil, this technique, termed oxidative bisulfite sequencing, allows hmC and C to be read as T when done in parallel with traditional bisulfite sequencing without oxidative treatment [Booth et al. 2012]. Related methods have been developed for sequencing fC uniquely [Booth et al. 2014] by selectively reducing fC nucleobases to hmC with sodium borohydride ($NaBH_4$) followed by trational bisulfite sequencing (approach overall termed reduced bisulfite sequencing) or by fC-Seal selective chemical labeling of fC after reduction to hmC with $NaBH_4$ [Song et al. 2013]. Sites of fC and carC can be sequenced together via methyatlion-assisted bisulfite sequencing (MAB-seq) in which treatment with the S-adenosyl-methionine-dependent CpG methyltransferase M.SssI converts all unmodified C to mC. Following bisulfite treatment, original C, mC, and hmC will all be read as C while fC and carC will be read as T as a result of deamination to uracil [Neri et al. 2015]. This is currently the best available option to attempt to map carC broadly as other methods rely on a low efficiency chemical labeling with 1-ethyl-3-[3-dimethylaminopropyl]-carbodimide hydrochloride (EDC), followed by reaction with a primary amine-biotin tag and affinity capture, which are not amenable to a genome wide scale without substantial biases [Lu et al. 2013].

Finally, hmC and fC can be tracked in cells by isotopic labeling methods [Globisch et al. 2010; Bachman et al. 2015; Bachman et al. 2014]. Cells or animals fed $^{13}CD_3$-S-adenosyl methionine incorporate this molecule via DNMTs to first add an extra 4 daltons of molecular weight to mC nucleobases relative to unmodified mC. When these labeled mC molecules

oxidized to yield hmC or fC, a mass difference of 3 daltons remains relative to unmodified hmC. Mass differences of 2 and 1 dalton for fC and carC nucleobases, respectively, allow these bases to be quantitated, albeit at very low detection levels. These mass differences are detected after fragmentation, and tandem HPLC purification and mass spectrometry. The degree to which a given modification state will become labeled depends on the dynamics of both installation enzymes (DNMTs and TETs) and modification turnover, such that a stable modification would show little labeling in non-proliferating cells.

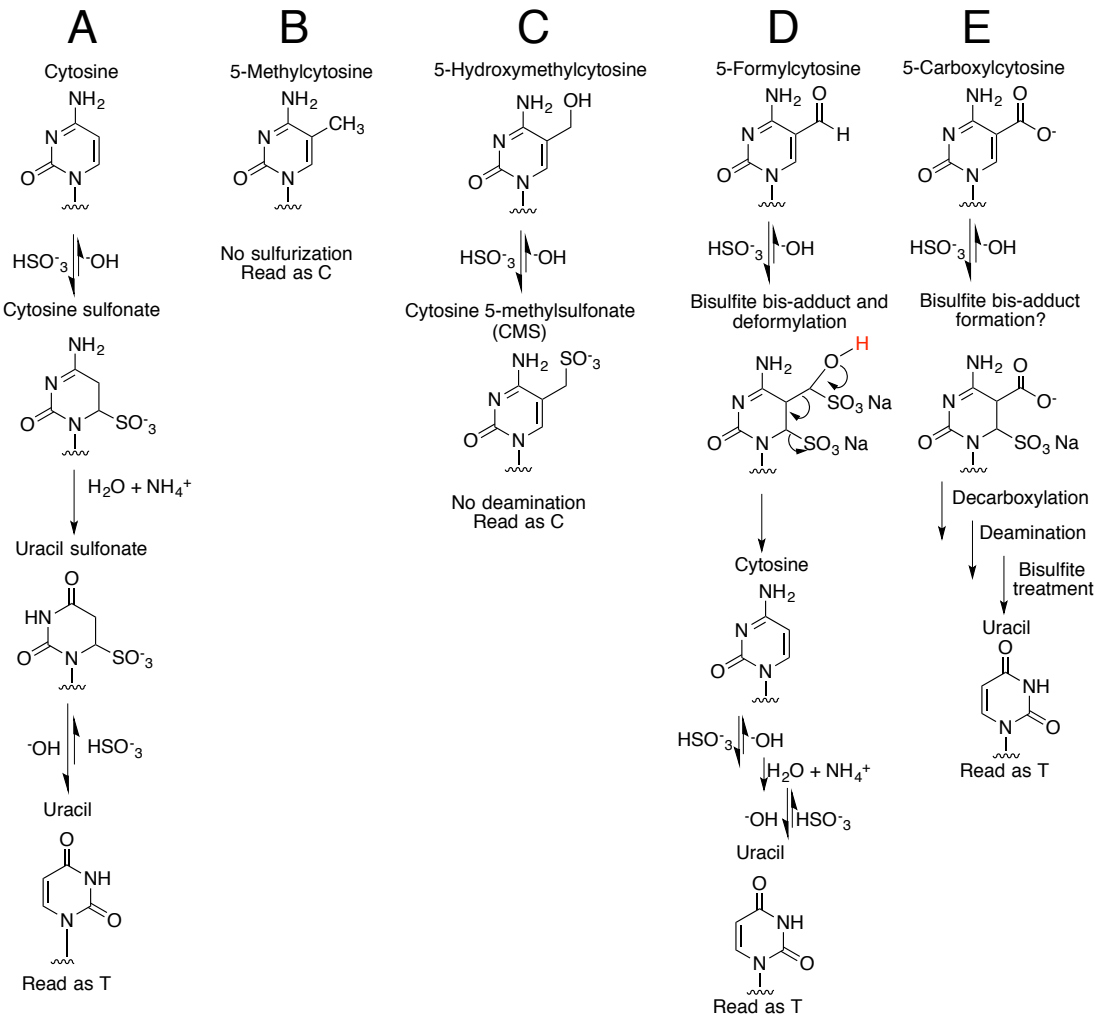


Figure 1.5: Bisulfite sequencing methods do not distinguish between methylcytosine and oxidized methylcytosine. (A) Bisulfite treatment converts unmodified cytosine to cytosine sulfonate, which can be deaminated to form uracil. During PCR amplification as part of the bisulfite sequencing workflow, this uracil is reverted to thymine to pair with adenine. When sequenced, the unmodified cytosines will be read as T. (B) mC is resistant to bisulfite treatment and will still be read as C. Comparison of bisulfite treated and untreated samples allows sites of mC to be mapped. (C) However, bilsufite treatment of hmC rapidly creates a bulky cytosine methylsulfonate (CMS) species that cannot be deaminated and will therefore be read as C in sequencing experiments. This CMS adduct can be prohibitive to PCR amplification, resulting in an underrepresentation of hmC sites in traditional bisulfite sequencing. (D and E) fC and carC are both sensitive to bisulfite treatment and converted to sulfonate adducts that are resolved through decarbonylation and decarboxylation under the acidic conditions of bisulfite treatment. Ultimately, this means that fC and carC will be deaminated to yield uracil, and be read as thymine when sequenced. (Summarized from work by: [Booth et al. 2012; Pastor et al. 2011; Huang et al. 2010; Booth et al. 2014])

1.4 Distribution and abundance of oxidized methylcytosine and phenotypes associated with TET enzymes

Collectively, these sequencing studies have been able to reveal several key features of the distribution and abundance of [ox]mC in mammalian genomes:

1.4.1 Oxidized methylcytosine species are rare, but most abundant in the brain and embryonic stem cells

Current quantitations indicate that the relative abundance of [ox]mC species in mouse embryonic stem cells are, as parts per million of total cytosine 1300 ppm for hmC, 20 for fC, and 5 ppm for carC, while mC is roughly 4% of cytosine [Ito et al. 2011; Globisch et al. 2010; Song, Yi, and He 2012]. Enrichment in the brain is greater for hmC, where 13% of CpGs are highly modified [Wen et al. 2014]. Enrichment of fC is age-dependent, with generally lower levels that decline with age in the brain, and at very low levels in other tissues (kidney, heart) [Bachman et al. 2015; Wagner et al. 2015].

1.4.2 Oxidized methylcytosine species and TET enzymes are enriched within particular genomic features and chromatin states.

In both human and mouse embryonic stem cells, hmC species are enriched at annotated enhancers elements and other distal regulatory elements including p300 and CTCF binding sites [Yu et al. 2012a; Booth et al. 2012]. hmC does not overlap these sites precisely, but is found in the sequence immediately adjacent to the motifs associated with these and other pluripotency transcription factors, including NANOG, OCT4, SOX2, and TCF4 [Yu et al. 2012a]. Several groups have noted the correlation between hmC modification sites, enhancers, and pluripotency transcription factor binding sites in mouse embryonic stem cells, and noted that mC is not observed at these sites [Stroud et al. 2011; Ficz et al. 2011].

hmC enrichment also overlaps with chromatin signatures of active enhancers: H3K4me1 and H3K27 acetylation [Stroud et al. 2011]. Perhaps the most exciting work to date on the involvement of TETs and hmC in the activation of cell type specific enhancers during a differentiation event is that of Serandour and colleagues [Sérandour et al. 2012], This work suggests that hmC and H3K27 acetylation and H3K4me2 are gained in concert at predicted transcription factor binding sites of cell type specific distal enhancers during neural differentiation of mouse embryonic stem cells.

hmC is also enriched to a lesser extent in promoter proximal regions and transcription start sites (TSS) of lowly expressed genes in mouse embryonic stem cells. While the relative abundance of hmC here is less than that of enhancer elements, it has a striking TSS-biased distribution across the gene body that is distinct from the profile of mC [Williams et al. 2011; Pastor et al. 2011]. This is consistent with the finding that the mC profile of TET1-bound transcription start sties in mouse embryonic stem cells is biased into the gene body relative to unbound transcription start sites, suggesting the accumulation of hmC close to the TSS. DNAseI hypersensitivity also is positively correlated with hmC and and negatively correlated with mC enrichment at these sites. This is indicative of a change in chromatin accessibility state at sites of hmC as compared to those with mC, and is further supported by evidence of histone modifications at sites of hmC and sites bound by TET enzymes. TET enzymes are found at regions of H3K4 methylation, but also within regions of coincident H3K27 and H3K4 methylation known as bivalent regions [Wu et al. 2011]. In the absence of TET1 in mouse embryonic stem cells, TET1-bound genes that decrease in their expression are largely bivalent, whereas genes that increase in their expression in the absence of TET1 exhibit only H3K4 methylation. Moreover, loss of TET1 also disrupts the occupancy of the polycomb group complex member EZH2, suggesting that the localization of this chromatin modifying complex may be dependent on TET1, [ox]mC species, or both [Wu et al. 2011]. Collectively, these results indicate that the role of TET1 activity can be context dependent, but that

its signature and the signature of hmC directly suggests that they function in concert with other chromatin modifications in particular patterns that are distinct from that of mC to impact gene expression.

The potential role of hmC within promoters and gene bodies is most striking in the brain, where hmC species are even more abundant and more positively correlated with gene expression [Szulwach et al. 2011; Jin et al. 2011b; Lister et al. 2013]. In the adult brain, similar to embryonic stem cells, hmC is enriched near the transcription start site while mC is found deeper within the gene body. This is particularly true of highly expressed genes in the brain, where hmC is distinctly more enriched on the sense strand of highly expressed genes in the mouse and human brain [Wen et al. 2014; Lister et al. 2013]. The same is not true of mC enriched genes, where the expression level is not correlated with strand-biased mC enrichment. When genes specific to several brain cell types are examined, the sense strand bias for hmC remains striking within each cell type, and a bias for mC on the antisense strand is observed. Finally, hmC is also enriched at the 5' splice site of exon-intron boundaries of alternatively spliced exons in the brain [Khare et al. 2012; Wen et al. 2014]. This enrichment is specific to neuronal tissues, but not neuronal-specific genes, however, mC is not otherwise found at this position in any cell type. The presence of hmC is significantly correlated with exon inclusion when compared to exon-intron structures with unmodified C at this position. As shown in fig. 1.6 on page 25. These results invite the possibility that hmC could regulate genes by directing alternative splicing decisions in a manner that has not been observed for mC at highly expressed genes

As described previously, two methods exist for quantifying the low abundance of fC. In a close technical examination of these fC sequencing methods, its appears that reduced bisulfite sequencing [Booth et al. 2014] may be slightly more sensitive than chemically assisted bisulfite sequencing [Song et al. 2013] in that it can detect rare sites where fC is found at comparable levels as mC and hmC. At such fC-enriched sites, the abundance of fC is inversely

correlated with that of hmC. One distinction between hmC and fC is the CpG island context in which each are captured. fC is more enriched immediately upstream of promoters of actively transcribed genes within dense CpG and of differentiating mouse embryonic stem cells, which are otherwise hypomethylated, while hmC enrichment is greater within non-CpG island promoters [Raiber et al. 2012]. This suggests that fC is an intermediate of epigenetic reprogramming at developmentally regulated CpG islands, a context in which hmC is not similarly implicated.

Deeper sequencing insight into the distribution of both fC and carC is possible through examination of stem cells lacking thymine deglycosylase (TDG), the base excision repair enzyme known to remove fC and carC and replace them with unmodified C. This serves to exaggerate the abundance of fC and carC, but enhance our ability to detect these low abundance marks and infer where they otherwise have a brief lifetime prior to removal.

In the absence of TDG, fC and carC are observed to accumulate at distal regulatory elements [Neri et al. 2015]. Whereas hmC is found at active enhancers, fC is found at poised enhancers demarcated with H3K4me1 but lacking H3K27 acetylation. These sites are otherwise low in mC and are specific to mouse embryonic stem cells (as compared to neural progenitor fC sites), suggesting a cell-type specific generation of these higher oxidation state derivatives. TDG-depletion-induced sites of fC and carC are also somewhat distinct from each other and are relatively well conserved as compared to flanking sequence. fC and carC tend to overlap with the binding sites of core pluripotency transcription factors, active enhancers, and sites of cohesin, p300, mediator, and CTCF-based promoter-enhancer looping interactions [Shen et al. 2013]. Like hmC sites, fC sites are similarly asymmetric (discussed later in more detail in section 2.5, *Key properties of the native oxidized methylcytosine specific binding activity*). Collectively, these results are consistent with the idea that the generation and turnover of higher [ox]mC species occurs in a cell type and gene specific manner at sites involved in the transcriptional regulation of stem cell identity.

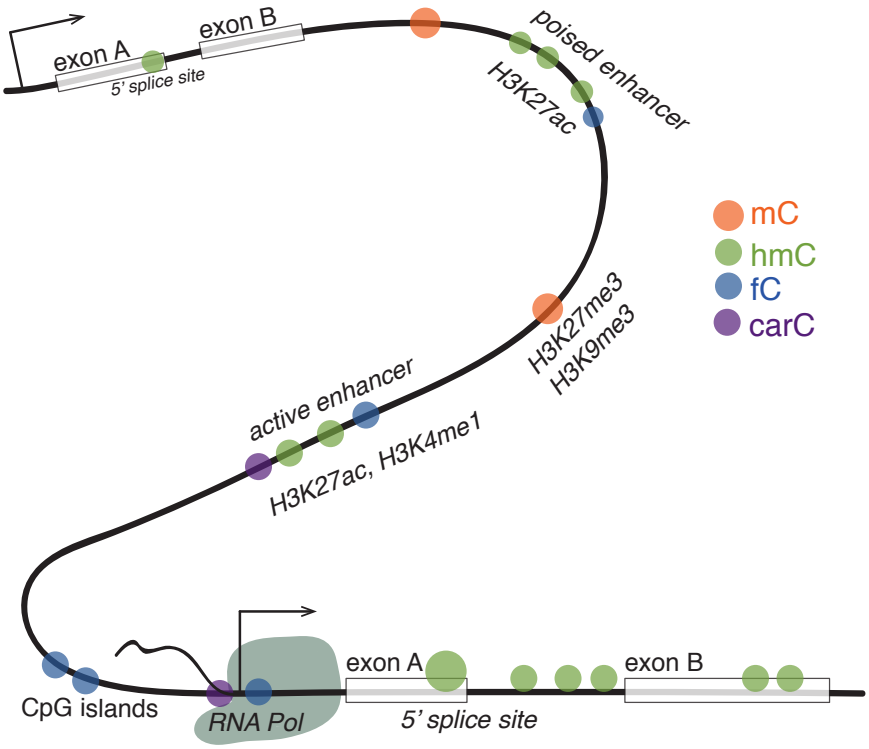


Figure 1.6: Oxidized methylcytosine species are enriched at particular genomic features and are distinct from the distribution of mC. [ox]mC species are abundant within active enhancers, active transcription start sites, pluripotency transcription factor binding sites in mouse embryonic stem cells, and highly expressed genes in mammalian brains. Enrichment at the 5' splice site of exon-intron boundaries, observed in the brain, and asymmetric coding sense strands, observed in stem cells and the brain, are noted. Relative enrichment at chromatin features such as histone modification states and CpG islands are also indicated. Adapted from [Wen et al. 2014; Neri et al. 2015; Wu et al. 2014; Song et al. 2013; Raiber et al. 2012; Yu et al. 2012a]

1.4.3 Isotopically labeled hmC is as stable as mC in cell culture and in mice.

For several years after the initial reports and preliminary sequencing studies of [ox]mC species in mammalian cells, their lifetime in cells and between cell divisions was unknown. It was shown that DNMT3a and 3b could install methyl groups on unmodified cytosines across from hmC with similar activity when presented opposite C or mC, however, DNMT1 greatly prefers mC/C over hmC/C substrates [Hashimoto et al. 2012] while TETs and TDG prefer to generate asymmetric sites ([ox]mC/C or [ox]mC/mC) [Wu et al. 2014]. Amidst the interplay

of [ox]mC generation, removal and ongoing maintenance methylation, it was not clear how stable or long-lived [ox]mC modifications could be.

Careful isotopic labeling experiments have helped resolve what actually happens to total [ox]mC abundance through multiple cell divisions. When cultured cells are fed $^{13}CD_3$-S-adenosyl methionine and the incorporation of this isotope into DNA quanitified, as described above, it was found that hmC is as stable as mC in proliferating cells, in that similar bulk measurements of each labeled species can be made up to ∼100 hours after labeling [Bachman et al. 2014]. This was found in cancer cell lines, undifferentiated mouse embryonic stem cells, and mouse embryonic stem cells induced to differentiate at the start of labeling. When adult mice are fed $^{13}CD_3$-S-adenosyl methionine, the degree of mC and hmC labeling is correlated across tissues. This observation of similar low labeling of both hmC and mC in slow dividing or non-proliferative tissues is consistent with a small amount of mC being incorporated, and an expected small amount of hmC being derived from that, but the two having similar lifetimes in the tissue. When fC is tracked by its isotopic signature, the degree to which fC accumulates in brain, heart, and liver over the developmental lifetime of mice can differ from that of hmC and mC [Bachman et al. 2015]. This suggests that fC has stability similar to mC and hmC in some tissues, but is lost, presumably by excision and replacement with C, at distinct rates in different tissues and as a function of mouse age. Cultured undifferentiated mouse embryonic stem cells exhibit similar fC isotopic labeling, but adult mice fed $^{13}CD_3$-S-adenosyl methionine are not found to incorporate this isotope into new fC within the brain at detectable levels. Collectively, these experiments suggest that hmC is a predominantly stable modification within the brain and stem cells with a lifetime similar to that of mC, while the stability of fC is similar in some tissues, but undergoes rapid turnover in others.

1.4.4 The distribution of oxidized methylcytosine species is distinct from that of mC even though they are derived from it.

In all of the sequencing experiments described above, it is worth noting that not only do the [ox]mC species have a genomic distribution that is enriched at regulatory elements of biological interest, their abundance at these sites is *independent of mC enrichment.* That is to say, a similar relative enrichment of mC at [ox]mC-enriched sites is *not* observed in sequencing experiments capable of distinguishing [ox]mC and mC species. [ox]mC species are of course derived from TET-mediated oxidation of mC, and no [ox]mC species are observed in the absence of all three TET enzymes. However, there appear to be many sites where short-lived mC is rapidly oxidized to [ox]mC species, either hmC or to a lesser extent fC, and then these modification states persist longer than their mC precursor at that site. Therefore, [ox]mC species have a distribution and lifetime at certain sites that is distinct from that of mC, suggesting that they are generated there not exclusively as part of an mC removal pathway, but for an equivalent [Bachman et al. 2014; Bachman et al. 2015] if not longer lifetime as an [ox]mC modified state. This is also consistent with the observation that TET1 and TET2 are more active in converting mC to hmC than performing the subsequent oxidation steps, thereby generating stable hmC that is not immediately processed to higher oxidation intermediates and removed [Hu et al. 2015].

It is also worth noting that while mC largely correlates with repressive histone modifications such as H3K9 and H3K27 methylation, [ox]mC species do not largely overlap with these marks [Wen et al. 2014; Pastor et al. 2011]. This indicates that mC and [ox]mC are also separated in terms of the chromatin contexts in which DNA methylation persists and those in which TET enzymes are active and [ox]mC modifications have some stable lifetime. One interesting exception to this is the observation of hmC at the edges of DNA methylation "canyons," regions of low DNA methylation that contain many developmentally regulated genes that are distinct from heavily methylated CpG islands that also bear canonically

repressive histone modifications of heterochromatin. At the edges of the hypomethylated canyon where it meets the hypermethylated island, hmC can be observed in close proximity to repressive histone modifications. This hmC is presumed to be involved in methylation turnover as the span of the neighboring methylated region is dynamically controlled by the action of the TET enzymes preventing the spread of mC into the canyon [Jeong et al. 2013].

1.4.5 hmC abundance in cancer contexts

Given the importance of [ox]mC species for both removal of mC and as modifications with their own distribution at enhancers and promoters, there is clearly potential for aberrant TET activity or [ox]mC levels to contribute to cancer development and progression. Genome-wide loss of hmC is noted in some cancer types, [Liu et al. 2013; Lian et al. 2012; Jin et al. 2011a] where the loss stems from down regulation of both TET enzymes and the enzymes that process TET co-factors (specifically isocitrate dehydrogenase [Figueroa et al. 2010], which produces α-ketoglutarate). In melanocytes, loss of hmC is a marker of early melanoma, and is correlated with its aggressiveness [Lian et al. 2012]. This loss of hmC is observed within the promoters and exons of genes associated with melanoma progression, and excessive mC is found in its place. Exogenous overexpression of TET2 in melanoma cell lines is correlated with reduced tumor aggressiveness and rescue of normal hmC and mC levels. The precise role of hmC in this case remains to be clarified as its loss is associated with both gene activation and repression. In this and many other solid tumor cancers, down regulation of the TETs is observed as a function of tumor aggressiveness.

However, over abundance of TET1 can contribute directly to leukemia progression. MLL fusion proteins present in MLL-rearranged leukemias promote the overexpression of TET1, the level of which is correlated with the progression of those leukemias [Huang et al. 2013]. In bone marrow transformation assays using MLL-fusions, downregulation of TET1 via shRNAs led to reduction in tumor aggressiveness whereas overexpression of TET1 exacerbates it. In

these leukemias, TET1 and MLL-AF9 fusions act together to upregulate classic oncogenic targets of MLL. This behavior is distinct to TET1 and to leukemia, suggesting a unique role for TET1 in gene regulation.

1.4.6 Developmental and neuronal phenotypes associated with TET enzymes

The loss of TET enzymes is associated with several interesting developmental phenotypes in embryonic stem cells, induced pluripotent (iPS) cells, and the brain. Loss of all three TET enzymes does not disrupt pluripotency of mouse embryonic stem cells, however, it does compromise normal differentiation of embryoid bodies with reduced expression of mesodermal and endodermal markers [Dawlaty et al. 2014]. Due to the lack of these tissues, TET-triple knockout embryoid bodies cannot contribute to teratomas or chimeric embryos. Abnormal promoter hypermethylation is observed that is correlated with misregulation of developmentally regulated genes. Collectively, these results indicate that TETs are required for proper differentiation of mammalian embryos. The loss of all three TET enzymes is required to see these full effects, suggesting that the TET enzymes can function with some redundancy in the absence of each other [Dawlaty et al. 2013; Dawlaty et al. 2011]. Similarly, loss of all three enzymes and TDG are prohibitive to iPS cell reprogramming, as methylation marks on cell type specific genes are not efficiently lost [Hu et al. 2014]. TET enzymes and an hmC-intermediate state are also important for imprint erasure during designation of primordial germ cells during embryonic development [Hackett et al. 2013], but from there, passive dilution predominates over active removal by TDG [Seisenberger, Peat, and Reik 2013; Kagiwada et al. 2013]. During neuronal development, TET2 and TET3 are needed for proper differentiation and hmC is as observed to stably accumulate the promoter regions of activated neuron-specific genes [Hahn et al. 2013]. Loss of TET1 has greater effects in the adult brain, where it negatively impacts neurogenesis, proper learning [Zhang et al. 2013], and memory, specifically disrupting long term depression ("unlearning," as opposed to long

term potentiation associated learning) and memory extinction of fear-based memories that are no longer reinforced by the environment (a phenotype not unlike post-traumatic stress disorder) [Rudenko et al. 2013; Kaas et al. 2013]. Several neurodegenerative diseases are associated with increased levels of TETs and hmC, including Alzheimers and ALS [Al Mahdawi, Virmouni, and Pook 2014].

1.5 Biological advantages and challenges to specific recognition of [ox]mC species for chromatin signaling

Sites of [ox]mC species are observed to demarcate enhancers, active promoters, binding sites for key chromatin regulators, and other sites undergoing cell-type specific regulation of gene expression. Moreover, [ox]mC species are found in a tissue-specific manner at sites where mC is not otherwise abundant, suggesting that they are generated for a distinct lifetime independent of mC. Given these unique signatures of [ox]mC, there are potentially several biological advantages to specific recognition of [ox]mC species as the basis of a chromatin based signaling pathway to regulate gene expression.

First, specific recognition of [ox[mC] species provides the cell with another layer or information about the state of genomic region. In this case, the presence of [ox]mC could indicate either a previously methylated locus undergoing demethylation, or demarcate independently of mC the location of a cell type specific enhancer, promoter, highly expressed gene, or alternatively spliced exon. Second, presence of [ox]mC could confer this information via specific binding partners that are exclusive to [ox]mC that recruit or are otherwise in complex with additional chromatin regulators, such as nucleosome remodeling, histone modifying enzymes, or the basal transcriptional machinery. In this way, [ox]mC species could direct the chromatin state or transcriptional activity of a genomic region. Third, and by extension, the existence of [ox]mC specific binding activities that recruit or are present in other histone modifying or modification-specific binding complexes could serve to couple readout

of DNA and histone modification states. Combinations of histone and DNA modifications could therefore allow more nuanced chromatin substates to be staged. This could in turn enable multivalent chromatin interactions to confer highly discerning binding patterns to achieve exceedingly targeted regulatory effects.

The regulatory potential of specific [ox]mC recognition is highly analogous to the signaling pathways already described for mC-specific binding proteins which achieve mC-specific chromatin regulatory effects via their localization to mC by MBD and SRA domains and their recruitment of other activities to these sites. Several research groups have noted this possibility and pursued such binding proteins through both candidate gene approaches and large scale proteomics screens for [ox]mC binding activities.

However, with the immense regulatory potential for [ox]mC comes the tremendous challenge of achieving specificity for these rare and subtle DNA modification states. Importantly, if [ox]mC species are to confer regulatory information distinct from that of mC, as they are poised to do based on their distinct distribution, any [ox]mC specific binding domain would need to be *very* specific for [ox]mC modifications, and discriminate highly between mC and [ox]mC states. This degree of specificity is essential in order to ensure that signaling that might emanate from relatively rare [ox]mC modifications is not conflated with signaling from other DNA binding activities derived from more abundant DNA states.

The subtlety of [ox]mC modifications as compared to mC, for example, is only the first issue in envisioning an [ox]mC specific interface. An [ox]mC specific binding protein needs to achieve a high degree of binding energy discrimination from slight biochemical differences in its substrate, and capitalize on that difference enough to account for the scarcity these rare states among more abundant C and mC. Critically, the relative specificity of any putative [ox]mC specific binding protein must be considered in light of the abundance of these modifications. If the degree of discrimination by a given binding protein between [ox]mC modified states and other DNA substrates is not commensurate with the relative abundance

of [ox]mC as compared to mC or C, for example, then this protein is not likely to be a specific binding protein in the cell. Even if some preference for [ox]mC is inherent in a given protein in an *in vitro* binding assay, if the fold difference in binding affinity between [ox]mC and the far more abundant C or mC is not similar to the fold difference in abundance of these modification states, then the likely binding occupancy of this protein at sites of [ox]mC in the cell is expected to be very low as compared to its occupancy at more abundance C and mC sites. Together, the chemical subtlety and low abundance of [ox]mC makes it very biochemically challenging to bind these modifications specifically and with a biologically meaningful probability in the cell. While it is certainly tempting to speculate that [ox[mC specific binding proteins exist, all searches for these proteins must be scrutinized with these biochemical realities of specific binding in mind. Ahead of demonstrating such specificity, it cannot be concluded that biologically meaningful [ox]mC specific binding proteins exist.

1.5.1 Overview of candidate gene and large scale proteomic searches for oxidized methylcytosine DNA binding proteins

Candidate gene approaches

Despite this challenge, binding proteins specific for [ox]mC modifications have been reported in the literature at a steady rate since the discovery of the TET enzymes and [ox]mC species in mammalian DNA. Many of these studies have reexamined known mC binding proteins and report some cross-specificity for hmC [Yildirim et al. 2011; Frauer et al. 2011; Mellén et al. 2012]. Each of these reports have included some*in vitro* binding assays that seek to demonstrate specificity of the protein of interest (MBD3, UHRF1, or MeCP2) for hmC modified DNA as compared to C or mC.

In each case, while some binding for hmC is observed, equivalent binding is observed for either C or mC. This means that while hmC is tolerated by theses proteins and (in isolation) bound equivalently to their described substrates, in the context of the cell, hmC will not be

bound at a high probability because it is far lower in abundance than C or mC.

Other published studies have discerned these binding affinities precisely using quantitative affinity assays for symmetric C, mC, hmC, and cross-annealed substrates. These measurements show that when the binding affinity appears qualitatively equivalent in other quantitative assays, it is not. The binding propensity of MeCP2 for hmC was examined even before the TET enzymes were described because hmC was a known form of oxidative damage, and ruled out by electrophoretic mobility assays as a specific substrate [Valinluck et al. 2004]. The fluorescence polarization binding studies reported by Hashimoto and colleagues [Hashimoto et al. 2012] for many MBD and SRA containing proteins are an incredibly useful contribution to the field that seems to have been overlooked by many pursuing the biology of hmC, and even though it partially redoubles on previous efforts [Jin, Kadam, and Pfeifer 2010]. Their examination of these described symmetric mC and mC/C binders again shows that these proteins have some *in vitro* capacity to bind hmC in various presentations (symmetric and asymmetric with C or mC). However, in all cases, the MBD and SRA domains have the greatest affinity, as described by their lowest saturation K_d, for their canonical substrates (symmetric mC and mC/C, respectively). The MBDs are observed to bind hmC/mC with the second lowest K_d, but between two fold and ten fold lower affinity than symmetric mC. The affinity for symmetric hmC is between three and one hundred fold lower depending on the MBD in question. The SRA domain containing protein UHRF1 binds mC/C containing DNA with an order of magnitude greater affinity than any other combination of C, mC, or hmC. As outlined before, in light of the relative abundance of C, mC, and hmC in the cell, these measurements rule out biologically meaningful binding occupancy by the MBD or SRA domain-containing proteins examined. Most recently, it has also been proposed that the CXXC domain of TET3 exhibits a binding preference for carC, [Jin et al. 2016], as does the C_2H_2 zinc finger Wilm's tumor protein (WT1) [Hashimoto et al. 2014]. Here again, the binding discrimination observed is not at all sufficient to afford specific localization to carC

given its exceedingly low abundance.

At the same time, these studies of canonical mC binding proteins do offer some insight into whether one role of [ox]mC at sites where the genomic distribution of mC and hmC overlaps is to intermittently tune down the binding of mC binding proteins. The binding profiles are consistent with this possibility, as the presence of hmC in CpG dinucleotides opposite mC disrupts the binding of MBD and SRA domain containing proteins by at least two fold. There are other interesting genetic observations among these candidate binding protein studies that should be followed up on. In the case of MeCP2, it is noted that mutations of MeCP2 associated with the neurodegenerative disease Rett syndrome have different binding activity toward hmC and modifications of cytosine within non-CpG contexts present in the brain [Guo et al. 2014; Mellén et al. 2012]. Secondly, in the case of MBD3 [Yildirim et al. 2011] there is also the curious observation that when TET1 levels are knocked down, the of MBD3 to transcription start sites in mouse embryonic stem cells is dramatically disrupted. While this observation by no means can speak to any direct TET1-MBD3 interaction or role of [ox]mC as determinant of MBD3 binding, it is nonetheless a curious connection between the two proteins that should be further investigated as part of the biology of MBD3.

Proteomics approaches

Two studies have taken broad proteomics based approaches to identify [ox]mC-specific binding activities. These studies [Iurlaro et al. 2013; Spruijt et al. 2013] each utilize DNA pull downs and quantitative mass spectrometry from mouse embryonic stem cell nuclear extracts, a cell type known to be enriched in [ox]mC, and therefore presumed to be enriched in [ox]mC specific activities, should they exist. In the studies of Spruijt and colleagues, the mouse embryonic stem cells were cultured in enriched media for SILAC quantitation, and pull down and mass spectrometry inputs also included neural progenitor cells, and adult brain. The DNA probes used in each study differ in that Spruijt et al. used a single twenty-eight basepair sequence symmetrically modified with C, mC, hmC or fC at four consecutive CpG

dinucleotides, while Iurlaro et al. used two ∼250 basepair promoter regions amplified by PCR containing either dCTP, dmCTP, dhmCTP or dfCTP.

The two studies do not overlap in their findings of hmC- and fC-specific binding proteins in mouse embryonic stem cells, but report many intriguing proteins. Their hits including previously described proteins involved in DNA repair, and chromatin regulator complexes, often capturing multiple complex members. Both studies also identify many uncharacterized proteins, and endeavored to validate the specificity of a few. Again, with two exceptions, these binding studies do not reveal binding affinities that are significantly different between C, mC, and [ox]mC modified DNA to confer biologically meaningful specificity for [ox]mC given its low abundance. The two exceptions may be MPG and UHRF2, but both of these proteins are more likely involved with the turnover of [ox]mC rather than a distinct signaling pathway emanating from these modifications. Moreover, cold competitor is not used in these studies to challenge their affinity, so the degree of specificity cannot be determined. As such, these proteomics studies did not conclusively demonstrate that [ox]mC specific binding proteins exist and that they might support a unique signaling pathway. However, I will return to these studies in the Discussion as they informed my own work using a different proteomics approach to identify [ox]mC specific binding proteins, and I believe there are both similarities and critical differences between our studies.

1.5.2 Motivation for more rigorous biochemical searches for highly specific oxidized methylcytosine binding activities

The importance of specificity in binding for biologically relevant function of [ox]mC cannot be overemphasized: given the extremely low abundance of [ox]mC relative to mC and unmodified C, any factor that binds [ox]mC with a biologically meaningful probability must exhibit an affinity that is at minimum commensurate with the relative abundance of [ox]mC. None of the candidate binding proteins reported to date have provided validation of a spe-

cific binding affinity capable of conferring unique localization to the rare sites [ox]mC in the cell in the presence of far more abundant C and mC. However, the regulatory potential of such a binding event remains enticing and worth pursuing with more biochemical rigor in order to detect more specific binding proteins. To improve upon these searches based on the nature of the binding event I believe must occur, I sought an approach that would screen for novel binding activities based on their specificity to the degree I believe must be met for biologically meaningful binding.

Chapter 2

BIOCHEMICAL FRACTIONATION OF THE MAMMALIAN BRAIN TO ISOLATE HIGHLY SPECIFIC [OX]MC BINDING ACTIVITIES

2.1 The principle of extract fractionation for discovery of new biochemical activities

In contrast to direct DNA pulldowns, another approach to identify new biological activities in cell types of interest is to assay for the activity in extract from that cell type. If the activity is present, the protein components of that extract can then be fractionated and the assay repeated to determine how the activity responds to fractionation, and narrow in on the minimal extract components responsible for the activity. This strategy, classically referred to as biochemical fractionation, allows novel activities to be detected and characterized in native extracts. As compared to candidate gene approaches, biochemical fractionation is relatively free of bias and allows the experimenter to detect activities that may be due to proteins never previously implicated in DNA binding. In essence, biochemical fractionation is only limited by the relative enrichment of activity in the extract chosen for fractionation, the sensitivity of the activity assay used to screen and characterize native activities, and the ability of fractionation steps to isolate the activity from all other components of the extract.

Biochemical fractionation has historically proven useful for delineating key pathways in nucleic acid biochemistry, including transcription [Geiduschek, Nakamoto, and Weiss 1961] and the first reports of mC specific binding proteins [Meehan et al. 1989]. In each case, the advantage of biochemical fractionation for characterizing the proteins responsible for novel activities is that the approach scrutinizes the activity foremost. The activity assay can be designed such that only activities that are specific enough (to a modification state,

a template, or under particular conditions) will be revealed by the assay. Other activities that are not sufficiently specific or robust under the experimenter's imposed conditions will not be detected or distract from other activities of interest. For these reasons, biochemical fractionation is a powerful approach for detecting new DNA binding conditions that are highly specific to a given modification state even when that modification is present at very low abundance, such as the case of [ox]mC. This degree of scarcity can be recapitulated in an electrophoretic mobility shift activity assay using five- to twenty-fold molar excess of C or mC cold competitor, and thereby mimicking the cellular context of low [ox]mC abundance and searching for activities that can operate in that context. Fractionation is well-suited to search for novel binding activities with a high degree of biochemical stringency, and (in light of many reports that overlook this requirement for specificity in the cell) is capable of resolving whether biologically meaningful, highly [ox]mC-specific binding proteins actually exist.

This approach is not without limitations. It requires the activity of interest to function under the fractionation and activity assay conditions that the experimenter has explored. It is limited by the ability of the fractionation procedure to separate the activity of interest from non-specific activities without massively diluting the activity such that it cannot be detected in the activity assay, or such that it overlaps with a native protease or nuclease that precludes detection of binding. Biochemical fractionation is also ultimately a proteomics approach using an input highly enriched in the activity of interest, and a major hurdle to activity discovery is confident mass spectrometry identification of proteins in highly fractionated (low total protein concentration) extract. Finally, biochemical fractionation also relies upon validation of direct binding and specificity with recombinantly expressed and purified protein. In the course of identifying [ox]mC-specific binding proteins, both mass spectrometry and the ability to isolate active recombinant protein were formidable challenges to my use of biochemical fractionation. However, the clear specificity of the activity in native extract in

the course of fractionation lends confidence to the pursuit of high quality mass spectrometry results, recombinant validation, and ultimately, characterization of the function of these specific activities in cells.

The following is an overview of a fractionation procedure of mammalian brain to isolate highly [ox]mC specific activities, and the properties of the native activity revealed by electrophoretic mobility shift assays in the course of fractionation. The mammalian brain, specifically the cortex of pigs, was chosen as a starting material because of its known enrichment in [ox]mC species, and therefore its presumed enrichment in [ox]mC specific binding activities, should they exist.

2.2 Nuclear extract preparation from cortex

2.2.1 Preparation of hemispheres for tissue processing

Whole brains were taken from recently sacrificed adult pigs of indeterminate agricultural breeds. The whole brains were briefly washed in PBS and then coarsely dissected to remove the brain stem and cerebellum from the cerebrum for separate processing. The meninges and external vasculature were removed before finely mincing the cerebral cortex. This tissue was suspended in 200 g batches with 2 L of homogenization buffer (20 mM Na·HEPESpH 7.9, 30 mM KCl, 1 mM EDTA, 1 M sucrose (34% w/v), 10% glycerol (v/v)). This suspension was then processed twice through a Yamato LH-21 continuous flow homogenizer operating at 150-180 RPM at a flow rate of 20 mL per minute. The cell homogenate was centrifuged in 500 mL bottles to pellet the intact nuclei (Sorvall RC5B with Fiberlite F10-6 x 500y rotor at 16,000xg for 20 minutes, 4° C). After decanting the supernatant, and sloughing off excess lipid residue with paper towels, recovered crude nuclei were resuspended by pipette in 200 mL of buffer per 500 mL of lysate pellet in a hypotonic reduced sucrose buffer (20 mM Na-HEPES pH 7.9, 10 mM KCl, 1 mM EDTA, 340 mM sucrose (10% w/v), 10%

glycerol (v/v)) and loaded onto a 30% (w/v) sucrose cushion (30% sucrose variation of the same hypotonic sucrose buffer, 100 mL of cushion in 750 mL Bio-Bottle, Thermo Scientific). Gentle centrifugation of the intact nuclei through the cushion (Sorvall Legend XTR with TX-750 rotor at 600 x g for 10 minutes, 4° C) serves to wash the nuclei and remove lipid carry-over from the cell lysis. The purified pelleted nuclei were resuspended in three packed nuclear pellet volumes of the hypotonic reduced sucrose buffer supplemented with 1 mM PMSF and 5 mM β-mercaptoethanol, and flash frozen for storage at -80° C.

