Readers of DNA and Histone modifications in Development

Cornelia Gijsbertha Spruijt

ISBN: 978-94-6295-055-9
Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress

Readers of DNA and Histone modifications in Development

Lezers van DNA en Histon modificaties in Ontwikkeling

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties in het openbaar te
verdedigen op woensdag 21 januari 2015 des middags te 2.30 uur

door

Cornelia Gijsbertha Spruijt

geboren op 8 maart 1987 te Amersfoort

Promotoren: Prof. dr. H.Th.M. Timmers
Prof. dr. M. Vermeulen

TABLE OF CONTENTS

Chapter 1

Introduction

EPIGENETICS

DNA contains the blueprint for all proteins in a cell. Even though all the cells in an organism contain the same DNA, over 200 different cell types are known to exist in humans. To achieve this specification, each cell type expresses a different subset of the known ~20.000 genes. All cells belonging to one cell type predominantly express the same set of genes and this state is passed on to daughter cells. The inheritance of gene expression profiles and phenotypic traits without changing the DNA sequence is called epigenetic: "on top of genetic". Epigenetic phenomena include DNA methylation and some of the post-translational modifications (PTM) on core histones.

In eukaryotic cells, DNA is packed in a structural polymer called chromatin. The nucleosome forms the basic repeating unit of chromatin and consists of two copies of each of the four core histones H2A, H2B, H3 and H4 with ~150 base pairs of DNA wrapped around them (Figure 1A and B). The globular domains of the histones form the core particle of the nucleosome while the N-terminal tails, which contain many positively charged residues, protrude from the nucleosome. In addition to serving as a means to store and compact DNA inside the nucleus, nucleosomes regulate nuclear processes such as replication and transcription. Particularly relevant in this context are PTMs on the histone tails such as lysine acetylation or methylation. Some histone modifications are only short-lived, such as serine phosphorylation during mitosis, while other modifications can be passed on to daughter cells to affect gene expression and are therefore called epigenetic modifications. In the next paragraph the biology of histone modifications will be explained in detail.

Histone modifications

As mentioned above, histones can be post-translationally modified in multiple ways. Common modifications include lysine acetylation, lysine and arginine methylation, lysine ubiquitination and threonine, serine and tyrosine phosphorylation. Each modification is denoted by histone name, residue and type of modification. Trimethylation of lysine 4 on histone H3, for example, is denoted as H3K4me3.

The total of all histone modifications and their functions is often referred to as the histone code. The proteins regulating this code are called 'writer' (adds a modification), 'eraser' (removes a modification) and 'reader' (interacts with a modification) (figure 1C)[1]. Each of the histone modifications has a specific functional effect that is achieved by a change in charge of the protein or through specific recognition of the modification by the readers. Furthermore, each modification is reversible, like many biological signals, and can be regulated in a positive and negative manner. Table 1 gives an overview of the different writers, erasers and readers for the most common types of histone modifications [1, 2].

Lysine acetylation neutralizes the positive charge of the amino acid, which results in reduced electrostatic interactions between the negatively charged DNA and the positively charged lysine. It thereby leads to decondensation of the chromatin. Decondensation of chromatin positively affects transcription and other processes which require access to the DNA. Furthermore, acetylated lysines can be recognized by proteins with a Bromo-domain, which are often associated with protein complexes that activate transcription. Acetylated histones are mainly found on promoters and in gene bodies of active genes, as well as on active enhancers.

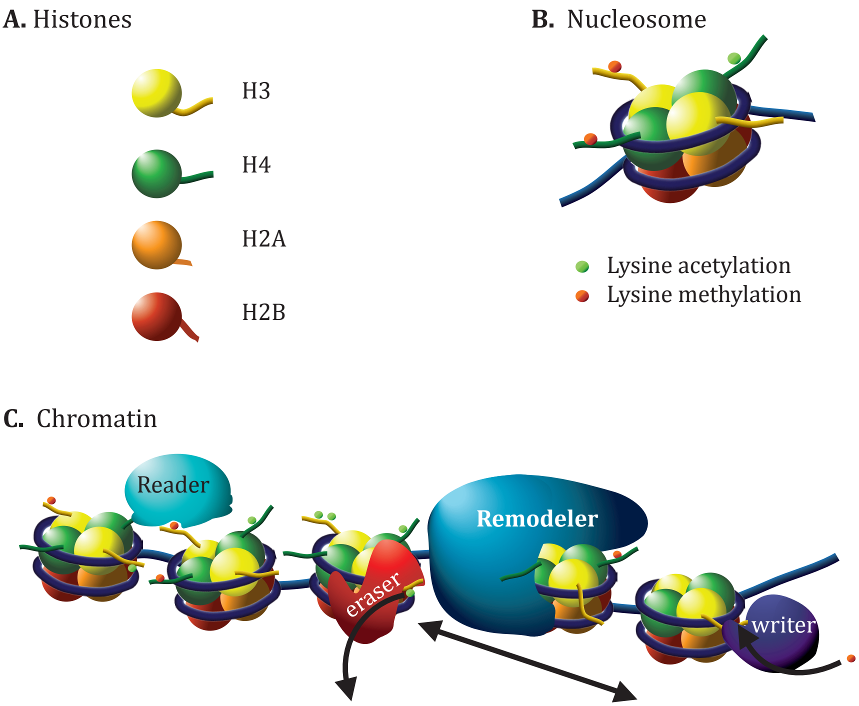


Figure 1: Schematic representation of chromatin. A. The four different core histones with a globular body and an extending N-terminal tail are represented by different colors. **B.** A nucleosome consists of an octamer of the four core histones, with 147 base pairs of DNA wrapped around it. **C.** Chromatin is the structure formed by arrays of nucleosomes including all the proteins that bind to them to modify or remodel them.