2.2.2 Preparation of nuclear extract from nuclei pellets

Nuclear extract was prepared as previously described with some modifications. Briefly, nuclear pellets in 50 mL conical tubes were thawed on ice, gently resuspended, and supplemented with 200x Protease Inhibitor Cocktail in DMSO (concentrations at 1x: 1mM AEBSF, 0.8 μM aprotinin, 20 μM leupeptin, 15 μM pepstatin A, 40 μM bestatin, 15 μM E-64). Saturated ammonium sulfate stock was added drop-wise to 400 mM final concentration and the mixture was incubated with gentle rocking at 4° C for 30 minutes. The nuclear extract was then clarified by ultracentrifugation (Beckman Ti70.1 rotor at 37,000 rpm for 1.5 hours, 4° C). The soluble nuclear extract was filtered through a 0.45 μm mixed cellulose syringe filter (Millipore) prior to dilution with 50 mM Bis-Tris pH 6.2, 1 mM EDTA, 10% glycerol (v/v), 5 mM β-mercaptoethanol (BTEG0) to lower the salt in preparation for ion exchange chromatography.

2.3 The competitive electrophoretic mobility shift assay (EMSA)

2.3.1 Standard EMSA conditions

Electrophoretic mobility shift assays (EMSAs) are a classic biochemistry method for visualizing nucleic acid binding interactions. An EMSA involves incubating a putative nucleic acid

binding protein or extract containing such proteins with a labeled nucleic acid of interest, and subjecting this mixture to native electrophoresis in a gel. The nucleic acid, if not bound, will migrate faster than most protein:nucleic acid complexes. However, if these complexes are present, they will be detected as a 'shift' in the mobility of the nucleic acid to a lower mobility. The stability of these protein:nucleic acid complexes to various binding challenges can also be visualized by EMSA. To do so, additional unlabeled nucleic acid can be provided as a 'cold competitor' to binding of the labeled nucleic acid. The cold competitor, provided in excess to the labeled species, can act as a non-specific competitor to binding in that it challenges the relative binding preference of the extract proteins for the labeled nucleic acid as compared to the unlabeled cold competitor. If binding to the cold competitor is preferred, this will diminish the shift observed. The strength of shift as a function of cold competitor therefore reveals important properties of the specific binding events observed, namely, their rank preference between different labeled substrates and unlabeled cold competitors. If a labeled species is vastly preferred to the unlabeled non-specific cold competitor, the shift observed may be enhanced by the presence of this off-target cold competitor. Alternatively, the cold competitor may be a specific competitor in that it specifically disrupts binding to the labeled nucleic acid. This is observed as a loss of shift as a function of specific cold competitor concentration. Both types of competitor are useful for determining the specificity properties of the protein or extract in question, and are critical for discerning the degree of specificity I believe must be inherent in any biologically meaningful [ox]mC specific binding activity.

Routine EMSAs were performed as follows. 10μL binding reactions were set up using: 2 μL 5x binding buffer (where the final 1x concentration is 20 mM Na·HEPESpH 7.8, 10 mM $(NH_4)_2SO_4$, 30 mM KCl, 1 mM EDTA, 0.2% Tween 20, supplemented with 0.1 mM PMSF, 5 mM β-mercaptoethanol, and including cold competitor as needed), 1-6 μL chromatographic fraction (remaining volume is current column input buffer), and 1 μL radiolabeled DNA,

added last. Typical concentrations of radiolabeled DNA are 50-200 nM depending on the stage of the prep and the strength of the activity. Cold competitor supplied in the 5x reaction buffer is therefore typically 1.25 μM to 5μM for a 5-fold molar excess depending on the amount of radiolabeled DNA used. The oligonucleotides used are all the same sequence of a 16 basepair duplex DNA containing a central CpG dinucleotide that is modified accordingly. Sequence and synthesis details are included in Appendix A.

The binding reactions are incubated at room temperature for 20 minutes, then 1 μL of loading buffer is added (20 mM Tris-HCl, pH 7.5, 20% glycerol (v/v), 0.02% bromophenol blue, 0.02% orange G). The binding reactions (4-10 μL) are then loaded into a 19:1 polyacrylamide gel pre-equilibrated and running at 4° C in 0.25x TBE for 30-60 minutes at 8 mA. The percentage of the gel varied between 4-7% depending on the mobility of the activity and the resolution desired. The volume of chromatographic fraction used was adjusted to allow for optimal gel loading (minimal well shift) and sensitivity (resolution of shifts with discrete mobility). The loaded gel was run at a maximum current of 20 mA (typically 225 volts during initial loading) until the free DNA, with a mobility similar to that of orange G, was within 2cm of the bottom of the gel (approximately 2 hours). Each gel was dried on Whatman blotting papers (Grade 3MM Chr) on a Bio-Rad gel drier before overnight exposure (12-18 hours) to a 'CR'phosphorimaging plate (Fujifilm), and then scanned on a Typhoon 9200 (GE Healthcare) at 100 nm resolution.

An example chromatogram of extract fractionation by protein chromatography is shown in fig. 2.1. The resulting fractions are screened by EMSA to detect hmC-specific binding activities in the presence of a five-fold molar excess unmodified CpG. Multiple shifts can be observed that require further characterization of individual fractions under different EMSA conditions to resolve. When fractions with similar specificity properties are identified, they are pooled for further fractionation.

2.3.2 Overview of column chromatography fractionation of nuclear extract

The chromatography steps that follow represent a highly refined and reproducible path to isolating the major [ox]mC-binding activity observed. This path was determined through much trial and error as many column types, column sequences, loading and buffer conditions were examined. Each fractionation step was optimized by experimenting with the column identity, volume, input, buffer, and gradient conditions to afford the greatest amount of purification and enrichment of [ox]mC specific activity in as few and distinct fractions as possible. After each column, EMSAs were used to assess the DNA binding activity and interrogate its properties through variation of the labeled oligonucleotide, as well as cold competitor ratio and identity. These assays reveal the properties of the fractions and allow distinctions in specificity between native activities to be visualized. Only fractions that exhibit specific shift for [ox]mC over mC and C cold competitors were then pooled and subjected to further chromatographic fractionation.

Herein lies one of the most crucial advantages of biochemical fractionation, which is that it enriches the fractions for the activity of interest while also discarding (leaving behind after fractionation) extract components that are not active to the degree desired. In the pursuit of highly specific native activities, it is equally important to selectively move forward material as it is to selectively omit material from the fractionated extract. The series of different chromatographic steps allows different chromatographic chemistry to partition the components of the extract into different 'bins' of fractions which can be scrutinized for activity in the presence of different background of extract components. In this way, the chromatographic steps act as distinct sieves or filters through which active proteins can be scrutinized and enriched, and non-specific proteins can be excluded.

The final, highly refined and reproducible column chromatography path to isolating [ox]mC-specific binding activities is shown in fig. 2.2. This sequence of five columns purifies a final activity that is highly specific for hmC and fC containing DNA in the presence

of excesses of C and mC cold competitor DNA that are on the order of that present in the cell.

The exact running conditions, namely the volume of the input and the volume of the gradient segments, may be proportionally modified from prep to prep in order to account for changes in the scale of the prep, new properties as visualized by EMSAs in boundary fractions, and the amount of active material carried forward from the previous step. The fractionation scheme outlined here represents the composite of many successful fractionation attempts and performs well for an input of approximately four porcine cerebrums (400 grams total cortex tissue, or 30 mL packed nuclei volumes).

From point of nuclear lysis onward, freezing the partially purified activity significantly diminished the strength and specificity of the ox-mC specific activity. As such, extract and fractions were never frozen and purification to mass spectrometry inputs was performed as rapidly as possible, typically within ten to fourteen days after nuclear lysis.

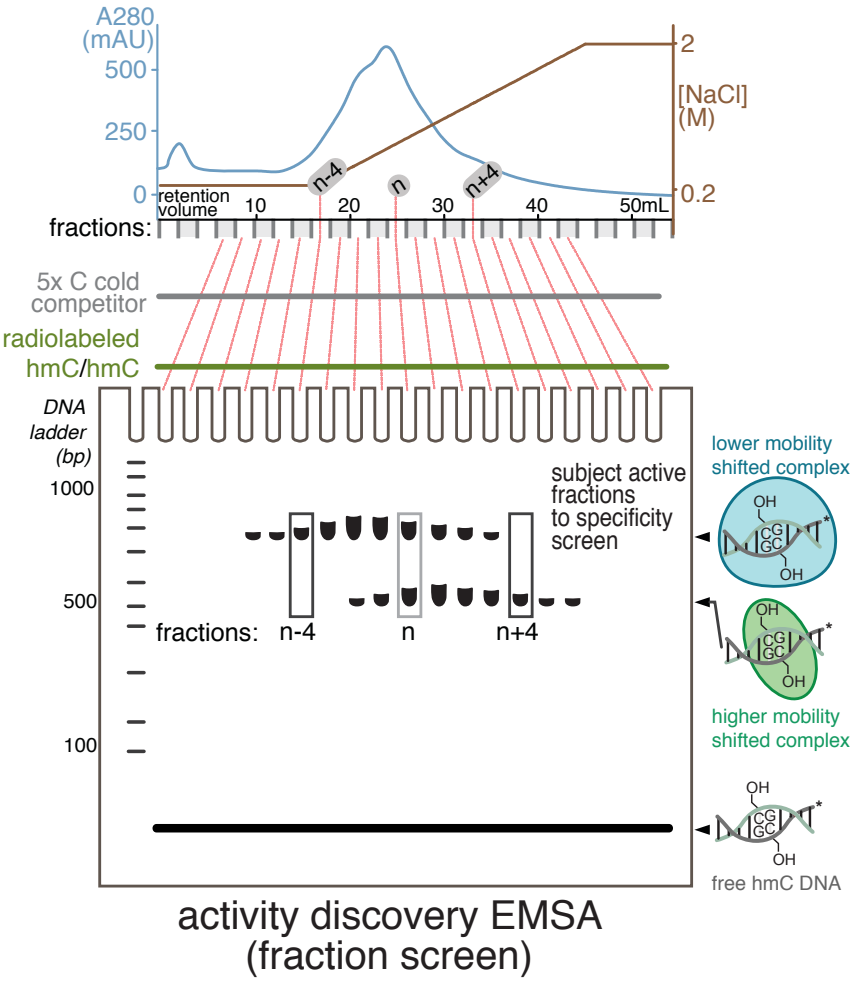


Figure 2.1: Fractions generated by column chromatography are screened in this example "activity discovery" gel shift. Shifts at two distinct electrophoretic mobilities are observed above the unbound free DNA. The two shifts observed could be due to two distinct hmC-specific proteins that have different chromatographic and native gel mobility properties. Alternatively, it could be that the two shifts represent one protein complex that the chromatography is partially separating into distinct specific complexes. Additional screening of individual fractions under different binding conditions (comparing the properties of n, n-4, and n+4 in the presence of different identities and molar ratios of cold competitor) are needed to discern the specificity properties of each fraction, and determine which fractions are similar in their [ox]mC-specific activity and should be pooled for further fractionation.

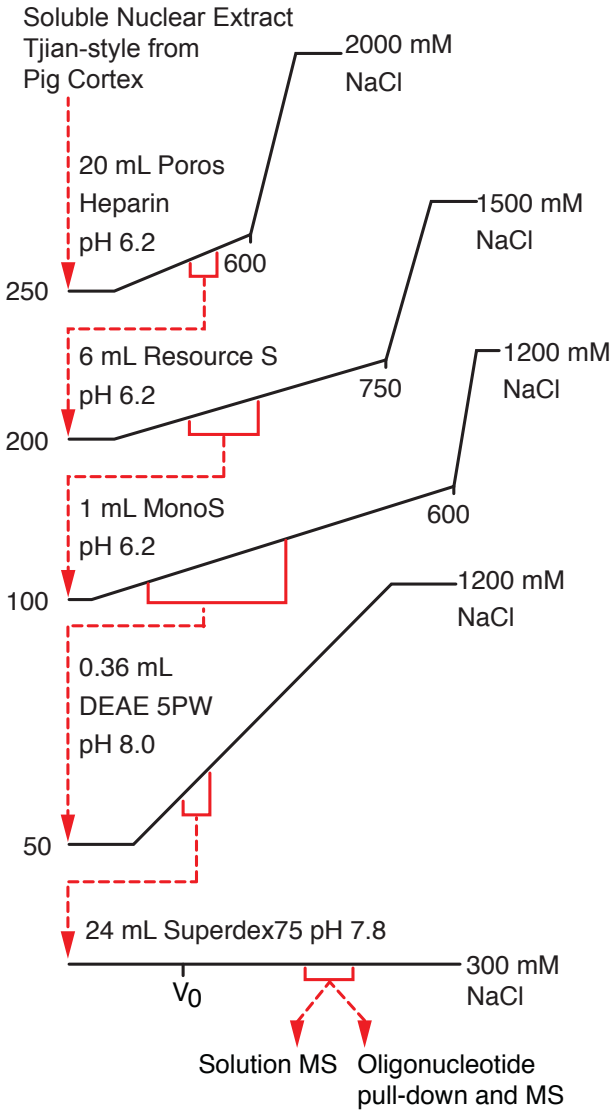


Figure 2.2: Optimized chromatography scheme for the isolation of the major [ox]mC-specific DNA binding protein activities from nuclear extract of porcine cerebrum. Shallow salt gradients over the various forms of ion exchange chromatography used here reproducibly elute the [ox]mC-specific activity in a small number of fractions that are bracketed along each column gradient. The movement of these fractions forward in the purification and onto the next column is indicated. After the final size exclusion chromatography step, the completely fractionated extract either proceeded directly to mass spectrometry of the solution fractions, or were subjected to a final affinity purification step using a modified DNA pulldown, followed by mass spectrometry examination of specifically recovered protein gel bands.

2.4 Chromatography approach for isolation of oxidized methylcytosine-specific binding activities from porcine brain

2.4.1 The 20 mL POROS Heparin isolates a [ox]mC-specific activity from bulk DNA binding activities

To load the 20 mL bed volume POROS Heparin (Applied Biosystems resin packed in XK-16 column, GE Healthcare), the salt concentration and pH of the soluble nuclear extract described above was first lowered by dilution in BTEG0 (50 mM Bis-Tris pH 6.2, 1 mM EDTA, 10% glycerol (v/v), 5 mM β-mercaptoethanol) until the extract matched the pH and conductivity of BTEG250 (50 mM Bis-Tris pH 6.2, 1 mM EDTA, 250 mM NaCl, 10% glycerol (v/v), 5 mM β-mercaptoethanol). After two column volumes of wash in BTEG250 on an Akta Purifier (GE Healthcare), the column was developed first with a linear gradient to 600 mM NaCl over four column volumes, then by a linear gradient to 2M NaCl over one column volume, and held at this salt for another column volume. 5 mL fractions were collected during the shallow gradient and the [ox]mC-binding activity in each was examined by a gel shift.

For this and other columns, the initial EMSA screened all column fractions, washes and input for preferential binding activity for radiolabeled hmC/hmC as compared to radiolabeled mC or C in the presence of mC or C cold competitors. The fractions that exhibit hmC-specific shift in this first EMSA are then subjected to further characterization by EMSA using other oligonucleotide presentations and other cold competitor identities and ratios.

An hmC specific activity elutes late in the two step gradient off of the POROS Heparin column. This specific activity is apparent by the intense shift for hmC in the presence of five fold molar excess unmodified C cold competitor observed with fractions 13 and 14 shown in fig. 2.3 panel B. While hmC-active fractions such as 13 do shift mC somewhat (far right of panel B gel), this shift is clearly less for mC. This indicates qualitatively that while more

fractionation remains to be done, activities that preferentially bind hmC over mC in the presence of an excess of unmodified C do exist in the mammalian brain. This fractionation step is particularly useful for the biochemical fractionation scheme as a whole because it so strongly separates the proteins responsible for the hmC-specific activity from the large peak of protein that elutes early. Though this earlier peak is very rich in protein, it is not mC- or hmC-active by EMSA under these conditions (fractions 6 and 7 in fig. 2.3 panel B).

Other types of heparin resins did not separate the [ox]mC-specific activities from the many other DNA binding proteins captured and eluted early on in fractionation. When heparin sepharose is used in the initial fractionation step, the [ox]mC-specific activity largely overlaps with an mC binding activity. While the [ox]mC-specific activity is somewhat enriched in late fractions from these columns, it remains difficult to separate these activities in subsequent chromatographic fractionation steps. My first observations of [ox]mC-specific shift in the presence of excess C and mC cold competitor came from fractions derived from heparin sepharose fast flow (GE Healthcare) resins. However, when it became necessary to scale up the fractionation scheme to isolate more material for mass spectrometry identification, it was difficult to reproducibly separate the blurred [ox]mC-specific activities from mC activities at this key lead-off fractionation step. To proceed with the purification, a higher capacity and higher resolution chromatographic step was needed, and the POROS heparin resin provides this.

While the POROS heparin does isolate the [ox]mC-specific activity away from other DNA binding activities, the activity drops off rapidly, as observed in the lack of shift by fraction 15 in fig. 2.3 panel B. This is reproducibly seen at this same point in the fractionation gradient in multiple preparations of this extract over this column, and could be due to the presence of a contaminating protease, or an inhibitory factor that disrupts the shift. Given this, under these chromatographic conditions the POROS Heparin only isolates a few fractions of [ox]mC specific material that can be carried forward, and I cannot know by this assay if the

activity is also found in later fractions, even though it seems reasonable to expect that the presence of the putative [ox]mC-specific factors would not chromatographically drop off so sharply when the elution gradient is not changing dramatically between fractions 14 and 15.

However, the activity that can be recovered by fractionation by the POROS Heparin is remarkably specific, as is revealed in further EMSAs to characterize the properties of the later fractions. These fractions were screened for binding to mC, hmC, and fC in the presence of increased amounts of C and mC cold competitor (fig. 2.3 panel C). The activity preferentially shifts both hmC and fC containing DNA in the presence of twice as much C cold competitor as previous. Interestingly, the activity is markedly stronger in terms of the amount of radioactive DNA shifted when mC is used as a cold competitor. This suggests that the presence of excess mC further shifts the binding equilibrium toward [ox]mC-containing DNA. This is consistent with mC being even less well tolerated than C by this activity. From the first column forward, it is also worth noting that the fC-specific activity is typically stronger than hmC specific activity, in that both a greater amount of oligonucleotide is shifted for the same amount on fraction input and labeled oligonucleotide, and that the fC-specific activity persists longer in solution than hmC-specific activity during the course of fractionation (7-10 days). The fC-specific shift also has a slightly higher electrophoretic mobility, as it is observed slightly lower on the gel than the hmC-specific shift. These early observations are consistent with there being multiple distinct [ox]mC-specific activities that may have preferences for the different [ox]mC states.

2.4.2 The 6 mL Resource S isolates multiple activities at several electrophoretic mobilities

Active fractions from the POROS Heparin column were individually dialyzed against 2 L of BETG300 for a total of approximately 16 hours with one buffer exchange after approximately 10 hours. These fractions were then diluted with BTEG0 to iso-conductivity with BTEG200

(50 mM Bis-Tris pH 6.2, 1 mM EDTA, 200 mM NaCl, 10% glycerol (v/v), and 5 mM β-mercaptoethanol), and loaded onto a 6 mL Resource S column (GE Healthcare), pre-equilibrated in BTEG200. After loading via superloop and washing for two column volumes with BTEG200, the column was developed with a linear gradient from 200 to 750 mM NaCl over six column volumes then washed by linear gradient over 1 column volume to BTEG1500 (BTEG with 1.5M NaCl) and held in this buffer for a column volume thereafter.

The fractions eluted in the fifth column volume of the Resource S exhibit [ox]mC-specific activity. This activity is observed as shifts of several mobilities for hmC, suggesting multiple activities or the decomposition of a complex activity into several components as a result of this chromatography. The major activity is the highest mobility activity, though lower mobility activities are also revealed by the Resource S column that are not enriched in the input. These activities were not found to be significantly different in their specificity for [ox]mC at this stage, and fractions containing both the high mobility activity and the low mobility activities were combined for further fractionation.

2.4.3 The 1 mL Mono S column reveals key properties of the [ox]mC-specific activities

Active fractions from the Resource S column were diluted with BTEG0 to iso-conductivity with BTEG200 (50 mM Bis-Tris pH 6.2, 1 mM EDTA, 100 mM NaCl, 10% glycerol (v/v), and 5 mM β-mercaptoethanol). The Mono S column was loaded and washed for two column volumes with BTEG200, then developed over sixteen column volumes to 600 mM NaCl, then over 2 column volumes to 1.2 M NaCl. 0.3 mL fractions were collected across the shallow gradient elution. In the gel shift assays, the binding activity of every other fraction for hmC was screened in the presence of C and mC cold competitor. The fractions eluted in the fifth to seventh column volumes (as shown in the chromatogram of fig. 2.5, panel A) of the gradient exhibit the strongest [ox]mC specific activity in the gel shift assay (fig. 2.5, panel

B).

Overall, the MonoS column greatly concentrates the [ox]mC-specific activity from the Resource S in that these fractions more strongly shift hmC symmetrically modified DNA in the presence of a greater amount of mC cold competitor, as shown in fig. 2.5, panel B. Some shift for symmetric mC is still observed, albeit weaker, here in the middle of the purification as some fractionation remains to be done to exclude this activity. At this stage in the preparation, however, I also became interested in whether the [ox]mC-specific activity had any preference for symmetric or asymmetrically modified DNA. If so, it may be that activity for asymmetric mC-containing substrates should also be monitored and excluded in the course of the purification. There is good biological reason to believe that asymmetric modification binding activities could exist. As previously introduced, hmC and fC modifications are primarily asymmetric within mammalian genomes [Booth et al. 2014; Wen et al. 2014]. There is also reason to believe that these asymmetric states are maintained because of the enzymatic preferences of the TETs and TDG to generate asymmetrically modified states. Finally, in tissues where [ox]mC species are highly abundant, namely the brain, hmC is preferentially found on the sense strand of highly expressed genes. Given these observations, it may be that the cell maintains an asymmetric distribution of [ox]mC species and that this contributes to its function in a chromatin based signaling pathway. This idea will be explored further in section 2.5, *Key properties of the native oxidized methylcytosine specific binding activity.*

I examined the binding activity for asymmetrically modified presentations of hmC and fC with the strong MonoS activity found in fraction 20, shown in the left side of the gel in panel C of fig. 2.5. I observed that the native activity does not shift substrates with [ox]mC opposite mC, but strongly shifts substrates with [ox]mC opposite unmodified C. The relative discrimination between these asymmetric presentations of [ox]mC is much greater than the discrimination observed for symmetric presentations at this stage in the prep and under

these binding conditions. This finding suggests that the binding of native [ox]mC specific activities may be tuned to discriminate most highly between these asymmetric presentations that are more commonly found in the cell. While this activity in bulk native extract must be validated for individual proteins, it may be that recognition of asymmetric [ox]mC is critical for a particular functional readout of these modifications.

The MonoS also partially separates the previously mentioned multiple mobility activities noted in the chromatogram shown in fig. 2.5 panel A. When these activities are isolated in distinct fractions some differences in their activities can be revealed. Near neighbor fractions across the shallow gradient exhibit striking differences in the their activity as shown in the EMSA in fig. 2.5, panel C. Here, fractions less than a column volume apart (0.3 mL) show differential enrichment in a lower mobility activity, observed in fraction 22 but not in fraction 20. This low mobility activity shows less discrimination between [ox]mC states and asymmetric presentations as compared to the high mobility activity. The low mobility activity also exhibits less sensitivity to hmC as a specific competitor at the concentrations presented here as compared to fraction 20. This suggests that the lower mobility activity is less specific for hmC, just as it seems to be less specific for [ox]mC modified states in general.

In pursuit of the major and stronger high mobility activity, fractions more cleanly enriched in the higher mobility activity were pooled together for further fractionation, whereas fractions with the low mobility activity were pooled separately and processed secondarily. This observation by EMSA of separation of the two activities along the gradient is very sensitive to the loading level of the column (more material conflates the two) and the percentage of the gel used in the EMSA and how it is run, just as is observed between fig. 2.5 panels B and C. Even with careful pooling, however, some lower mobility activity is typically found in the next column (DEAE-5PW) but is separated further by the final column (size exclusion). It may be that the bulk high mobility native activity continues to generate lower mobility activity during the next chromatography step or over time, but as we will see,

careful separation of the two does reveal distinct properties of each.

2.4.4 The 0.36 mL DEAE-5PW concentrates the high mobility activity

Fractions from the Mono S column enriched in the more specific high mobility activity were carried forward to the next column in the purification scheme. These fractions were diluted in 100 mM Tris pH 8.0, 1 mM EDTA, 50 mM NaCl, 5% glycerol (v/v), and 5 mM β-mercaptoethanol (TEG0) to iso-conductivity with TEG50 (50 mM NaCl). A 0.36 mL DEAE-5PW column (TSKgel DEAE-5PW, Tosoh Biosciences) was equilibrated in TEG50, loaded, and eluted over 15 column volumes to 1.2 M NaCl. 0.2 mL fractions were collected over the gradient.

The fractions eluted in the first four column volumes of the gradient exhibit the strongest [ox]mC specific activity in the gel shift assay, as well as a remarkable concentration of the activity by this column. However, depending on how this column is loaded and developed, the DEAE-5PW is only partially chromatographically separates the high and low mobility activities, and some activity is left behind in the column loading flow through as observed for fractions 4 and 5 in fig. 2.7. The low and high mobility activities again show slightly different specificity for [ox]mC, with the high mobility activity being qualitatively stronger and more specific. Typically, fractions containing predominantly the high mobility activity were carried forward to the size exclusion column and ultimately to the oligonucleotide pull down and mass spectrometry experiments. However, the low mobility activity was also further characterized through size exclusion chromatography and EMSAs.

2.4.5 A carC-specific activity can be isolated from some DEAE-5PW preparations of the high mobility activity

During one preparation, I attempted to minimize overloading of the DEAE-5PW and further separate the upper and lower activities. To do so, I loaded divided the active material from

the MonoS column and ran a shallower gradient to segment 600 mM NaCl over 30 column volumes. This differs from my standard use of a linear to 1.2M NaCl with the DEAE-5PW. However, in this prep, I had a great deal of strong high mobility activity and wanted to avoid leaving it behind in the flow through or conflating the residual lower and major higher mobility activities in the DEAE-5PW elution, as I had seen before. The resulting fractions were screened by EMSA, and found to achieve greater separation of the lower and higher mobility activities, but the degree of discrimination between symmetric and asymmetric substrates was now less striking in both activities. However, in subsequent characterization EMSAs, I also observed a very strong symmetric carC specific activity in fractions spanning the elution portion noted on the chromatogram in fig. 2.8 panel A and as shown in the EMSA for fraction B9 in fig. 2.8 panel B. This is the fractionation space in which the hmC and fC specific high mobility activity resides, but in this preparation when the upper and lower activities are more clearly separated, a carC activity is revealed.

When initially developing the biochemical fractionation of porcine brain for [ox]mC specific activities, I routinely would look for overlapping or distinct carC-specific activities, but typically would not find them. The activity, when present, was quite weak and not distinct from the stronger hmC and fC activities that were usually better indicators of how the prep was proceeding. At worst, the carC activity was poorly behaved in that it would become aggregated in the wells of the gel. In addition to not being productive for assessing any carC-specific activity, this well aggregation would sometimes distort imaging of other activities present on the gel, in addition to carC activity precipitating onto glass plates, creating radioactive messes. Given the relatively weak activity that was not well resolved by the EMSA conditions I had come to favor for examining hmC and fC specific activities, I screened for carC binding less often. However, when I observed less binding preference for hmC and fC as compared to mC after the shallow run DEAE-5PW experiment, I ventured a lane to checking for carC activity.

To confirm that this carC activity I observe here is distinct and real, I selected the fractions containing the carC specific activity and loaded these onto a Superdex 200 size exclusion column. These size exclusion fractions contained both the greatest enrichment of the lower mobility activity that I have seen, but also a very clean separation of this and the higher mobility activity. Here again, the high mobility activity again exhibits a preference for symmetric carC, and less discrimination between mC, hmC, and fC. However, some asymmetric substrate discrimination is observed as before for [ox]mC/C over [ox]mC/mC in both the lower and higher mobility activity. The carC activity is again unique to the high mobility activity only. The intensity of the binding in terms of the amount of carC DNA present in this shift is even stronger after the size exclusion, suggesting relative enrichment of the activity. However, this strong carC-specific activity is difficult to work with because it was not observed consistently in every preparation. It remains to be tested whether there is asymmetric binding discrimination between carC/mC and carC. Notably, in oligonucleotide pulldown experiments using symmetric carC and hmC/C containing DNA, I observe specific enrichment of a recovered gel band when pulling down with symmetric carC, indicated at the purple arrow in fig. 2.9, panel C. Together, these results support the idea that in this preparation, several specific activities for the different [ox]mC states, including carC, exist in native brain nuclear extract.

2.4.6 Superdex 75/200 size exclusion chromatography polishes the final high mobility [ox]mC-specific activity

Active fractions from the DEAE-5PW were finally purified by size exclusion chromatography. Two different size exclusion columns were interchangeably used and yield slightly different resolutions of the active fractions from co-purifying proteins, as shown in fig. 2.10. The Superdex 75 (S75) column concentrates the activity into fewer fractions, but largely does not separate the low and high mobility activities. The Superdex 200 (S200) can separate

these activities at the cost of mildly diluting them more so than the S75. In either case, the size exclusion columns were equilibrated and developed in 50 mM Na·HEPESpH 7.8, 1 mM EDTA, 300 mM NaCl, 5% glycerol (v/v), and 5 mM β-mercaptoethanol. Fractions from the previous column enriched in the high mobility activity were concentrated on a 10 kD MWCO Ultrafree centrifugal concentrator (Millipore) prior to loading. 0.3 mL fractions were collected.

It's relevant to comment here on the existence of two [ox]mC specific mobility activities in general. Throughout the fractionation and activity assay process, it is formally possible that DNA binding activities observed at distinct electrophoretic mobilities are either (1) distinct protein molecules that happen to co-fractionate, or (2) the same protein in distinct complexes that may exhibit different binding properties as it is disassociating in the course of the chromatography, or (3) some conflation of both of those scenarios for different contributing proteins that happen to migrate at the same electrophoretic mobility. This assay cannot absolutely distinguish these interpretations. However, the recurring coincidence of the high and low mobility activities across the last 3 columns of the purification scheme (strong evidence from MonoS, DEAE-5PW, and size exclusion) lends some support to the idea that the two shifts represent the same activity falling apart rather than two activities traveling together over three types of different chromatography. That said, differences in the specificity of this activity are observed at the level of EMSA in terms of the slight but consistent difference in the electrophoretic mobility of the hmC and fC specific shifts within the higher mobility activity, and for carC, though not in every prep. The chromatographic separation of carC activity from hmC and fC asymmetric specificity, though not the predominant activity, is still very compelling evidence that some degree of activity separation exists (fig. 2.9). Given these observations, it seems that distinct activities for each of the [ox]mC modification states are present in these preparations, but that they chromatographipally overlap a great deal. It may also be that [ox]mC specific proteins partially co-exist in common complexes that can

fall apart in the course of purification to reveal different specificities. Given the differences in the distribution and abundance of the different [ox]mC modification states, it stands to reason that distinct binding activities for each could be useful for giving rise to different outcomes in the different contexts in which the modifications are found.

Due to its reproducible specificity for hmC and fC, I optimized the fractionation scheme to isolate the major high mobility activity with the final S200, as shown binding symmetric hmC in fig. 2.11. The [ox]mC specific activity elutes at ∼12.5 mL retention volume on the S200. Within the major high mobility shifted complex bands, there are still two sub-bands visible at the beginning of the elution of the [ox]mC specific activity, but they are not as enriched here as previously shown in fig. 2.10 panel B when the DEAE-5PW column prior was run differently. This preparation of the major high mobility activity was then subjected to affinity purification by DNA pulldowns and mass spectrometry for identification of the proteins responsible. These results will be described in Chapter 3, after a discussion of the key properties of this bulk are first reviewed.

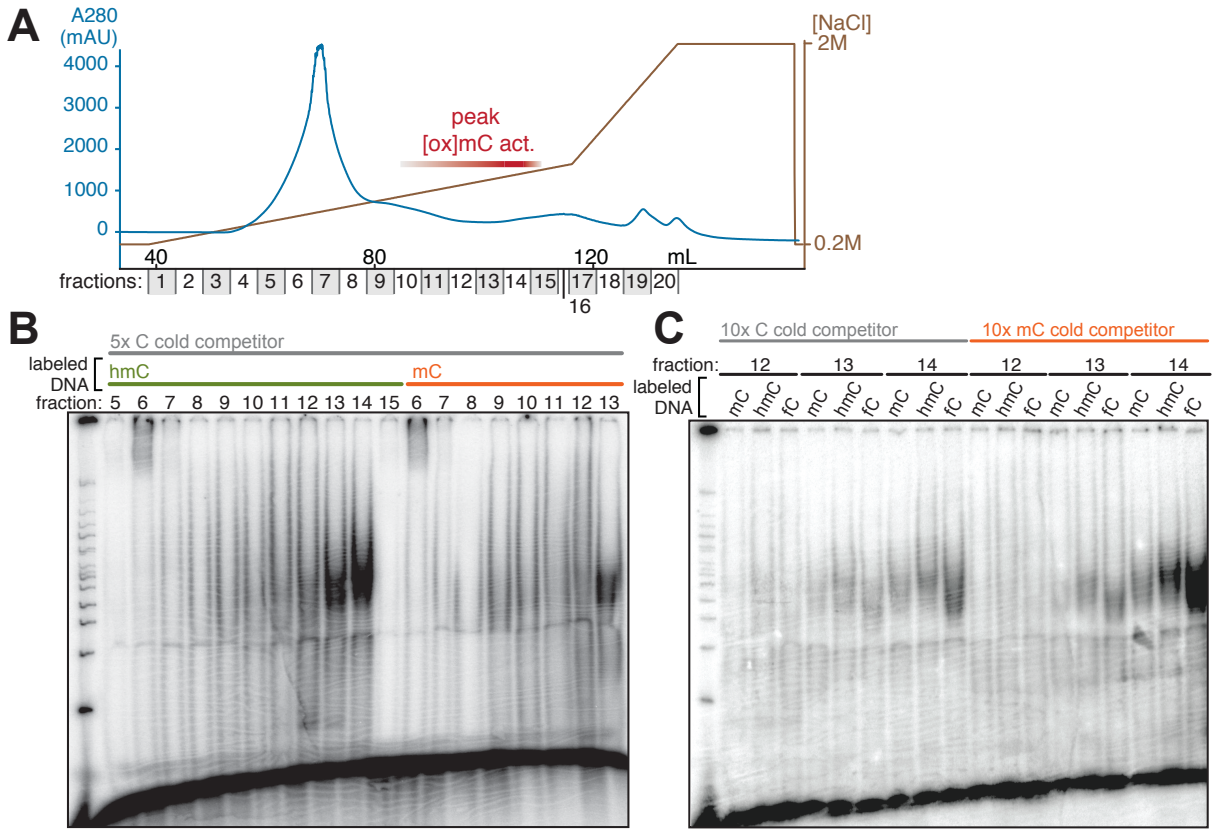


Figure 2.3: Fractionation of crude soluble extract over the POROS heparin reveals a late-eluting [ox]mC-specific activity. (A) Soluble nuclear extract loaded onto a 20 mL POROS Heparin column and resolved with a two-segment linear gradient (flow through and wash portions of the chromatogram are cropped so that the relevant individual fractions can be seen). The blue chromatogram indicates UV absorbance, brown trace indicates salt gradient, and gray lines with alternating gray shading indicate relevant fractions collected. (B) hmC-specific fractions also shift mC to a lesser extent under these conditions. While the major hmC-specific activity needs to be chromatographically resolved other activities, this assay demonstrates that there are factors with a strong preference for hmC over mC present in the mammalian brain. (C) A specificity screen of fractions in the peak of [ox]mC-binding activity with radiolabeled symmetric mC, hmC and fC duplexes in the presence of 10-molar equivalents of C (right) or mC (left) cold competitor. Activities specific for hmC and fC are present in fractions 13 and 14 (perhaps with slightly different electrophoretic mobility). In the presence of mC cold competitor, the shifted bands are more intense as compared to C cold competitor, suggesting that mC is an even less preferred binding partner than C

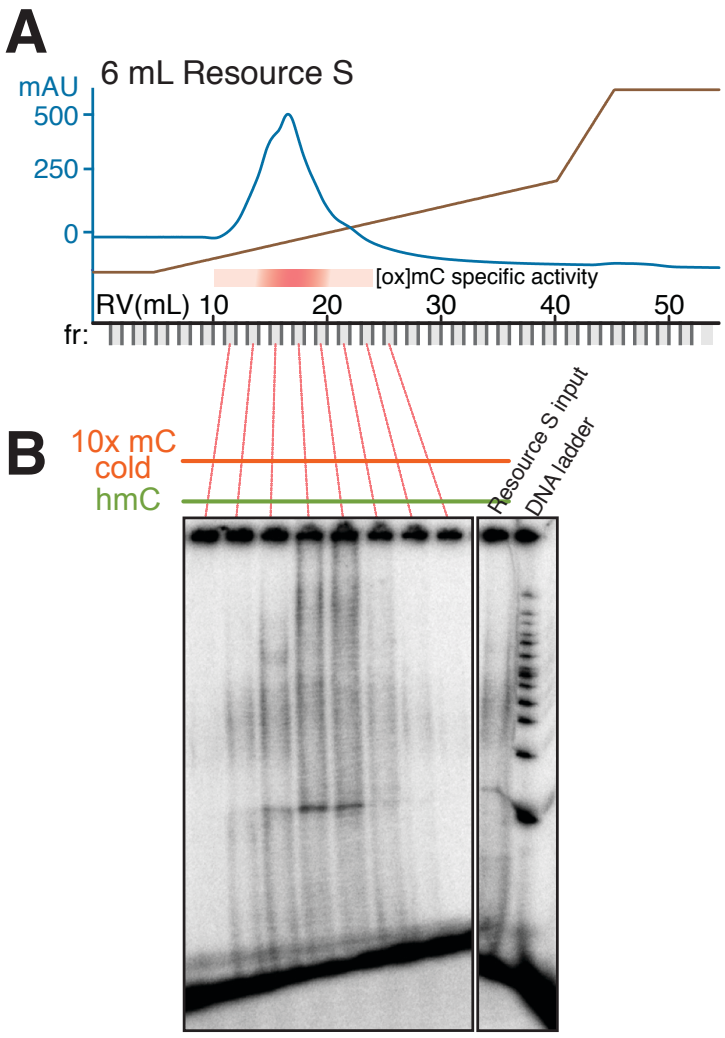


Figure 2.4: The Resource S column separates [ox]mC-specific activities of several distinct electrophoretic mobilities. In the second purification step, pooled [ox]mC-binding fractions from the POROS Heparin column were subjected to 6 mL Resource S chromatography (flow through and wash portions of the chromatogram (A) are cropped so that the relevant individual fractions can be seen). Alternating fractions were screened by EMSA with hmC radiolabeled and 10 molar equivalents of cold mC DNA (B). The fractions spanning the A280 and hmC-specific activity peak are shown in the EMSA. The activities observed in the Resource S elution represent a relative enrichment of lower mobility activities relative to the column input. This could suggest that the column concentrates a co-purifying low mobility activity not previously detect in the POROS Heparin, or that the column separates the input activity into several activities of different mobilities. In subsequent EMSAs, the activities observed here are not significantly different in their binding preferences for symmetric or asymmetric [ox]mC, and were therefore carried forward together for further purification.

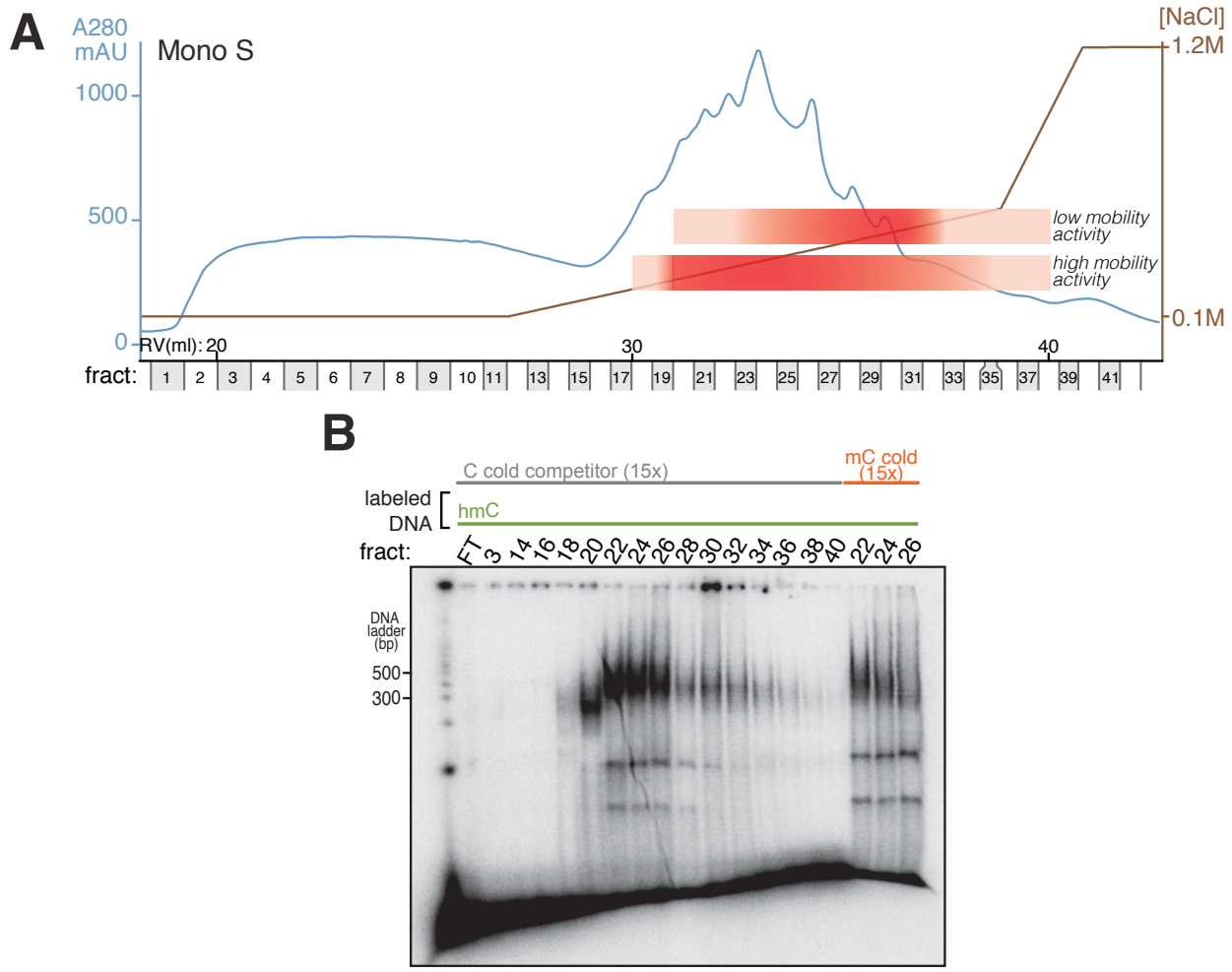


Figure 2.5: The Mono S column separates upper and lower electrophoretic mobility activities and reveals asymmetric binding preferences. (A) The third column in the native purification scheme, a 1 mL Mono S, was loaded with the [ox]mC-specific fractions from the Resource S column, washed for 10 column volumes, and eluted with a linear gradient to 600 mM NaCl over 16 column volumes. The elution profiles of the higher and lower mobility activities are indicated across the chromatogram. (B) EMSA of the indicated Mono S fractions and flow-through from column loading (FT), examining hmC-binding in the presence of 15-fold molar excess C or mC cold competitor. The hmC-binding activity first appears in fraction 18 and the bulk of this mobility activity tails into the next few fractions, appearing with a lower mobility activity. This shift is more robust in the presence of C cold competitor as compared to mC, which while different from the activity previously observed at the POROS Heparin stage, may indicate a change in the preferences of the protein pool as the composition and complexity of the protein pool has changed.

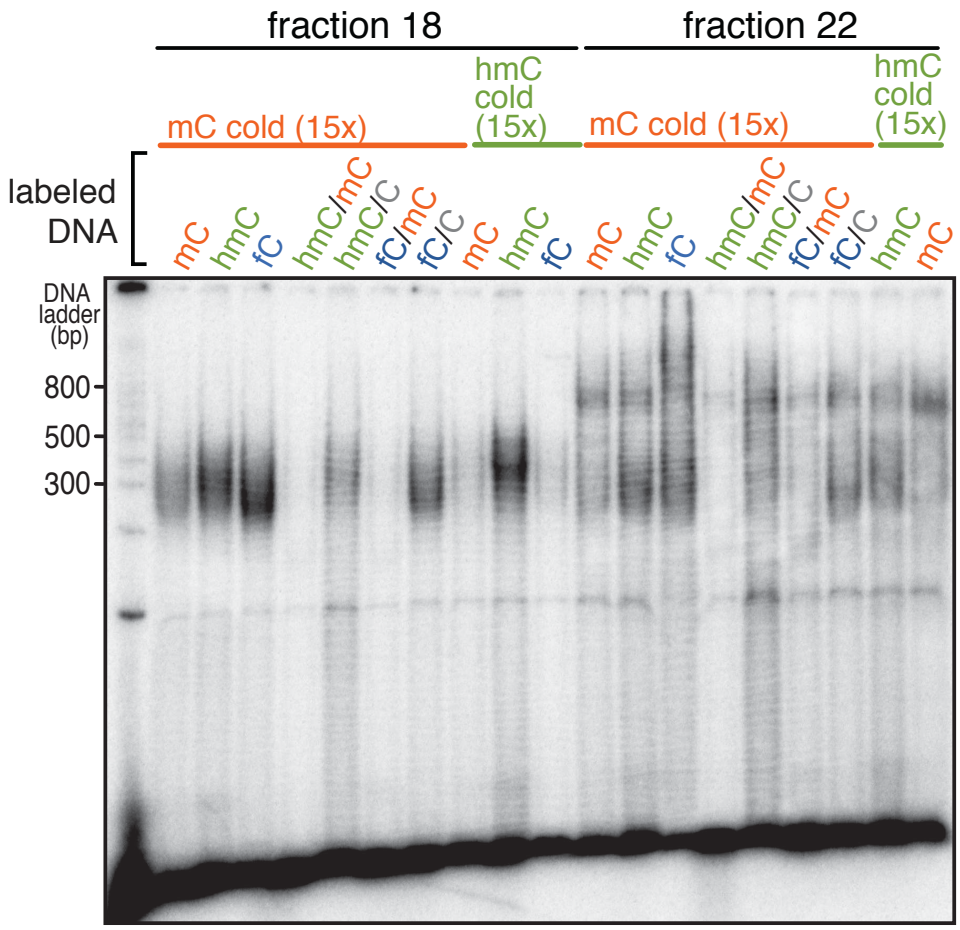


Figure 2.6: The specificity of the two distinct electrophoretic mobility activities isolated by the MonoS reveals specificity differences. Two fractions from the MonoS fractionation step shown in fig. 2.5 were further investigated in a lower percentage gel with a variety of oligonucleotide presentations and cold competitors. In both fractions 20 and 22, the higher mobility complex exhibits strong binding for symmetric hmC, fC, and to a lesser extent, mC, in the presence of 15-fold molar excess of mC cold competitor DNA. This higher mobility activity is remarkably sensitive to the symmetry of [ox]mC: [ox]mC juxtaposed with unmodified C is effectively bound (hmC/C and fC/C), whereas little binding of [ox]mC/mC counterparts (hmC/mC and fC/mC) is detected under these conditions. When symmetric hmC is provided as a cold competitor in 15-fold excess, it acts as a specific competitor to both the mC and fC shifts, yet the radiolabeled hmC-binding persists at this concentration (higher concentrations can disrupt it, however). Fraction 22 displays an additional lower mobility activity that is less discriminating between the symmetric mC, hmC, and fC DNA, and shows similar properties with asymmetric presentations as the higher mobility activity exemplified by fraction 18. Curiously, this lower mobility activity is far less sensitive to hmC as a specific cold competitor at this concentration point, indicating lower affinity for symmetric hmC, but also consistent with its less striking specificity for [ox]mC overall. Fractions containing predominantly the high mobility activity were pooled together for further purification.

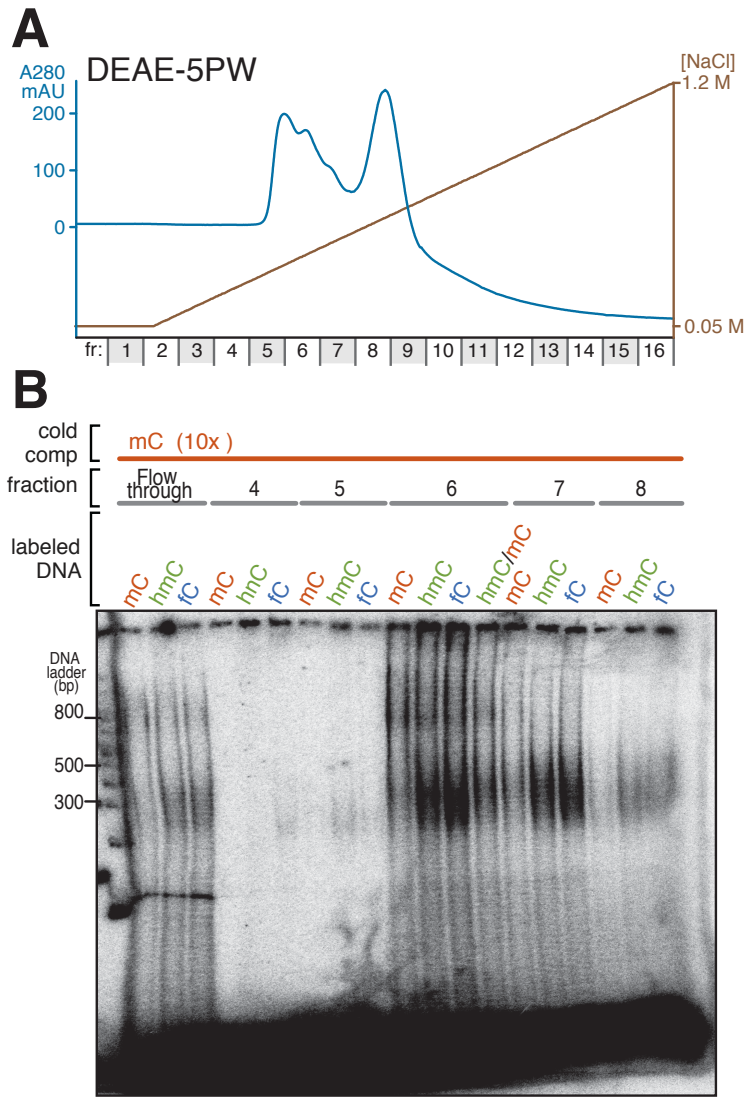


Figure 2.7: In the fourth column purification step, DEAE-5PW chromatography concentrates the activity. (A) The column was developed with a linear gradient from 0.05 to 1.2 M salt over 15 column volumes and fractions for each column volume (cv, 0.33mL) collected. The portion of the chromatogram with the active fractions is shown. (B) EMSA shows that both the low and high mobility activities co-elute in fraction 6, although some residual activity still present in the flow-through (fractions 4 and 5) is consistent with column saturation under these loading conditions. The high mobility activity, which retains previously apparent [ox]mC-specificity, is isolated in fractions 7 and 8. These two fractions were pooled for further purification, while fractions such as 6 are processed separately, either by reloading the DEAE-5PW for a second round of capture and separation, or by using the S200 to separate the low and high mobility activities.

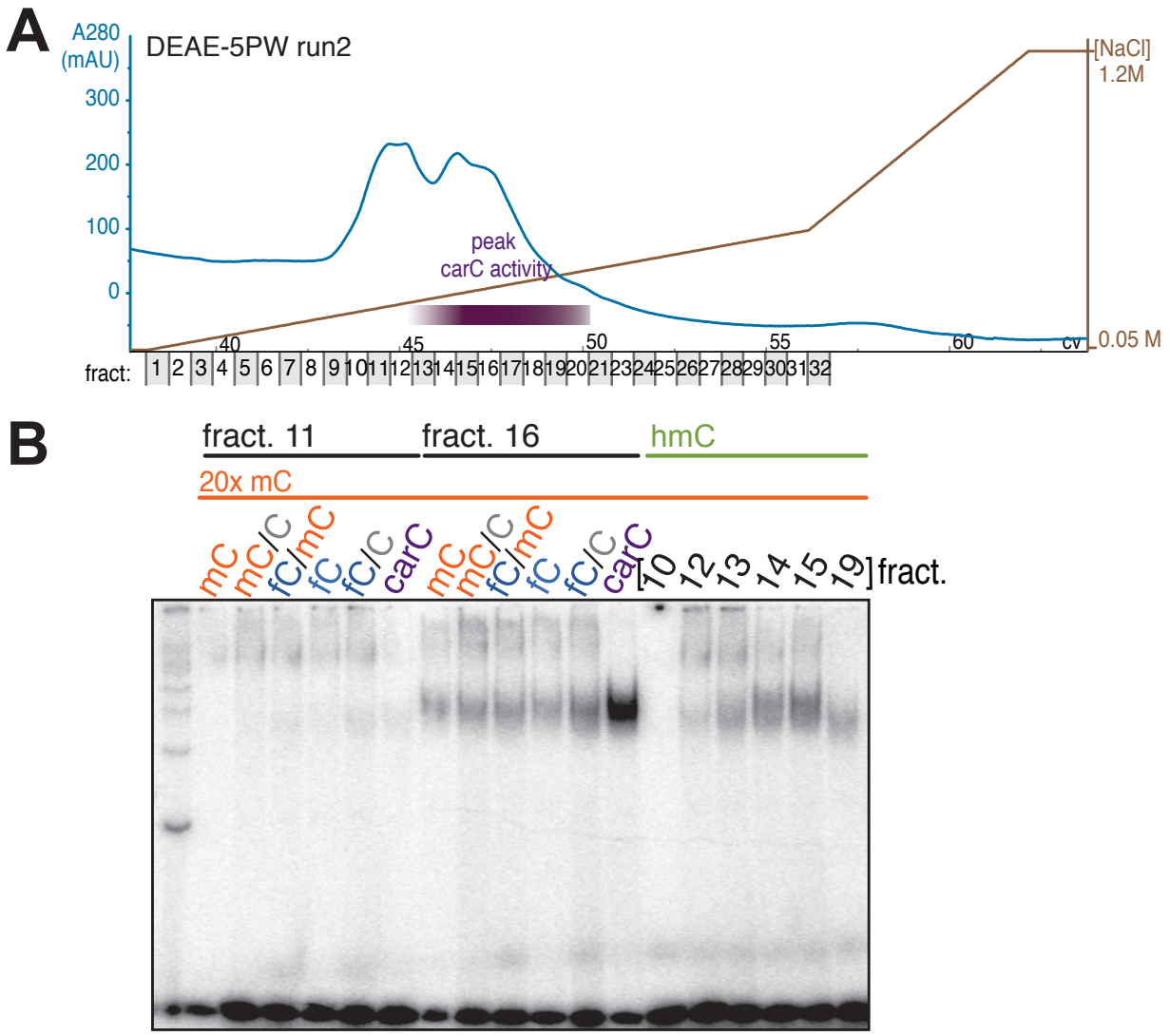


Figure 2.8: A strong carC-specific activity can be separated from the major mobility activity. (A) When the DEAE-5PW column is loaded with half of the activity from the previous column and eluted over a longer and shallower gradient, it can partially separate the low and high mobility activities. In the high mobility activity, a potent carC-specific shift is observed in this preparation (B). However, this activity is otherwise less discriminating between mC, hmC, and fC, which is *not* a property of the major activity, suggesting that this carC activity is distinct.

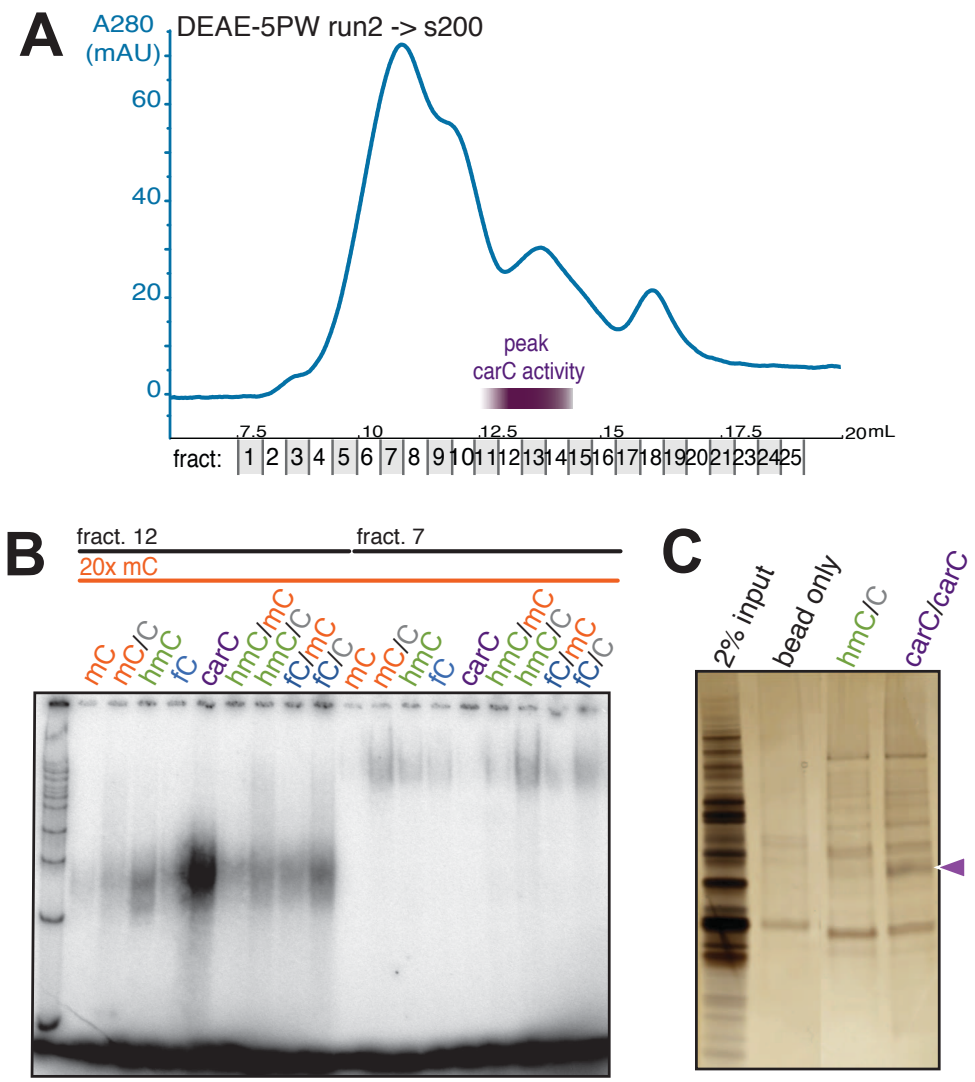


Figure 2.9: The carC-specific activity from the shallow underloaded run of the DEAE-5PW was purified over the Superdex 200 yielding the chromatogram shown in (A). This chromatographic step revealed further enrichment of the carC-specific activity in the higher mobility activity (B). There is also clean separation of the lower and higher mobility activities in the S200 run, and both activities display some asymmetric substrate discrimination for [ox]mC/C, but only the higher mobility activity strongly shifts symmetric carC. (C) A protein band specific to carC is observed in the oligonucleotide pull down assay. Together, these observations suggest that multiple specific activities for the different [ox]mC states, including carC, exist within the compound [ox]mC-specific activity I observe in brain. However, this carC-specific activity, while strong, is difficult to work with because it was not observed consistently in every preparation.

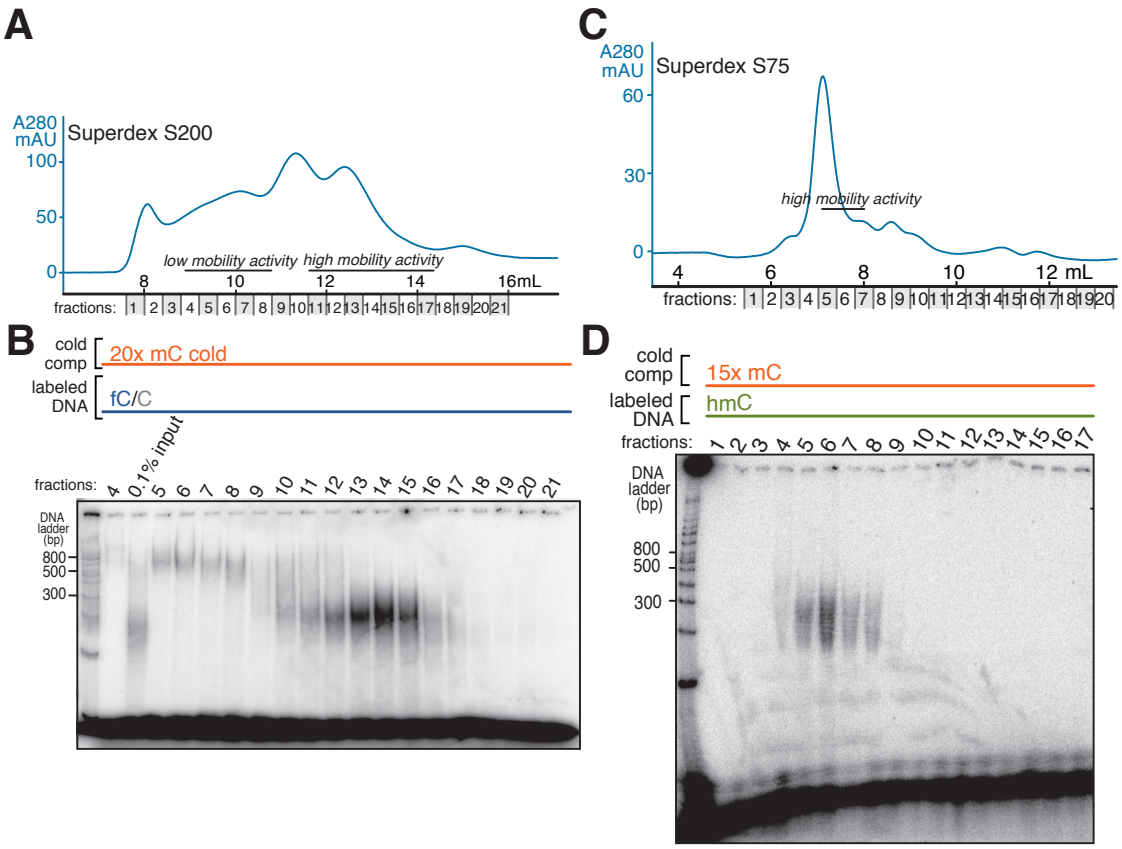


Figure 2.10: The final chromatographic step, either a Superdex 200 or 75 (S200 or S75), affords size-based isolation of the [ox]-mC specific activity. (A) The S200 chromatogram with the elution profile of the two distinct electrophoretic mobility activities noted. (B) By EMSA, in this preparation, the lower and higher mobility activities are largely separated by S200 size exclusion following a relative under loading and shallow gradient elution of the DEAE-5PW. (C) With the smaller sizing range of the S75, the higher and lower mobility activities elute together in a more concentrated elution profile as assessed by EMSA (D).

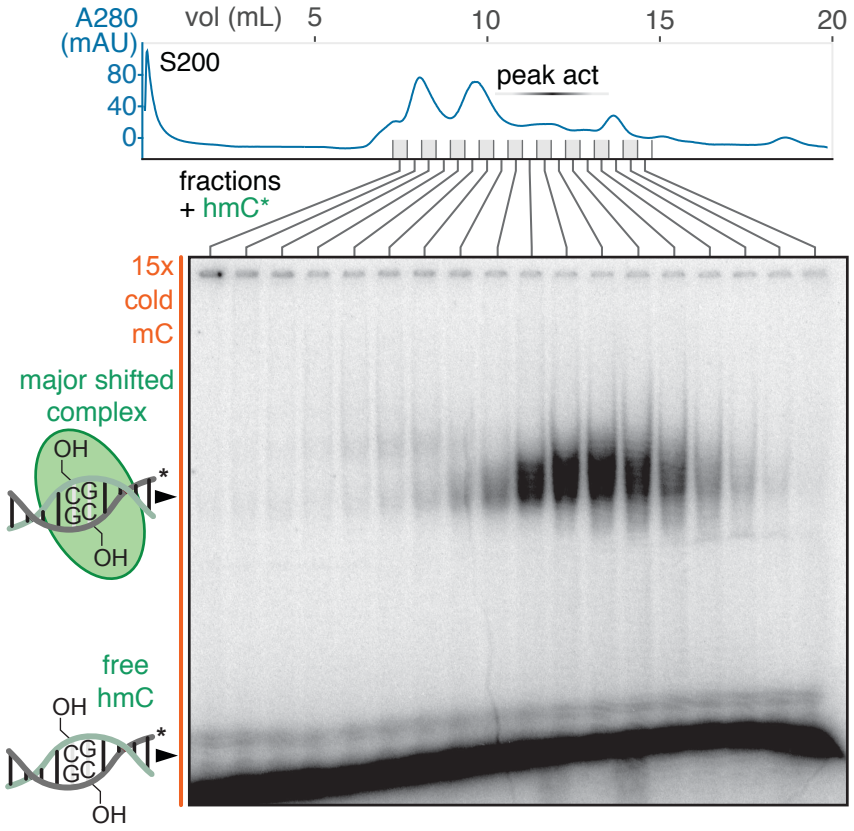


Figure 2.11: The S200 column resolves the major [ox]mC-specific high mobility activity. Fractions from the final Superdex 200 column (S200) are shown shifting symmetric hmC in the presence of 15-fold excess unlabeled mC DNA by EMSA. The major hmC-binding activity is resolved from two flanking minor activities of differing mobility.

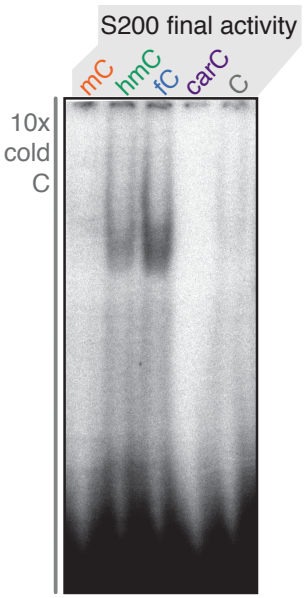


Figure 2.12: The S200-purified final major activity is specific for hmC- and fC- containing DNA in the presence of ten-fold molar excess unmodified C cold competitor. The carC activity is not observed in the major high mobility activity, but can be derived from it as previously shown. The differential mobility of hmC and fC specific activities is seen again in the final activity.

2.5 Key properties of the native oxidized methylcytosine specific binding activity

The final fractionated [ox]mC-specific activity exhibits several key properties that speak to its high degree of specificity and nuances in its substrate preferences. In light of other reports of binding proteins in the literature, it is worth reviewing the novel attributes of this biochemically fractionated activity as compared to other studies.

2.5.1 Specificity for oxidized methylcytosine in the presence of biologically-relevant excesses of non-specific cold competitor

The low abundance of [ox]mC modifications makes uniquely recognizing these marks with a specificity that can confer binding localization in the cell a tremendous biochemical challenge. To do so, any biologically meaningful binding protein will have to bind [ox]mC modifications with an affinity that is commensurate with their relative abundance, and not perturbed by the far more abundant C and mC. The specificity for hmC, fC, and carC that I observe in fractionated brain nuclear extract strongly supports the existence of such activities in the mammalian brain. In the extant literature of candidate gene studies and large scale proteomics searches, the problem of more abundant non-specific substrates has either been blatantly ignored in the interpretation of candidate binding studies, or not addressed in the design and interpretation of 'specificity assays' used to validate candidate proteins. I would therefore argue that the existence of proteins with meaningful specificity could not be concluded from these previous studies.

My biochemical fractionation approach accounted for the challenge of specific binding in the design and implementation of the competitive EMSA as a screening and characterization tool. As a result, I believe that my studies provide the first conclusive evidence that these activities exist *with the specificity required for biologically relevant binding*, that is, that

[ox]mC-specific binding can occur in the presence of a large excess of C and mC that is comparable with that present in the cell. This specificity is clearly manifest in the following experiments. Using the final S200 purified activity, I performed a side by side titration of hmC and mC in an EMSA using unmodified C as a cold competitor. The purified native activity is capable of shifting hmC at a hmC to cold competitor ratio of 1:140, whereas the shift for mC persists only to a mC to cold competitor ratio of 1:28. This represents a five-fold greater affinity for hmC as compared to mC under these cold competitor conditions. While this is not a quantitative K_ given that the native activity is still comprised of many proteins that do not shift the DNA to completion, it speaks qualitatively to an affinity difference that is commensurate with the relative abundance of hmC. No previously published candidate binding protein as reported binding discrimination of this magnitude.

2.5.2 Sensitivity to specific cold competitor

The ability of the native activity to bind [ox]mC modified DNA in the presence of large molar excesses of unmodified C and mC cold competitor is a critical hallmark of its affinity for these marks at low abundance. Another metric of specificity that supports this interpretation is the sensitivity of the EMSA activity to [ox]mC as a cold competitor. When a preferred substrate is supplied in excess of the radioactive species as a cold competitor, it is expected that this will disrupt the observed shift, and this is what I observe.

The potency of [ox]mC as a cold competitor has already been presented at the MonoS stage where hmC is a specific competitor to mC- and fC-specific shift, but only slightly tunes down hmC-specific shift (see fig. 2.5, panel C). In a side by side comparison of all modified cytosine cold competitors, it is clear that (1) the major [ox]mC-specific native activity is sensitive to these cold competitors, and (2) the degree of sensitivity suggests that the major activity has a rank affinity preference among the [ox]mC modification states. In fig. 2.14, panel A, [ox]mC-specific shift by the high mobility activity isolated at the MonoS stage is

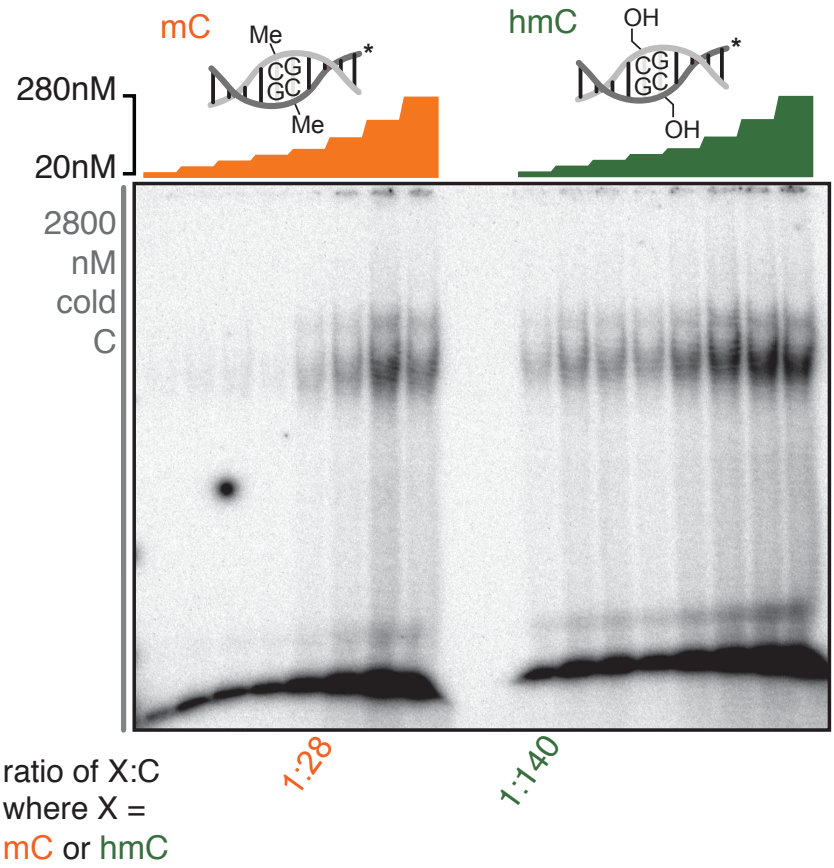


Figure 2.13: The S200 peak activity fraction is highly-specific for hmC over mC or C. Fixed volumes of this fraction titrated with either mC or hmC radiolabeled DNA from 20-280 nM in the presence of 2.8 μM final concentration throughout (10 molar equivalents with respect to the highest concentration point). The final S200-purified activity specifically binds symmetric hmC containing DNA at concentrations five-fold lower than the last comparable shift for mC containing DNA.

completely disrupted by fC as a cold competitor provided at two ratios. As compared to the strong binding observed with hmC and fC in the presence of C and mC cold competitors, hmC cold competitor disrupts residual shift for mC, abolishes shift for fC, and diminishes and smears shift for hmC. Collectively, these results support the idea that hmC and fC are specifically preferred over mC and C. The potency of fC cold competitor also suggests that between the [ox]mC states, fC is preferred over hmC at this purification point. Together, this binding profile suggests a rank preference among the modified states such that the major native activity has the highest affinity for fC, followed by hmC, then mC, and finally C.

When the cold competitor sensitivity of carC-specific s200 high mobility activity is examined, the result is somewhat different. As a reminder, this activity is derived from the major hmC- and fC- specific activity at the DEAE-5PW stage by underloading the DEAE-5PW and running a shallower gradient. This reveals a carC-specific activity in the high mobility activity, which is more chromatographically distinct from the low mobility activity in this purification scheme. The S200 size exclusion column exacerbates these differences yet further and enhances the carC specific shift. As seen in fig. 2.14, panel B, this activity preferentially binds symmetric carC in the presence of symmetric mC cold competitor, with some binding for hmC that is not very distinct from residual mC/C binding. This activity does not bind fC under these conditions, and using fC as a cold competitor allows mC/C and hmC to be bound similarly, suggesting it is not as good a competitor as symmetric mC.

The distinctions between the hmC and carC competitor cases are more nuanced. Providing carC as a cold competitor diminishes residual mC/C shift such that it is weaker than hmC-specific shift. This hmC-specific shift is diminished relative to mC cold competitor conditions, but still clear. Importantly carC as a cold competitor diminishes carC-specific shift, but not entirely. hmC cold competitor more thoroughly disrupts carC specific shift under these conditions. Collectively these results suggest that this activity prefers the symmetric [ox]mC substrates hmC and carC over mC and fC. However, titration of more cold competitor concentration points for hmC and carC cold competitors is needed to more precisely compare the rank affinity between them. Since the total shift for carC is greater than that for hmC, this somewhat obscures my ability to quantitatively compare them as purified extract by this method. Once the proteins responsible for the predominant carC-specific activity are identified, their binding preference for hmC and carC can be assessed more exactingly.

The susceptibility of the major purified activities to [ox]mC as a cold competitor further speaks to the preference of these activities for [ox]mC containing DNA. The differential sensitivity of the activities to each of the [ox]mC modified states as cold competitors also

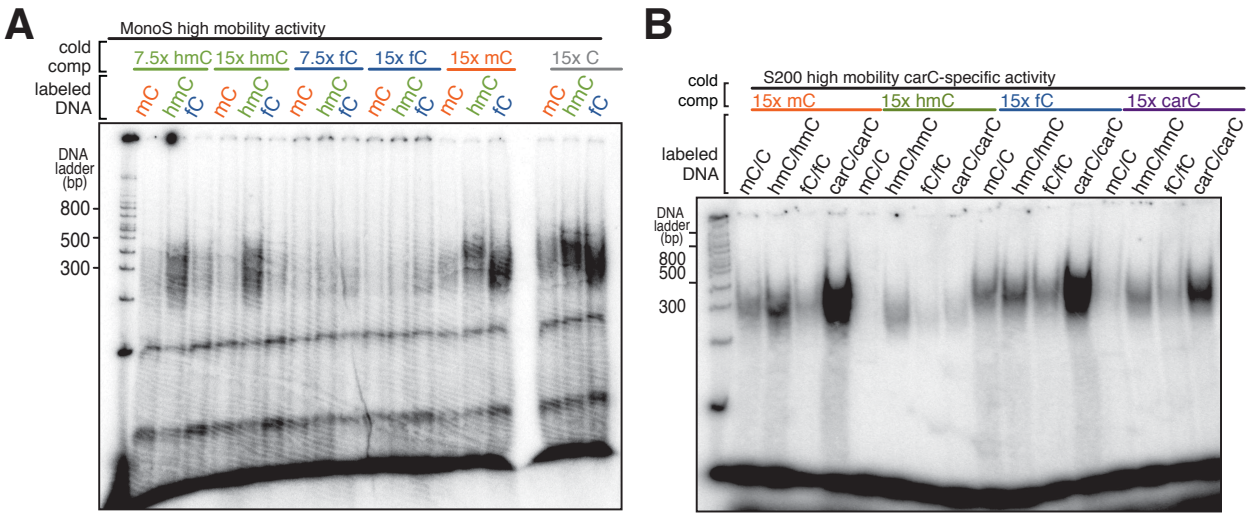


Figure 2.14: The major [ox]mC specific activities exhibit sensitivity to [ox]mC provided as a cold competitor. (A)The Mono S-purified high mobility activity preferentially binds hmC and fC in the presence of mC and C cold competitors, with some tolerance for mC in the presence of C cold competitor. hmC and fC cold competitor disrupt all residual binding for mC, indicating that these [ox]mC substrates are preferred over mC. While fC cold competitor completely disrupts all binding to hmC and fC radiolabeled DNA, hmC cold competitor disrupts mC and fC specific shift, but only partially smears and diminishes shift for hmC at these two concentration points. (B) This preparation of native brain extract is one of two that exhibited a clear carC-specific activity that is stronger than that for hmC and fC. This activity also exhibits strong sensitivity to hmC and carC as cold competitors, while fC is less potent. This again suggests a specificity for and rank preference among [ox]mC modification states over mC in this purified native activity.

suggests that there may exist activities unique to each modification state.

2.5.3 Asymmetric binding preferences

Given that no prior search for [ox]mC-specific binding proteins has reported the binding potential of asymmetric substrates, I was interested to examine these substrates in the competitive gel shift as either preferred or excluded substrates of the major native activity.

There is a strong precedent in the literature to believe that mC and [ox]mC species differ in their symmetry, and that symmetry status is dynamically regulated. Asymmetric mC has been hypothesized as a source of both epigenomic plasticity and instability in normal

development and disease states [Bird 2002]. Asymmetry is a precursor of methylation remodeling during differentiation, and is correlated with aberrant gene expression in cancer. The origin of such asymmetry, and its fate, is not known. The frequency of mC asymmetry can be estimated by hairpin bisulfite sequencing [Laird et al. 2004], a variation on bisulfite sequencing that ligates short bisulfite treated DNA fragments together with a short hairpin loop so that the methylation information carried on each strand is kept together through PCR amplification. This allows asymmetry to be estimated for a single clonal reads, but has found that mC is largely symmetric. However, these classic bisulfite sequencing studies have conflated the presence of mC with [ox]mC species, and so new strategies were needed.

Asymmetry in [ox]mC is clearly observed in several [ox]mC-sensitive sequencing studies [Ficz et al. 2011; Yu et al. 2012a; Wen et al. 2014; Booth et al. 2014]. Current estimates suggest that 20-40% of fC and hmC are asymmetric with mC, as opposed to the 92% reported for mC from bisulfite sequencing experiments. Two previous cases of striking hmC asymmetry have already been mentioned: the flanking bimodal asymmetry of hmC on the strand opposite a CTCF site in mESCs [Yu et al. 2012a], and the asymmetry of hmC to the sense strand of highly expressed genes in the mouse brain [Wen et al. 2014]. It is also noteworthy that both TET and TDG enzymes are observed to be asymmetric in their activity, in that TET enzymes tend to generate and TDG enzymes tend to leave behind asymmetric modification states [Wu et al. 2014]. It seems likely then that the enzymes that maintain [ox]mC levels generate asymmetry, and that the cell may use this to some functional effect.

The observed discrimination between [ox]mC/mC and [ox]mC/C states during fractionation of native extract suggests that [ox]mC-specific binding proteins inherently distinguish between these two presentations more so than between their corresponding symmetric presentations (C/C and M/M versus [ox]mC/[ox]mC). It is tempting to speculate that the binding interfaces of [ox]mC specific binding proteins achieve some of their specificity for [ox]mC not only by recognizing the oxidized chemical groups on one or both strands but also

by somehow excluding mC opposed [ox]mC. Given that mC on its own is highly symmetric, and [ox]mC is not, it seems biochemically useful to further constrain the binding energy of any [ox]mC specific binding protein by also scrutinizing the presentation context any [ox]mC. The asymmetric binding preferences of the final purified activity lend support to the idea that asymmetric [ox]mC is not only coincidently generated because of the asymmetric activities of TETs and TDG, but that any follow on signaling pathway that emanates from specific recognition of [ox]mC has harnessed the biochemical subtleties of these asymmetries.

2.5.4 Salt tolerance, and chemical and enzymatic sensitivity

An additional notable attribute of the bulk native [ox]mC specific activity is its resilient binding in the presence of high ionic strength and various chemical and enzymatic additives. Many nucleic acid binding proteins are very sensitive to salt given that some of their binding energy is derived from ionic contacts with the phosphodiester backbone that are compromised at high salt, and hydrogen bonding interactions are also compromised by high salt. However, given the observation that the native activity elutes at relatively high salt from the first purification column, the POROS Heparin, I was interested in how ionic strength might impact its binding. [ox]mC-active POROS Heparin fractions are dialyzed to 300 mM NaCl prior to EMSA screening. I supplemented the typical 5x EMSA reaction buffer with additional salt and subjected this pool of active dialyzed fractions to a titration of salt conditions in the EMSA experiment, shown in fig. 2.15. The ionic strength of the gel and the running buffer remained at 0.25x TBE. Shift is observed up to 1.2M NaCl, where the ionic strength of the loaded binding reaction distorts the running of the gel, but some binding preference for hmC and fC persists.