Table 1

Modification	Writer	Eraser	Reader
Lysine methylation	KMT, HMT	KDM, LSD	PHD, Chromo, Tudor, MBT, Ankyrin, PWWP
Arginine methylation	PRMT		WD40
Lysine acetylation	HAT	HDAC	Bromo
Lysine Ubiquitination	E1,E2,E3	DUb	UIM, UBA, UEV
S, T, Y phosphorylation	Kinase	Phosphatase	14-3-3
Unmodified histone tails			PHD, WD40

Lysine methylation occurs in three gradations: mono-, di- or trimethylation. Arginines can be mono-, symmetrically di- or asymmetrically dimethylated. Histone methylations can be recognized by a number of domains (Table 1) and the position of the methylation on the histone tails determines the biological outcome. H3K4me3 is found on promoters of actively transcribed genes and H3K36me3 marks gene bodies of active genes. In contrast, some histone methylations are repressive to transcription, such as H3K9me3 and H3K27me3 [1, 2]. These different functions for trimethylated lysines on histone H3 can be explained by the fact that each of these modifications is 'read' by different readers. H3K4me3, for example is read by transcriptional activators such as TFIID and SAGA [3], whereas H3K27me3 is read by repressive Polycomb proteins [4]. Readers for different lysine methylations have been identified in HeLa cells or other commonly used cancer cell lines using mass spectrometry-based screenings [4-6]. In addition, binding specificities of candidate chromatin-associated domains for epigenetic modifications have been determined using protein micro-arrays [7]. Some reports have described cell-type specific chromatin readers, such as the H3K4me3 reader protein RAG2, which is specifically expressed in cells of the immune system and which regulates VDJ recombination [8]. Whether many of such cell-type specific readers for common epigenetic modifications exist remains unclear.

Histon modifications exist in many different forms. However, gene expression is not only epigenetically regulated by histone modifications but also by DNA methylation, which will be introduced in the next section.

DNA methylation

In vertebrates, DNA methylation is a common epigenetic modification. It entails the addition of a methyl (CH3) group to the carbon on the fifth position of the cytosine base (mC) (figure 2A). DNA methylation mostly occurs in the context of palindromic CpG dinucleotides and can exist in a hemi- (half) or fully methylated state. In mammals, the genome is globally methylated with some patches being hypo-methylated. These so-called CpG islands (CGI) are very GC-rich, are found at about 80% of all mammalian promoters and are highly conserved [9].

DNA methylation is catalyzed by DNA methyltransferases (DNMTs). Mammals contain three DNMTs: 'maintenance' DNMT1, and the '*de novo*' methyltransferases DNMT3A and 3B [10]. DNMT1, which is associated with Uhrf1, recognizes the hemi-methylated DNA that is formed after DNA replication. Since only the parental strand is methylated, DNMT1 maintains genomic methylation patterns by symmetrically methylating the daughter strand (Figure 2B). The exact mechanisms of DNMT3A and B recruitment to non-methylated DNA are still unclear, but they are thought to be guided by non-coding RNA. Whether an enzyme exists that can actively demethylate DNA is still unclear. However, in 2009 the enzymatic conversion of methylcytosine to hydroxymethylcytosine by TET proteins was observed [11, 12].

The family of TET proteins catalyzes the iterative oxidation of methylcytosine into hydroxymethylcytosine (hmC), formylcytosine (fC) and carboxylcytosine (caC) (Figure 2C) [13]. These modifications are particularly abundant in mouse embryonic stem cells (mESC) and brain, whereas they are generally low abundant in immortalized cell lines [14]. Hydroxymethylcytosine is about an order of magnitude more abundant than fC and two orders of magnitude more abundant than caC [15]. The function of hmC is still unknown, although it has been suggested to have a function in transcription regulation. Since the further oxidation products, fC and caC, can serve as substrates for Thymine-DNA glycosylase (TDG), these oxidized cytosine derivatives most likely are part of an active DNA demethylation pathway [16, 17]. The high levels of hmC in ESC and brain suggests additional functions for this modification that may be regulated by (tissue-)specific readers. However, only a few proteins, such as MeCP2 and MBD3, have so far been described to recognize this modification [18, 19].

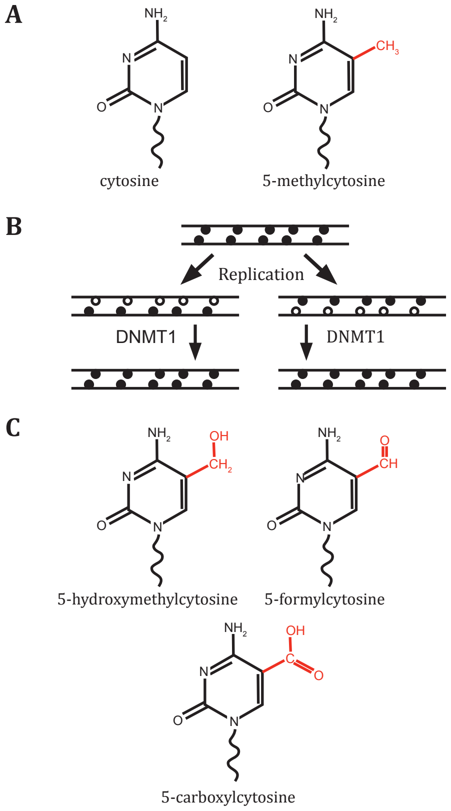


Figure 2: DNA modifications. A. Chemical structure of non-modified and 5-carbon methylated cytosine. **B.** After DNA replication of fully methylated DNA, two hemi-methylated DNA duplexes are formed. These are recognized by DNMT1, which subsequently methylates the non-methylated daughter strands. **C.** Chemical structures of hydroxymethylcytosine, formylcytosine and carboxylcytosine.

DNA methylation is associated with repression of transcription initiation and plays an important role in X-chromosome inactivation and during cellular differentiation [20]. Female mammals have two X-chromosomes, whereas males have only one. To compensate the expression of genes located on the X chromosome, one of the two X chromosomes in females is inactivated by epigenetic modifications, including DNA methylation. A similar silencing is observed in differentiating cells, in which genes that are not required for differentiation into the target lineage are silenced by DNA methylation [21]. These phenomena, in addition to results obtained using reporter systems [22], indicate that DNA methylation represses gene expression *in cis*.

The methyl-group on the fifth carbon of cytosine is positioned in the major groove of the DNA, and neither affects the DNA sequence significantly nor does it affect the efficiency of transcription by RNA polymerase II. So how can the repressive effect of DNA methylation on transcription be explained? Two major mechanisms have been proposed. The first is DNA methylation-mediated inhibition of transcription factor binding to DNA. Transcription factors whose binding to DNA is known to be negatively affected by DNA methylation include MYC, CXXC1 (CFP1) and MLL [23, 24]. CXXC1 and MLL are incorporated in transcriptional activating COMPASS complexes. These proteins bind to DNA via a so-called CXXC domain and binding of this domain to CG-rich DNA is inhibited by DNA methylation. The second mechanism through which DNA methylation silences transcription is by recruiting methyl-CpG binding proteins (MBPs). Three MBP families have been described so far: the methyl-CpG Binding domain (MBD)-containing proteins [25], the Kaiso-like proteins that bind to DNA through zinc fingers [26], and the Set-and-Ring-Associated (SRA) domain-containing proteins: Uhrf1 and Uhrf2 [27, 28]. Whether additional proteins exist that can recognize methylated DNA remains unclear. Furthermore, the question is whether these three MBP families cover all putative mCpG-binding domains.