The resilience of the [ox]mC specific binding activity to such high concentrations of salt suggests that in bulk, the binding is not dependent on ionic interactions that should be compromised at this ionic strength. Rather, it invites the possibility that the [ox]mC-

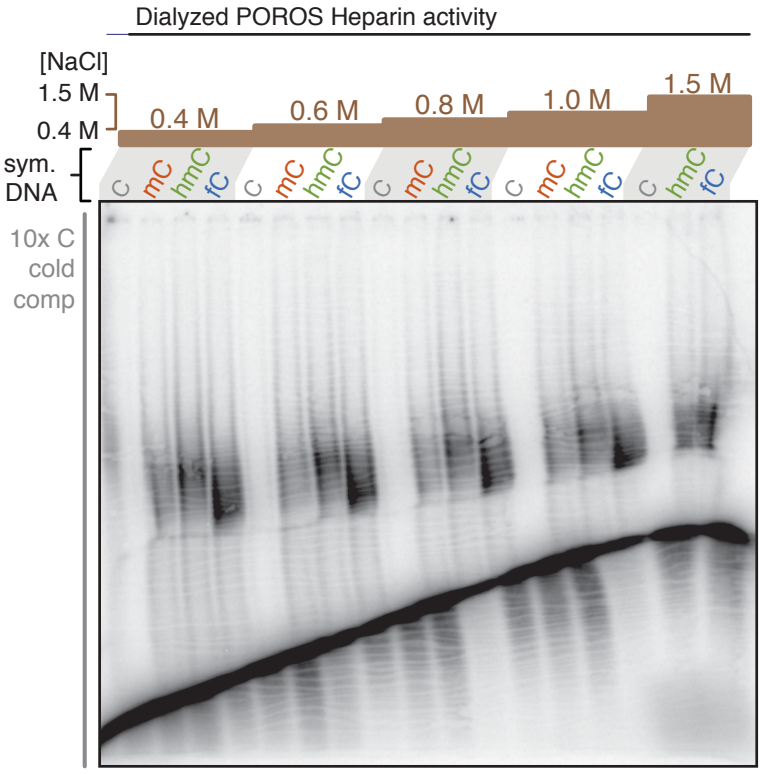


Figure 2.15: The major [ox]mC-specific activity is remarkably insensitive to high concentrations of salt. Active dialyzed POROS Heparin fractions were subjected to an EMSA with a titration of salt provided in the binding/incubation buffer. The binding reactions were allowed to equilibrate prior to electrophoresis at normal 'native'ionic strength of 0.25x TBE. Despite severe gel running artifacts at this supplemented ionic strength, the shifted [ox]mC-specific complex appears unperturbed by salt concentrations up to 1.5M.

specific binding is primarily hydrophobic in nature. This could be because the [ox]mC-specific interfaces are somewhat buried in the proteins responsible for the activity, or because the proteins read out the [ox]mC modifications of DNA by primarily hydrophobic means, such as flipping the base out of the double helix to interact with the aromatic surfaces of the nucleobase. This mode of binding would be similar to that of the SRA domain containing proteins such as UHRF1, which reads out asymmetric methylation states by flipping the mC of mC/C asymmetrically modified CpG dinucleotides out of the helix [Avvakumov et al. 2008; Arita et al. 2008]. Given my observation of asymmetric binding preferences in the native [ox]mC specific activity, it seems possible that the proteins responsible engage the DNA in an

asymmetric fashion that scrutinizes one strand of the helix for [ox]mC modifications, while tolerating C and rejecting mC on the other strand. These ideas are purely speculation at this point, but the resistance of the native activity to such high salt is still remarkable and invites exploration of these binding mechanisms.

I was also curious whether I could discern relevant cofactors or chemical dependencies of critical residues for the native [ox]mC-specific binding activity in extract in the competitive EMSA. To do so, I supplemented the 5x EMSA binding buffer with various chemical additives, including metals, metal chelators, reducing and oxidizing agents. I also challenged the activity with RNAse A and the protein denaturing detergent SDS.

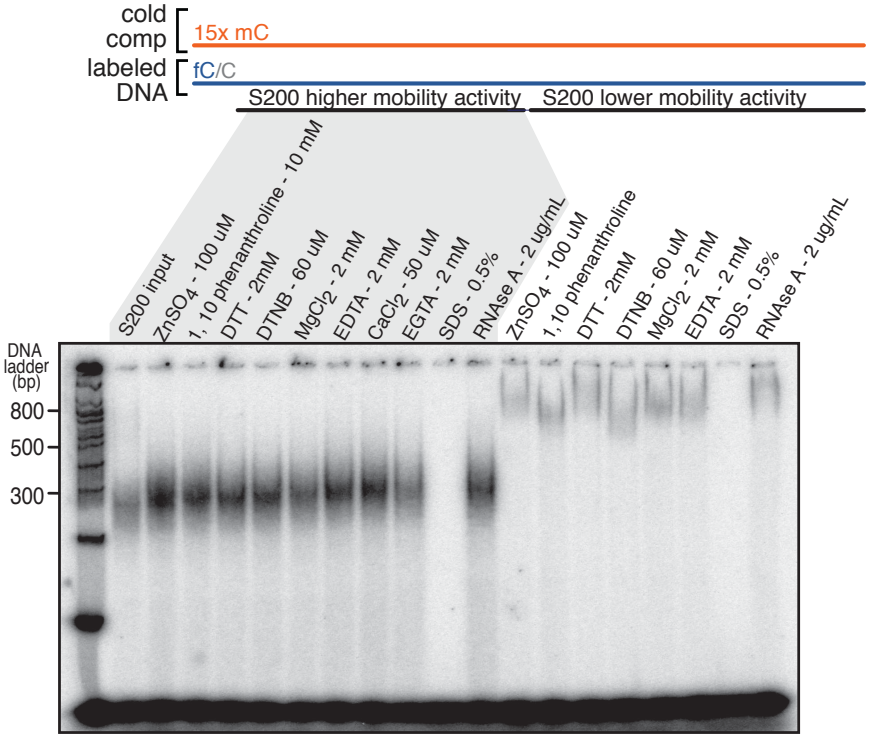


Figure 2.16: The native [ox]mC-specific activity in bulk is resistant to various chemical challenges, but is non-covalent. The [ox]mC activity of the final purified lower mobility and higher mobility activities is not sensitive to various chemical additives, including metals, metal chelators, reducing agents, oxidizing agents, and RNAse A treatment. The shift is sensitive to the protein denaturant SDS. Collectively, this suggests that both activities represent non-covalent complexes with protein factors that lack disulfides, RNA or typical metal cofactors relevant to the DNA-binding interface.

As shown in fig. 2.16, only SDS was observed to disrupt the shift, which it does completely, supporting the idea that the both the upper and lower [ox]mC-specific binding activities are non-covalent protein complexes. While I might suspect that individual [ox]mC-specific binding proteins present in these purified fractions might be sensitive on their own to the redox state of disulfides or metal cofactors within the [ox]mC binding interaction, such dependencies are not detected in bulk extract. While the exact mobility of the low mobility activity does vary in the presence of these different chemical challenges, when this experiment is repeated using a 1% lower percentage gel, these mobility differences are not observed. Given that the fC-specific activity is very strong in most preparations and has the longest lifetime in solution over the course of a seven to ten day biochemical fractionation prep, I wondered if some of the strong shift I observe for fC was due to covalent linkage of its aldehyde group to a protein. The complete sensitivity of the fC/C-specific shift to SDS suggests that this is not the case and that all of the shift I see is non-covalent.

2.5.5 Separation of activities for each oxidized methylcytosine species

The two other large scale proteomic searches for [ox]mC activities have reported candidate binding proteins for both hmC, fC, and carC modifications, with some bias towards fC-specific binders in their pulldown assays [Spruijt et al. 2013; Iurlaro et al. 2013]. Candidate gene reports for WT1 and TET3 suggest that these are carC specific binders [Hashimoto et al. 2014; Jin et al. 2016]. It is also known that TDG preferentially binds and excises fC and carC [He et al. 2011; Maiti and Drohat 2011], but this is part of the processing of [ox]mC, and is not presumed to be the basis of modification-specific chromatin signaling from these states. Given that both the candidate gene studies and the large proteomic reports of [ox]mC specific binding proteins are not accompanied by particularly convincing binding validation data, it is not clear that distinct binding proteins exist for each of the modification states. If such proteins did exist however, then one can imagine that recognition of each of the

[ox]mC modifications uniquely (which already are known to exist in slightly different genomic contexts) could give rise to signaling outcomes particular to each modification state.

Within the native [ox]mC-specific activity fractionated from brain, hmC, fC, and occasionally carC-specific activities largely coexist. The initial activity isolated by the POROS Heparin column has strong activity for hmC and fC, and a carC-specific activity can be derived from that in subsequent purification steps. However, there is evidence that the native activity, or more appropriately, *activities*, do differentiate among the [ox]mC modification states. hmC and fC show some distinction in their relative strength and mobility in that the of the amount of DNA shifted in an fC-specific shift is typically greater and observed at a slightly higher mobility than an hmC-specific shift with the same fraction. These differences are apparent at POROS Heparin, MonoS, and s200 stage (see figs. 2.3, 2.5 and 2.11). There are also distinctions among these activities in terms of their susceptibility to [ox]mC as a cold competitor, as shown in fig. 2.14, panel A. Finally, chromatographic separation of carC-specific activity from hmC and fC specific activity is possible, and also supported by the cold competitor sensitivity of this activity (see fig. 2.14, panel B) and the unique retrieval of silver stain protein gel bands using symmetric carC pulldowns (see fig. 2.9).

While a conclusive separation and assignment of hmC, fC and carC specific activities to distinct proteins remains elusive in native extract, I continued to bear this hypothesis and my biochemical fractionation evidence for it in mind as I began to characterize the affinity purification pulldowns of the final fractions using different [ox]mC modified DNAs, and the mass spectrometry results found from these pulldowns and fractionated extract.

Chapter 3

MASS SPECTROMETRY IDENTIFICATION OF PROTEINS RESPONSIBLE FOR OXIDIZED METHYLCYTOSINE DNA BINDING ACTIVITY

From very early on in my attempts to fractionate [ox]mC specific activities from brain nuclear extract, it was clear that these activities exist. But, it became clear that in order to confidently identify the proteins responsible for the activity in the final fractionated extract, the purification would need to be performed on a large scale to greatly enrich the presence of these proteins, and while also further isolating them from other activities. Many improvements to the purification scheme described above improved my ability to isolate more [ox]mC-activity enriched material for mass spectrometry identification. However, the development of an oligonucleotide pulldown strategy, and many replicates of this final affinity purification, were needed to obtain useful mass spectrometry results from the fractionated extract.

3.1 Initial approach using mass spectrometry services

At the end of one of the first biochemical fractionation attempts that revealed clear [ox]mC specific activity, I sought to identify the proteins potential responsible for the activity by examining the total protein composition of the S200 elution, spanning both active and inactive fractions, on a silver stain gel. I excised proteins that appeared unique to the active fractions, and sent them to a mass spectrometry collaborator at Harvard Medical School.

These first experiments identified some interesting putative candidate proteins, including the first identification of interleukin binding factors 2 and 3, ILF2/3, proteins that have lurked in many mass spectrometry experiments. Worthy candidates were considered based on their predicted homology or predicted domain structure potential to bind DNA, known

function in the nucleus, or other implication in a characterized active demethylation process (such as is known for ILF2/3 in transcriptional control of the β-globin locus control region [Karmakar et al. 2010]). However, efforts to validate these two proteins and three others (CMP-N Act, CIP29, and BLZW1) when expressed and purified from *E. coli* did not reveal specific binding activity.

Given these results, it was necessary to consider how abundant I might expect an [ox]mC active protein to be within the final purified fractions. None of the EMSAs I have run with fractionated extract show shift of the radiolabeled DNA (typically less than 50 nanomoles) to completion such that no free DNA is observed at the bottom of the gel. Since such a small portion of the DNA is shifted, the proteins responsible for it are present at very low abundance (though the protein's total abundance may be greater than the concentration of it which is active). As such, it may be that there isn't enough of the [ox]mC active protein for it to be seen in the final purified fractions by silver stain gel, and therefore there is no hope of excising them by this method. It is also apparent that there many proteins that are common to both active and inactive fractions having co-eluted with the activity over several columns.

To work around this limitation, I attempted to enrich the gel inputs for [ox]mC specific proteins by using a DNA pulldown to capture these proteins from the purified fractions. This affinity purification procedure has two advantages over protein identification directly from the size exclusion fractions. Foremost, it allows for enrichment of low abundance proteins within the active fractions that may contribute greatly to EMSA activity, but that might not otherwise be detected in solution mass spectrometry samples without affinity capture and washing. It is clear that there many proteins that are common to both active and inactive fractions that have co-eluted with the activity over several columns, but are not specific binding proteins. Selecting for active binding proteins that are captured by the DNA pulldown therefore increases the likelihood of identifying the proteins responsible for

the activity in the same way that the biochemical fractionation procedure: by allowing active proteins to be visualized while inactive proteins are discarded.

The enrichment of these proteins through several washing steps also represents a more stringent protein identification as compared to direct fractions, as the pull-down represents multiple captures through dilution and re-equilibration between the binding partners. Secondarily, by performing pull-downs with different substrates, I can differentiate symmetric, asymmetric, hmC, fC, and carC binding activities at the level of a captured polypeptide in a way that the EMSA cannot allow.

3.2 Oligonucleotide pull down as a final affinity purification step

As a final enrichment step, I performed comparative affinity capture with immobilized duplex oligonucleotides from highly purified size exclusion fractions. Symmetric and asymmetric presentations of C, mC and [ox]mC were used as immobilized affinity reagents on streptavidin-coated M280 Dynabeads (Life Technologies). Following pre-equilibration in 20mM Tris-HCl pH 7.8, 1M NaCl, 2 mM DTT, this paramagnetic resin was incubated with 600 pmol of biotin-labeled DNA per mg of beads (this amount represents at least 3-fold the manufacturer's calculated binding capacity of the resin for short oligonucleotides). These duplex biotinylated oligonucleotides were annealed as described in Appendix A but with a 10% molar excess of the non-biotinylated strand to ensure that all the immobilized DNA is double-stranded.

This oligonucleotide-capture incubation proceeded for 1 hour at room temperature with gentle rotation in siliconized tubes. The beads were then washed six times with three tube changes for ten minutes in five times the original bead slurry volume with 20mM Tris-HCl pH 7.8, 1M NaCl, and 2 mM DTT. These loading conditions yield equivalent oligonucleotide attachment across preparations as assessed by PNK radiolabeling of 5ug of beads and denaturing gel electrophoresis. Prior to use, these immobilized duplex beads were resuspended

at the original concentration of 10 mg/mL in 50 mM Na·HEPESpH 7.8, 300 mM NaCl, 5% glycerol (v/v), and 5 mM β-mercaptoethanol.

Size exclusion fractions with peak [ox]mC binding activity were pooled and incubated with 100 μg prepared bead slurry. All comparisons were performed with side-by-side handling controls, for example, a given active fraction pool was compared by probing with equivalent amounts of immobilized hmC/hmC- and mC/mC-resin in side by side experiments with the same volume of fractionated activity input and the same washing regimen. The beads were incubated with the active fractions for 30 minutes in siliconized tubes and washed by separating the beads from the solution by setting the tubes in a neodymium magnetic rack (Life Technologies) and removing the supernatant. Typically, the beads were washed three times for ten minutes with two tube changes with 0.5 mL of 50 mM Na·HEPESpH 7.8, 400 mM NaCl, 5% glycerol (v/v), and 5 mM β-mercaptoethanol. Washing conditions were screened, varying salt concentration, number, and length of washes, in order to find conditions that yielded preferentially enrichment of proteins to the DNA pulldown conditions of interest with low background binding. For a given input, all of the beads were washed in the same manner. Using more or more concentrated input would necessitate more washes to reveal specificity.

Proteins retained by the beads were eluted with 12 μL of 4x SDS-PAGE loading buffer (1x = 50 mM Tris-HCl pH 6.8, 2% SDS (w/v), 10% glycerol (v/v), 0.2% bromophenol blue and 5 mM β-mercaptoethanol) and boiled for five minutes before loading onto a Bis-Tris 4-20% NuPAGE (Life Technologies) gradient SDS-PAGE gel in 1x MOPS-SDS running buffer. The gels were then stained by standard procedures using the Pierce Silver Quest silver staining kit. Protein bands unique to [ox]mC pull-downs of interest were excised for in-gel trypsin digestion and mass spectrometry. Corresponding gel slices in parallel bead only, C/C, mC/mC and [ox]mC/mC pull-downs were also excised to use as negative controls in the mass spectrometry analysis.

3.2.1 *Limitations on [ox]mC-specific protein recovery and MS detection*

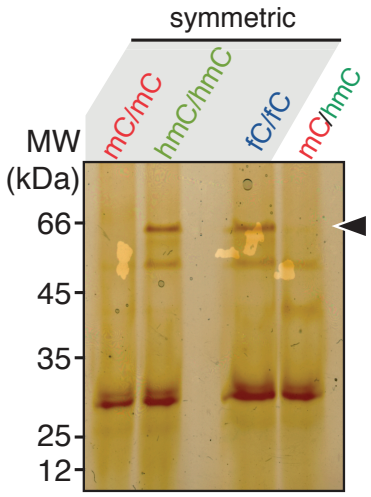


Figure 3.1: Recovery of protein bands specific to [ox]mC modified DNA. A protein band is uniquely observed in pulldowns using the '[ox]mC-active'oligonucleotides of hmC/hmC and fC/fC and is not observed for mC/mC and mC/hmC, modifications that are not favored by the native activity. Bands such as this were excised for identification of the protein composition my mass spectrometry.

Protein bands specific to [ox]mC pulldowns were retrieved with varying enrichment at the end of every biochemical fractionation attempt and were sent to mass spectrometry collaborators at Harvard Medical School and later to the commercial mass spectrometry services offered by MS Bioworks. The results from several rounds of this were largely discouraging and did not lead to many candidates. Given this, I considered what other data I could possibly use to leverage an identification of the proteins responsible. Identification of proteins from specific pulldown bands offers very compelling confirmation that the proteins recovered from them have demonstrated a capacity to bind DNA through multiple washes. However, silver stained pulldown bands have very little protein in them. Additionally, the ability to identify protein from these bands depends on efficient extraction and ionization of the peptides in the presence of the residual staining agent, which is not ideal. Another source of information about what proteins may or may not be specific is the distribution of these proteins between active and inactive solution fractions from the final size exclusion column. These fractions

are still relatively complex, as shown in fig. 3.2, and may be more complex than the silver stain can reveal. In order to bring more data into my mass spectrometry identification of [ox]mC-specific proteins, I decided to survey these solution fractions without gel assessment, both active and inactive, and correlate the peptides observed here with peptides observed in [ox]mC specific pulldowns. This experiment allows me to be more confident in what minimal peptide detection I can obtain from pulldown gel bands by connecting these hits to proteins that are also present in active solution fractions and not inactive solution fractions or inactive pulldowns. Given the scale of this mass spectrometry experiment, I wanted to be very much in control of how these samples were processed, data acquired, and analyzed. To do this, I pursued access to mass spectrometry instruments and resources in order to perform this experiment myself.

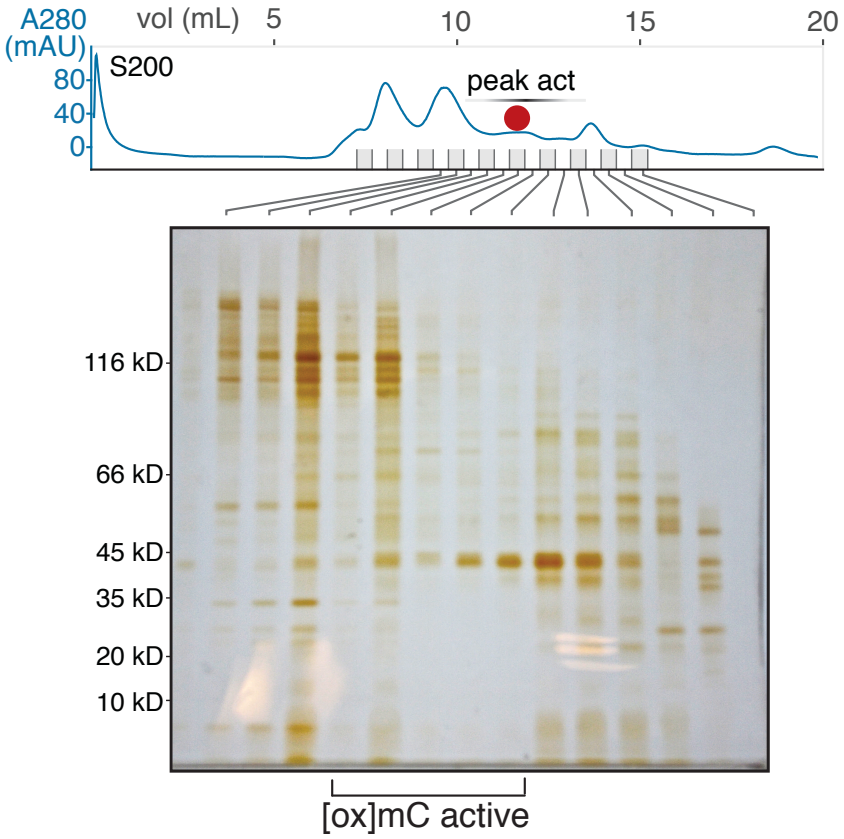


Figure 3.2: Chromatogram and composition of final purified fractions from S200. Both active and inactive are noted. EMSA of these fractions shown in fig. 2.11. Fractions such as these were analyzed by mass spectrometry to substantiate identifications of proteins from [ox]mC-specific pulldowns.

3.3 Development of self-serve mass spectrometry resources and techniques

3.3.1 UIC proteomics workshop and training for external users

In order to pursue my own mass spectrometry assessment of the proteins present in the final purified fractions, I contacted the Research Resources Center at the University of Illinois at Chicago (UIC RRC). However, training, technical assistance and instrument access was intermittent and difficult to set up until I participated in the Chicago Biomedical Consortium's proteomics workshop at the UIC RRC in the summer of 2013. This workshop and

subsequent individual training session covered mass spectrometry experimental design, sample preparation, instrument calibration and operation for LC, electrospray, and MS, data acquisition, and data processing. After the workshop, the following contacts were more responsive and assisted in my training: Roderick David, now staff scientist in Neil Kelleher's lab at Northwestern (roderick.davis@northwestern.edu) and Jerry White, now working in LC-MS industry. Future external users to the UIC RRC should strongly consider participating in this workshop to make contacts with the current staff and gain training and access. The workshop also provided access to copies of and training in the use of key software programs, namely Xcalibur Qual Broswer, Chromeleon, Mascot and Scaffold, with local help from Don Wolfgeher (senior research specialist in the Kron lab, donw@uchicago.edu). The following protocols and analysis were all informed by my training by the UIC RRC staff and consulting with Don Wolfgeher.

3.4 Preparation of gel and solution samples

20 size exclusion fractions were prepared for mass spectrometry analysis: 10 fractions from each size exclusion column, composed of 5 [ox]mC-active fractions and 5 [ox]mC-inactive fractions. Individual size exclusion fractions were digested using reagents from the In-Gel Tryptic Digestion kit from Thermo Fisher Scientific. Briefly, disulfide bonds were reduced by incubation with 3 mM TCEP for 20 minutes at 30° C, followed by alkylation with 12 mM iodoacetamide for 15 minutes at 30° C shielded from light. 20uL of size exclusion fraction (<1 μg total protein) was digested with 0.1 μg trypsin for 3 hours at 37° C, then for 16 hours at 30° C. This material was filtered in 0.45 μm Ultrafree-MC HV Centrifugal Filter (Millipore) and protonated by addition of 5% formic acid prior to LC-MS loading.

Individual gel bands were digested with the In-Gel Tryptic Digestion kit from Thermo Fisher Scientific according to the manufacturer's instructions. Briefly, disulfide bonds were reduced by incubation with 3 mM TCEP for 20 minutes at 30° C, followed by alkylation

with 12 mM iodoacetamide for 15 minutes at 30° C. Gel bands were then shrunken and dried with addition of 50% acetonitrile. Each gel band was digested with 0.1 μg trypsin for 4 hours at 37° C, then for 16 hours at 30° C. This material was filtered in 0.45 μm centrifugal spin filters (Millipore), and protonated by addition of 5% formic acid prior to LC-MS loading.

3.5 LC-MS protocols for fractionated DDA runs

Liquid chromatography-mass spectrometry and data analyses of the digested samples were carried out as previously described [Kalli and Hess 2012]. Briefly, all experiments were performed on a Dionex Ultimate 3000 Nano-HPLC, coupled to a linear ion trap Orbitrap Velos Pro mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) equipped with a nanoelectrospray ion source mounted with New Objective uncoated silica tip electrospray emitters, 260/75 μm OD/ID, 8 μm tip internal diameter. For HPLC separations, buffer A consisted of 5% acetonitrile, 0.1% formic acid in water and buffer B consisted of 95% acetonitrile and 0.1% formic acid in water. For the LC-MS/MS experiments, digested peptides were directly loaded at a flow rate of 500 nL/minute onto an Acclaim PepMap 100 C18 LC trapping cartridge (Thermo Scientific) and then onto a Zorbax 300 SB C18 3.5 μm x 150 mm x 75 μm LC column (Agilent Technoligies). The column was enclosed in a column heater operating at 30° C. After 30 minutes of loading time, the peptides were separated with a 90-minute gradient at a flow rate of 350 nL/minute. The gradient was as follows: 0-5% buffer B (5 minutes), 5-45% buffer B (60 minutes), and 100% buffer B (5 minutes).

The Orbitrap was operated in data-dependent acquisition (DDA) mode to automatically alternate between a survey scan (m/z 5, 400-1,600) in the Orbitrap and 20 collision-induced dissociation (CID) MS/MS scans in the linear ion trap. DDA involves populating a top 20 list of MS1 precursor ions observed in the first linear ion trap within the survey scan parameters listed above, and then tuning the second linear ion trap to collect MS2 fragmentation data on those top 20 hits. This list is repopulated with every scan (500 ms). CID was performed with

helium as the collision gas at a normalized collision energy of 35% and 10 ms of activation time.

In total, all of the native brain purification gel shift assays, the oligonucleotide pull down experiments, and the mass spectrometry analyses presented herein reflect biochemical preparation of three different batches of porcine brain, six distinct preparations, 20 solution mass spectrometry experiments, and 48 oligonucleotide pull down gel bands digests arising from two of the fractionation replicates through two different final size exclusion columns.

3.6 Data processing, Mascot searches, and data analysis using Scaffold software

Thermo RAW files were converted to MGF file for searches with Mascot (Matrix Science). Spectra were searched against all NCBInr mammalia taxonomy database (version 20131113, 2037278 entries). Allowed variable modifications, as defined by Unimod included (followed by monoisotopic mass): oxidation of methionine (115.9949), carboxyamidomethylation (57.0215), deamination of asparagine and glutamine (0.984016), and N-terminal formylation (27.994915).

The mass accuracy for ESI-FT-ICR by the Orbitrap Velos Pro in Mascot searches was set as 10 ppm/0.02 Da for MS1 precursors and the fragment mass tolerance was set at +/- 0.6 Da, with peptide charges of 2^+, 3^+, and 4^+ allowed for the monoisotopic mass. Trypsin cleavage tolerance was set at +/- 1 missed cleavages. Individual mass spectrometry runs were inspected in XCalibur Qual Browser (Thermo Scientific) to assess the quality of gradient profiles and the reproducible elution and retention time of parent ion species for proteins of interest.

Mascot search output DAT files were analyzed using Scaffold (version 4, Proteome Software) to compare protein identifications between the different oligonucleotide pull down and solution mass spectrometry experiments with peptide identification threshold set at 90% (3%

FDR) in Scaffold (calculated according to Peptide Prophet algorithm [Keller et al. 2002]).

3.7 Summary of full data set - lead and future candidates

The following table presents the top candidates from the full mass spectrometry data set. These candidates are presented based on their 90% confidence (3% FDR) peptide identification threshold using Peptide Prophet spectral matching statistics, statistically significant Mascot ion and Mascot peptide scores, and biological reasonableness by homology or known function (in the nucleus, annotated DNA binding domains, and implicated in gene regulatory activity). The number of unique spectral matches is reported, followed by the total number of unique observations of those spectra in parentheses, for each type of experiment.

Table 3.1: Lead candidate proteins - Mass spectrometry results

protein name	active soln. frac.	inactive soln. frac.	active PD	inactive PD	active PD type	inactive PD type	soln. frac. type
ACIN1	2(19)	0	0	0	-	-	S200, S75
ANXA1	0	0	2(6)	0	h/f	-	-
ANXA2	0	0	7(13)	0	h/f	-	-
Brg1	2(16)	0	0	0	-	-	S200
BRWD1	0	0	1(1)	0	f/C	-	-
C11ORF35	0	1(1)	1(1)	0	f/C	-	S75
C4ORF36	0	0	2(6)	0	f/f	-	-
CDC2	0	0	2(3)	-	h/h		-
DEK	5(69)	0	0	1(6)	-	h/m	S200, S75
DR1	8(63)	2(1)	0	0	-	-	S200, S75
DRAP1	6(33)	0	0	0	-	-	S200, S75
hnRNPU	0	0	1(25)	1(6)	f/C	h/m	-
ILF2	8(16)	0	1(2)	0	h/f	-	S75
ILF3	1(7)	0	0	0	0	0	S200, S75
LEO1	0	0	1(8)	0	f/f	-	-
MBD3	0	0	3(6)	1(2)	f/f	h/m	-

Table 3.1: (continued)

protein name	active soln. frac.	inactive soln. frac.	active PD	inactive PD	active PD type	inactive PD type	soln. frac. type
Med16	1(2)	0	0	0	-	-	S200
MTA3	0	0	2(3)	1(2)	h/f	h/m	-
p300	0	0	1(1)	0	h/f	-	-
PUF60	1(6)	0	0	0	-	-	S200
RNF224	2(13)	1(3)	0	0	-	-	S200
RPA49	1(2)	0	1(5)	0	h/h, f/f, h/f	-	S75
Sirt1	0	0	1(1)	0	h/f	-	-
SNF2L	0	0	3(6)	1(2)	f/f	h/m	-
Taf3	0	0	1(1)	0	f/C	-	-
UHRF1	0	1(1)	1(1)	0	f/C	-	S200
WDHD1	0	0	1(1)	0	f/C	-	-
WDR10	0	0	1(4)	0	f/f, f/C, h/h	-	-
WDR33	0	0	1(2)	0	f/f	-	-
WDR47	0	0	1(1)	0	f/C	-	-
WDR76	6(8)	0	8(27)	1(4)	h/h, f/f, f/C, h/f	h/m	S200, S75
WDTC1	0	0	1(2)	0	h/f	-	-
ZNF212	0	0	2(5)	1(1)	h/f, f/C	h/m	-
ZNF35	0	0	1(2)	0	h/f, f/C	-	-
ZNF367	2(2)	0	0	0	-	-	S200
ZNF525	1(6)	1(1)	0	0	-	-	S75, S200
ZNF772	0	1(4)	1(6)	1(1)	h/f	h/m	S75

Table 3.1: Mass spectrometry identification of candidate proteins from pulldowns and solution fractions. The number of unique peptide spectra is noted followed by the total number of observations in parentheses, as well as the type of input. In the case of pulldown observations, the oligonucleotide presentation used to generate the protein band analyzed is noted. In the case of solution fractions, the type of size exclusion column that yielded the fraction with peptide observations is noted.

3.8 WDR76 is a lead candidate [ox]mC-specific protein

These mass spectrometry experiments strongly identify WDR76 in both [ox]mC-active solution and [ox]mC-specific pulldown experiments with multiple oligonucleotide presentations, as summarized in fig. 3.5. The enrichment of WDR76 in active solution fractions and [ox]mC specific pulldowns is clear, such that WDR76 was a striking lead candidate protein in this dataset. WDR76 was observed as a single MS2 peptide in active solution samples from both the S200 and S75 final size exclusion columns (10 active fractions sampled of each). WDR76 was not observed in inactive solution fractions (10 flanking fractions sampled from each). WDR76 was also observed in gel bands excised from oligonucleotide pull down elutions using [ox]mC DNA in both symmetric hmC and fC and asymmetric hmC/C and fC/C presentations. WDR76 was observed in one hmC/mC pull down assay. Collectively, the mass spectrometry evidence for WDR76 as a candidate [ox]mC specific protein is the strongest evidence in my dataset.

However, it is very important to note that WDR76 is among the candidate proteins reported by Spruijt and colleagues [Spruijt et al. 2013] six months prior to my observation of this protein by mass spectrometry. This report (in combination with my own findings) absolutely played a role in my decision to pursue WDR76, perhaps at the expense of other interesting candidates with slightly less compelling mass spectrometry evidence. I will return to how the work of Spruijt and colleagues compares to my own in the Discussion, and in our collaborative binding studies of other candidates reported in their work, found in Appendix B.

While this mass spectrometry approach report has yielded several interesting candidate [ox]mC-specific binding proteins outlined in table 3.1, it could certainly be improved upon to yield more confident identifications. Material limitations remain an issue, and given the low amount of signal, it is difficult to confidently compare pulldowns and conclude that proteins are unique to a particular modification state. At this resolution and degree of enrichment,

there is simply too much overlap to be sure. A quantitative strategy for comparing pulldowns with different oligonucleotides would allow more clear assessment of whether proteins are specific hmC, fC, or carC, and if so, by how much relative to each other or to C or mC pulldowns. To pursue this in the future, I am interested in di-methyl labeling of material recovered from pulldowns using different C modification state oligonucleotides in order to allow for comparison of proteins recovered by each. This may allow a more quantitative and confident assessment of the separation of [ox]mC activities in the mass spectrometry inputs.

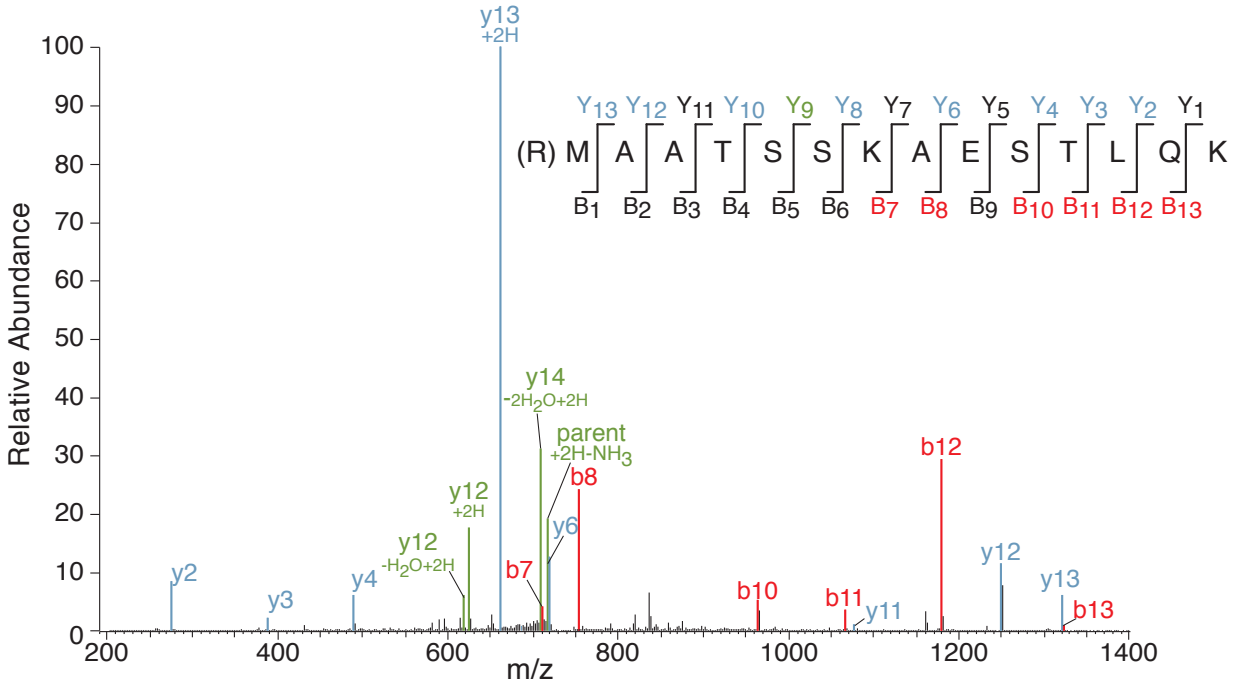
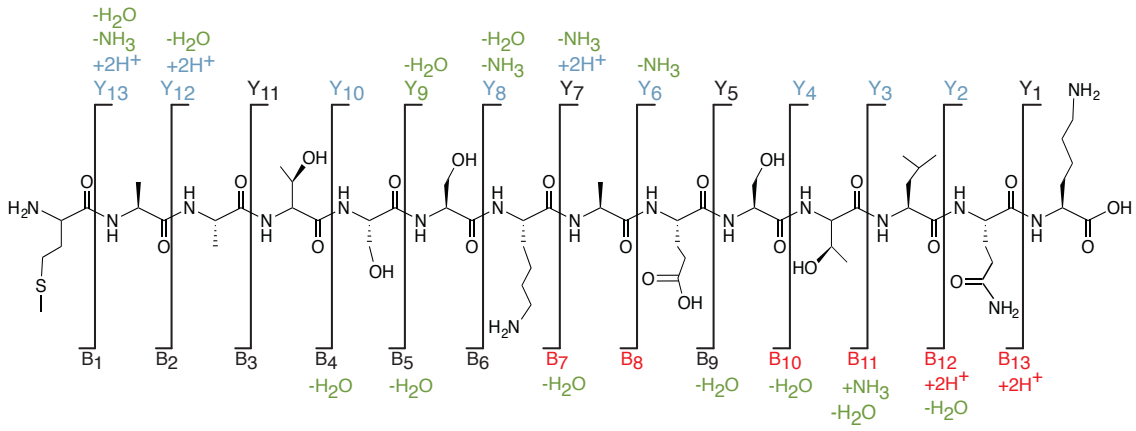


Figure 3.3: The protein WDR76 was identified in mass spectrometry experiments by a single MS2 peptide, whose spectrum is shown in (A).) A representative MS2 peptide spectrum for the WDR76 peptide found predominantly in [ox]mC-specific pull downs and active size exclusion fractions (Mascot ion score of 44.1 ($p < 4$ x10-5) and a Mascot identity score of 51.5. The fragmentation of the peptide is shown in the inset with the observed b and y ions noted in blue and red, respectively. Ions that are observed with additional protons, ammonia (-17) from the amino acids R, K, or Q, or loss of water (-18) from fragmentation of the amino acids S, T, or E, are labeled green. This MS2 peptide was observed independently in 15 samples at 90% peptide confidence (Peptide Prophet) and was observed 31 times in total within those samples (multiple observations at distinct LC-MS retention times, indicating multiple copies of the peptide within the sample). At 95% confidence (1% FDR), 11 observations of WDR76 are made in three [ox]mC-specific pulldowns using hmC/hmC, fC/fC, and hmC/fC. No observations were made at 95% confidence in [ox]mC-inactive pulldowns.

A



B

B	B Ions	B+2H$^+$	B-NH$_3$	B-H$_2$O	AA sequence	Y Ions	Y+2H$^+$	Y-NH$_3$	Y-H$_2$O	Y
1	132				M	1452.7	726.9	1435.7	1434.7	14
2	203.1				A	1321.7	661.4	1304.7	1303.7	13
3	274.1				A	1250.7	625.8	1233.6	1232.6	12
4	375.2			357.2	T	1179.6	590.3	1162.6	1161.6	11
5	462.2			44.2	S	1078.6	539.8	1061.5	1060.6	10
6	549.2	275.1		531.2	S	9991.5	496.3	974.5	973.5	9
7	677.3	339.2	660.3	659.3	K	904.5	452.8	887.5	886.5	8
8	748.4	74.7	731.3	730.4	A	776.4	388.7	759.4	758.4	7
9	877.4	439.2	860.4	859.4	E	705.4	353.2	688.4	687.4	6
10	964.4	482.7	947.4	946.4	S	576.3		559.3	558.4	5
11	1065.5	533.2	1048.5	47.5	T	489.3		472.3	471.3	4
12	1178.6	589.8	1161.5	1160.6	L	388.3		371.2		3
13	1306.6	653.8	1289.6	1288.6	Q	275.2		258.1		2
14	1452.7	726.9	1435.7	1434.7	K	147.1		130.1		

Figure 3.4: The detailed fragmentation pattern of the WDR76 MS2 peptide shown in fig. 3.3. Observed ion species are indicated on the peptide (A) and their masses are noted in the table (B). B and y ions noted in blue and red, respectively. Ions that are observed with additional protons, ammonia (-17) from the amino acids R, K, or Q, or loss of water (-18) from fragmentation of the amino acids S, T, or E, are labeled green. This peptide maps to mammalian proteome databases, but not uniquely to available proteome database for pig, *Sus scrofa* - it varies by one amino acid (second alanine is observed as a glutamate). This amino acid position is generally not well conserved across mammals.

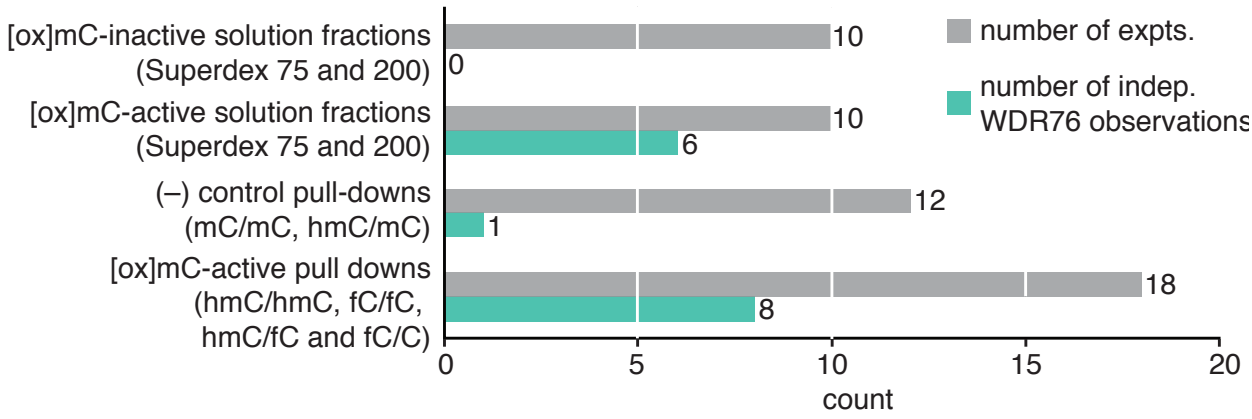


Figure 3.5: Quantification of mass spectrometry observations WDR76. WDR76 was observed as a single MS2 peptide in active solution samples from both the S200 and S75 final size exclusion columns (10 active fractions sampled). WDR76 was not observed in inactive solution fractions (10 flanking fractions sampled). WDR76 was also observed in gel bands excised from oligonucleotide pull down elutions using [ox]mC DNA in both symmetric hmC and fC and asymmetric hmC/C and fC/C presentations. WDR76 was observed in one hmC/mC pull down assay.

Chapter 4

BIOCHEMICAL STUDIES OF WDR76

In order to investigate whether WDR76 is an [ox]mC specific protein, I sought to validate my mass spectrometry evidence through biochemical studies of the protein's capacity to bind DNA. Such validation with isolated protein in a qualitative affinity assay with cold competitor, or ideally in a quantitative affinity assay, is absolutely necessary in order to determine whether [ox]mC-specific binding activity is resident in any proposed candidate protein. Ahead of this, it is entirely possible that my observation of WDR76 within mass spectrometry samples is due to its close association with another protein that actually encodes the specific interface for [ox]mC specific recognition.

4.1 Characterization of WDR76 expressed in *E. coli*

Human WDR76 was cloned by Gibson assembly of 3 codon-optimized synthetic fragments based on the sequence of the long isoform (Genbank Accession BC025247). I attempted to express this protein in *E. coli*, and observed little soluble full length expression, and specific but limited activity by EMSA. Four C-terminal constructs encompassing the WD repeat region were also designed and pursued in various *E. coli* hosts, none of which were soluble under multiple expression conditions, including expression-optimized hosts and chaperone-assisted expression hosts. Three N-terminal constructs were also designed and found to be soluble from *E. coli*, but did not exhibit full specificity for [ox]mC in filter binding experiments. In insect cell expression systems, I made both the full length and a complete WD repeat region C-terminal construct from both human and mouse sequences, all with various affinity and solubility tags on the N- or C-terminus. The best-behaved soluble protein was made from the full length human WDR76 sequence with an N-terminal FLAG-HA tag, the properties of which are reported here.

4.2 Expression by baculoviral transduction of *S. frugiperda* and *T. ni*

4.2.1 Purification and characterization of protein expressed in insect cells

To produce recombinant protein for these studies, the human WDR76 construct was cloned into pFastBac with an N-terminal FLAG-His6-Arg6-tag cleavable with human rhinovirus R3C protease. This construct was transformed into DH10Bac cells and clones containing recombinant bacmid DNA were isolated and screened by PCR for proper incorporation of the construct into the bacmid. *S. frugiperda* (Sf9) cells were transfected with validated bacmid DNA using FuGENE HD (Promega), and baculovirus was amplified and isolated for large scale infections of *T. ni* Hi5 (Invitrogen) cells. Sf9 cells were grown in Sf-900 II SFM media (Gibco) supplemented with 10% FBS (v/v), 2 mM L-glutamine, 50 mg/L gentamicin, 10 units/mL penicillin, and 10 mg/L streptomycin. Hi5 cells were grown in Insect-XPRESS Protein-Free Insect Cell Medium with L-glutamine (Lonza) supplemented with an additional 2 mM L-glutamine, 50 mg/L gentamicin, 10 units/mL penicillin, and 10 mg/L streptomycin. Both cell types were grown in 2L fernbach flasks in a platform shaker (Kuhner) rotating at 100 rpm at 27° C.

Cells were collected by centrifugation 72 hours post-infection with baculovirus (Sorvall Legend XTR with TX-750 rotor at 500 x g for 10 minutes, 4° C) and each liter of culture was resuspended in 60 mL HEGN600 (50 mM Na·HEPESpH 7.8, pH 7.8, 1 mM EDTA, 10% glycerol (v/v), 0.02% NP-40, 600 mM NaCl) and supplemented with protease inhibitor cocktail. Cell slurry was flash-frozen with liquid nitrogen and stored at -80° C until purification.

Thawed cell slurries were lysed in a Dounce homogenizer with 30 tight pestle strokes. Lysis was confirmed under the microscope with Trypan Blue staining. The lysate was clarified by centrifugation (Sorvall RC5B with SS-34 rotor at 30,000xg for 25 minutes, 4° C), the salt was lowered to 200 mM by dilution with supplemented HEGN0 (no NaCl), and the

centrifugation step was repeated. This clarified extract was incubated with 250 μL of FLAG M2 agarose affinity gel (Sigma-Aldrich) per liter of culture input for 1 hour, rotating at 4° C. The flow-through was collected by centrifugation at 500 x g for 5 minutes. The resin was washed in 10 resin volumes of HEGN600 for 10 minutes, and then in HEGN300 for 10 minutes, rotating at 4° C. The FLAG-fusion protein was eluted by incubating the resin in 0.5mL of HEGN300 supplemented with 150 μg/mL 3x FLAG peptide for 30 minutes, rotating at 4° C. Six resin volumes of elution were collected.

Pooled elutions were visualized by SDS-PAGE before pooling for further purification on a 1mL POROS Heparin column (Applied Biosystems resin packed in Tricorn 10/50 column, GE Healthcare). The pooled elutions were diluted in BTEG0 to iso-conductivity with BTEG150, loaded onto the column, the column washed with four column volumes BTEG150, then developed in a step gradient to 20% BTEG1000 with a four column volume hold, and then finally to 100% BTEG1000 over six column volumes. This purification step serves to remove the excess FLAG peptide and reduce the amount of co-purifying nucleic acid carried with the protein from the insect cells. The eluted material was visualized by SDS-PAGE and quantified by A280 (ϵ=42,860 L$\cdot$ M^{-1} cm^{-1}) for use in gel shift assays.

In an effort to enhance the limited specific activity of the recombinant WDR76, protein purified by the 1 mL POROS Heparin was also purified over the 0.36 mL DEAE 5-PW. The input was diluted in TEG0 (100 mM Tris pH 8.0, 1 mM EDTA, 50 mM NaCl, 5% glycerol (v/v), and 5 mM β-mercaptoethanol) to iso-conductivity with TEG50 (Buffer A). The input was loaded by superloop, washed with 10 column volumes of 20% TEG1000 (Buffer B) and eluted over a 5 column volume linear gradient to 100% Buffer B. The eluted material is again visualized by SDS-PAGE and quantified by A280 (ϵ=42,860 L$\cdot$ M^{-1} cm^{-1}) for use in gel shift assays at a final concentration of 1μM. The specific activity of WDR76 in EMSAs was not largely enhanced by this or other types of chromatography (MonoQ and Superdex 75 were also tested).

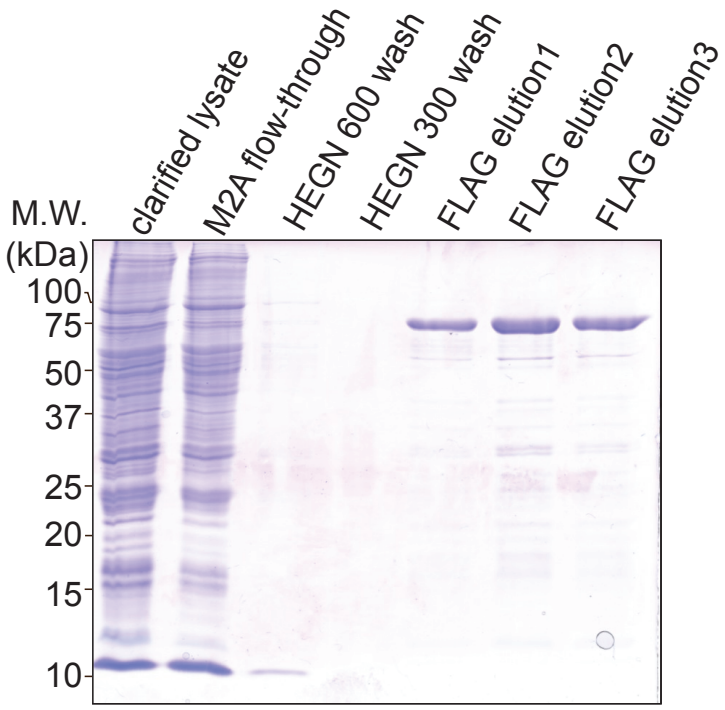


Figure 4.1: Purification of N-terminally tagged WDR76 from baculovirally transduced *T. ni* cells. FLAG-HA-WDR76 expressed in these cells was purified by FLAG affinity purification of whole cell lysate from infected *T. ni* cells. The captured protein was washed on resin using 600 mM NaCl buffer, and eluted using FLAG peptide. Pooled elutions were further purified by ion exchange chromatography.

4.2.2 Qualitative fingerprint gel shift procedure

In order to assess the specificity of WDR76, I subjected the purified protein to the same competitive EMSA as before for brain extract. These assays reveal that WDR76 is an hmC-specific binding protein, that importantly, can shift hmC in the presence of a large excess of mC and C cold competitors (fig. 4.2). Asymmetric presentations of hmC are also bound equivalently as symmetric hmC under these conditions. Interestingly, WDR76 does not bind fC in the presence of any cold competitor. I have consistently observed both hmC-specific binding and a lack of binding for fC for different WDR76 constructs purified from insect cells (FLAG-HA-WDR76 and GST-WDR76) and *E. coli* (His6-WDR76). Separation of hmC and fC binding activities was not clearly seen in native extract. However, shift mobility and cold

competitor sensitivity were noted for hmC and fC that led me to hypothesize that distinct activities for each of the [ox]mC modification states might exist. Under these conditions, it is clear that fC is not well tolerated by WDR76 because when it is used as a cold competitor, shift for hmC increases (fig. 4.2 panels A and B) as does non-specific binding to mC and C (fig. 4.2 panel A). However, WDR76 was observed by mass spectrometry in pulldown experiments using fC. It may be that in pulldown experiments where only one oligonucleotide type is presented to a pool of proteins, binding to a less well preferred substrate is possible, particularly when fC is paired opposite hmC. Similarly, the SILAC experiments of Spruijt and colleagues [Spruijt et al. 2013] report that WDR76 binds to hmC and fC, perhaps because the pulldown captures WDR76 from a complex pool with no other oligonucleotide choices. My EMSA experiments (and later pulldown experiments from extracts taken from cells expressing WDR76) clearly show that purified WDR76 does not bind fC under equilibrium binding conditions or in the presence of biologically relevant cold competitors. While it remains possible that another factor in fractionated extract and in the cell could alter the activity of WDR76, my assessments with the isolated purified protein indicate that it only binds hmC.

Every WDR76 construct I have attempted to shift has shown hmC specificity, but observing strong shift does require careful tuning of the gel percentage (typically higher than what I used for brain extract, at 8-10%) and cross-linker ratio (37.5:1 performing better than 19:1, which is otherwise preferred for brain). These conditions were varied in order to maximize the resolution of the shift within the gel. Lower gel percentages and ratios produced more smeary shift. Under these higher percentage conditions, a very high mobility shift is observed that is both protein and oligonucleotide identity dependent fig. 4.2. However, the activity mobility does not match the relative mobility of the major [ox]mC activity (as read out by the DNA ladder) in native extract or the shift mobility of other constructs of WDR76 (GST-WDR76, His6-WDR76) examined under other EMSA conditions.

4.2.3 Semi-quantitative titration gel shift procedure

As previously described for purified extract, WDR76 was subjected to a semi-quantitative titration EMSA to visualize its relative affinity for C, mC, and hmC. When radiolabeled oligonucleotides of symmetric C, mC, and hmC are titrated in the presence of a constant amount of cold competitor, shift is observed for hmC at lower concentrations than for C and mC. In this experiment shown in fig. 4.3, the hmC oligonucleotide is relatively under-labeled as compared to mC or C, even though the same molar concentration is used in side by side experiments. Given this, the amount of shift at low concentrations is likely higher than can be visualized by this diminished amount of radioactivity per mole of hmC. Setting this caveat aside, the last concentration at which comparable shift is observed for C, mC, and hmC are 100nM, 100nM, and 20nM respectively. Given that this shift occurs in the presence of 40μM cold competitor, the ratio of radiolabeled C, mC and hmC DNA to cold competitor at these low concentration shifts is 1:28, 1:28, and 1:140, respectively. This suggests that WDR76 has a five fold greater affinity for hmC than mC and C under these conditions, and perhaps would have more when the degree of labeling is comparable. This degree of specificity is likely sufficient to confer unique localization of WDR76 to sites of hmC even when it is present at five fold lower abundance than mC.

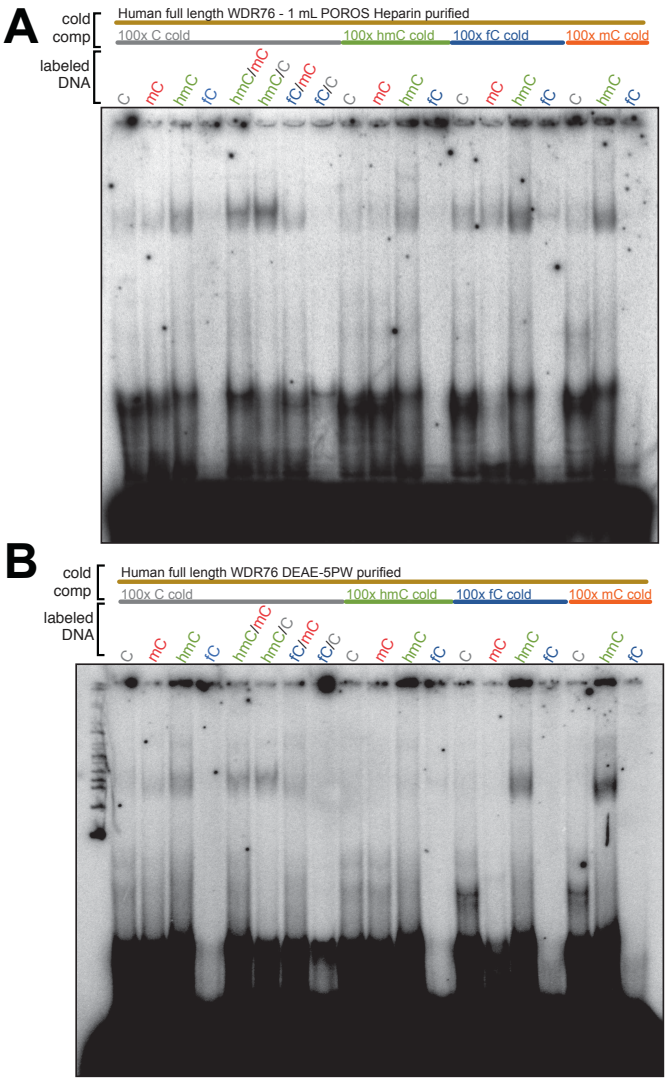


Figure 4.2: WDR76 is an hmC specific binding protein. WDR76 was subjected to the same competitive EMSA as used before for brain. Specific shift for hmC and asymmetric presentations of hmC are observed. Two EMSAs are presented that use differently purified protein inputs from the same source: N-terminally tagged WDR76 was purified from baculovirally transduced *T. ni* cells using FLAG-affinity resin followed by one the 1mL POROS Heparin column (A) or additionally with the 0.3mL DEAE-5PW column (B). Both preparations have limited specific activity, but show the same specificity for hmC as compared to C, mC and fC. Importantly, this shift persists in the presence of 100 fold molar excess of unmodified C and mC cold competitor, supporting tremendous specificity of WDR76 for hmC. hmC provided as a cold competitor also diminishes shift for radiolabeled hmC, particularly in (B). Interestingly, WDR76 does not shift fC, and fC cold competitor increases shift for radiolabeled hmC in (B) and increases background shift for mC and C in (A). This suggests that fC is not well tolerated as a substrate.

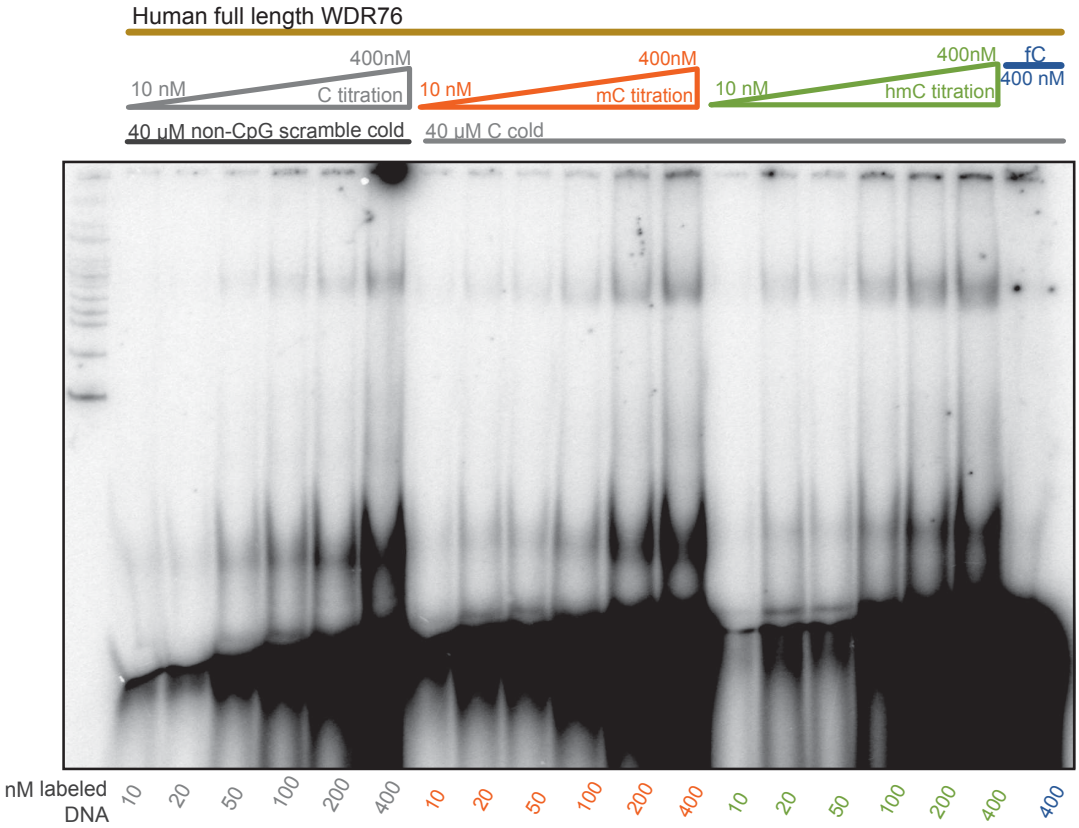


Figure 4.3: WDR76 shows specificity for hmC at lower concentrations than for mC or C. Titration of C, mC and hmC in an EMSA with WDR76 allow approximation of its relative affinity for each modification state. Despite radiolabeling in side-by-side reactions, the hmC duplex incorporated less radioactivity per mole than the mC and C duplexes, and thus the slightly greater apparent shift observed here, is an underestimate of the hmC-specificity. In order to assess the specificity for unmodified C in a CpG dinucleotide context, shifts with radiolabeled unmodified CpG containing DNA were challenged with a cold competitor in which the same sequence had been scrambled to not include any CpG dinucleotides. The major shift has an electrophoretic mobility of approximately 600 bp, similar to that of the native activity. A higher mobility shift is also observed that is both protein and oligonucleotide identity dependent (this shift is notably not observed with oligonucleotides containing fC modifications).

4.3 Expression of WDR76 by stable transfection of HEK293 cells

In order to assess the specificity of WDR76 by an orthogonal means, I generated stable cell lines expressing tagged fusion proteins of WDR76. Protein recovered from these cells could be used to assess both a distinct preparation of WDR76 and determine whether WDR76 is active in a cell extract. The WDR76 construct described above was subcloned into a pCDNA5-derived mammalian expression vector to yield an N-terminal FLAG-HA tag fusion under control of a CMV promoter, followed by a hygromycin resistance gene, and flanked by FRT donor sites. The construct was co-transfected into HEK293 FRT cells with Flp recombinase (pOG44) using FuGENE HD (Promega). Single clones were isolated and amplified under hygromycin selection at 150 μ/mL and assessed by blotting described below.

In order to investigate the direct contribution of WDR76 to hmC-specific binding in whole cell extract, I sought to design a mutant of WDR76 that was likely to disrupt DNA binding (albeit perhaps not hmC-specific binding exclusively), while still allowing the protein to fold and associate similarly with chromatin. To design such a mutant, the similarity between WDR76 and a close characterized homolog, DDB2, was analyzed. Human DDB2 bears approximately 55% homology and 25% identity to human WDR76 within the C-terminal WD repeat domain, including regions DNA recognition and basic residues forming phosphodiester contacts with the DNA backbone. In order to design a WDR76 mutant with altered DNA binding capacity, the conserved lysine and arginine residues within the DDB2 recognition loops were targeted. There are certainly many types of DNA binding null mutants that one could envision creating based on this homology, ranging from dramatic truncations of putatively involved domains, to subtle modifications that I might hope could disrupt hmC-specificity while leaving other non 5-position DNA binding capacity intact. The approach I took sought to disrupt residues that likely contributed to the binding energy of WDR76 to DNA in a non-sequence or modification specific fashion, but rather through phosphodiester contacts predicted by DDB2.

Mutants of the full-length WDR76 construct were generated by site-directed mutagenesis or through reassembly of the Gibson assembly construct using alternative gene blocks to encode the mutations. The mutant constructs were individually transfected into HEK293 FRT with FLP recombinase and screened with hygromycin as described above. All HEK293 cell lines were grown in DMEM supplemented with 10% FBS, 10 units/mL penicillin, and 10 mg/L streptomycin at 37° C in 5% CO_2. All HEK293 FRT lines were validated by blotting whole cell lysates for the FLAG tag epitope with either monoclonal α-FLAG-HRP, Sigma-Aldrich) or 12CA5 monoclonal antibody serum and observation of the FLAG-HA-fusion constructs at the appropriate molecular weight. Each mutant was examined for expression levels comparable to wild type as assessed via loading control and subcellular fractionation control levels as visualized by Rbbp5 and H3 blotting in lower sections of the same blot.

4.3.1 Oligonucleotide pulldown assay

HEK293 cell lines stably transfected with FLAG-HA-WDR76 constructs, both wild type and mutant, were grown under hygromycin selection at a concentration of 150 μg/mL. To perform oligonucleotide pull down assays, a modified Dignam-Roeder nuclear extract and whole cell extract were prepared. Equal cell numbers of wild type and mutant cell lines were counted, washed in PBS, and collected by centrifugation at 500xg (Sorvall Legend XTR with TX-750 rotor) for 10 minutes at 4° C. The cells were then resuspended 3x the packed cell volume in 10 mM Na·HEPESpH 7.8, 10 mM KCl, 1.5 mM MgCl2, 340 mM sucrose (33% w/v), 10% glycerol (v/v), 0.5 mM PMSF, 0.5 mM DTT, and 1x protease inhibitor cocktail (hypotonic buffer: HB). The cells were lysed by addition of an equal volume of HB supplemented with 0.2% Triton-X100 (v/v). The lysis proceeded during an incubation at 4° C with gentle end-over-end rotation for 12 minutes. Cell lysis was confirmed under the microscope using Trypan Blue staining.

To prepare whole cell extract, the salt concentration of the cell lysate was increased

to 0.6M NaCl and the mixture incubated for 30 minutes at 4° C with gentle end-over-end rotation. The whole cell lysate was then clarified by centrifugation at 15,000xg for 30 minutes at 4° C (Beckman Coulter 22R microfuge). The clarified lysate was collected and the salt concentration lowered to 200 mM NaCl by dilution with HEGN0 (no NaCl). The adjusted lysate was again clarified by centrifugation at 15,000xg for 30 minutes at 4° C (Beckman Coulter 22R microfuge). This whole cell extract was the input for oligonucleotide pull down assays.

To perform oligonucleotide pull down assays with HEK293 cell lines, 800 μL of whole cell lysate was added to 100 μg M280 streptavidin beads (Life Technolgies) pre-loaded with biotinylated modified oligonucleotides as described previously for the duplex oligonucletotide pull-down assays from fractionated brain nuclear extract. The lysate was incubated with the loaded beads in siliconized tubes for rotating for 1 hour at 4° C and the flow through was collected by isolation of the beads on a neodymium magnetic rack. The beads were washed with 500 μL of HEGN400 for five minutes, three times, with two intervening tube changes. The proteins retained on the beads were eluted with 10 μL of 4x SDS-PAGE buffer and resolved by gel electrophoresis. The retained proteins were transferred onto 0.22 μm Immobilon-PSQ membrane (Millipore) with a semi-dry electrobloting apparatus (Bio-Rad) in Towbins buffer and visualized by blotting with antibodies against the FLAG (M2-HRP, Sigma-Aldrich) or HA (12CA5) epitope tags using chemiluminescent detection within the linear range by ECL Ultra (Lumigen) on an LAS 4000 Imager (Fuji).

To confirm fair loading of biotinylated DNA onto the streptavidin beads used in the pull down assay, prepared beads were subjected to radiolabeling. For each oligonucleotide type, 5 μL of prepared bead slurry was incubated with T4 PNK (NEB) overnight at 37° C with 50 pmoles of [α-^{32}P] ATP. The labeling reaction was then heated to 90° C for 5 minutes with 1:1 addition of 95% formamide denaturing loading buffer. This reaction was then clarified by centrifugation and placement of the reaction tube against the neodymium magnetic tube

rack. 1 μL of the 20 μL labeling reaction was then loaded onto a thermally equilibrated 22% polyacrylamide gel with 6M urea and 1x TBE. After electrophoresis, the gel was dried and exposed to a Fujifilm 'CR' phosphorimaging plate for one hour.

4.3.2 WDR76 binds symmetric hmC by DNA pulldown

The blotting results show that FLAG-HA-WDR76 expressed in HEK293 cells specifically binds symmetric hmC. Very little binding is detected for symmetric C and mC under these washing conditions, and for hmC/C and fC/C. This assay confirms WDR76's preference for hmC over C, mC, and fC.

However, the asymmetric binding potential of WDR76 is not entirely clear. In the EMSA, shift is detect for hmC/mC and hmC/C in the presence of a large excess of unmodified C cold competitor. In this format, the shift for hmC/C is arguably the strongest observed with POROS Heparin purified protein. In the pulldown assay, the binding to symmetric hmC predominates in three replicates, with minor binding to asymmetric [ox]mC/C. To interpret these seemingly conflicting results, it is necessary to note the critical differences between the assays. The input to the EMSA is a single highly purified protein, albeit of limited specific activity. The EMSA uses cold competitor to challenge the binding of WDR76 to radiolabeled substrates and help solubilize it during the assay. The pulldown assay uses a crude whole cell lysate from a WDR76 overexpressing cell line, and no cold competitor is present. A consequence of this is that during the pulldown assay only one type of oligonucleotide is present, so the binding is not challenged so much by excess of a competitor as by the rate of re-binding equilibration during each bead wash. Importantly, the whole cell lysate also provides many other proteins (albeit at lower concentrations and therefore presumably not at stoichiometric ratios to the overexpressed WDR76) that may modulate the activity of WDR76, therefore making the two assays less comparable.

As such, it may be that in the EMSA format the presence of excess unmodified symmetric

C, an excluded substrate, drives the equilibrium binding toward less well favored but tolerated modification states, such as hmC/mC and hmC/C. This hypothesis could be tested by supplying cold competitor in the pulldown assay, or using different identities and amounts of cold competitor in the EMSA (although a great deal is already needed for solubility of the EMSA reaction, so the amount likely cannot be far less). These experiments provide a qualitative and relative assessment of the binding preferences of WDR76 as a function of cold competitor. ChIP-seq studies in combination with high resolution TAB-seq data may also help reveal whether the binding of WDR76 within the genome is at predominantly symmetric or asymmetric sites of hmC. Finally, affinity purification of WDR76 from a cell type in which it is expected to function normally, such as mouse embryonic stem cells, may reveal co-purifying partners that can be identified by mass spectrometry. These factors may modulate the activity of WDR76, or improve its specific activity. Ideally, the recombinant expression and purification of WDR76 as a single protein, and perhaps with a relevant co-factor identified from cells, should be improved to allow quantitative affinity measurements of its preferences for isolated substrates.

4.3.3 WDR76 triple positive charge mutant does not bind symmetric hmC by DNA pulldown

In order to confirm that the binding of WDR76 to symmetric hmC observed from HEK293 cells is attributable solely to WDR76, I repeated the DNA pulldown using the positive charge mutants described above. I found that the triple positive charge mutant of K, K, R to serine disrupted specific binding of WDR76 to symmetric hmC containing DNA. Single and double mutants examined in the pulldown did not disrupt binding.

It is worth noting that longer exposure of the triple mutant pulldown side by side with the wild type pulldown does show faint but equal binding to symmetric C, mC, and hmC by triple mutant WDR76. This suggests that the mutant protein is capable of weakly binding

DNA, but does not show preference for cytosine modification state. Importantly, the WDR76 triple mutant is comparable expressed, soluble, and associated with chromatin as the wild type protein in sub-cellular fractionation blotting experiments. Collectively, these results suggest that the triple charge mutant WDR76 protein is available in the cell for associating with other proteins tethered to chromatin, and may itself bind DNA with specificity for cytosine modifications. This makes the WDR76 triple charge mutant a useful comparison point for studies of WDR76 localization by ChIP-seq. In experiments investigating the function of WDR76's hmC-specific binding, the WDR76 triple charge mutant should not rescue a phenotypic effect that is due to hmC-specific localization, but should be available to bind with other partners in the nucleus or as a scaffold for other proteins on chromatin.

One reaction to the WDR76 positive charge mutant results that I'd like to address is the sense that it is somewhat incongruous with the observation that the bulk native [ox]mC-specific activity is highly resistant to salt. Some audiences have commented that they therefore do not expect a charge mutant that presumably disrupts ionic contacts to be in line with the native activity (which in bulk doesn't seem to rely upon ionic contacts given that it functions at high salt). However, it is important to note that the native activity almost certainly conflates many different protein activities, and that the compound shift might not be effected by salt concentration even if a few proteins within the shift are sensitive to salt. It would be interesting to subject wild type WDR76 to a salt challenge to determine if it is as resilient as the native activity. Similarly, the effect of the triple mutant may not be as drastic under low salt wash conditions.

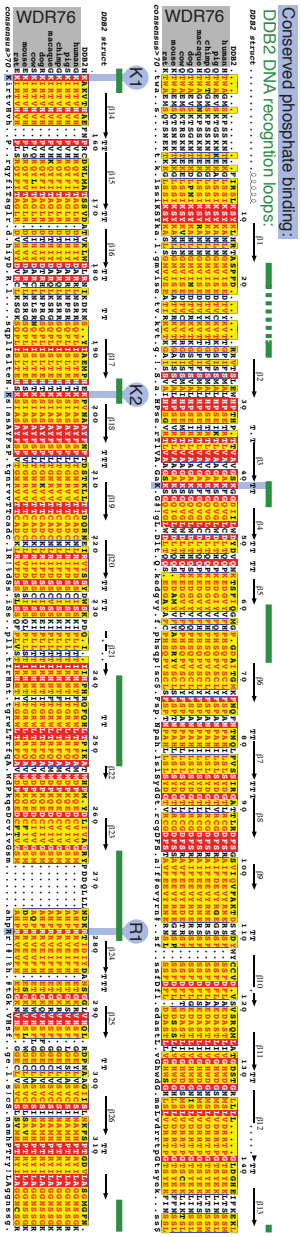


Figure 4.4: Alignment of the protein sequence of zebrafish DDB2 with the protein sequence of WDR76 from various mammals. The WD repeat 'blades' are relatively well conserved but the loops between them, including the loops which recognize DNA specifically in the case of DDB2 [Scrima et al. 2008] are divergent. Positively charged residues targeted in mutagenesis are noted.

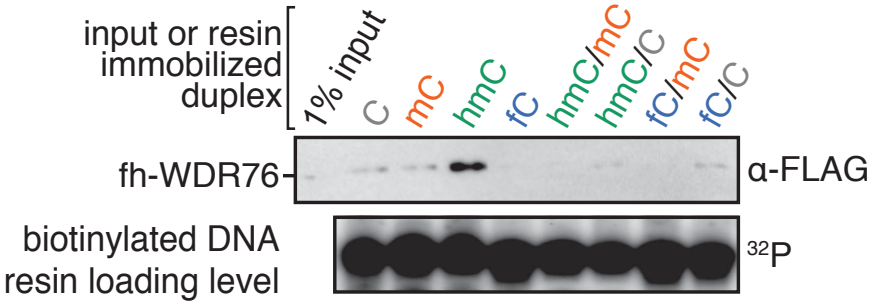


Figure 4.5: WDR76 specifically binding symmetric hmC-containing DNA. Weak binding is detected for symmetric C and mC, and for asymmetric hmC/C and fC/C. This western blot used M2-FLAG antibody to visualize recovered FLAG-HA-WDR76 after oligonucleotide pulldown from whole cell lysate of HEK293 cells stably transfected with FLAG-HA-WDR76. Loading levels of immobilized DNA were assessed by post hoc radiolabeling of bead slurry.

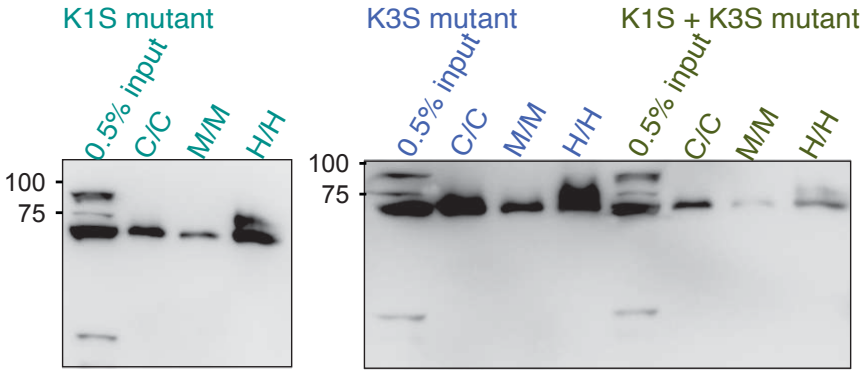


Figure 4.6: Select single and double positive charge mutants of WDR76 do not disrupt hmC-specific binding in the DNA pulldown assay. The mutations of K1S, K3S, and K1S+K3S were generated by site-directed mutagenesis into the mammalian expression constructFLAG-HA-WDR76 full length, and stably transfected into HEK293 cells. Whole cell lysates were subjected to the same DNA pulldown assay using symmetric C, mC, and hmC oligonucleotides and an HA immunoblot using mouse 12CA5 serum. Binding to hmC is not disrupted by these mutations.

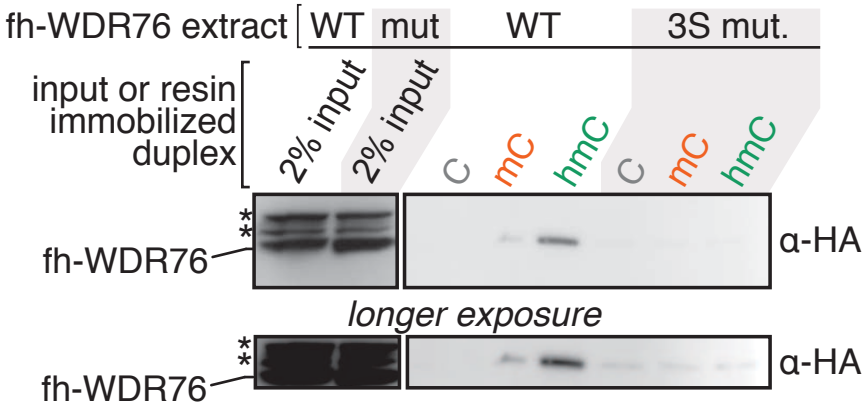


Figure 4.7: Mutation of three positively charged residues on the predicted DNA binding interface of WDR76 to serine (mut) disrupts DNA binding without altering expression level, solubility, or degradation pattern relative to WT protein in the same experimental format as panel fig. 4.5 but blotted with mouse 12CA5 serum. Three nonspecific bands in the HA immunoblot input are indicated(*). Weak but equivalent nonspecific binding is detected for C, mC, and hmC in the longer exposure of the same blot.

Chapter 5

FUNCTIONAL STUDIES OF WDR76

5.1 Biological questions of focus for assessment of [ox]mC-specific binding functions

The biology of [ox]mC has been greatly informed by sequencing studies of its distribution and abundance, and studies of the TET and TDG enzymes that generate and process these modifications. Having found an hmC-specific binding protein in WDR76, two critical questions remain to be answered before we can truly make sense of the presence [ox]mC species in the genome. The first is whether a subset of these modifications are specifically bound by unique factors. The second is whether such binding events are important for biological functions.

The biochemical fractionation results and validation of WDR76 strongly support the idea that [ox]mC species can be specifically bound by unique factors with affinity that could localize these factors to rare sites of [ox]mC. It remains to be seen whether the ability of [ox]mC-specific proteins such as WDR76 to bind to sites of [ox]mC in cells confer important gene regulatory effects that are dependent on the act of binding. While TET and TDG phenotypes have been reported that are largely due to accumulation of methylation in the absence of these enzymes, no evidence currently speaks to the potential functions that could emanate from specific recognition and binding of [ox]mC. Having identified WDR76 as an hmC-specific binding protein, I sought to understand its function and investigate whether this binding event has important consequences in the cell, and if so, does it act at particular sites of hmC. Moreover, it would be useful to determine whether other chromatin or sequence features of these sites (beyond hmC) promote WDR76 binding. Such results would greatly advance the field's thinking about [ox]mC modifications as signaling molecules for gene regulation, and not merely intermediates in the turnover of mC.

5.2 Known biology of WDR76

Little is known about the function of WDR76. It is well conserved across vertebrates but less well so across all metazoans. The WD repeat region, as previously mentioned, shares homology with the DNA repair protein DDB2 fig. 4.4. However, the N-terminal region, which is predicted to be predominantly alpha-helical, is less well conserved. WDR76 may associate with the DNA helicase HELLS and the H3K4me3 binding protein Spindlin-1, albeit this is taken from co-immunoprecipitation data from HeLa [Spruijt et al. 2013]. WDR76 may also act as an adapter for the cullin family of ubiquitin ligases involved in DNA damage responses [Higa et al. 2006]. A final curious signature of WDR76 is that it is found enriched at sites of H3K27 acetylation and H3K4 trimethylation in mouse ESCs [Ji et al. 2015]. This reason for this signature is unknown but invites the possibility that it overlaps with hmC abundance at actively transcribed enhancers. Interestingly, in MLL-rearranged leukemias in which TET1 overexpression is a known target of MLL-fusions and critical driver of leukemia progression [Huang et al. 2013], WDR76 is also highly expressed (Jianjun Chen, personal analysis of publicly available data). This correlation invites the possibility that the hmC-specific binding activity of WDR76 contributes to oncogenesis, and that understanding the binding of WDR76 at altered sites of hmC could help explain gene expression perturbations in these leukemias.

Though I identified WDR76 through purification of a brain extract, there is good reason to believe that hmC-specific proteins like WDR76 would also function in embryonic stem cells, a cell type also enriched in [ox]mC. Given that WDR76 is moderately well expressed in brain and ESCs, the accessibility of ESC culture for biochemical and functional studies, and interest in the biology of [ox]mC in ESC, I elected to work in mouse embryonic stem cells to begin to discern the function of WDR76.

5.3 Studies of WDR76 in mouse embryonic stem cells

5.3.1 Culture of E14 mouse embryonic stem cell lines

E14 mouse embryonic stem cells were obtained from Bruce Lahn's lab. The original aliquots I received had undergone 18 passages since isolation from a mouse blastocyst. The cells are grown under conditions that preserve their pluripotency, namely, in presence of the leukemia inducible factor LIF, to promote Jak/STAT3 signaling and high expression levels of the pluripotency transcription factors Oct4 and Nanog [Williams et al. 1988], and with '2i' inhibitors of MEK/ERK1/2, FGF4, and GSK3, preventing activation of Wnt and β-catenin signaling.