Many of the MBD-containing proteins and other chromatin readers associate with multi-subunit protein complexes that contain multiple reader domains as well as enzymatic activities. By combining reader domains and enzymatic activities, the biological effect of an epigenetic mark can be influenced. The enzymatic activities may be writer or eraser functions for other histone modifications (both called chromatin modifiers) or chromatin remodeling activity. Chromatin remodeling activity is defined as ATP-dependent sliding of nucleosomes on DNA or evicting nucleosomes from DNA to create higher or lower compaction of the chromatin (see Figure 1). Chromatin remodeling therefore can have either an activating or a repressive effect on transcription. Several ATP-dependent chromatin remodeling complexes have been described during the last decade. One of these complexes combining multiple reader proteins, such as MBDs, and enzymatic activities is the Mi-2/NuRD complex, which will be described in detail below.

THE NURD COMPLEX

One of the histone and methyl-DNA reading complexes that contains chromatin remodeling activity is the Nucleosome Remodeling and Deacetylase or Mi-2/NuRD complex. Many different compositions of NuRD exist, of which the MBD2-containing MBD2/NuRD is best known for its methyl-DNA binding capability and was therefore first described as MeCP1 [29]. Figure 3 shows a schematic representation of the different paralogues including their stoichiometries in this multi-subunit protein complex that has a molecular weight of around 1 MDa [30]. The NuRD complex is conserved throughout the animal kingdom, even though invertebrates lack DNA methylation. Table 2 gives an overview of the different NuRD subunits in different species. From this table it is clear that the NuRD complex in invertebrates is much less complex than the mammalian NuRD complex, which harbours many paralogues.

The core of the NuRD complex is formed by the large chromatin remodeling subunits Mi-2α and β, also called Chromodomain-helicase and DNA-binding protein

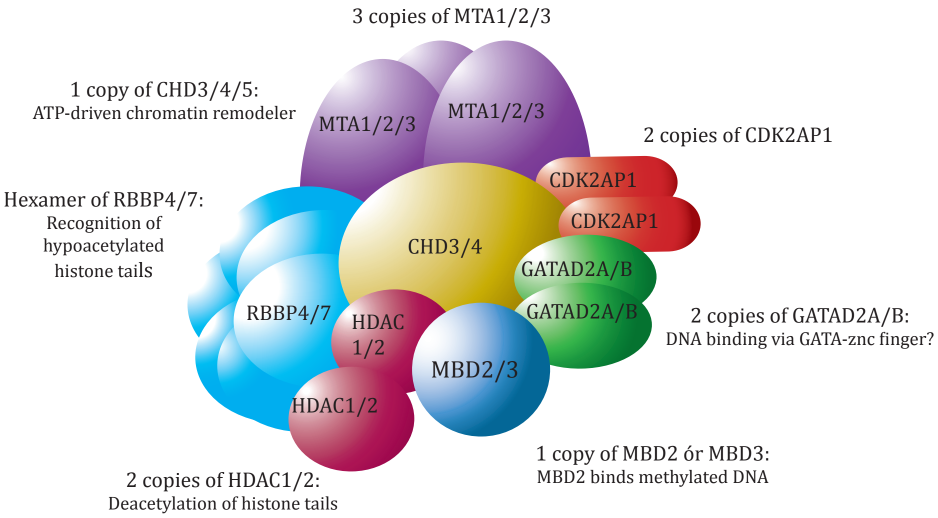


Figure 3: The nucleosome remodeling and histone deacetylase complex. The stoichiometry as well as (putative) function of each of the subunits is indicated.

(CHD) 3 and 4. With a molecular weight of around 220 kDa, these proteins are the largest subunits of the complex. As the name implies, these proteins contain 2 Chromodomains, an N- as well as a C-terminal helicase domain and two PHD-type zinc fingers. The PHD fingers recognize H3-tails that are unmodified at lysine 4 in combination with H3K9me2 or H3K9me3 [31, 32]. In specific tissues such as the central nervous system (CNS) and testis, CHD5 is incorporated in NuRD [33].

As mentioned above, the NuRD complex contains histone deactylase activity. Two copies of HDAC 1 and/or 2 can be incorporated in the complex. HDACs are rather promiscuous enzymes and are also active on non-histone proteins. Furthermore, HDAC1 and 2 are incorporated into many different co-repressor complexes.

In addition to one of the CHD proteins and two HDACs, the NuRD complex contains 4-6 copies of Retinoblastoma Binding Protein (RBBP) 4 and/or 7. These proteins, also called Retinoblastoma Associated protein (RbAp) 48 and 46 respectively, are WD40-repeat proteins of about 48 and 46 kDa. They are part of many co-repressor complexes. The RBBPs are the most dynamic subunits within the NuRD complex and they are thought to function as histone chaperones [34, 35].

Table 2

Function	***Homo sapiens***	***Mus musculus***	***Xenopus leavis***	***Danio rerio***	***Drosphila melanogaster***
ATPase remodeler	CHD3 (MI-2α) CHD4 (MI-2β)	Chd3 Chd4	chd3 chd4	chd3 chd4	Mi-2
Histone deacetylase	HDAC1 HDAC2	Hdac1 Hdac2	hdac1 hdac2	hdac1	Rpd3
	MTA1 MTA2 MTA3	Mta1 Mta2 Mta3	mta1 mta2	mta2	mta1-like
	RBBP4 (RbAp48) RBBP7 (RbAp46)	Rbbp4 Rbbp7	rbbp4a rbbp4b rbbp7	rbbp4 rbbp7	CAF-1
	GATAD2A (p66α) GATAD2B (p66β)	p66α p66β	gatad2a gatad2b	gatad2ab	Simjang
Methyl-CpG binding	MBD2 MBD3	Mbd2 Mbd3	MBD2 MBD3	MBD2 MBD3a MBD3b	MBD2/3
	CDK2AP1 (DOC-1)	Cdk2ap1	cdk2ap1		CG18292

Furthermore, the NuRD complex contains three copies of the Metastasis Tumor Associated protein (MTA) paralogues 1, 2 and 3, which are around 60-80 kDa in size. These proteins contain a BAH, an ELM and a SANT domain in addition to a GATA-type zinc finger that might bind to GATA-like DNA sequences. These proteins have also been shown to interact with non-modified histone tails [36]. MTA2 is ubiquitously expressed, whereas MTA1 and MTA3 are expressed in a tissue-specific way. Both *MTA1* and *MTA3* genes result in two different proteins by alternative splicing. Since *MTA2* is most homologous with MTA-like proteins in invertebrates, this was probably the ancestral gene. A gene duplication event may have created the *MTA1/3* gene, which after another duplication event created two distinct genes that may have gained tissue-specific functions [37, 38]. Each of the MTA proteins pulls down the other two MTA paralogues in affinity purification experiments, indicating that these proteins can form a heterotrimer (unpublished observations). However, Fujita *et al.* did not identify the other MTA proteins in their MTA3 purifications [38].

One copy of either Methyl-CpG binding domain-containing protein (MBD) 2 or 3 is present in each NuRD complex [39]. MBD2 is around 43 kDa, whereas MBD3, which lacks a part of the N-terminus, is around 33 kDa. Both proteins evolved from a single ancestral *MBD* gene capable of binding to methylated DNA [40]. Although both proteins have an MBD, only MBD2 is capable of binding to methylated DNA in mammals [25]. Following a gene duplication event, the affinity of the mammalian MBD3 protein for methylated DNA decreased significantly due to some crucial mutations in the MBD that were apparently not selected against during evolution. In *Xenopus leavis*, both MBD2 and MBD3 are enriched in affinity pull-downs with methylated DNA [41]. Even the *Drosophila melanogaster* orthologue MBD2/3 is capable of binding to methylated DNA [42]. *Danio rerio* contains two *mbd3* genes in addition to its *mbd2* gene. These genes encode MBD3a and MBD3b of which only MBD3a binds methylated DNA, whereas MBD3b lacks a functional MBD [40]. Since MBD2 and MBD3 are mutually exclusive within the NuRD complex, MBD2/NuRD and MBD3/NuRD define functionally distinct NuRD complexes with different DNA binding characteristics. Among others, a recent study by Baubec *et al.* clearly revealed two clusters of MBD2 loci in ESC and neuronal precursor cells [43-45]. One cluster correlated with high levels of DNA methylation. The other, much smaller cluster was present at non-methylated promoters and showed a remarkable overlap with MBD3 and CHD4. How the NuRD complex is recruited to such non-methylated promoters or other functional DNA elements remains to be solved.

Via a coiled-coil domain, MBD2 and 3 can interact with the Conserved Region (CR)1 domains of p66α and β, which are also called GATAD2A and B [46, 47]. The GATAD2A and B proteins, which are around 66 kDa, contain one zinc finger of the GATA-type in addition to two CR regions. The second CR domain binds to hypoacetylated histone tails [48].

The last putative NuRD subunit identified was Deleted in Oral Cancer 1 (DOC-1), which is also called Cyclin-Dependent Kinase 2 Associated Protein 1 (CDK2AP1). This protein has a long (14 kDa) and a short (12 kDa) isoform. It contains an intrinsically disordered N-terminus and seems to lack any functional domains. This protein was identified in MBD2 and MBD3 purifications by Le Geuzennec *et al.* [39]. However, since only a few peptides of the protein were sequenced, it received a low probability score. In *Drosophila melanogaster* an orthologue of this protein was identified as a subunit of the NuRD complex [49].

Table 2 shows the different paralogues of each subunit in different species. Most subunits have only one paralogue in fruit fly, while two to three paralogues of each subunit are present in mammals. As described above, MBD2 and MBD3 have diverged functions. Whether genome duplication events also resulted in specialization of the other paralogues within NuRD is currently largely unclear. Some of the proteins are described to have many different functions, while for others their functions remain unclear. The possibility that so many different combinations of subunits exist that may result in a large number of different NuRD subcomplexes compromises a comprehensive characterization of the complex.

In addition to the core subunits, a number of interactors for NuRD is known. These include FOG-1 (Friend of GATA), SALL4 and Lysine specific demethylase 1 (LSD1 or KDM1A), which demethylates H3K4me2 and H3K4me1 [35, 50-52]. FOG-1 and SALL proteins interact with NuRD via a conserved motif of ~20 amino acids at their N-terminus [53]. Presumably, different sequence-specific transcription factors may also transiently interact with the NuRD complex. Some of the interactions are cell type- or tissue-specific. Tissue-specific expression and paralogue-specific roles of NuRD subunits during development will be described in the next section.

Role in development

Like many chromatin associated protein complexes, the NuRD complex is important during development. Some of the mammalian paralogues are expressed at specific stages during differentiation or in specific tissues. For example, the previously described MTA1 and 3 that are tissue-specific, while MTA2 is ubiquitously expressed [38]. Another tissue specific NuRD paralogue is CHD5, which is only expressed in brain and testis [33]. The structure of this protein is very similar to its paralogues CHD3 and CHD4. Why exactly brain and testis require a different CHD protein with the same nucleosome binding characteristics is unclear.

Furthermore, during development, MBD2 and 3 expression seem to anti-correlate: in mESC, MBD2 levels are low and a splice-variant that might be incapable of binding to methylated DNA may be expressed [54], whereas MBD3 levels are high in these cells. MBD3 is important for maintenance of the differentiation potential. MBD3 knock-out mESCs can self-renew, but they are unable to differentiate [55]. Additionally, knock-down of different NuRD subunits enables faster and more efficient reprogramming, whereas overexpression of MBD3/NuRD blocks reprogramming by silencing pluripotency genes [56, 57]. The fact that an MBD3 knock-out is embryonic lethal, whereas MBD2 knock-out mice only show defects in maternal behaviour, further strengthens the role of MBD3 in development [58]. In addition, this illustrates that MBD2 and MBD3 are not functionally redundant.