Cells were cultured on gelatinized plates using DMEM supplemented with 15% heat-inactivated fetal bovine serum, 2 mM L-glutamine, 0.1 mM non-essential amino acids, 50 units/mL penicillin, 50 μg/mL streptomycin, 0.2 mM β-mercaptoethanol, 1000 units/mL LIF, and 2i components at 1 μM (MEK inhibitor PD0325901) and 3 μM (GSK inhibitor CHIR99021). Cells were incubated at 37°C and 5% CO%_2 in a passively humidified incubator. I consulted general protocols for the care and use of mouse E14 embryonic stem cells available from the Mouse ENCODE Consortium. Cells were passaged every day to preserve their health and limit differentiation at high confluence, which should not exceed 50%. Briefly, every morning, 80% the media from the night before was replaced with fresh complete media. The removed media was collected and filtered through a 0.1μm filter and saved as 'conditioned media'. This media contains factors secreted by the cells when they are growing well. After 3-6 hours, the cells were passaged by dissociating colonies to single cells using 0.05% trypsin. Cells were re-plated at between 30-50% confluence depending on the next desired application. Cells were plated in small volumes of fresh completely supplemented media and 25% conditioned media in order to promote pluripotency. The small volumes ensure the concentration of the factors secreted by the cells can increase rapidly

and promote attachment and organization into colonies. The cells are observed to attach to plates and form colonies 4-5 hours after plating. At the end of the day, 2-4 additional volumes of completely supplemented media are added to keep the cells fed overnight. Mouse embryonic stem cells should not be left alone for more than 14 hours without feeding in order to preserve their health and pluripotency.

I also obtained a mESC E14 cell line into which an FRT site has been introduced. This allows for stable transfection via site specific recombination at this site when co-transfecting plasmids containing a gene of interest flanked by FRT sites and a plasmid encoding Flp recombinase. This cell line was provided by the Helin lab at the University of Copenhagen. The aliquots of these cells that I received had undergone between 30 and 35 passages and are cared for in the same way as unmodified mESC E14 cells.

5.3.2 Transfection procedures for generation of tagged FRT-based mES cell lines

To transfect FRT-modified mESC E14 cells, 2 million cells were prepared per transfection on 10cm diameter gelatinized tissue culture plates. While plating after a passage, plasmids were transfected using 60 μL of Lipofectamine 2000 (Life Technologies) provided in reduced serum media (Optimem, Life Technologies). The pOG44 plasmid encoding Flp recombinase was provided at a ratio of 40:1 with a plasmid encoding the WDR76 constructs of interest in a pCDNA5-derived mammalian expression vector as an N-terminal FLAG-HA tag fusion under control of a either a EF1 or CMV promoter, followed by a hygromycin resistance gene, and flanked by FRT donor sites. Typical transfections used 24 μg of pOG44 and 0.6 μg of WDR76-encoding plasmid. Wild type full length WDR76 and triple positive charge mutant full length WDR76 were transfected by this method under either a CMV or EF1 promoter. The cells are not split the next day, but media is replaced during feedings twice per day, morning and evening.

Two days after transfection, the cells are split to 15 cm diameter gelatinized tissue culture plates. Selection begins upon re-plating y adding 150 μg/mL of hygromycin. This drug concentration is maintained in morning and evening feedings for 10-14 days. Many cells are dying during this phase but are removed prior to feedings. After 14 days, isolated colonies can be observed growing in the presence of drug. These colonies are removed from the 15 cm by careful pipetting off of the plate while viewed under the microscope. The picked colonies are briefly trypsinized and re-plated into a single well of a 96 well gelatinized tissue culture plate. Isolated colonies are slowly expanded into vessels of increasing surface area. Hygromycin is omitted until the cells are growing well in a 24 well plate. Once colonies are expanded to 2 million cells, they can be frozen down in complete media (no drug added) supplemented with 10% DMSO in an isopropanol-insulated freezing chamber to -80°C and stored long term in a cryogenic freeze (-140°C).

Colonies can be analyzed by blotting for the protein of interest, and by genomic PCR and TOPO cloning, as described in the next section, to confirm proper insertion. By this method, I isolated stable mESC E14 lines encoding WDR76 wild type and triple positive charge mutant under the control of either a EF1 or CMV promoter.

5.3.3 Design of CRISPR-based tag knock-in and gene knockout constructs, transfection and validation

mESC E14 cells were subjected to CRISPR-based genome editing to (1) introduce a FLAG-HA tag at the endogenous WDR76 locus within the start of the protein coding sequence within exon 2, and (2) to create truncated 'knockouts' by targeting exon 11 to introduce stop codons such that the WD repeat and its folding are disrupted.

Several guide RNAs were designed and cloned as described [Cong et al. 2013] into pX330 at BbsI sites. Guide RNA constructs were transfected into cells using Lipofectamine 2000 at a ratio to 6:1 for 2μg of DNA per million cells, or using mFECT reagent (Clontech).

Transfected cells were plated at low density and single colonies picked, or sorted by FACS for GFP signal and re-plated in gelatinized 96 well plates as single cells, or the bulk sorted material was plated at low density and individual colonies were picked manually after 2-3 days, and re-plated in gelatinized 96 well plates.

When cells in 96 well plates are confluent, the entire plate is passaged into a new 96 well plate and half of the cell material is collected for analysis by PCR. After heating, sonicating, and vortexing the cell material to lyse the cells, genomic DNA is isolated from the cells using SPRI beads (Sera-mag speed beads, Fisher) at a ratio of 1:1 (v/v). This purified genomic DNA is used as a template for PCR of short regions of exons of interest and analysis of individual clones by agarose gel. Tag-in edited clones of interest are identified by insertions of an expected molecular weight based on the design of the tag, and the sensitivity of this amplicon to a restriction enzyme digest specific to the desired edited sequence. 'Knockout'clones are identified by insertions or deletions into exon such that no wild type amplicon is present. To confirm these manipulations by sequencing, I used a TOPO-cloning approach to clone amplicons after purification of the PCR reactions.

5.3.4 CRISPR-editing yields two biallelic WDR76 'knockouts' that disrupt expression of full length protein

By this method, after screening ∼600 colonies, I isolated two biallelic 'knockouts' of WDR76 The biallelic 'knockouts' are each composed to two distinct deletion alleles that shift the WDR76 open reading frame such that multiple premature stop codons are introduced in the middle of the WD repeat encoded by exon 11. The truncated gene protein product for each of the alleles of one of these knockout lines is shown in fig. 5.1. These truncated protein products are not expected to fold well given that several blades of the WD repeat are omitted, leaving the protein susceptible to proteolysis. The absence of the full length WDR76 protein product in mESC knockout lines was confirmed by Western blot using an

antibody targeted to the c-terminal region of the protein.

5.3.5 Preparation of RNA libraries for gene expression studies of WDR76 biallelic knockouts

In order to assess a potential effect of WDR76 on the expression of genes that are enriched in hmC, I prepared RNA libraries from wild type and WDR76 'knockout'cells grown in parallel and isolated in triplicate. Briefly, cells were grown side by side at similar confluence over the course of five days. Twelve million cells of each genotype were collected. Total RNA was isolated in triplicate for 4 million cells each by Trizol-chloroform extraction and purification using Zymo RNA columns (Zymo Research). For each sample, 2 μg of total RNA were processed using Ribo Zero (Illumina) ribosomal RNA depletion beads to remove abundant rRNA. Prior to rRNA depletion, however, four *in vitro* transcribed RNA standards were added to act as an internal calibration ladder. These RNA standard transcripts are derived from yeast and bacterial genes that will not map to mammalian genomes, but provide a way to calibrate the reads from sequencing of the entire library and estimate relative abundance between samples. The reagents and methods for this RNA standard doping method were developed and prepared by a fellow graduate student in the lab, Michael Werner. I added his prepared standards to my RNA libraries at 40 copies per cell equivalent (yRAD51) 200 copies per cell equivalent (RNL2), 1000 copies per cell equivalent (MBP), and 5000 copies per cell equivalent (ySUMO). Assuming the typical range of transcript reads measured by an RNA sequencing experiment on this scale, these RNA standards should provide a calibration ladder within the linear range of reads measured.

RNA libraries were prepared from rRNA depleted RNA using the NEB Next Ultra Directional RNA Library Kit and NEB Next Multiplex index oligonucleotide primers for use with Illumina sequencing platforms. The library was prepared according the manufacturer's instructions and amplified using 15 cycles of PCR. Indices for each library sample were

chosen to minimize the Hamming distance between index reads in order to ensure proper assignment of reads to samples after sequencing.

At the same time, fellow graduate student, Michael Werner, also pursued knockdown of WDR76 in K562 cells using CRISPRi. K562 cells are an immortalized myelogenous erythroleukemia line, of potential interest given the known role of TET1 in acute myelogenous leukemia [Huang et al. 2013]. Michael Werner designed five CRISPRi guide RNAs and three non-targeting controls, cultured the cells, performed the transfections, collected the cells, and purified total RNA. This CRISPRi (or CRISPR-interference) approach only yields knockdown and not knockout of targeted loci, as the catalytically-inactive cas9:gRNA complex is bound to homology sites such that it precludes transcription, but does not cut or alter the DNA sequence. Michael has observed that WDR76 is moderately well expressed in K562 cells in his own experiments (to a degree on par with its expression in mESCs), and we were both interested in using a knockdown strategy to investigate its function. I prepared cDNA from the CRISPRi transfected cells and assessed the degree of knockdown by qPCR. I found that the guide RNAs knocked down WDR76 expression by 4.4 fold on average compared to non-targeting controls. I also prepared these RNA samples as libraries for sequencing as described above for the three most potent guide RNAs and the three non-targeting controls.

These RNA libraries were submitted for single end 50 basepair next generation sequencing at the University of Chicago Functional Genomics Core Facility. The mouse embryonic stem cell and K562 experiments were each process in an individual lane of the Illumina HiSeq4000 instrument.

5.3.6 Analysis of gene expression in WDR76 knockout mouse embryonic stem cells

RNA-seq fastq result files were processed using the Tuxedo Suite of software for alignment and differential expression analysis. Bowtie2.2 was used to concatenate fasta files for the

RNA standards to the mouse mm9 FaMasked genome obtained from the UCSC genome browser. Reads were then aligned to this genome assembly using Tophat 2.1.0 [Trapnell et al. 2012; Kim et al. 2013]. Transcriptomes were then assembled de novo for each replicate using Cufflinks2 [Trapnell et al. 2012] with an rRNA, snoRNA, tRNA, and 7SK mask. The three wild type mESC biological replicate transcriptome assemblies and the three Wdr76-/- biological replicate transcriptome assemblies were then merged to a single de novo transcriptome for differential expression analysis using Cuffdiff. The Cuffdiff analysis was performed using per condition cross-replicate dispersion estimation and either classic FPKM normalization in order to allow the standards to be uniquely mapped without further normalization, or using geometric mean normalization. The performance of the RNA standards in the sequencing experiment was assessed by counting the number of reads mapping to the standards in the BAM file output of Tophat using by using Samtools [Li et al. 2009]. These measurements revealed consistent mapping of the standards during data processing; the number of reads observed for each standard was plotted against the number of molecules added for each standard to yield a calibration curve for each of the samples (Fig. S16A). The average slopes of the standard curves for wild type and knockout samples were calculated separately and applied as a scalar to the FPKM values calculated by Cuffdiff (or raw counts per gene), resulting in scaled FPKM values that reflect a back-calculation of actual number of molecules present in the library input. Due to modest changes in RNA quantification between the internal standard-spiked library normalized analyses versus the geometric mean normalized analyses (Fig. S16C), I conclude that the effects of Wdr76 knockout are not global [Lovén et al. 2012], and therefore well-treated using the latter method. Accordingly, we performed subsequent mESC and K562 analyses using geometric mean normalization dataset. In the latter, we treated the knockdowns with three distinct gRNAs as a single condition as compared to the combined off-target gRNA replicates for per-condition dispersion estimation using the GRC37 reference genome.

A combination of Cuffdiff [Roberts et al. 2011; Trapnell et al. 2010], Bedtools 2.25 Quinlan:2010km, CEAS [Shin et al. 2009] and shell/Awk scripting were deployed to interrogate the connections between altered gene expression in the Wdr76 depletion experiments and the previously reported hmC TAB-seq in the same E14 cell line [Yu et al. 2012a]. Refseq coding genes were assessed for hmC density in subsets defined by significant log2 fold changes (FDR < 0.05; p < 0.003), as compared to random shuffles of the KO down coordinates in bedtools restricted to be within the set of all coding genes or defined expression quantiles. The two different E14 TAB-seq datasets used generally recapitulated the same trends - either the all-reported hmC set or the FDR < 0.05 subset [Yu et al. 2012a]. For each, the hmC density was computed as the %hmC for each site summed over a given interval (either genes or TADs), divided by its span then mapped onto the indicated genes using bedtools. CEAS was used to contour hmC abundance over metagene sets representing significantly down-regulated or up-regulated genes in the *Wdr76-/-* versus the E14 WT line, or all genes. The Hind III two-replicate combined Hi-C dataset representing E14 contact domains [Jin et al. 2013] or K562 Hi-C contact domains filtered to be smaller than 10Mb [Rao et al. 2014] were used for TAD calculations. Briefly, the density of hmC within the whole interval of each TAD was computed and rank-ordered for division into the indicated quantiles. Bedtools intersect was used to count the number of overlapping genes from those significantly altered in the knockout , or bedtools map was used to compute the TAD hmC density in which the indicated gene sets reside. Plots of average WT FPKM of genes within TAD as a function of genes significantly up- or down- regulated by knockout or CRSIPRi depletion, versus the genes defined by the middle log2 fold change percentile of expressed genes (>0.25 FPKM in both conditions) with (approximately number matched to the up and down gene sets in each case) were prepared in R.

Preliminary analysis of the mouse embryonic stem cell WDR76 knockout RNA sequencing results suggests that in the absence of WDR76, but that the cells remain pluripotent, as

was also observed in culture. While there are several individual genes of interest within the significantly differentially expressed set, some of which are secondary pluripotency regulators, ahead of validating the expression levels of these genes by quantitative PCR from reverse transcribed RNA, I do not seek to assert any of these genes as direct targets of WDR76 or a WDR76-based gene expression response. ChIP-seq analysis of WDR76 tagged mESC lines will be required to establish direct connections between WDR76, differentially regulated genes of interest, and particular sites of high hmC modification.

As an initial analysis, I sought to connect the differentially regulated genes to hmC levels in a general sense. Initial assessments of the hmC density of genes that are differentially expressed in *Wdr76-/-* mESCs indicates that these genes are preferentially enriched in hmC as compared to genic shuffles of the same spans, as shown in fig. 5.2. While this effect is slightly more significant for genes which are down regulated in the absence of WDR76, upregulated genes have a similar signature. While ChIP-seq studies will greatly inform what genes are directly responsive to WDR76, this trend suggests that the effects of WDR76 are correlated with high hmC levels, consistent with its specific binding of these modifications. The hmC signature is particularly localized to the 3' end of these differentially regulated genes, as shown in fig. 5.3, again more so in the case of up regulation as opposed to down regulation. Without knowing what regulatory processes WDR76 might specifically impact by way of hmC, this is an preliminary indicator of where hmC might localize WDR76.

It is worth noting that hmC within gene bodies is not as abundant as hmC within enhancer elements or proximal to promoters. Therefore, while it is promising to find hmC within differentially regulated genes in *Wdr76-/-* mESCs, ChIP-seq experiments with WDR76 will indicate whether this hmC signature is correlated with WDR76 occupancy. Ahead of that, however, it is also striking the degree to which the differentially regulated genes in *Wdr76-/-* mESCs tend to cluster together. This prompted an assessment as to whether these genes are concentrated in known chromatin structures that are known to be co-regulated,

such as topologically associated domains (TADs, [Jin et al. 2013]). TADs are chromatin structures formed from loops of 0.5-1 Mb of DNA containing 3-5 genes and typically an enhancer element. These loops are formed by the structural proteins CTCF and cohesin which scaffold the extruded loop and allow the enhancer element to make local contacts with the promoters within the loop. As such, genes in TADs tend to be coregulated at the transcriptional level. From analysis of TADs annotated in mESCs by Hi-C methods, I observed that the genes that are differentially regulated in *Wdr76-/-* mESCs are disproportionately found within TADs that have high levels of hmC (the top 20%) as compared to random genic shuffles of the same spans. These TADs also tend to be highly expressed, a trend which is recapitulated in the K562 knockdown RNA-seq results. This enrichment, shown in fig. 5.4 suggests that the absence of WDR76 impacts gene regulation at the chromatin level, focused on highly expressed TADs that are enriched in hmC.

These observations of (1) genic hmC density in differentially expressed genes and (2) overlap of differentially regulated genes with highly expressed TADs enriched in hmC both strongly suggest involvement of enhancer elements in WDR76-mediated gene regulation. Enhancer elements have been annotated in E14 mESCs based on various signatures, namely H3K27 acetylation, H3K4 monomethylation, and occupancy by core pluripotency transcription factors (OSK), Mediator components, and responsiveness to LSD1 decommissioning [Whyte et al. 2013]. In order to assess a possible connection between genes that are differentially regulated in *Wdr76-/-* mESCs and enhancers, I compared the relationship between the annotated set of "biochemically-responsive" enhancers recently described by the Young lab [Whyte et al. 2013] and genes of interest in the *Wdr76-/-* mESCs. This analysis revealed that there are many enhancers within these differentially regulated genes, and that the enhancer set also overlaps greatly with promoter elements. This made the analysis of hmC density within close enhancers difficult, as the closest enhancers often actually overlap these genes or their promoter elements. As a result, no clear unique relationship between differentially

regulated genes in *Wdr76-/-* mESCs, enhancers, and the hmC content of each emerged. Further analysis with ChIA-PET defined RNA Pol II enhancer-promoter looping [Kieffer-Kwon et al. 2013] also did not reveal significant unique trends. Here again, ChIP-seq studies will greatly inform the analysis by indicating which enhancer elements and enhancer-promoter contacts are most directly implicated by WDR76 occupancy. However, ahead of further studies, collectively these data suggest that Wdr76 plays a positive role in hmC-driven gene expression in mESCs, and suggest that it acts in a local chromatin context to achieve both gene activation and repression.

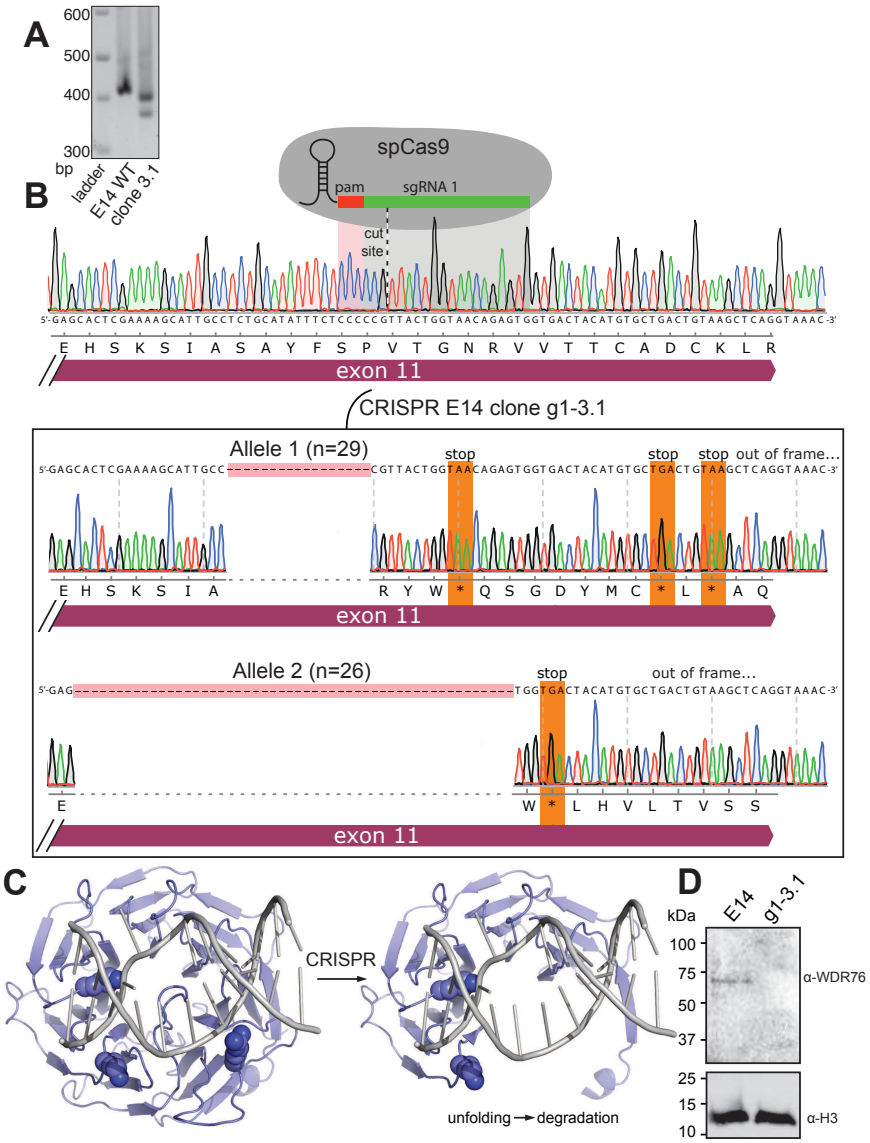


Figure 5.1: CRISPR-based gene editing was used to create knockout WDR76 mESC cell lines. (A) PCR amplification of exon11 within clone 3.1 yields two short products as compared to wild type E14s. (B) Sequencing confirms that this clone contains biallelic deletions within WDR76 exon 11. The span of the guide RNA sequence is shown across exon 11 of the mouse sequence of WDR76. Two deletions (number of sequencing observations denoted by n) within exon 11 have caused frameshift within this coding sequence, giving rise to premature stop codons. (C) These deletion alleles are predicted to give rise to a truncated WDR76 protein product. The structure of DDB2 bound to DNA [Scrima et al. 2008] is used as a homology model for WDR76 deletion alleles. Sites of the positive charge residues previously targeted in the hmC-binding null mutant are highlighted. (D) A full length WDR76 protein product is not produced in the 'knockout' mouse embryonic stem cell line. As compared with wild type E14 mESCs, no protein is observed using a native WDR76 antibody in cell line 3.1. Histone H3 loading levels are provided as a control.

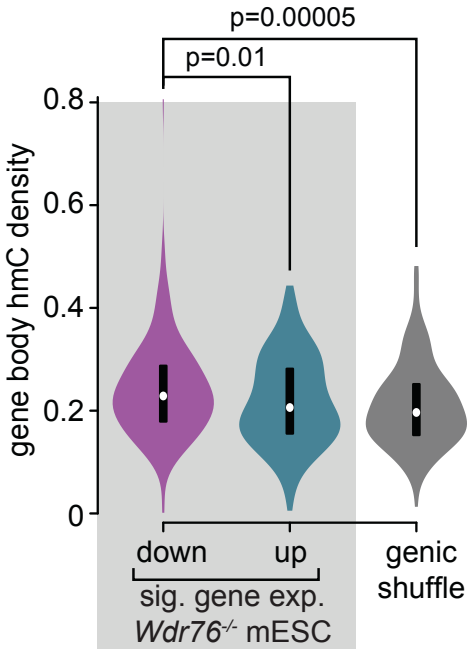


Figure 5.2: The coding genes which are significantly differentially regulated in the absence of WDR76 are preferentially enriched in hmC. Violin plot of hmC density (abundance values [Yu et al. 2012a] normalized by the length) within genes that are significantly down- or up-regulated in *Wdr76-/-* mESCs (FDR < 0.05; $p < 0.003$), as compared to the hmC-density within randomly shuffled genic coordinates with identical spans to the down-regulated set. Median values are rendered as white dots, the central two quartiles are spanned by black rectangles, and p-values computed via Mann-Whitney-Wilcoxon test.

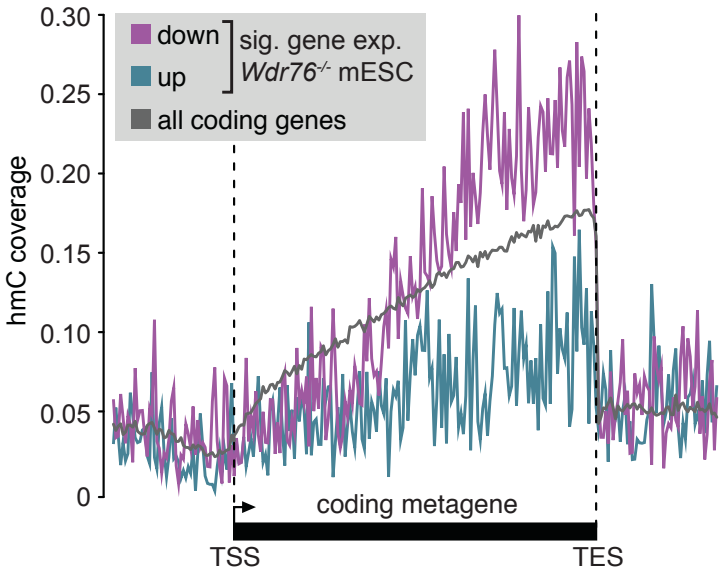


Figure 5.3: The abundance of E14 hmC contoured over coding metagenes for the indicated gene sets. The differentially regulated genes which are upregulated in the absence of WDR76 are particularly enriched in hmC toward the 3' end of the gene.

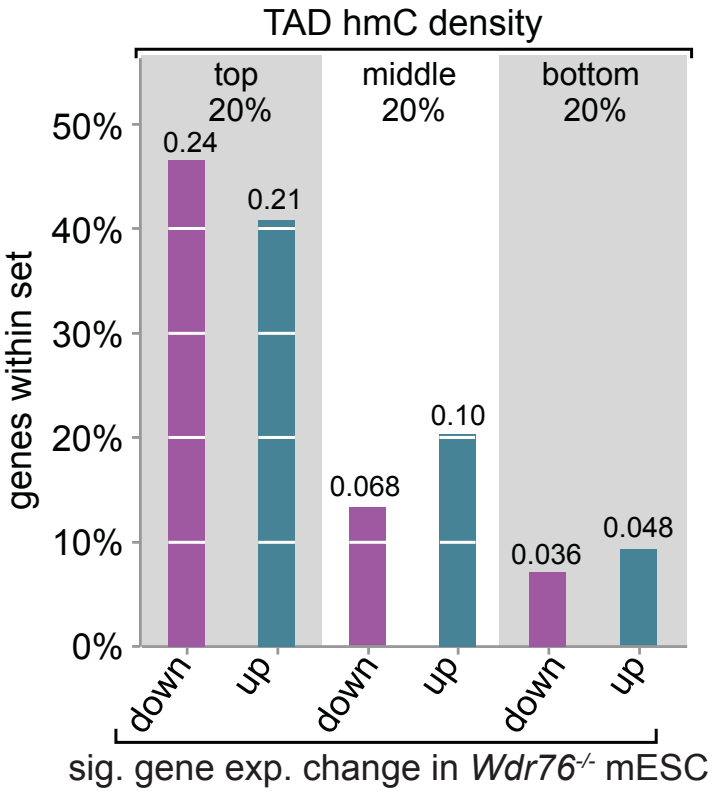


Figure 5.4: Genes which are differentially expressed in the absence of WDR76 are disproportionately in gene neighborhoods enriched in hmC. The percentage of significantly up- or down- regulated genes in *Wdr76-/-* mESCs within topologically associating domains (TADs, [Jin et al. 2013]) with the top, bottom, and middle quintiles of hmC-density. The average TAD hmC density for each set is indicated atop each histogram bar.

Chapter 6

CONCLUSIONS, IMPLICATIONS, AND FUTURE DIRECTIONS

The molecular basis of cell identity is in large part derived from chromatin modifications that determine the transcriptional activity of a given cell. Molecular biology, through genetic and biochemical studies, has revealed how many chromatin modifiers and modification-specific binding proteins determine how this physiological form of the genome is used. Recent reports of oxidized methylcytosine in mammalian genomes at gene regulatory features has inspired much work to define the distribution and abundance of these marks in cell types and disease states of interest. Critical to understanding these subtle and rare modification states is determining whether they are merely transient intermediates in the turnover of methylcytosine, or whether they are both *exist* and *are recognized in specific binding events* as molecular states distinct from the unmodified cytosine and methylcytosine from which they are derived. Sequencing technologies have advanced a great deal in their ability to map sites of [ox]mC modifications across the genome and suggest that these marks do have a stability within the genome that could support signaling, and are somewhat supported by the molecular phenotypes of TET knockouts. A key missing piece of our view of these modifications is a specific binding partner intermediary of such signaling.

Identifying such a binding partner requires attention not only to the potential of a candidate protein to bind [ox]mC modifications with some affinity, but critically, the ability of such a protein to bind these rare modifications with tremendous specificity that is commensurate with their relative abundance. While efforts to identify such proteins have been reported, all of these studies have been largely inattentive to the importance of affinity relative to fold abundance. No previous report of [ox]mC binding proteins has provided validation of the protein's intrinsic specificity for [ox]mC such that the protein could be expected to bind to [ox]mC modifications uniquely in the presence of the more abundant C and mC in the genome. Therefore, it has not been possible to conclude from these studies that binding pro-

teins exist which are specific enough to [ox]mC modified DNA that they could be expected to occupy sites of [ox]mC with a high enough probability that some signaling function could be derived from this unique localization.

My work provides crucial evidence of the existence of highly specific binding partners for each of the [ox]mC states in the mammalian brain, a tissue enriched in [ox]mC. The design of my biochemical fractionation approach used cold competitor to scrutinize nuclear extract for activities that could be as specific as the abundance of [ox]mC indicates is needed to be in order to bind these marks uniquely. Ahead of knowing what factors are responsible for this activity, my observation of [ox]mC-specific activities that can operate in the presence of biologically relevant excesses of cold competitor is important for simply showing that such activities exist. In a biophysical sense, the activities revealed by biochemical fractionation support remarkable discrimination between C, mC, [ox]mC and among the [ox]mC species, suggesting a collection of exacting molecular interfaces that afford meaningful distinctions in binding energy among very minor differences in substrates.

The relatively unbiased biochemical fractionation approach also allowed me to interrogate the properties of [ox]mC binding proteins in bulk extract before the particular identities of the proteins were known. In doing so, I uncovered binding preferences of the native activities that greatly informed my ability to isolate them. Namely, I discovered that native [ox]mC-specific activities can discriminate between asymmetric presentations of [ox]mC, C, and mC, and that they preferentially bind [ox]mC/C and not [ox]mC/mC. The degree of binding discrimination between these asymmetric forms is markedly greater than that observed for the symmetric presentations of these modifications. The ability of a given extract fraction to discriminate between these forms has been as useful hallmark as their ability to bind [ox]mC in the presence of C and mC cold competitors. This property guided both the fractionation procedure and the DNA pulldowns used to recover the proteins responsible for the activity as input to mass spectrometry.

The asymmetric binding preference of the native activities invites speculation as to whether these are the relevant modification states used within the cell for [ox]mC specific binding. This observation should inform future sequencing studies of [ox]mC and its distribution. During the course of my research, several groups have reported that hmC and fC are predominantly found in asymmetric contexts, and that the activities of the TET enzymes and TDG are such that asymmetric [ox]mC modifications are favored *in vitro* [Wu et al. 2014; Booth et al. 2014]. While WDR76 seems to prefer symmetric symmetric hmC above asymmetric presentations, I expect that other [ox]mC-specific proteins yet to be validated will discriminate based on these asymmetric modification states and that maintenance of asymmetric [ox]mC modifications may be important to any potential signaling pathway that emanates from [ox]mC-specific binding events. Some possibility of this is already suggested by the observation that hmC is preferentially found at the 5' splice site of alternatively spliced exon-intron junctions, and is enriched on the sense strand of highly expressed genes in the brain [Khare et al. 2012; Wen et al. 2014]

To advance this view of gene regulation by asymmetric [ox]mC, it remains to be investigated how the TET and TDG enzymes are targeted selectively to process sites of mC, outright, and given that mC is highly symmetric, how the activities of TET and TDG are regulated to yield asymmetric modification states. Generating particular asymmetric sites of [ox]mC requires careful regulation of both of these activities that is sensitive to the CpG context in which they occur. My results suggest that [ox]mC/mC and [ox]mC/C states should be created with a great deal of discretion because in bulk native extracts they give rise to different binding outcomes. How the TETs and TDG achieve this discretion remains to be seen, but strikes me as a challenging enzymology problem. While it is known that there is some cell type specific division of labor among the TET enzymes at the level of their expression and activity, there may be other factors that modify this activity at specific sites within the genome (or it may be intrinsic or particular to each of the TETs, which are

expressed at different levels throughout development and in different tissues). Such a factor might recruit TET or TDG activity to a particular site, and/or then modify their activity to rise to asymmetric sites. No known sequence motif preference has yet been noted for the TET enzymes [Shen et al. 2013; Yu et al. 2012a; Song et al. 2013], and they do rely upon where DNA methylation is already present, but given the particular distribution of [ox]mC in the genome, TET activity must be responsive to some elements of chromatin. In isolation, TET1 and TET2 are known to perform the first oxidation reaction converting mC to hmC faster than subsequent oxidation steps to fC and carC [Hu et al. 2013], so factors may also be needed to favor further oxidation reactions by the TETs in order to promote higher oxidation states that can be excised by TDG and generate asymmetric sites. I predict that in addition to delineating [ox]mC-specific binding-based signaling pathways for gene regulation, the field will also move to characterize the determinants of TET and TDG activity in cells. It could very well be that the [ox]mC-specific binding factors that I and others have pursued contribute to this process.

I am also interested to see whether certain TET phenotypes are also explained by [ox]mC binding events being sensitive to asymmetry. In cases where TETs are overexpressed, it could be that aberrant modification symmetry is as much a problem as simply elevated levels of [ox]mC modifications. This would be the case if the level of TETs were not in tune with the availability of factors that modify their activity. Based on my observations in native extract and other known binding characterizations of MBDs, excessive TET activity that generates aberrant [ox]mC/mC sites could change localization of symmetric MBDs, while aberrant [ox]mC/C and [ox]mC/C sites would attract binding proteins. In the absence of TETs, these states could also decay in ways that attract unwanted binding. This should be investigated as a potential mechanism of TET-based gene regulation in the contexts where TET activity is known to be abnormal, namely in leukemias and in some brain disease. I am particularly interested in whether TETs could be involved in the many neurological diseases

that same to derive from disruption of imprinted genes that are normally monoallelically expressed [Gregg et al. 2010; Santiago et al. 2014].

The biochemical fractionation strategy I pursued certainly has limitations that restricted the sorts of proteins that I could find, and may present other shortcomings. Confident mass spectrometry identification of the proteins responsible was certainly a significant challenge that I will return to. Another limitation during the fractionation and activity assay procedure however, is the use of a single duplex oligonucleotide sequence for binding studies. I have been asked about this choice many times during public talks and would like to address its implications.

Due to time and material limitations, I did not synthesize more sequence variants of [ox]mC containing oligonucleotides. Producing more sequences would have limited the amount I could make of the main sequence of interest, and for fC and carC containing DNA, the expense of large quantities of phosphoramidite discouraged diversification of the sequences pursued. However, using only one sequence is very biochemically reasonable given the design of the activity assay and its use of cold competitor. The gel shift assays binding to a radiolabeled sequence in the presence of an excess of a cold competitor of the same sequence differing only in its modification state at a central CpG dinuleotide, therefore exclusively interrogating the contribution of the modification state of that CpG dinucleotide to binding equilibrium. Any contribution from the flanking sequence, which remains the same between the cold competitor and the radiolabeled DNA, is negated. Binding to the flanking sequence should not promote or exclude modification-specific binding events, so long as the flanking sequence is *tolerated* to the degree that it allows [ox]mC-specific binding energy to prevail over any binding hindrances present by the flanking sequence.

This approach is also somewhat supported by fact that multiple different sequencing studies do not detect a clear preferred sequence context in which mC [Meissner et al. 2005] or [ox]mC species are found genome wide [Shen et al. 2013; Yu et al. 2012a; Song et al. 2013]

- these sites are simply CG rich and lack other sequence motifs. Similarly, structural analysis of the TET1 and TET2 enzymes bound to mC and hmC containing DNA do not show any base specific recognition outside of the CpG dinucleotide, suggesting no particular preferred sequence context outside of the CpG. Regardless, I fully expect that some [ox]mC-specific binding proteins do have preferences for the flanking sequence, as this is biologically useful for directing their binding to particular (rather than all) sites of [ox]mC within the cell for regulatory or signaling function. Given that, the sequence context I chose may have been poorly tolerated such that these types of proteins would not be active in my activity assays, and would not be included in fractionation. This is a limitation of the approach, but it does not seem to preclude identifying [ox]mC binding proteins outright.

More proteins implicated by mass spectrometry analysis of the brain purification should be tested for [ox]mC binding activity. Among those on the list presented in table 3.1, I am particularly interested in the ILF2/3 proteins, as they appear in a complex and are found with multiple types of oligonucleotide pulldowns, as noted in table 3.1, and were also found in the carC pulldown band shown in fig. 2.9. I am also interested in the number of WD repeat proteins, and the possibility that [ox]mC specific binding is something that is particular to this fold that has perhaps has evolved multiple times or in one common ancestor of these proteins. There is also tempting implication of [ox]mC specific binding capacity in other chromatin regulators and the basal transcription machinery. This is an enticing next direction as well, though the precise assignment of the activity to a direct binding partner within these larger complexes may be challenging. The mass spectrometry experiment can certainly be improved using a quantitative strategy such as dimethyl labeling or use of stable isotope-labeled peptides particular to each sample type. I hope to pursue such strategies during my remaining time in the lab.

Having validated that WDR76 is an hmC-specific binding protein, we now can ask many more questions about its function. Foremost among these are experiments within reach in

mouse embryonic stem cells and in leukemia models, through collaboration with the lab of Jianjun Chen at the University of Cincinnati College of Medicine. Here, the role of WDR76's hmC-specific binding in gene regulation, chromatin structure, and cell identity in normal development and disease can be investigated. My initial studies in *Wdr76-/-* mESCs and WDR76 knocked down K562 leukemia cell lines collectively suggest that Wdr76 plays a positive role in hmC-driven gene expression in a local chromatin context. Further studies could finally elucidate the molecular signaling properties of [ox]mC species, demonstrating fully that they are not merely intermediates in a demethylation pathway. There is also more biochemical insight to be gained from examining the molecular interface with which WDR76 exquisitely discriminates between cytosine modifications. I am also particularly interested in factors WDR76 with which may associate or interact with upon hmC binding. Identification of these factors will be critical for reaching the next important horizon in the study of [ox]mC biology: the functional effects that recognition of [ox]mC species confers or localizes to modified sites within the genome.

Given that the biochemical fractionation strategy I report is much more labor-intensive than the SILAC studies of Spruijt and colleagues [Spruijt et al. 2013], and we both identify WDR76 (among other proteins), it is worth reflecting on why this overlap may have occurred despite many of their other candidate proteins not being [ox]mC-specific (see Appendix B and [Hashimoto et al. 2012]). Both experiments utilize a DNA pulldown step to capture binding proteins based on their affinity, but the input to the two experiments is quite different.

Through biochemical fractionation, the input I used in pulldown assays was greatly enriched with [ox]mC specific binding protein relative to the nuclear extract that it was originally derived from. These proteins had become far more abundant in the extract with each fractionation step as they were selectively moved forward in the purification while many other proteins in inactive fractions were discarded. By comparison, the input to the SILAC experiments of Spruijt and colleagues is very similar to the input into my first column. Its

complexity has not been reduced through preferential enrichment of active proteins and exclusion of inactive proteins. This greatly impacts what the subsequent pulldown captures. Such an input, derived from the nucleus but now lacking significant nucleic acid content, is full of proteins that are strongly attracted to nucleic acids through charge based interactions. When a single nucleic acid is presented during the pulldown experiment, it is easy to capture many proteins that bind DNA tightly but lack specificity for the particular sequence or modification state of the DNA. If a given [ox]mC protein is not very abundant, it will likely not compete well with the many other abundant DNA binding proteins even if it has higher affinity for the modified DNA presented. For this reason, proteomic approaches that proceed directly from nuclear extract to affinity purification are liable to capture proteins for reasons that are not specific to the modification state of the DNA used. While it can happen, it does not seem to be a frequent result based on our examination of candidate proteins presented in Appendix B. The probability of capturing such proteins is greatly enhanced by fractionation and relative enrichment of these proteins through some extract purification.