Evidence for DNA-methylation independent roles of the NuRD complex during development comes from research in invertebrates that lack DNA methylation. Like mammalian MBD3, the planarian MBD2/3 ortholog is required for adult stem cell pluripotency. In the study by Jaber-Hijazi *et al.* MBD2/3 was not expressed in the regeneration blastula during the first days after truncation of the planarian head [59]. However, at day 5 of regeneration, MBD2/3 was detectable in these tissues. Furthermore, MBD2/3 is not required for stem cell maintenance, but is required for homeostasis of the animal by adult stem cell differentiation, which seems to be impaired in MBD2/3 knock-down worms. These processes are independent of DNA methylation. Furthermore, *Drosophila* MBD2/3 knock-out embryos are viable and fertile, similar to MBD2 knock-out mice [60]. This is surprising, since mice express the additional MBD3 protein that is required for development, whereas Drosophila does not have an additional paralogue. In summary, differentiation seems to require MBD3 in a DNA methylation-independent manner.

The MBDs are not the only NuRD subunits that are important during development. Deletion of the *Drosophila* GATAD2A/B ortholog Simjang results in developmental defects like shortened or bent legs and wings [61]. Furthermore, the animals die in late larval or early pupal stages. This is caused by defects in downstream

Wnt and Ecdysone signalling. Another study showed defects in neuronal development in Drosophila expressing Simjang loss-of-function mutants that were originally identified in patients with neuronal defects [62]. In contrast to mouse GATAD2B mutations that cause slight developmental defects, loss of function mutants of mouse GATAD2A are embryonic lethal [63]. The specific effects of the mutations are not known, but all embryonic tissues are affected and the embryos disintegrate by day 10.5 of embryogenesis.

In contrast to the above described studies that focus on a single NuRD subunit, a recent study in zebrafish established a role for the NuRD complex in regeneration [64]. The amputation of the caudal fin was used to show that knock-down of chd4, rbbp4 and mta2 decreases the regeneration outgrowth. In agreement with these results, treatment of the amputated fin with HDAC inhibitors also slowed down regeneration.

In short, the requirement of some NuRD subunits during development and the observed tissue-specifc expression of some NuRD core subunit paralogues hints at different functions. In addition, the recruitment to different target genes in different cell types caused by differential interactions with DNA binding proteins may also contribute to specialization of NuRD complexes. For example, the mESC-specific NuRD interactors SALL1 to 4 show decreased expression levels in differentiated tissues. Since these factors are most likely sequence-specific DNA binding factors, the NuRD complex may be recruited to different target genes in mESCs versus differentiated cells. Thus, differential interactions with tissue- and cell type-specific transcription factors may affect target gene specificity and therefore the molecular pathways affected by the NuRD complex in different cell types.

Role in DNA repair and aging

NuRD has not only been described in relation to early development, it has also been implicated in the DNA damage response (DDR) and aging. The fact that CHD4, MTA1/2 and HDAC1 have all been described in relation to DNA damage repair presents quite strong evidence for the entire NuRD complex being involved [65-67]. However, many different compositions of NuRD exist and some of the subunits, for example CHD4, can function independently of NuRD [68]. The exact function of these proteins in relation to DNA damage is not known, but they are recruited rapidly to sites of DNA damage and are likely silencing the damaged genes until DNA repair is completed.

Accumulation of DNA damage eventually results in aging, a process that NuRD might also be involved in. The levels of a number of NuRD subunits decrease in aging brain cells, while the levels of other co-repressor complexes are stable [69]. This reduction seems to be caused by Progerin, a protein that upon overexpression reduces the NuRD levels in healthy cells. A number of neuronal diseases are linked to reduced NuRD levels, and in these cases a partial loss of heterochromatin formation is also observed. Overexpression of NuRD subunits in a model cell line seems to rescue the loss of heterochromatin phenotype seen upon RbAp48 or 46 knock-down.

The levels of some NuRD subunits also decrease in diseases that are characterized by early aging, such as Huntington's and Parkinson's disease [33]. When NuRD is lost or reduced, for example in an aging brain, DNA damage may accumulate, given the apparent role for NuRD in the DDR. This accumulation of DNA damage in its turn could cause a phenotype associated with aging.

Role in cancer

Since regulation of transcription is such a fundamental process in cells, deregulation of transcription is likely to cause diseases. Indeed, many NuRD subunits are associated with cancer, as summarized in Table 3 (and reviewed in [70, 71]).

Table 3

Subunit	Up- or down-regulation	Type of cancer	Reference
MTA1	Up	Gastrointestinal carcinomas	[72, 73]
MTA2	Up Up	Breast cancer Non-small cell lung cancer (NSCLC)	[74] [75]
MTA3	Down	Metastasis of breast cancer	[76]
RBBP4	Up	Thyroid cancer	[77]
RBBP7	Up	Leukemia	[78]
CDK2AP1	Down Down Down	Oral cancer Lung cancer Prostate cancer	[79] [80] [81]
CHD5	Down	Neuroblastoma, gliomas, breast, colon, lung, ovary and prostate cancers	[82]
MBD2	Down	Protection against Colon cancer in a mouse model	[83]

For many of these mutations, the molecular mechanism of cancer development is unknown. However, for RBBP4, Pacifico *et al.* showed that expression is enhanced by the hyperactive NFKB signalling in thyroid cancer. Both the inhibition of NFKB signalling, as well as the downregulation of RBBP4, decreased growth in an anaplastic thyroid cancer cell line [77]. In contrast, the expression levels of CDK2AP1 have been shown to decrease in more progressive stages of this oral cancer by immunohistochemistry (IHC) [79]. Whether the loss is a consequence of deregulation of another protein or whether the loss has any causative role in development and progression of the disease is unknown. In addition to the roles of (individual) subunits of the NuRD complex, many substoichiometric interactors of NuRD are also involved in cancer.

Summarizing all the above about the NuRD complex, one could say a lot is known about it. On the other hand, one could say: much is known about its subunits, since many studies have only focused on a single subunit within the complex. Paralogues, which exist for almost every subunit, may specify distinct NuRD complexes, each having a slightly different function. These NuRD complex(es) are involved in tightly regulated processes such as DNA repair as well as transcription regulation during development. Therefore, mutation, deletion or overexpression of only a single subunit is likely to disrupt the function of different NuRD complexes and is potentially pathogenic. Understanding the molecular mechanisms of NuRD functioning and malfunctioning thus will be a key step towards the development of anti-cancer drugs.

MASS SPECTROMETRY

During the last decade, the field of mass spectrometry-based proteomics has evolved very rapidly. The current state-of-the-art allows the identification of some 4000 proteins or so in a one-hour LC-MS/MS run [84]. These developments have been made possible both by novel, more sensitive and fast scanning instrumentation as well as computational progress. In the context of the research field of interaction proteomics, this technology enables an unbiased and high-throughput identification of protein-protein and protein-DNA interactions using single affinity purifications from crude lysates. However, the majority of identified proteins in such affinity purifications will be formed by high abundant background proteins that bind non-specifically to the beads while only a small fraction of the identified proteins consists of specific interactors. A quantitative filter is needed to distinguish these specific interactors from the nonspecific background proteins. In the following paragraphs, basic sample preparation methods and the mass spectrometry set-up used in this thesis will be described, followed by a detailed description of methods that were applied to identify specific protein-protein and protein-DNA interactions.

Sample preparation and mass spectrometry set-up

The workflow of a typical proteomics experiment is schematically depicted in Figure 4. To analyse a protein sample by mass spectrometry the protein mixture is first denatured, for example using urea. Subsequently, all disulfide bonds and cysteines are reduced and alkylated to prevent (formation of) disulfide bonds between peptides that would make it impossible to identify them. The proteins are then digested using a well-characterized protease, such as trypsin. This protease is stable in 2M urea and cleaves C-terminal of lysines and arginines, which means the resulting peptides will all have a double positive charge at acidic pH, which is useful at a later stage. The peptide mixture then needs to be desalted, which is achieved via Stop-and-Go-extraction tips (STAGE-tips) [85]. These pipette tips contain a plug of C18 material, to which the peptides bind due to hydrophobic interactions. The small columns are then washed with a mass spectrometer-compatible solvent to remove all salts. The samples are eluted using acetonitrile, which is evaporated before the samples are analyzed by LC-MS/MS.

The mass spectrometry set-up used for the research described in this thesis, consists of an LTQ Velos orbitrap mass spectrometer (Thermo) connected online to a nano-HPLC system (Proxeon). The peptides are loaded onto the analytical C18 column (25-30 cm long, 75 µm inner diameter with a fused-silica emitter and packed with 3 µm beads) and eluted using a segmented gradient (usually 2 hours) mixed of solvent A (0.1% fromic acid in H_2O) with increasing solvent B (80% acetonitrile / 0.1% formic acid in H_2O). This set-up enables a gradual release of peptides based on their hydrophobicity, which reduces the number of peptides that enter the mass spectrometer simultaneously. Since a mass spectrometer detects peptide ions, separation of peptides by a C18 column is combined with a technology called Electro-Spray Ionisation (ESI) [86]. When arriving at the tip of the emitter, which is positively charged, the peptides fly towards the orifice of the mass spectrometer, which is negatively charged. The heath caused by the electrical field evaporates the solvent, while positively charged protons remain on the peptides as they enter the gas phase.

The peptides are guided through the mass spectrometer in an electrical field. Detection of the peptide-ions is based on their mass (m) divided by their charge (z),

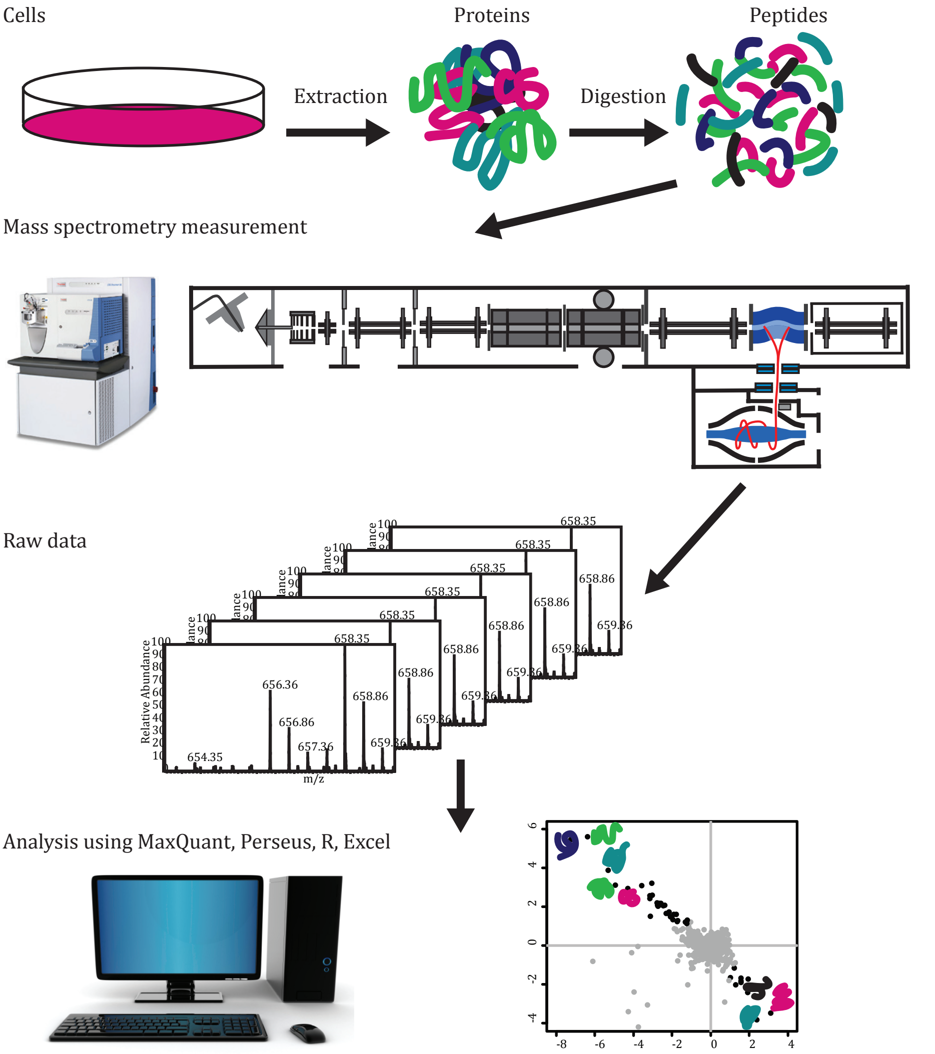


Figure 4: Schematic workflow of a typical proteomics experiment. Proteins are extracted from cells, digested to form peptides. The peptide mixtures are measured on the mass spectrometer, after which peptides and proteins can be identified and quantified by specialized software.