It is worth noting that if a stem cell biologist, for example, wanted to study the function of [ox]mC specific binding events in the cell by looking to a list of proteins generated by other large proteomics experiments, they would likely not choose a specific protein by working from the top candidates of Spruijt and colleagues. Several of these top candidates are ruled out by the experiments I performed with Matthew Sullivan described in Appendix B. WDR76 is statistically not as compelling a candidate as several proteins that we ruled out in binding studies, and would not be an obvious lead candidate in this set. From my identification and validation of WDR76, functional studies of hmC-specific binding in cells can move forward. In the case of MLL-rearranged leukemias examined in collaboration with Jianjun Chen, these studies have already yielded a functional role for WDR76 in TET1-driven oncogenesis. In my own gene expression analyses in WDR76 knockout mouse embryonic stem cells, there is also implication of this protein in the expression of genes enriched in symmetric hmC.

For these reasons and many others noted throughout the fractionation scheme, my biochemistry, though taxing, provided a far more stringent, insightful, and ultimately productive way to identify highly specific [ox]mC binding proteins, the function of which can now be studied in cells.

Appendix A

SYNTHESIS AND PREPARATION OF MODIFIED OLIGONUCLEOTIDES

A.1 Overview of synthesis conditions

All ox-mC modified oligonucleotides were synthesized on an Expedite 8909 DNA synthesizer (Applied Biosystems) using conventional solid phase 2-cyanoethyl (CE) phosphormidite synthesis reagents (Glen Research). The synthesis, deprotection, and purification were according to the manufacturers instructions, with some modifications. Nucleoside phosphoramidites with 'Ultramild' protecting groups (Pac-dA-CE, Ac-dC-CE and iPr-Pac-dG-CE) were used to synthesize the oligonucleotides containing C, mC, hmC and fC. Regular phosphoramidites (N-benzoyl-dA-CE, N-benzoyl-dC-CE and N-isobutyryl-dG-CE) were used to synthesize oligonucleotides containing carC, in order to accommodate its deprotection conditions without side reactions.

Modified cytosines were included at the underlined positions in the sequences below using 5-methyl-dC-CE, 5-hydroxymethyl-dC-CE (5-hydroxymethyl-dC II CE (N_2O-carbamoyl protecting group), 5-formyl-dC-CE (diacetyl protecting group), and 5-carboxy-dC-CE (ethyl-carboxy protecting group). Oligonucleotides bearing C, mC, hmC and fC were synthesized on Ac-dC 1000 angstrom controlled pore glass (CPG) support on a 1 μmol scale. Oligonucleotides bearing carC were synthesized on dC-CPG 1000 angstrom support in micromole scale syntheses.

The duplex sequence composition was based on motif analysis of hmC-specific sequencing data available at the time [Yu et al. 2012a] and is designed to present a central CpG dinucleotide for binding, and a distal CpG dinucleotide as a relevant binding context (potentially for further modification in future iterations of this sequence.

16 nucleotide sequence "top" - underlined C is subject to 5-position modification.

5'GCG GCT GCG TGG GTC C 3'

(ϵ = 141,400 L· M^{-1} cm^{-1})

16 nucleotide sequence "bottom" - underlined C is subject to 5-position modification.

5'GGA CCC ACG CAG CCG C 3'

(ϵ = 145,200 L· M^{-1} cm^{-1})

Oligonucleotides bearing C and mC at the indicated central CpG dinucleotide were deprotected with fresh aqueous 30% ammonium hydroxide, shaking at 55° C overnight in benchtop microfuge ThermoMixer (Eppendorf). The recovered material was then lyophilized and purified from truncated, degraded, or partially deprotected material by preparative 20% denaturing polyacrylamide gel in 1x TBE. Prior to electrophoresis, each sample was heated to 95° C for 5 minutes in denaturing nucleic acid loading buffer (95% formamide, 5mM EDTA pH 8.0, 0.025% bromophenol blue, 0.025% xylene cyanol) and then loaded onto a thermally equilibrated gel running at 250 mA/gel. The product band was visualized by UV shadowing on a fluorescent TLC plate, excised, crushed, and extracted with 100 mM triethyl ammonium acetate pH 7 (TEAA) shaking at 30° C overnight in 2 x 25mL extractions. This material was then loaded onto a Waters Sep-Pak C18 cartridge, washed in 100 mM TEAA, then MilliQ water and eluted in 50% aqueous acetonitrile.

A.2 Synthesis, deprotection and purification of oxidized methylcytosine-containing oligonucleotides

Oligonucleotides containing hmC were deprotected using 20 mM K_2CO_3 in methanol for 4 hours shaking at 800 rpm at 30° C in a benchtop microfuge ThermoMixer (Eppendorf). The deprotected material was then purified as described above.

Oligonucleotides containing fC were deprotected initially in a 1:1 solution of 40% CH_3NH_2 and 30% NH_4OH at 65° C shaking at 800 rpm for 2 hours in a benchtop microfuge ThermoMixer (Eppendorf) to reveal the 1,2 diol. The gel extracted, Sep-Pak purified material

was then lyophilized and resuspended in cold ddH2O at 4° C. To oxidize the 1,2 diol to generate the aldehyde of fC, aqueous $NaIO_4$ at 4° C was added to a final concentration of 50 mM, and the mixture incubated for 30 minutes at 4° C rotating end over end. This reaction was then quenched with 10 volume equivalents of 100 mM TEAA pH 7 and loaded onto a second C18 Sep Pak for purification as above.

Oligonucleotides containing carC were deprotected and cleaved from the resin using 0.4 M NaOH in 4:1 methanol:water solvent overnight at room 800 rpm at 30° C in a benchtop microfuge ThermoMixer. The deprotected material was then purified as described above.

Biotinylated oligonucleotides were synthesized with the same 'bottom'sequence with a 3'T8-TEG extension, yielding a 24 nucleotide 3'biotinylated oligonucleotide. These were synthesized on a 3'biotin CPG using the synthesis and deprotectection conditions dictated by the modified cytosine incorporated. This oligonucleotide was annealed with the 'top'sequence to form 3? biotinylated duplexes.

5'GGA CCC ACG CAG CCG CTT TTT TTT - TEG- 3'biotin

(ϵ = 209,700 L· M^{-1} cm^{-1})

Finally, a non-CpG containing scramble oligonucleotide duplex was designed to test the contribution of the CpG dinucleotide step to specific binding. The single stranded oligonucleotides of this duplex were purchased from Integrated DNA Technologies.

nonCpG16 nucleotide sequence "top"

5'GGC CTG GCC TGG GTC C 3'

nonCpG 16 nucleotide sequence "bottom"

5'GGA CCC AGG CCA GGC C 3'

Purified oligonucleotides were then analyzed by MALDI-TOF mass spectrometry to confirm identity and complete deprotection/conversion. Oligonucleotide purity was assessed by gel to be $>$ 90% in all cases. Finally, the oligonucleotides were lyophilized, and resuspended in 20 mM Tris HCl pH 7.5, 1 mM EDTA at 400 μM final concentration, and stored at -80°

C.

Duplex oligonucleotide stocks were annealed by combining equimolar top and bottom strands at a final duplex concentration of n 200 μM and heating this mixture to 90° C in TE300 (20 mM Tris, pH 7.5, 1 mM EDTA, 300 mM NaCl) for 5 minutes in a heating block, followed by passive slow cooling of the block to room temperature over ∼1 hour. Small aliquots were then stored at -80° C. Annealed oligonucleotides were analyzed by Sybr Green melts to assess stability of the duplex in a real time PCR instrument (BioRad CFX96). The calculated melting temperature of these 16 base pair oligonucleotides to be 63.5° C, (IDT Oligo Analyzer web tool) however, the observed melting temperature using Sybr Green is closer to 75° C at the concentrations of NaCl used in labeling, annealing and EMSA procedures (typically ∼300 mM NaCl).

Freshly thawed, annealed oligonucleotides were radiolabeled with T4 PNK (NEB) overnight using 50 pmoles of [α-^{32}P]ATP per 50 pmoles of 5'ends (6000 Ci/mmol, Perkin Elmer) in 1x PNK buffer. The PNK reaction was heat-denatured for five minutes at 95° C, supplemented with 300 mM NaCl, and the mixture slow cooled to reanneal as before. The concentrated labeled oligonucletotide stock were then diluted in TE300 as needed for use in gel shift assays.

Appendix B

ASSESSMENT OF OTHER CANDIDATE PROTEINS REPORTED IN THE LITERATURE

B.1 Examination of candidates from a large scale proteomics searches for oxidized methylcytosine-specific binding proteins

All binding studies of candidate [ox]mC specific binding proteins found in Spruijt et al., (2013) were completed in collaboration with Alex Ruthenburg, who contributed to construct design and cloning, and Matthew Sullivan, who preformed the purification of THY28, C3ORF37 constructs, and performed all filter binding assays in consultation with Kate Malecek. Matthew Sullivan also performed the ZHX1 gel shift assay in close consultation with Kate Malecek. Kate Malecek completed all other cloning, the purification of THAP11 and ZHX1, and other pilot gel shift experiments.

B.1.1 Expression and Purification of THY28

Human THY28 cDNA (residues 54-221), identified by SILAC as a pan-[ox]mC reader, was cloned into expression vector pMCSG7 from MGC IMAGE 30717693 (NCBI Accession BC093074) to form an N-terminally 6x-His-tagged construct with a TEV protease cleavage site and SNG linker. This construct was designed by reference to prior THY28 expression and structural biology. The resulting plasmid was transformed into E. coli BL21(DE3) + pRARE2. Expression cultures were grown at 37° C and induced at OD600 = 0.5 with 0.4 mM IPTG for 3 hours at 37° C. Cell pellets were collected via centrifugation (Sorvall RC-3B with H-6000A rotor, 4000 rpm for 20 minutes at 4° C), resuspended in Ni-NTA lysis buffer (600 mM NaCl, 50 mM NaxH2-xPO4 pH 8.0, 10% glycerol (v/v), 5 mM imidazole, 0.5 mM PMSF, 5 mM β-mercaptoethanol), and stored at -80° C. Freshly thawed cell pellets were

lysed with an Avestin EmulsiFlex-C3 homogenizer and lysate was clarified via two sequential centrifugation steps at 30,000xg (Sorvall RC-5B SS-34 rotor) at 4° C for 25 minutes each. Clarified lysate was incubated with pre-equilibrated Ni2+-NTA resin (Qiagen; 0.25 mL final bed volume per 1 L of culture) for 30 minutes, rotating at 4° C. After collection of flow-through, the resin was washed sequentially with 4 column volumes (CV) each of 1 M NaCl, 600 mM NaCl, and 400 mM NaCl wash buffers (otherwise identical in composition to Ni-NTA lysis buffer). Bound proteins were eluted over multiple fractions in Ni-NTA elution buffer containing 300 mM imidazole pH 7.5 and 400 mM NaCl (otherwise identical in composition to Ni-NTA lysis buffer). Peak fractions enriched for THY28 were pooled, combined with 1/100 mass equivalents of TEV protease, and dialyzed overnight at 4° C against 100 mM NaCl, 50 mM TrisůHCl pH 7.0, 5% glycerol (v/v), 5 mM β-mercaptoethanol. The dialysis output was clarified via centrifugation and diluted approximately 1.5-fold with Buffer TG100 (100 mM NaCl, 20 mM TrisůHCl pH 7.0, 5% glycerol (v/v), 5 mM β-mercaptoethanol). The pH and conductivity of the clarified output was adjusted with 1 M TrisůHCl pH 6.5 and salt free dilution buffer (identical in composition to Buffer TG100 except for the absence of NaCl) to match that of Buffer TG50 (50 mM NaCl, otherwise identical in composition to Buffer TG100). This sample was loaded onto a 6 mL RESOURCE S column (GE Healthcare Life Sciences) pre-equilibrated in Buffer TG50; the column was washed with 2 CV of Buffer TG50 and eluted with a linear gradient over 5 CV to 1 M NaCl. Desired peak eluate fractions were pooled and concentrated with an Amicon Ultra-4 10 kDa NMWL Centrifugal Filter Unit (EMD Millipore) according to manufacturer instructions. This concentrate was filtered with a 0.45 μm Ultrafree-MC HV Centrifugal Filter (EMD Millipore) and loaded onto a Superdex 75 10/300 GL column (GE Healthcare Life Sciences) equilibrated and eluted in 20 mM Na·HEPESpH 7.8, 150 mM NaCl, 5% glycerol (v/v). Peak eluate fractions were pooled and stored at 4° C for use in filter binding assays (designated T̈HY28 FBAIb̈elow). The THY28 concentration of this combination as measured by NanoDrop was approximately 320

μM, with an A260:A280 ratio of 0.58. All column chromatography steps were performed at 4° C.

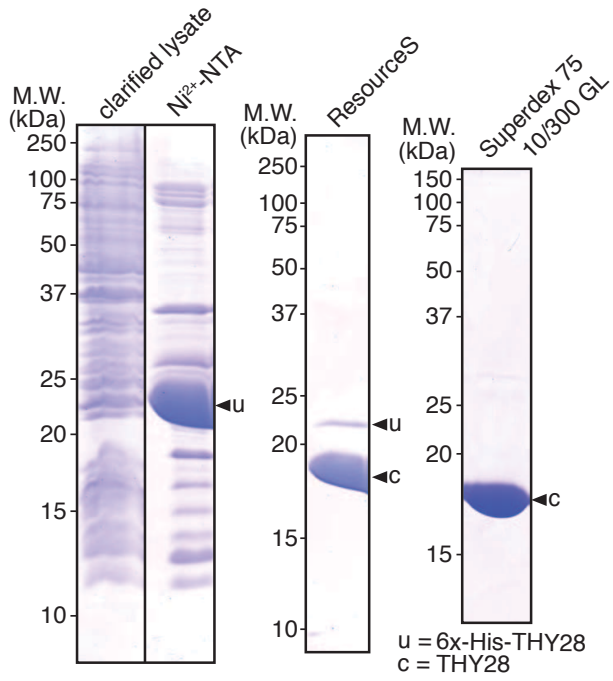


Figure B.1: The candidate [ox]mC binding protein THY28 was purified to homogeneity as an N-terminal 6x-His fusion protein (*completed in collaboration with Matthew Sullivan*).

B.1.2 Expression and Purification of C3ORF37/HMCES constructs

C3ORF37 is an uncharacterized protein that has been renamed ḦMCES̈(5-hydroxymethylcytosine-binding, embryonic stem cell-specific) based on mass spectrometry enrichment for [ox]mC. Human C3ORF37 cDNA (full length protein designated 'FL', or residues 1-275 designated 'NTD'that comprise the predicted SRAP domain was amplified from MGC IMAGE 30398247 (NCBI Accession BC088363) and cloned into the pMCSG20 expression vector to form N-terminally S- and GST-tagged constructs with a TEV protease cleavage site and SNG linker. FL and NTD constructs were expressed in E. coli BL21(DE3) + pRARE2 by induction at OD600 = 0.5 with 1 mM IPTG for 12 hours at 16° C. Cell pellets were collected via centrifu-

gation at 4000 rpm (Sorvall RC-3B H-6000A rotor) at 4° C for 20 minutes, resuspended in GST lysis buffer (500 mM NaCl, 50 mM Tris-HCl pH 7.5, 10% glycerol (v/v), 0.5 mM PMSF, 5 mM β-mercaptoethanol), and stored at -80° C. Freshly thawed cell pellets were lysed and clarified as described above for THY28. Clarified lysate was loaded onto a 5 mL GSTrap HP column (GE Healthcare Life Sciences) pre-equilibrated in GST lysis buffer; the column was washed with 2 CV of GST lysis buffer and eluted over 3 CV linear gradient to GST lysis buffer + 20 mM reduced glutathione. This and all subsequent column chromatography steps were performed at 4° C.

For FL, the desired GST column peak eluate fractions were pooled and combined with 1/50 mass equivalents of TEV protease, and dialyzed overnight at 4° C against 250 mM NaCl, 50 mM Tris-HCl pH 8.45, 10% glycerol (v/v), 5 mM β-mercaptoethanol. The dialysis output was clarified via centrifugation and adjusted to match the pH and conductivity of Buffer TG100 (100 mM NaCl, 20 mM Tris-HCl pH 7.0, 10% glycerol (v/v), 5 mM β-mercaptoethanol) by sequential addition of the appropriate volumes of 1 M Tris-HCl pH 6.0 and TG0 (identical in composition to Buffer TG100 except for the absence of NaCl). This sample was loaded onto a 5 mL HiTrap Heparin HP column (GE Healthcare Life Sciences) pre-equilibrated in Buffer TG100; the column was washed with 2 CV of Buffer TG100 and eluted over a linear 6 CV gradient to 550 mM NaCl, then to 1 M NaCl over 1 CV. Desired peak eluate fractions were pooled, supplemented with NaCl to achieve a final approximate concentration of 800 mM, and concentrated with an Amicon Ultra-4 10 kDa NMWL Centrifugal Filter Unit (EMD Millipore) according to manufacturer instructions. This concentrate was filtered as described above for THY28 and loaded onto a Superdex 75 10/300 GL column (GE Healthcare Life Sciences) equilibrated and eluted in 20 mM Na·HEPESpH 7.8, 150 mM NaCl. Desired peak eluate fractions were pooled and concentrated via centrifugation, and this concentrate was stored at 4° C for use in filter binding assays (designated ?C3ORF37 FL FBAI?). The FL concentration of this sample as measured by NanoDrop was approximately

8.7 μM, with an A260:A280 ratio of 0.62.

The NTD was purified using a similar strategy as that described above for FL, but, as was also the case for the FL preps, despite large expression scales, relatively inefficient and incomplete GST tag cleavage combined with various solubility issues throughout the preps (of greater severity for FL than for NTD) and loss to the membrane of the centrifugal concentrators resulted in significant yield losses between the GST, Heparin, and size exclusion chromatography steps that did not permit the recovery of a sufficiently large amount of protein following size exclusion to perform the desired number of filter binding replicates at final concentrations higher than approximately 4-6 μM. Accordingly, we elected to also purify NTD without cleaving the GST tag (designated 'GST-C3ORF37 NTD'), which ultimately permitted a final concentration of 10.9 ?M to be achieved in a repeat filter binding assay, though the overall yield of the prep remained lower than desired. For this prep, the desired GST column peak eluate fractions were pooled and adjusted to match the pH and conductivity of Buffer BTG100 (100 mM NaCl, 20 mM Bis-Tris pH 6.0, 10% glycerol (v/v), 5 mM β-mercaptoethanol) by sequential addition of the appropriate volumes of 769 mM Bis-Tris pH 6.0 and salt free dilution buffer (identical in composition to Buffer BTG100 except for the absence of NaCl). This sample was loaded onto a 5 mL HiTrap Heparin HP column (GE Healthcare Life Sciences) pre-equilibrated in Buffer BTG100; the column was washed with 3 CV of Buffer BTG100 and eluted over a linear 2 CV gradient to 1 M NaCl. Desired peak eluate fractions were pooled and split into two portions: the first (approximately 1/3 of the prep) was directly dialyzed overnight at 4° C against 20 mM Na·HEPESpH 7.8, 150 mM NaCl with one buffer change; the remaining portion was carried forward through a concentrating ion exchange chromatography step and subsequent size exclusion chromatography to, ideally, avoid the use of a centrifugal concentrator in sample preparation both before and after size exclusion. To this end, this remaining portion from the Heparin column was diluted approximately 2.5-fold with BTG300 (otherwise identical in composition to Buffer BTG100,

but with 300 mM NaCl) and then adjusted with the aforementioned salt free dilution buffer to match the conductivity of Buffer BTG100. The resulting sample was loaded onto a Mono S 5/50 GL column (GE Healthcare Life Sciences) pre-equilibrated in Buffer BTG100; the column was washed with 3 CV of Buffer BTG100 and eluted over a linear 2 CV gradient to 1 M NaCl. Desired peak eluate fractions were pooled, filtered as described above for THY28, and loaded onto a Superdex 75 10/300 GL column (GE Healthcare Life Sciences) equilibrated and eluted in 20 mM Na·HEPESpH 7.8, 150 mM NaCl. Desired peak eluate fractions were pooled, and due to unexpectedly low yield from this last chromatography step, it was still necessary to concentrate these pooled fractions via centrifugation in a 10 kD MWCO Ultrafree centrifugal concentrator (Millipore). Filter binding assays were performed with both protein obtained from direct dialysis (designated G̈ST-C3ORF37 NTD FBAI-1¨) and size exclusion chromatography (designated ?GST-C3ORF37 NTD FBAI-2?). The GST-C3ORF37 NTD concentration of FBAI-1 and FBAI-2 as measured by NanoDrop was approximately 14.4 μM and 12.2 μM, with A260:A280 ratios of 0.65 and 0.67, respectively.

B.1.3 Expression and Purification of THAP11 subdomain of Ronin

THAP11 or Ronin is a protein reported to bind hmC specifically by SILAC enrichement of material captured from adult mouse brain nuclear extracts. Moreover, Ronin has been reported to be required for pluripotentcy during early embryogenesis in mouse embryonic stem cells. Residues 2-89 of full length human THAP11 encode a THAP domain responsible for DNA binding activity, which was cloned from MGC IMAGE 4554554 (NCBI Accession BC012182) into the expression vector pMSCG20 to form an N-terminally S- and GST-tagged construct with a TEV protease cleavage site and a SNG linker. The fusion protein was expressed in E. coli BL21(DE3) + pRARE2 by induction at OD600 = 0.5 with 1 mM IPTG for 16 hours at 16° C. Cell pellets were collected via centrifugation at 4000 rpm (Sorvall RC-3B H-6000A rotor) at 4° C for 20 minutes, resuspended in GST lysis buffer (500 mM NaCl,

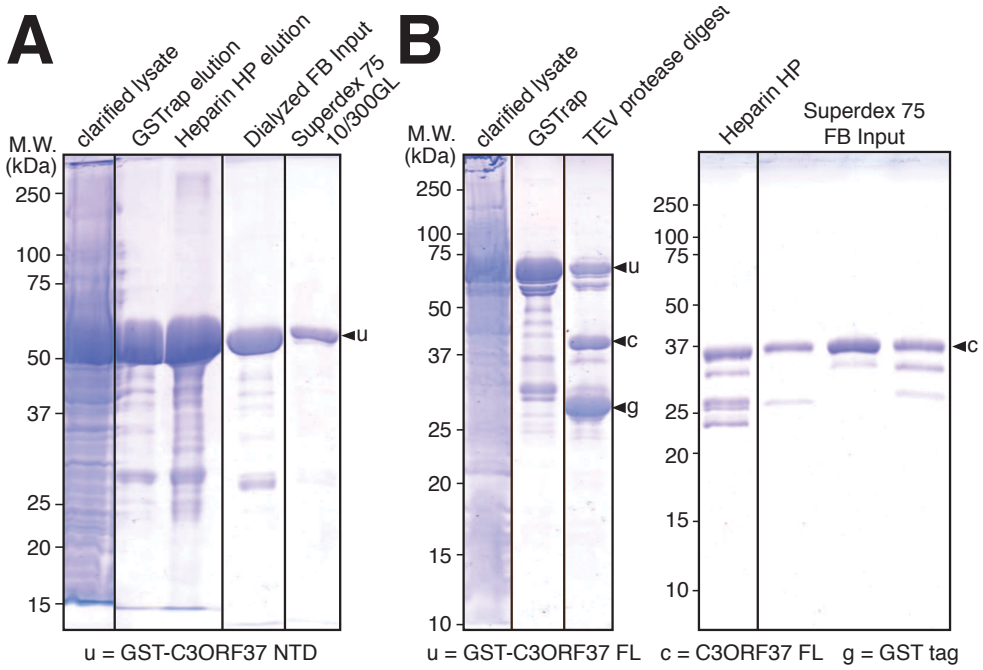


Figure B.2: Purification of GST-tagged full length C3ORF37 (HMCES) (A) and an N-terminal sub-construct of the predicted SRAP domain (B) via GST-Trap, ion exchange (heparin), and size exclusion chromatography. (*completed in collaboration with Matthew Sullivan*).

50 mM Tris-HCl pH 7.5, 10% glycerol (v/v), 0.5 mM PMSF, 5 mM βmercaptoethanol), and stored at -80° C. Freshly thawed cell pellets were lysed and clarified as described above for THY28. Clarified lysate was loaded onto a 5 mL GSTrap HP column (GE Healthcare Life Sciences) pre-equilibrated in GST lysis buffer; the column was washed with 2 CV of GST lysis buffer and eluted over a 3 CV linear gradient to GST lysis buffer + 20 mM reduced glutathione. This and all subsequent column chromatography steps were performed at 4° C.

The GST column peak eluate fractions were visualized by SDS-PAGE, pooled and combined with 1/50 mass equivalents of TEV protease to remove the GST fusion tag over 16 hours of incubation at 4° C. Complete removal of the tag was confirmed by SDS-PAGE, and as the purified protein showed no nucleic acid contamination at this point (A260:A280 0.63), an ion exchange column was not used as for the other protein purifications described here. The TEV-digested material was concentrated in a 3K MWCO centrifugal concentrator (Vi-

vaspin 2 Polyethersulfone concentrator, Sartorius). The concentrated material was loaded onto a HiLoad 16/600 Superdex 75 prep grade size exclusion column equilibrated in 20 mM Na·HEPESpH 7.8, 150 mM NaCl. The peak eluate fractions were inspected by SDS-PAGE and the desired fractions pooled and quantified as described below with a final concentration of 21.1 μM and an A260:A280 ratio of 0.63.

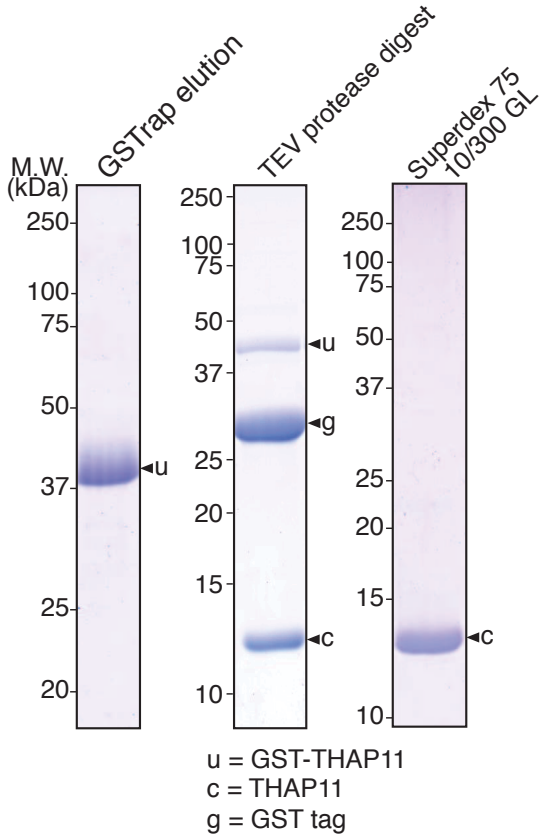


Figure B.3: Purification of GST-tagged THAP domain of THAP11 (Ronin) via GST-Trap,TEV digestion to remove the GST tag, and size exclusion chromatography.

B.1.4 Expression and Purification of ZHX1 construct

ZHX1 is a protein reported by Spruijt et al., 2013 to bind hmC in nuclear extracts from murine neuronal precursor cells. The largest previously described fragment (55-830) spanning two zinc- fingers and five homeodomains was cloned from MGC IMAGE 5271344 (NCBI

Accession BC040481.1) into pET30a to form an N-terminally 6x-His-tagged construct with a TEV protease cleavage site and SNG linker. The fusion protein was expressed in E. coli BL21(DE3) + pRARE2 by induction at OD600 = 0.5 with 1 mM IPTG for 4 hours at 37° C. At OD600 = 0.25, 50 μM ZnSO4 was added to the cultures to support folding of the zinc fingers during protein expression. Cell pellets were collected via centrifugation (Sorvall RC-3B with H-6000A rotor, 4000 rpm for 20 minutes at 4° C), resuspended in Ni-NTA lysis buffer (600 mM NaCl, 50 mM NaxH2-xPO4 pH 8.0, 10% glycerol (v/v), 5 mM imidazole, 0.5 mM PMSF, 5 mM β-mercaptoethanol), and stored at -80° C. Freshly thawed cell pellets were lysed with an Avestin EmulsiFlex-C3 homogenizer and lysate was clarified via ultracentrifugation (Beckman Ti70.1 rotor at 37,000 rpm) at 4° C for 40 minutes. Clarified lysate was incubated with pre-equilibrated Ni2+-NTA resin (Qiagen; 0.25 mL final bed volume per 1 L of culture) for 60 minutes, rotating at 4° C. After collection of flow-through, the resin was washed sequentially with 4 column volumes (CV) each of 1 M NaCl, 600 mM NaCl, and 200 mM NaCl wash buffers (otherwise identical in composition to Ni-NTA lysis buffer). Bound proteins were eluted over multiple fractions in Ni-NTA elution buffer containing 300 mM imidazole pH 7.5 and 200 mM NaCl (otherwise identical in composition to Ni-NTA lysis buffer). Elution fractions enriched in the desired full length protein were combined and diluted approximately 1.5-fold with Buffer TG100 (100 mM NaCl, 20 mM Bis-Tris pH 6.5, 10% glycerol (v/v), 5 mM β-mercaptoethanol and 50 uM ZnSO4). The pH and conductivity of the fraction pool was adjusted with salt free dilution buffer (identical in composition to Buffer TG100 except for the absence of NaCl) to match that of Buffer TG50 (50 mM NaCl, otherwise identical in composition to Buffer TG100). This sample was loaded onto a 1 mL POROS Heparain S column (GE Healthcare Life Sciences) pre-equilibrated in Buffer TG50; the column was washed with 2 CV of Buffer TG50 and eluted with a linear gradient over 12 CV to 1 M NaCl. Desired peak eluate fractions were pooled and concentrated with an Amicon Ultra-4 30 kDa NMWL Centrifugal Filter Unit (EMD Millipore) according

to manufacturer instructions. This concentrate was filtered with a 0.45 μm Ultrafree-MC HV Centrifugal Filter (EMD Millipore) and loaded onto a Superdex 200 10/300 GL column (GE Healthcare Life Sciences) equilibrated and eluted in 20 mM Na·HEPESpH 7.8, 150 mM NaCl, 5% glycerol (v/v). Peak eluate fractions were pooled and stored at 4° C for use in gel shift assays, as described above. All column chromatography steps were performed at 4° C.

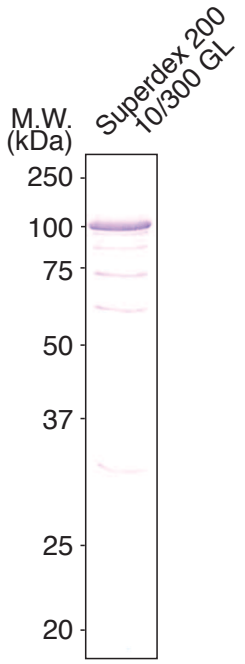


Figure B.4: Final purified protein of full length N-terminally 6His-tagged ZHX1 after Ni-NTA, POROS Heparin and S200 size exclusion chromatography.

B.1.5 Column Purification of Radiolabeled DNA

Frozen stocks of annealed, radiolabeled oligonucleotides stored at -20° C (see 'Synthesis and Preparation of Oligonucleotides' were thawed and purified from remaining [α-32P] ATP using Illustra ProbeQuant G-50 Micro Columns (GE Healthcare Life Sciences) according to manufacturer instructions. Eluted DNA was then supplemented with 300 mM NaCl, heat-denatured for five minutes at 95° C, and re-annealed by slow cooling to room temperature. Column purified, annealed, radiolabeled oligonucleotides were stored at -20° C.

B.1.6 THY28, C3ORF37, and THAP11 Binding Reaction Setup

For THY28, 70 μL binding reactions were setup in a 96-well microplate format as follows: 30 μL of THY28 FBAI was added to 20 ?L of binding buffer (20 mM Na·HEPESpH 7.8, 150 mM NaCl) and mixed by repeated pipetting (r̈ow 2̈); the resulting solution was serially diluted into 9 additional rows via the same 30 μL removal volume / 20 μL initial buffer volume scheme; following dilution, 10 μL of binding buffer was added to rows 2-11; rows 1 and 12 were prepared by addition of solely 30 μL of THY28 FBAI or binding buffer, respectively; the microplate was then briefly spun down at 1500 rpm (Sorvall RC-3B H-6000A rotor); lastly, 40 μL of radiolabeled DNA, diluted from frozen stock to achieve a final concentration of 1 nM in the binding reaction, was added to each row, and the resulting solution was mixed by pipetting. After a 30 minute incubation at room temperature, reaction mixtures were transferred to the assembled FBA apparatus, as described below.

For C3ORF37 FL, 50 μL binding reactions were setup as described above for THY28 with some differences, as follows: rows 1 and 2 were prepared by addition of 36 μL of C3ORF37 FL FBAI to 4 μL of binding buffer, or 17 μL of FBAI to 23 ?L of binding buffer, respectively; rows 3-11 were prepared via a 20 μL removal volume / 10 μL initial buffer volume serial dilution scheme; radiolabeled DNA was delivered in 10 μL to achieve a final concentration of 1 nM.

For GST-C3ORF37 NTD, 80 μL binding reactions were setup as described above for THY28 with some differences, as follows: rows 1 and 2 were prepared by addition of solely 70 μL of GST-C3ORF37 NTD FBAI-1, or 55 μL of FBAI-1 to 15 μL of binding buffer, respectively; rows 3-11 were prepared via a 110 μL removal volume / 70 μL initial buffer volume serial dilution scheme; radiolabeled DNA was delivered in 10 μL to achieve a final concentration of 1 nM.

GST-C3ORF37 NTD (sized), 50 μL binding reactions were setup as follows: due to limited availability of GST-C3ORF37 NTD FBAI-2, only two protein concentration points

were assayed; rows 1 and 2 were prepared by addition of solely 45 μL of this FBAI-2, or 29 μL of FBAI-2 to 16 μL of binding buffer, respectively; row 3 was prepared by addition of solely 45 μL of binding buffer; these samples were then briefly spun down at 1500 rpm (Sorvall RC-3B H-6000A rotor); radiolabeled DNA was delivered in 5 μL to achieve a final concentration of 1 nM.

For THAP11, 100 μL binding reactions were setup as described above for THY28 with some differences, as follows: rows 1-11 were initially prepared via a 55 μL removal volume / 33 μL initial buffer volume serial dilution scheme, after which an additional 15 μL of THAP11 FBAI diluted with 15 μL of binding buffer was added to row 1 and 30 μL of binding buffer was added to rows 2-11; radiolabeled DNA was delivered in 37 μL to achieve a final concentration of 1 nM.

B.2 Filter Binding Assay

Filter binding assays (FBA) were performed using the 96-well Bio-Dot Microfiltration Apparatus (Bio-Rad) with Amersham Protran 0.1 μ NC nitrocellulose membrane (GE Healthcare Life Sciences) on top of a Zeta-Probe membrane (Bio-Rad). Prior to apparatus assembly, membranes were well-equilibrated in binding buffer (20 mM Na·HEPESpH 7.8, 150 mM NaCl). The apparatus was assembled by sandwiching, in the order encountered by applied sample, the nitrocellulose membrane, Zeta-Probe membrane, and 2 dry Whatman blotting papers (Grade 3MM Chr). Vacuum was briefly applied to the assembled apparatus to ensure desired seal efficiency. Immediately prior to sample application, 100 μL of binding buffer was applied to each position, and a vacuum of approximately $\sim$ 20 inHg was temporarily applied to facilitate efficient drainage through the apparatus and removed when drainage was complete. Reaction mixtures (preparation described above) were then similarly applied and drained, after which a single 100 μL wash with binding buffer was performed. Following complete drainage of the wash, vacuum was applied for an additional 2-3 minutes before

disassembly of the apparatus to facilitate drying of the membranes. Disassembled membranes and Whatman filter papers were exposed to a Fujifilm 'CR' phosphorimaging screen, which was subsequently scanned at 100 nm resolution using the Typhoon 9200 system (GE Healthcare Life Sciences). The exposure times for THY28, C3ORF37 FL, C3ORF37 NTD, THAP11 and C3ORF37 NTD (sized) were overnight and 14 days, respectively.

B.3 Quantitation, Fraction Bound Calculation, and Curve Fitting

Signal quantitation of nitrocellulose and Zeta-Probe membranes was performed using TotalLab Quant Array Analysis of .GEL files generated by the Typhoon 9200 system. For each spot on both membranes, automatic background subtraction was applied to the observed signal using the spot edge average method to generate a raw counts value. Fraction DNA bound *Fb* was calculated for each spot in a given replicate as follows:

$$Fb = \frac{\frac{S_N}{X}}{\frac{S_N}{X} + \frac{S_Z}{Y}} - \frac{\frac{S_{\text{DNA only},N}}{X}}{\frac{S_{\text{DNA only},N}}{X} + \frac{S_{\text{DNA only},Z}}{Y}} \tag{B.1}$$

where S is the raw counts value and the indices N and Z correspond to the nitrocellulose and Zeta-Probe membranes, respectively; X and Y are normalizing factors given by $X = \mathrm{S}_{max,N}$ - $\mathrm{S}_{DNAonly,N}$ and $Y = \mathrm{S}_{DNAonly,Z}$ -$\mathrm{S}_{min,Z}$ for the nitrocellulose and Zeta-Probe membranes, respectively, where S_{max} and S_{min} are the maximum and minimum raw counts value observed within a given replicate for the respective membrane, and S_{DNA} only is the raw counts value of the DNA only spot (row 12) of that replicate for the respective membrane. Plots of *Fb* vs. protein concentration were generated using KaleidaGraph v4.5.2 (Synergy Software) from averaged *Fb* values across all replicates of a single FBA experiment for a given DNA modification state, and vertical error bars correspond to +/- 1 standard deviation of *Fb* values. For THAP11, unaveraged *Fb* values for a given DNA modification state obtained from all replicates of a single FBA experiment were plotted versus final protein concentration

and fit to the Hill function $m3 \cdot m_0^{m2}/(m_0^{m2})+(m_1^{m2})$ using the General Curve Fit functionality of KaleidaGraph v4.5.2 (Synergy Software), where m_0 is the concentration of free protein (assumed to be equal to the concentration of total protein), m_1 is the dissociation constant, K_d, m_2 is the Hill coefficient, and m_3 is the saturation point.

B.4 Results of candidate binding protein studies

B.4.1 THY28 is not a DNA binding protein

Under highly permissive DNA binding conditions, THY28 is not found to be a DNA binding protein. No binding is observed at high concentrations of highly purified THY28 for any substrate or cold competitor condition tested by EMSA or filter binding studies

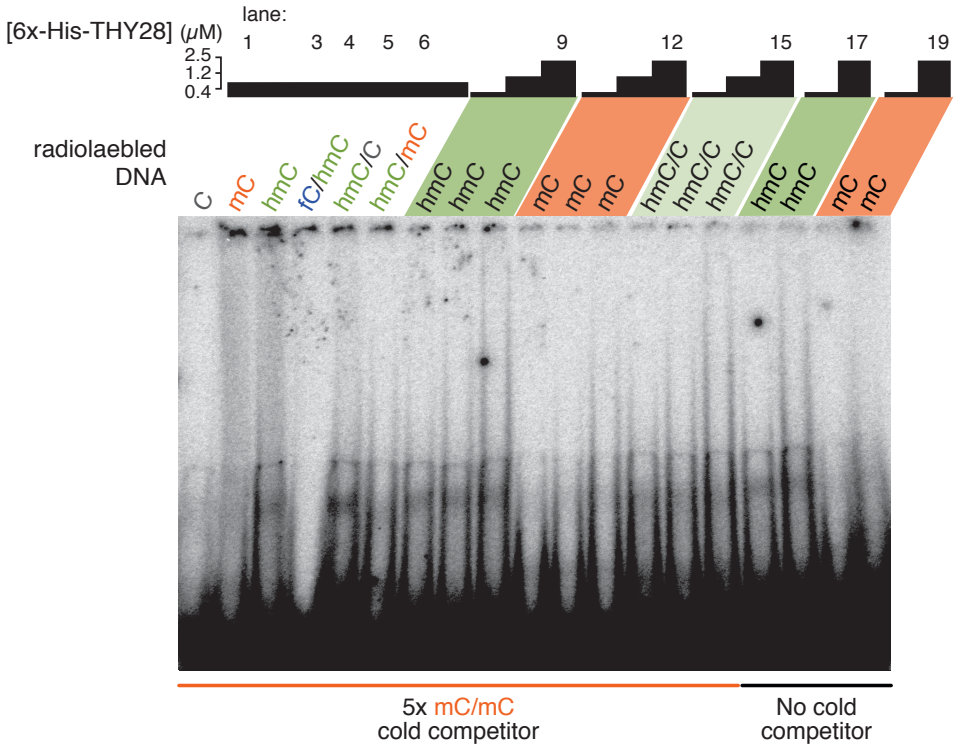


Figure B.5: The candidate [ox]mC binding protein THY28 is not a DNA binding protein by EMSA (*EMSA completed by Kate Malecek in collaboration with protein purification by Matthew Sullivan*).