and therefore is depicted as m/z. In the Orbitrap cell, peptide ions are oscillating around a central electrode and an image current is detected. The oscillation frequency is directly proportional to their m/z resulting in mass measurement with sub-ppm accuracy for each peptide. Multiple ions can be detected simultaneously using Fourier transformation of overlapping frequencies. After a ‘full scan’, which measures all ions present in the Orbitrap detector at a certain moment, ions within a mass window of a few dalton are selected for fragmentation using varying voltages in the ion trap. Fragmentation of these ‘parental’ ions occurs via collisions with an inert gas and is called collision induced dissociation (CID) [87]. CID causes peptides to preferentially break on their peptide bonds. Due to the distribution of the double positive charge of the peptide ions, fragmentation results in a *b* and a *y* fragment ion. Thousands of peptide ions fragmenting at different peptide bonds thus lead to a range of *b* and *y* ions from which amino acid sequence information can be derived. This spectrum is called an MS/MS, tandem-MS or MS2 spectrum. In the LTQ Orbitrap Velos, a CID-based MS/MS spectrum is recorded in the linear ion trap.

In the mass spectrometry-based proteomics field, a protein is identified in a database search based on the accurately determined precursor mass of a peptide derived from that protein in combination with the (partial) sequence information based on the fragmentation spectrum of this peptide. The accurate mass gives an indication for the amino acid composition of the peptide. The database used for the search contains all possible protein sequences and splice isoforms from a single species. The analysis software, in this case MaxQuant [88], makes an *in silico* digest of these proteins with the protease used in the actual experiment. In addition, it makes a pseudo-reversed decoy database, in which all peptides from the normal database have the reverse sequence. This decoy database thus contains peptides with exactly the same amino acid compositions and length distribution, which can be used to estimate false positive scores (FDR) [89]. A match between the combination of precursor mass and fragment masses with a peptide in the database is reported and scored. The more peptides of a single protein are identified in the sample (coverage), the higher is the confidence that this protein is a true positive. When two similar proteins, such as paralogues of NuRD subunits, 'share' peptides, this is also reported, but the intensity of the peptide is assigned to the protein with the most 'unique peptides'. When a match with the decoy database is found, this is a known false positive. Counting the number of known false positive hits helps to calculate the number of unknown false positive hits and thus can be used to determine the FDR cut off for the entire identification list.

Although modern mass spectrometers are able to sequence thousands of peptides in a short period of time, the instruments are usually not able to sequence all peptides in a complex mixture. Fractionation is therefore important to reduce sample complexity, allowing identification of more proteins in a particular sample of interest. This, however, comes at the cost of increased measurement time. Fractionation can either be performed at the protein level or the peptide level. In this thesis, SDS-PAGE was used for fractionation at the protein level. Proteins are separated based on their size and the gel lane can be divided into multiple slices (8-10), which are then analyzed by LC-MS/MS separately. The proteins in each slice are digested in-gel with trypsin. For sample fractionation at the peptide level, strong anion exchange (SAX) was used. The proteins are first digested into peptides and then loaded onto a SAX-column at basic pH. In multiple elution steps with decreasing pH, peptides are eluted from the column. After desalting, the peptides are analyzed by LC-MS/MS. Another way to reduce the complexity of the sample is to perform an affinity enrichment step for a protein or PTM of interest.

In this thesis, a number of affinity enrichment strategies combined with quantitative mass spectrometry-based proteomics technology are applied to answer a variety of questions in the research field of epigenetics. Immobilized *in vitro* synthesized modified histone peptides or stretches of DNA containing (hydroxy)methylated cytosines are incubated with nuclear extracts to identify readers for these epigenetic modifications. Furthermore, protein tagging and purifications are performed to identify interaction partners for proteins of interest. As mentioned before, such affinity purifications result in the detection of a large number of non-specific (high abundant) background binders. To distinguish these background proteins from the specific interactors, several quantitative methods can be used. In this thesis two quantitative filtering techniques were used to distinguish true interactors from background binders. Both of these techniques will be described in detail below.

Quantitative mass spectrometry

Mass spectrometry is not inherently quantitative. The summed peptide peak intensities, on which some quantification applications in the field of mass spectrometry rely [90, 91], is not comparable for every protein. First of all, because of their different physical properties, such as length, some peptides are ionized much easier than others, which also influences the observed intensity. Second, the size of the protein determines how many tryptic peptides, and thus how much intensity, can be observed. Last but not least, due to technical issues not every mass spectrometry measurement is identical, making it difficult to compare intensities of the same protein in different mass spectrometric samples. The two methods used in this thesis to compare protein levels in different samples are SILAC and label-free quantification.

SILAC

When protein abundances in two different samples need to be compared by mass spectrometry, stable isotope labeling approaches can be used. Stable isotopes can be introduced at different stages during the sample preparation, both at the protein and at the peptide level. Peptide level-based labeling strategies include di-methyl, iTRAQ, TMT and ICAT [92, 93]. The most commonly used metabolic labeling technique is Stable Isotope Labeling by Amino acids in Cell culture, or SILAC, which is used to incorporate isotope labels at the protein level [94].

In short, cells are grown in growth medium containing either 'light' or 'heavy' amino acids, in which variable numbers of ^{13}C and ^{15}N isotopes are incorporated. The amino acids used for stable isotope labeling need to be compatible with the protease used for the digestion of the proteins. Trypsin, which cleaves C-terminally of lysines and arginines, was used in this thesis in combination with labeled lysine and arginine. The use of this combination results in labeling of every single peptide except for the most C-terminal one of a protein, enabling thorough quantification of all the peptides and thus proteins. Labeled lysine has a molecular weight of 4 or 8 Dalton more than the naturally occurring lysine (K4 and K8, respectively). For arginine, the mass differences with the naturally occurring variant are 6 and 10 Dalton (R6 and R10, respectively). When both amino acids are used for labeling, three conditions can be tested in a single experiment: 'light' (K0R0, the naturally occurring isotopes), 'medium' (K4R6) and 'heavy' (K8R10). The used ^{13}C and ^{15}N isotopes have no effect on the physio-chemical behaviour of the peptides during chromatography or in the mass spectrometer.