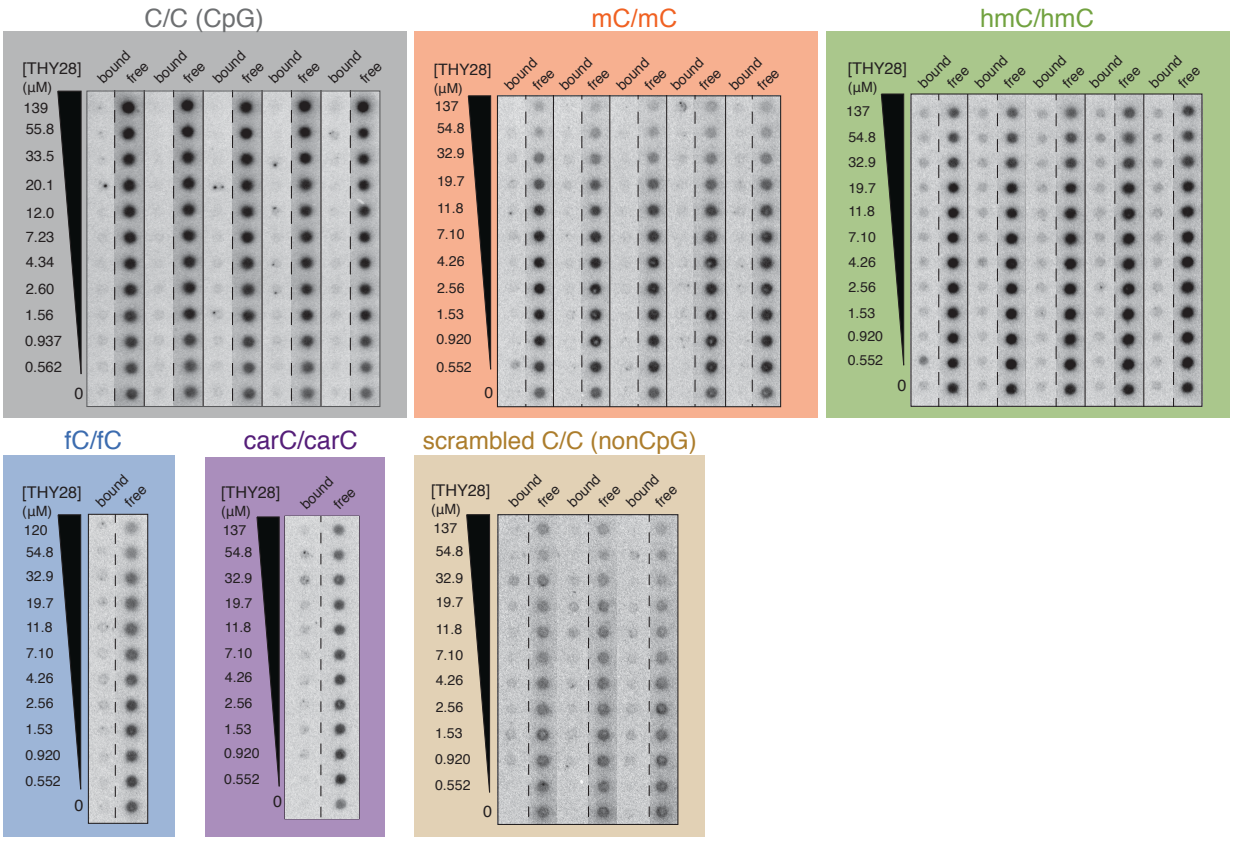


Figure B.6: The candidate [ox]mC binding protein THY28 is not a DNA binding protein in extensive filter binding assays with up to ~140μM purified THY28 (*completed in collaboration with Matthew Sullivan*).

B.4.2 C3ORF37 is not an hmC-specific DNA binding protein

Both full length C3ORF37 (HMCES) and an N-terminal SRAP domain sub-construct (NTD) do not specifically bind hmC modified DNA. Though this measurement is not made to full binding saturation to limiting amounts of pure protein, the trend of the binding curves suggests no meaningful fold difference between the *Kd* for mC and hmC symmetrically modified DNA. The binding curve trend for symmetric unmodified C suggests it is less preferred. It is worth noting that binding to half saturation requires low micromolar concentrations of C3ORF37, which is a relatively weak binding regime for a specific DNA binding protein.

B.4.3 THAP11 (Ronin) is not an hmC-specific DNA binding protein

The THAP domain of THAP11/Ronin binds to DNA but does not exhibit specific binding for hmC-modified DNA. While the *Kd* for hmC is the lowest measured in this filter binding experiment, this *Kd* is not significantly different from that of unmodified C within the error of this experiment. Moreover, a several fold difference in *Kd* commensurate with the relative abundance of hmC relative to that of C is required for any protein to bind rare hmC uniquely within the cell.

B.4.4 ZHX1 is not a DNA binding protein

Purified ZHX1 was subjected to an EMSA as described previously. No binding was observed at the available concentrations even in the absence of cold competitor, which given that the shifts appear no different in terms of the amount of well-shifted material between cold competitor conditions and not, alleviating severe solubility concerns, should be a highly permissive condition for binding.

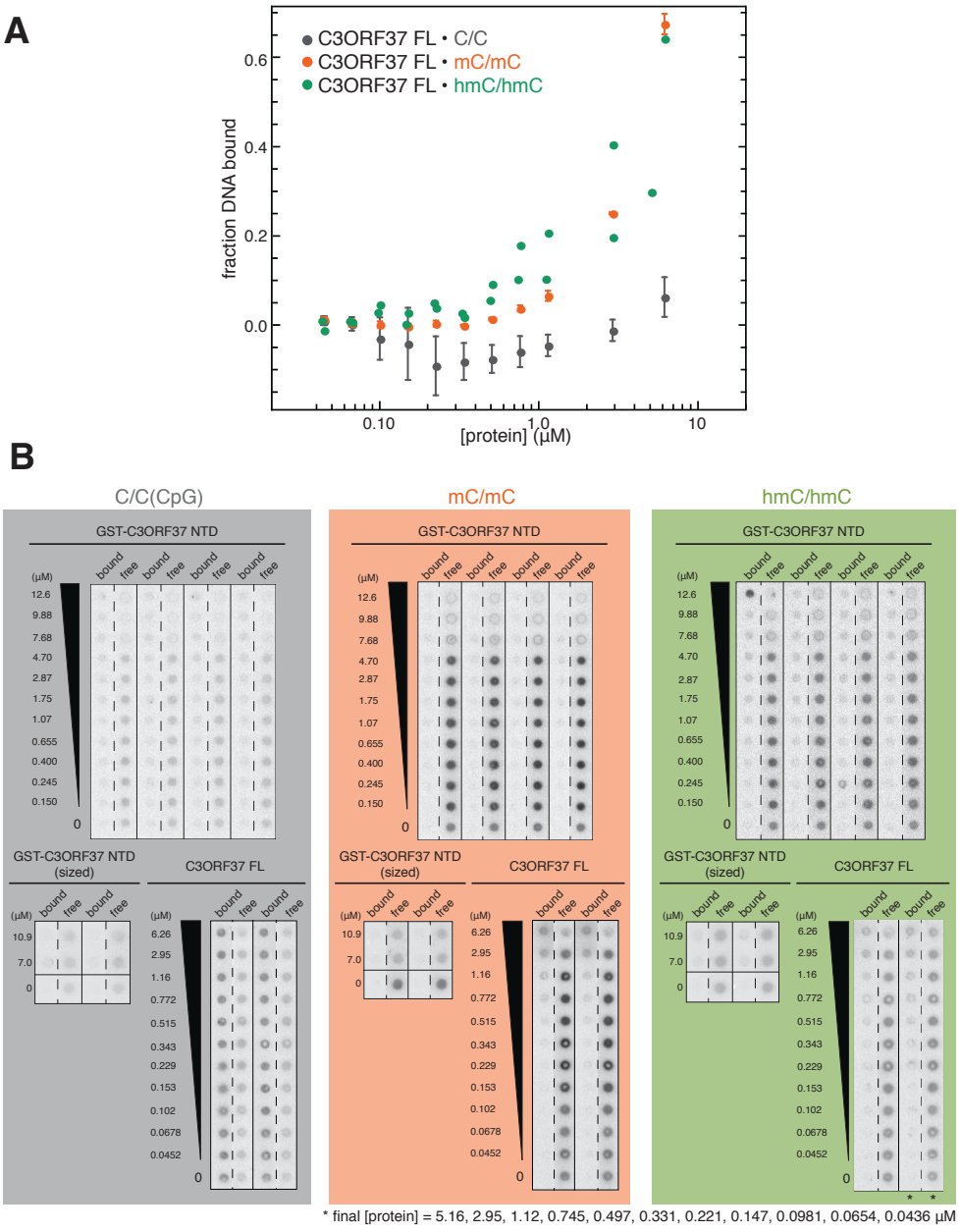


Figure B.7: C3ORF37 is not an hmC-specific DNA binding protein. Full length and an N-terminal sub-construct of the predicted SRAP domain (NTD) were subjected to filter binding studies with C, mC and hmC modified DNA. The protein does not show meaningful discrimination between mC and hmC, as quantified for the full length construct (A) and shown as raw filter binding membranes for full length and NTD. (*completed in collaboration with Matthew Sullivan*).

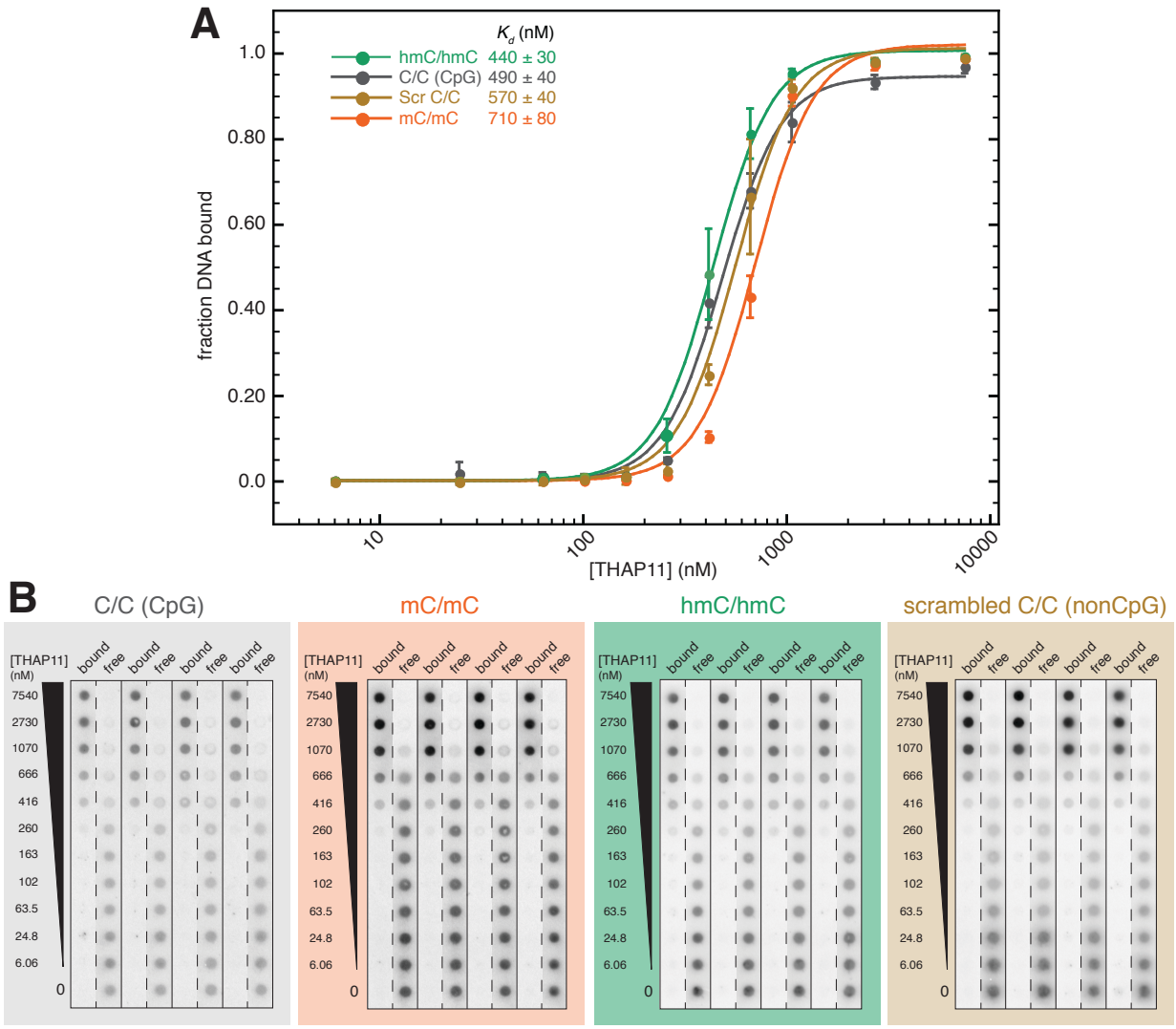


Figure B.8: THAP11 is not an hmC-specific DNA binding protein. The THAP domain of THAP11 (Ronin) was subjected to filter binding studies. The *Kd* measured for hmC and for unmodified C are not significantly different, quantified in (A) and shown as raw filter binding membranes in (B), and the binding hmC is not enhanced relative to the low abundance of this modification in the genome. (*completed in collaboration with Matthew Sullivan*).

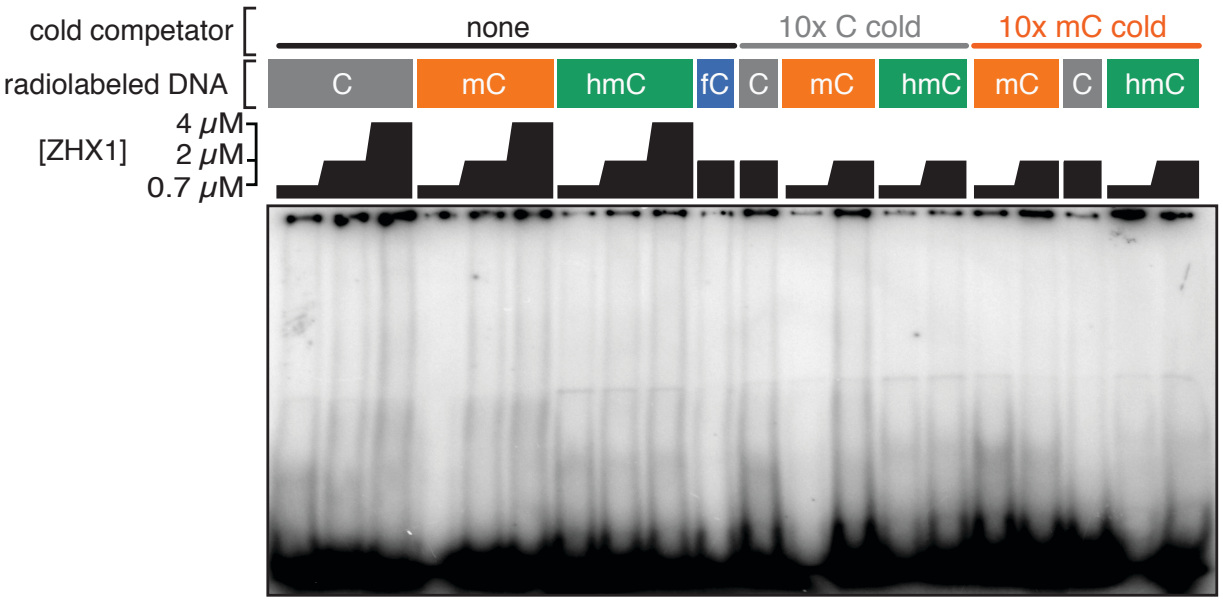


Figure B.9: ZHX1 is not a DNA binding protein under the concentrations of purified protein readily obtainable from bacterial expression. No distinct shift is observed even in the absence of cold competitor challenge, indicating that the protein does not meaningfully bind DNA under these permissive conditions (*completed in collaboration with Matthew Sullivan*).

REFERENCES

Al Mahdawi, Sahar, Sara Anjomani Virmouni, and Mark A Pook (2014). "The emerging role of 5-hydroxymethylcytosine in neurodegenerative diseases." In: *Frontiers in Neuroscience* 8.46, p. 397.

Arita, Kyohei et al. (2008). "Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism." In: *Nature* 455.7214, pp. 818–821.

Avvakumov, George V et al. (2008). "Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1." In: *Nature* 455.7214, pp. 822–825.

Bachman, Martin et al. (2014). "5-Hydroxymethylcytosine is a predominantly stable DNA modification." In: *Nature Chemistry* 6.12, pp. 1049–1055.

Bachman, Martin et al. (2015). "5-Formylcytosine can be a stable DNA modification in mammals." In: *Nature Chemical Biology* 11.8, pp. 555–557.

Barreto, Guillermo et al. (2007). "Gadd45a promotes epigenetic gene activation by repair-mediated DNA demethylation." In: *Nature* 445.7128, pp. 671–675.

Bergman, Yehudit and Howard Cedar (2013). "DNA methylation dynamics in health and disease". In: *Nature Structural & Molecular Biology* 20.3, pp. 274–281.

Bestor, T H (2000). "The DNA methyltransferases of mammals." In: *Human Molecular Genetics* 9.16, pp. 2395–2402.

Bhutani, Nidhi et al. (2010). "Reprogramming towards pluripotency requires AID-dependent DNA demethylation." In: *Nature* 463.7284, pp. 1042–1047.

Bird, Adrian (2002). "DNA methylation patterns and epigenetic memory." In: *Genes & Development* 16.1, pp. 6–21.

Bird, Adrian et al. (1998). "Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex." In: *Nature* 393.6683, pp. 386–389.

Bird, Adrian P and Alan P Wolffe (1999). "Methylation-Induced Repression— Belts, Braces, and Chromatin". In: *Cell* 99.5, pp. 451–454.

Booth, M J et al. (2012). "Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution". In: *Science* 336.6083, pp. 934–937.

Booth, Michael J et al. (2014). "Quantitative sequencing of 5-formylcytosine in DNA at single-base resolution." In: *Nature Chemistry* 6.5, pp. 435–440.

Bostick, Magnolia et al. (2007). "UHRF1 Plays a Role in Maintaining DNA Methylation in Mammalian Cells". In: *Science* 317.5845, pp. 1760–1764.

Brenner, Carmen et al. (2005). "Myc represses transcription through recruitment of DNA methyltransferase corepressor." In: *The EMBO Journal* 24.2, pp. 336–346.

Brinkman, Arie B et al. (2012). "Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk." In: *Genome Research* 22.6, pp. 1128–1138.

Bruniquel, Denis and Ronald H Schwartz (2003). "Selective, stable demethylation of the interleukin-2 gene enhances transcription by an active process." In: *Nature Immunology* 4.3, pp. 235–240.

Castanotto, Daniela et al. (2005). "Short hairpin RNA-directed cytosine (CpG) methylation of the RASSF1A gene promoter in HeLa cells." In: *Molecular Therapy* 12.1, pp. 179–183.

Cedar, Howard and Yehudit Bergman (2009). "Linking DNA methylation and histone modification: patterns and paradigms." In: *Nature Reviews Genetics* 10.5, pp. 295–304.

Chan, Simon W-L et al. (2004). "RNA silencing genes control de novo DNA methylation." In: *Science* 303.5662, pp. 1336–1336.

Clark, S J et al. (1994). "High sensitivity mapping of methylated cytosines." In: *Nucleic Acids Research* 22.15, pp. 2990–2997.

Cliffe, Laura J et al. (2009). "JBP1 and JBP2 are two distinct thymidine hydroxylases involved in J biosynthesis in genomic DNA of African trypanosomes." In: *Nucleic Acids Research* 37.5, pp. 1452–1462.

Cong, Le et al. (2013). "Multiplex genome engineering using CRISPR/Cas systems." In: *Science* 339.6121, pp. 819–823.

Dawlaty, Meelad M et al. (2011). "Tet1 Is dispensable for maintaining pluripotency and its loss is compatible with embryonic and postnatal development". In: *Cell Stem Cell* 9.2, pp. 166–175.

Dawlaty, Meelad M et al. (2013). "Combined deficiency of Tet1 and Tet2 causes epigenetic abnormalities but is compatible with postnatal development." In: *Developmental Cell* 24.3, pp. 310–323.

Dawlaty, Meelad M et al. (2014). "Loss of Tet enzymes compromises proper differentiation of embryonic stem cells." In: *Developmental Cell* 29.1, pp. 102–111.

Day, Jeremy J and J David Sweatt (2010). "DNA methylation and memory formation." In: *Nature Neuroscience* 13.11, pp. 1319–1323.

Di Croce, Luciano et al. (2002). "Methyltransferase recruitment and DNA hypermethylation of target promoters by an oncogenic transcription factor". In: *Science* 295.5557, pp. 1079–1082.

Engel, Nora et al. (2009). "Conserved DNA methylation in Gadd45a(-/-) mice." In: *Epigenetics* 4.2, pp. 98–99.

Falnes, Pål Ø, Rune F Johansen, and Erling Seeberg (2002). "AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia coli". In: *Nature* 419.6903, pp. 178–182.

Ficz, Gabriella et al. (2011). "Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation." In: *Nature* 473.7347, pp. 398–402.

Figueroa, Maria E et al. (2010). "Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation." In: *Cancer Cell* 18.6, pp. 553–567.

Frauer, Carina et al. (2011). "Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain." In: *PloS One* 6.6, e21306.

Gabel, Harrison W et al. (2015). "Disruption of DNA-methylation-dependent long gene repression in Rett syndrome." In: *Nature* 522.7554, pp. 89–93.

Geiduschek, E P, T Nakamoto, and S B Weiss (1961). "The enzymatic synthesis of RNA: complementary interaction with DNA." In: *Proceedings of the National Academy of Sciences* 47.9, pp. 1405–1415.

Globisch, Daniel et al. (2010). "Tissue distribution of 5-hydroxymethylcytosine and search for active demethylation intermediates". In: *PloS One* 5.12, e15367.

Gregg, Christopher et al. (2010). "High-resolution analysis of parent-of-origin allelic expression in the mouse brain." In: *Science* 329.5992, pp. 643–648.

Gu, Tian-Peng et al. (2011). "The role of Tet3 DNA dioxygenase in epigenetic reprogramming by oocytes". In: *Nature* 477.7366, pp. 606–610.

Guo, Junjie U et al. (2011). "Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain." In: *Cell* 145.3, pp. 423–434.

Guo, Junjie U et al. (2014). "Distribution, recognition and regulation of non-CpG methylation in the adult mammalian brain." In: *Nature neuroscience* 17.2, pp. 215–222.

Hackett, J A et al. (2013). "Germline DNA demethylation dynamics and imprint erasure through 5-hydroxymethylcytosine". In: *Science* 339.6118, pp. 448–452.

Hahn, Maria A et al. (2013). "Dynamics of 5-hydroxymethylcytosine and chromatin marks in Mammalian neurogenesis." In: *Cell reports* 3.2, pp. 291–300.

Hashimoto, Hideharu et al. (2009). "UHRF1, a modular multi-domain protein, regulates replication-coupled crosstalk between DNA methylation and histone modifications." In: *Epigenetics* 4.1, pp. 8–14.

Hashimoto, Hideharu et al. (2012). "Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation." In: *Nucleic Acids Research* 40.11, pp. 4841–4849.

Hashimoto, Hideharu et al. (2014). "Wilms tumor protein recognizes 5-carboxylcytosine within a specific DNA sequence." In: *Genes & Development* 28.20, pp. 2304–2313.

He, Yu-Fei et al. (2011). "Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA". In: *Science* 333.6047, pp. 1303–1307.

Hendrich, B and A Bird (1998). "Identification and characterization of a family of mammalian methyl-CpG binding proteins." In: *Molecular and Cellular Biology* 18.11, pp. 6538–6547.

Higa, Leigh Ann et al. (2006). "CUL4-DDB1 ubiquitin ligase interacts with multiple WD40-repeat proteins and regulates histone methylation." In: *Nature Cell Biology* 8.11, pp. 1277–1283.

Hu, Lulu et al. (2013). "Crystal Structure of TET2-DNA Complex: Insight into TET-Mediated 5mC Oxidation." In: *Cell* 155.7, pp. 1545–1555.

Hu, Lulu et al. (2015). "Structural insight into substrate preference for TET-mediated oxidation". In: *Nature* 527.7576, pp. 118–122.

Hu, Xiao et al. (2014). "Tet and TDG Mediate DNA Demethylation Essential for Mesenchymal-to-Epithelial Transition in Somatic Cell Reprogramming". In: *Cell Stem Cell* 14.4, pp. 512–522.

Huang, Hao et al. (2013). "TET1 plays an essential oncogenic role in MLL-rearranged leukemia." In: *Proceedings of the National Academy of Sciences* 110.29, pp. 11994–11999.

Huang, Yun et al. (2010). "The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing." In: *PloS One* 5.1, e8888.

Ito, Shinsuke et al. (2010). "Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification". In: *Nature* 466.7310, pp. 1129–1133.

Ito, Shinsuke et al. (2011). "Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine." In: *Science* 333.6047, pp. 1300–1303.

Iurlaro, Mario et al. (2013). "A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation." In: *Genome Biology* 14.10, R119.

Jeong, Mira et al. (2013). "Large conserved domains of low DNA methylation maintained by Dnmt3a." In: *Nature Genetics* 46.1, pp. 17–23.

Ji, Xiong et al. (2015). "Chromatin proteomic profiling reveals novel proteins associated with histone-marked genomic regions." In: *Proceedings of the National Academy of Sciences* 112.12, pp. 3841–3846.

Jin, Fulai et al. (2013). "A high-resolution map of the three-dimensional chromatin interactome in human cells." In: *Nature* 503.7475, pp. 290–294.

Jin, Seung-Gi, Cai Guo, and Gerd P Pfeifer (2008). "GADD45A does not promote DNA demethylation." In: *PLoS Genetics* 4.3, e1000013.

Jin, Seung-Gi, Swati Kadam, and Gerd P Pfeifer (2010). "Examination of the specificity of DNA methylation profiling techniques towards 5-methylcytosine and 5-hydroxymethylcytosine." In: *Nucleic Acids Research* 38.11, e125–e125.

Jin, Seung-Gi et al. (2011a). "5-Hydroxymethylcytosine is strongly depleted in human cancers but its levels do not correlate with IDH1 mutations." In: *Cancer research* 71.24, pp. 7360–7365.

Jin, Seung-Gi et al. (2011b). "Genomic mapping of 5-hydroxymethylcytosine in the human brain." In: *Nucleic Acids Research* 39.12, pp. 5015–5024.

Jin, Seung-Gi et al. (2016). "Tet3 Reads 5-Carboxylcytosine through Its CXXC Domain and Is a Potential Guardian against Neurodegeneration." In: *Cell Reports* 14.3, pp. 493–505.

Jones, P L et al. (1998). "Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription." In: *Nature Genetics* 19.2, pp. 187–191.

Kaas, Garrett A et al. (2013). "TET1 Controls CNS 5-Methylcytosine Hydroxylation, Active DNA Demethylation, Gene Transcription, and Memory Formation". In: *Neuron* 79.6, pp. 1086–1093.

Kagiwada, Saya et al. (2013). "Replication-coupled passive DNA demethylation for the erasure of genome imprints in mice." In: *The EMBO Journal* 32.3, pp. 340–353.

Kalli, Anastasia and Sonja Hess (2012). "Effect of mass spectrometric parameters on peptide and protein identification rates for shotgun proteomic experiments on an LTQ-orbitrap mass analyzer". In: *Proteomics* 12.1, pp. 21–31.

Kanellopoulou, Chryssa et al. (2005). "Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing." In: *Genes & Development* 19.4, pp. 489–501.

Karmakar, Subhradip et al. (2010). "A multiprotein complex necessary for both transcription and DNA replication at the β-globin locus." In: *The EMBO Journal* 29.19, pp. 3260–3271.

Kawasaki, Hiroaki and Kazunari Taira (2004). "Induction of DNA methylation and gene silencing by short interfering RNAs in human cells." In: *Nature* 431.7005, pp. 211–217.

Keller, Andrew et al. (2002). "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." In: *Analytical Chemistry* 74.20, pp. 5383–5392.

Khare, Tarang et al. (2012). "5-hmC in the brain is abundant in synaptic genes and shows differences at the exon-intron boundary." In: *Nature Structural & Molecular Biology* 19.10, pp. 1037–1043.

Kieffer-Kwon, Kyong-Rim et al. (2013). "Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation". In: *Cell* 155.7, pp. 1507–1520.

Kim, D et al. (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In: *Genome Biology* 14.4, R36.

Klose, Robert J and Adrian P Bird (2006). "Genomic DNA methylation: the mark and its mediators." In: *Trends in Biochemical Sciences* 31.2, pp. 89–97.

Ko, Myunggon et al. (2011). "Ten-Eleven-Translocation 2 (TET2) negatively regulates homeostasis and differentiation of hematopoietic stem cells in mice." In: *Proceedings of the National Academy of Sciences* 108.35, pp. 14566–14571.

Kriaucionis, Skirmantas and Nathaniel Heintz (2009). "The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain." In: *Science* 324.5929, pp. 929–930.

Laird, Charles D et al. (2004). "Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules." In: *Proceedings of the National Academy of Sciences* 101.1, pp. 204–209.

Li, Heng et al. (2009). "The Sequence Alignment/Map format and SAMtools." In: *Bioinformatics* 25.16, pp. 2078–2079.

Li, Xiang et al. (2014). "Neocortical Tet3-mediated accumulation of 5-hydroxymethylcytosine promotes rapid behavioral adaptation." In: *Proceedings of the National Academy of Sciences* 111.19, pp. 7120–7125.

Lian, Christine Guo et al. (2012). "Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma." In: *Cell* 150.6, pp. 1135–1146.

Lin, I G et al. (2000). "Modulation of DNA binding protein affinity directly affects target site demethylation." In: *Molecular and Cellular Biology* 20.7, pp. 2343–2349.

Lister, Ryan et al. (2011). "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells." In: *Nature* 471.7336, pp. 68–73.

Lister, Ryan et al. (2013). "Global epigenomic reconfiguration during mammalian brain development." In: *Science* 341.6146, pp. 1237905–1237905.

Liu, Chungang et al. (2013). "Decrease of 5-hydroxymethylcytosine is associated with progression of hepatocellular carcinoma through downregulation of TET1". In: *PloS One* 8.5, e62828.

Liutkeviciute, Zita et al. (2014). "Direct decarboxylation of 5-carboxylcytosine by DNA C5-methyltransferases." In: *Journal of the American Chemical Society* 136.16, pp. 5884–5887.

Lorsbach, R B et al. (2003). "TET1, a member of a novel protein family, is fused to MLL in acute myeloid leukemia containing the t(10;11)(q22;q23)." In: *Leukemia* 17.3, pp. 637–641.

Lovén, Jakob et al. (2012). "Revisiting global gene expression analysis". In: *Cell* 151.3, pp. 476–482.

Lu, Xingyu et al. (2013). "Chemical modification-assisted bisulfite sequencing (CAB-Seq) for 5-carboxylcytosine detection in DNA." In: *Journal of the American Chemical Society* 135.25, pp. 9315–9317.

Maiti, Atanu and Alexander C Drohat (2011). "Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites." In: *Journal of Biological Chemistry* 286.41, pp. 35334–35338.

Martinowich, Keri et al. (2003). "DNA methylation-related chromatin remodeling in activity-dependent BDNF gene regulation." In: *Science* 302.5646, pp. 890–893.

Meehan, Richard R et al. (1989). "Identification of a mammalian protein that binds specifically to DNA containing methylated CpGs". In: *Cell* 58.3, pp. 499–507.

Meissner, Alexander et al. (2005). "Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis." In: *Nucleic Acids Research* 33.18, pp. 5868–5877.

Meissner, Alexander et al. (2008). "Genome-scale DNA methylation maps of pluripotent and differentiated cells." In: *Nature* 454.7205, pp. 766–770.

Mellén, Marian et al. (2012). "MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system." In: *Cell* 151.7, pp. 1417–1430.

Métivier, Raphaël et al. (2008). "Cyclical DNA methylation of a transcriptionally active promoter". In: *Nature* 452.7183, pp. 45–50.

Mette, M F et al. (2000). "Transcriptional silencing and promoter methylation triggered by double-stranded RNA." In: *The EMBO Journal* 19.19, pp. 5194–5201.

Moran-Crusio, Kelly et al. (2011). "Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation." In: *Cancer Cell* 20.1, pp. 11–24.

Morselli, Marco et al. (2015). "In vivo targeting of de novo DNA methylation by histone modifications in yeast and mouse." In: *eLife* 4, e06205.

Muramatsu, M et al. (2000). "Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme." In: *Cell* 102.5, pp. 553–563.

Nabel, Christopher S et al. (2012). "AID/APOBEC deaminases disfavor modified cytosines implicated in DNA demethylation." In: *Nature Chemical Biology* 8.9, pp. 751–758.

Nady, Nataliya et al. (2011). "Recognition of multivalent histone states associated with heterochromatin by UHRF1 protein". In: *Journal of Biological Chemistry* 286.27, pp. 24300–24311.

Neri, Francesco et al. (2015). "Single-base resolution analysis of 5-formyl and 5-carboxyl cytosine reveals promoter DNA methylation dynamics." In: *Cell Reports* 10.5, pp. 674–683.

Niehrs, Christof and Andrea Schäfer (2012). "Active DNA demethylation by Gadd45 and DNA repair." In: *Trends in Cell Biology* 22.4, pp. 220–227.

Okada, Yuki et al. (2010). "A role for the elongator complex in zygotic paternal genome demethylation." In: *Nature* 463.7280, pp. 554–558.

Okano, M et al. (1999). "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development." In: *Cell* 99.3, pp. 247–257.

Ono, Ryoichi et al. (2002). "LCX, leukemia-associated protein with a CXXC domain, is fused to MLL in acute myeloid leukemia with trilineage dysplasia having t(10;11)(q22;q23)." In: *Cancer Research* 62.14, pp. 4075–4080.

Ooi, Steen K T et al. (2007). "DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA." In: *Nature* 448.7154, pp. 714–717.

Oswald, J et al. (2000). "Active demethylation of the paternal genome in the mouse zygote." In: *Current Biology* 10.8, pp. 475–478.

Pastor, William A et al. (2011). "Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells." In: *Nature* 473.7347, pp. 394–397.

Pichler, Garwin et al. (2011). "Cooperative DNA and histone binding by Uhrf2 links the two major repressive epigenetic pathways". In: *Journal of Cellular Biochemistry* 112.9, pp. 2585–2593.

Popp, Christian et al. (2010). "Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency". In: *Nature* 463.7284, 1101–U126.

Qin, Su and Jinrong Min (2014). "Structure and function of the nucleosome-binding PWWP domain." In: *Trends in Biochemical Sciences* 39.11, pp. 536–547.

Raiber, Eun-Ang et al. (2012). "Genome-wide distribution of 5-formylcytosine in embryonic stem cells is associated with transcription and depends on thymine DNA glycosylase". In: *Genome Biology* 13.8, R69.

Ramsahoye, B H et al. (2000). "Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a." In: *Proceedings of the National Academy of Sciences* 97.10, pp. 5237–5242.

Rao, Suhas S P et al. (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." In: *Cell* 159.7, pp. 1665–1680.

Revy, P et al. (2000). "Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2)." In: *Cell* 102.5, pp. 565–575.

Roberts, Adam et al. (2011). "Improving RNA-Seq expression estimates by correcting for fragment bias." In: *Genome Biology* 12.3, R22.

Robertson, Keith D (2005). "DNA methylation and human disease." In: *Nature Reviews Genetics* 6.8, pp. 597–610.

Rudenko, Andrii et al. (2013). "Tet1 is critical for neuronal activity-regulated gene expression and memory extinction." In: *Neuron* 79.6, pp. 1109–1122.

Santiago, Mafalda et al. (2014). "TET enzymes and DNA hydroxymethylation in neural development and function - how critical are they?" In: *Genomics* 104.5, pp. 334–340.

Scharer, Christopher D et al. (2013). "Global DNA methylation remodeling accompanies CD8 T cell effector function." In: *Journal of Immunology* 191.6, pp. 3419–3429.

Schiesser, Stefan et al. (2012). "Mechanism and stem-cell activity of 5-carboxycytosine decarboxylation determined by isotope tracing." In: *Angewandte Chemie (International ed. in English)* 51.26, pp. 6516–6520.

Scrima, Andrea et al. (2008). "Structural basis of UV DNA-damage recognition by the DDB1-DDB2 complex." In: *Cell* 135.7, pp. 1213–1223.

Seisenberger, Stefanie, Julian R Peat, and Wolf Reik (2013). "Conceptual links between DNA methylation reprogramming in the early embryo and primordial germ cells." In: *Current Opinion in Cell Biology* 25.3, pp. 281–288.

Seisenberger, Stefanie et al. (2012). "The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells". In: *Molecular Cell* 48.6, pp. 849–862.

Sérandour, Aurélien A et al. (2012). "Dynamic hydroxymethylation of deoxyribonucleic acid marks differentiation-associated enhancers." In: *Nucleic Acids Research* 40.17, pp. 8255–8265.

Shen, Li et al. (2013). "Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics." In: *Cell* 153.3, pp. 692–706.

Shin, Hyunjin et al. (2009). "CEAS: cis-regulatory element annotation system." In: *Bioinformatics (Oxford, England)* 25.19, pp. 2605–2606.

Song, Chun-Xiao, Chengqi Yi, and Chuan He (2012). "Mapping recently identified nucleotide variants in the genome and transcriptome." In: *Nature Biotechnology* 30.11, pp. 1107–1116.

Song, Chun-Xiao et al. (2011). "Selective chemical labeling reveals the genome-wide distribution of 5-hydroxymethylcytosine." In: *Nature Biotechnology* 29.1, pp. 68–72.

Song, Chun-Xiao et al. (2012). "Sensitive and specific single-molecule sequencing of 5-hydroxymethylcytosine
In: *Nature Methods* 9.1, pp. 75–77.

Song, Chun-Xiao et al. (2013). "Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming." In: *Cell* 153.3, pp. 678–691.

Spruijt, Cornelia G et al. (2013). "Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives." In: *Cell* 152.5, pp. 1146–1159.

Statham, A L et al. (2012). "Bisulfite sequencing of chromatin immunoprecipitated DNA (BisChIP-seq) directly informs methylation status of histone-modified DNA". In: *Genome Research* 22.6, pp. 1120–1127.

Stroud, Hume et al. (2011). "5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells." In: *Genome Biology* 12.6, R54.

Szulwach, Keith E et al. (2011). "5-hmC–mediated epigenetic dynamics during postnatal neurodevelopment and aging". In: *Nature Neuroscience* 14.12, pp. 1607–1616.

Tahiliani, Mamta et al. (2009). "Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1." In: *Science* 324.5929, pp. 930–935.

Tan, Li et al. (2013). "Genome-wide comparison of DNA hydroxymethylation in mouse embryonic stem cells and neural progenitor cells by a new comparative hMeDIP-seq method." In: *Nucleic Acids Research* 41.7, pp. 84–96.

Trapnell, Cole et al. (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation". In: *Nature Biotechnology* 28.5, pp. 511–515.

Trapnell, Cole et al. (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." In: *Nature Protocols* 7.3, pp. 562–578.

Trewick, Sarah C et al. (2002). "Oxidative demethylation by Escherichia coli AlkB directly reverts DNA base damage". In: *Nature* 419.6903, pp. 174–178.

Valinluck, Victoria et al. (2004). "Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2)." In: *Nucleic Acids Research* 32.14, pp. 4100–4108.

Viré, Emmanuelle et al. (2006). "The Polycomb group protein EZH2 directly controls DNA methylation." In: *Nature* 439.7078, pp. 871–874.

Wagner, Mirko et al. (2015). "Age-dependent levels of 5-methyl-, 5-hydroxymethyl-, and 5-formylcytosine in human and mouse brain tissues." In: *Angewandte Chemie (International ed. in English)* 54.42, pp. 12511–12514.

Wen, Lu et al. (2014). "Whole-genome analysis of 5-hydroxymethylcytosine and 5-methylcytosine at base resolution in the human brain." In: *Genome Biology* 15.3, R49.

Whyte, Warren A et al. (2013). "Master transcription factors and mediator establish super-enhancers at key cell identity genes." In: *Cell* 153.2, pp. 307–319.

Williams, Kristine et al. (2011). "TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity". In: *Nature* 473.7347, pp. 343–348.

Williams, R L et al. (1988). "Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells." In: *Nature* 336.6200, pp. 684–687.

Wu, Hao et al. (2011). "Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells." In: *Nature* 473.7347, pp. 389–393.

Wu, Hao et al. (2014). "Single-base resolution analysis of active DNA demethylation using methylase-assisted bisulfite sequencing." In: *Nature Biotechnology* 32.12, pp. 1231–1240.

Wu, Susan C and Yi Zhang (2010). "Active DNA demethylation: many roads lead to Rome." In: *Nature Reviews Molecular Cell Biology* 11.9, pp. 607–620.

Yildirim, Ozlem et al. (2011). "Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells." In: *Cell* 147.7, pp. 1498–1510.

Yu, Miao et al. (2012a). "Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome." In: *Cell* 149.6, pp. 1368–1380.

Yu, Miao et al. (2012b). "Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine". In: *Nature Protocols* 7.12, pp. 2159–2170.

Zhang, Run-Rui et al. (2013). "Tet1 regulates adult hippocampal neurogenesis and cognition". In: *Stem Cell* 13.2, pp. 237–245.

Zhao, Lei et al. (2014). "The dynamics of DNA methylation fidelity during mouse embryonic stem cell self-renewal and differentiation." In: *Genome research* 24.8, pp. 1296–1307.