SILAC labeled cells (or the extracts of the cells) can be used to compare two experimental conditions. Examples are: the effect of a drug on the protein content of cells, DNA or peptide affinity purifications with and without modification, or GFP-affinity purifications for which a control purification is performed from the differentially labeled cell extracts (Figure 5A) [95, 96]. After the differential steps, the samples are combined (the 'light' control with the 'heavy' experiment), trypsin digested and measured in a single mass spectrometry measurement. This will result in pairs of peptide peaks (light and heavy peptide) from which the ratio can be determined based on integrated peak volume. A one-to-one ratio means that the protein was a background protein, whereas a high heavy/light ratio shows enrichment of a protein in the specific experiment. This experimental set up is called the 'forward' experiment. In addition, a label-swap or 'reverse' experiment can be performed in which the control experiment is performed with heavy extract and the specific experiment with light extract. All the proteins for which the ratios invert in this experiment are enriched (or depleted) in the specific experiment, and thus of interest. To visualize the proteins, a scatterplot can be made with the log2(heavy/light) in the forward experiment on the X-axis and the log2(heavy/light) in the reverse experiment on the Y-axis (Figure 5A). The background proteins, which will have one-to-one ratios, will cluster around the origin of the figure, whereas proteins enriched in the specific experiment show a high forward and a low reverse ratio. Proteins depleted in the specific experiment are visualized in the opposite quadrant. Contaminants, such as keratins that are often identified in mass spectrometry studies, are only present in the naturally occurring 'light' state, so they have both a low forward and a low reverse ratio and can be easily distinguished.

SILAC is applicable to many cell lines that are commonly used and even a number of organisms, such as *C. elegans* and mice, can be labeled [97, 98]. However, SILAC mouse tissues are very expensive, and for some research questions that require comparison of more than three conditions, the use of SILAC is more complicated. To compare whole proteomes, for example of tumor samples, a heavy SILAC labeled spike in can be used, referred to as super-SILAC [99]. Alternatives are chemical stable isotope labeling methods on peptide level, such as dimethyl (cheaper) or iTRAQ (enabling higher multiplexing) [92]. However, stable isotope labeling has more disadvantages, such as doubling of the complexity of the sample. Having a light and a heavy peak for each peptide may compromise the measuring depth and number of identified proteins of your experiment. This is not a problem when iTRAQ is applied, since the intensities of the different labels are only visible after fragmentation. However, this approach is less accurate since quantification of a peptide is based on a single fragmentation event only. Furthermore, iTRAQ is very expensive. In this thesis we thus applied label-free quantification in a number of experiments as a cheap alternative that enables comparison of multiple samples. Furthermore, this method enables comparison of samples derived from tissues or cell lines that are difficult to label metabolically.

Label-free quantification

The label-free quantification (LFQ) method, as the name implies, does not make use of stable isotopes. Peptide intensities obtained in different mass spectrometry runs cannot be compared directly due to possible differences in sample concentration or retention time, for example. This analysis method, which is present in the MaxQuant analysis software, normalizes the intensity of all peptides in a fraction by assuming that the abundance of the majority of peptides will be the same in the different samples. After normalization, pairwise ratios between two samples are determined for each protein, based on the ratios of peptides that are present in both samples. Finally, the abundance profile of a protein, which is based on the determined pairwise protein ratios, is scaled to preserve the total summed intensity of a protein over all samples. The resulting values are the LFQ intensities [90]. The intensity of each peptide in each

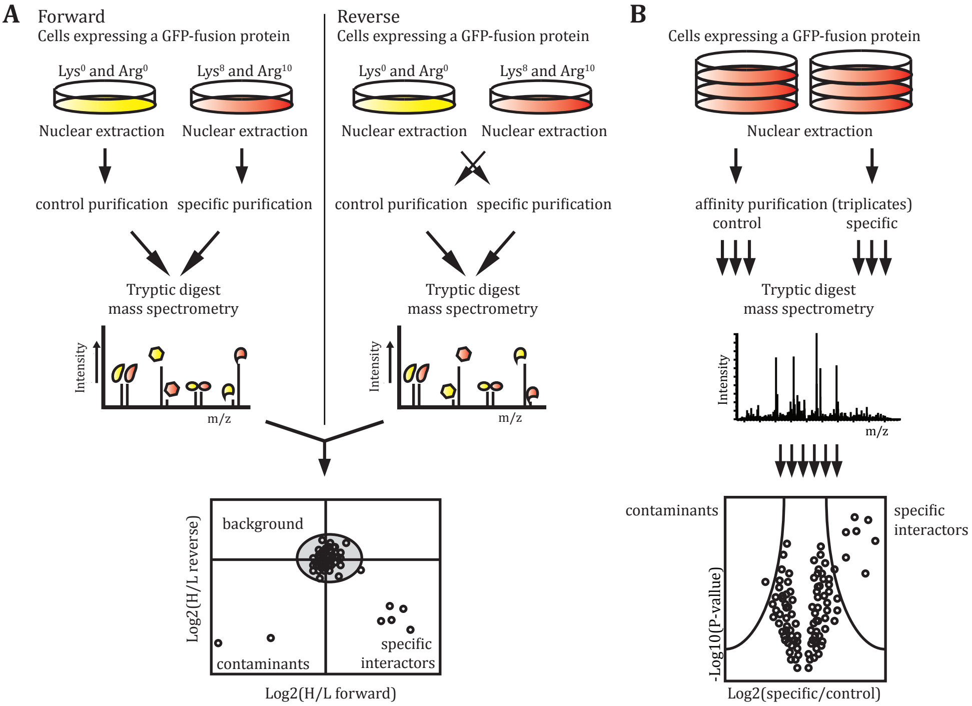


Figure 5: Quantitative mass spectrometry. The workflow for SILAC-based (**A**) or LFQ-based (**B**) GFP-affinity purifications, as applied in this thesis, is depicted.

sample is thus corrected for all fluctuations introduced before (i.e. by differential fractionation) or at the measurement step. Three or more replicates for each condition are performed to enable statistical analysis by *t*-test or ANOVA. Figure 5B shows a schematic representation of the LFQ workflow for GFP affinity purifications.

One of the advantages of this method is that multiple conditions can be compared, although the number of samples equals three times the number of conditions and this results in an increase in mass spectrometry time. Furthermore, this method is applicable to all cell lines or tissues because they can be grown in their regular medium. Finally, the extra measurement time in combination with lower complexity may lead to identification of more proteins. Not only is the measuring depth per sample higher than for stable isotope labeling approaches, the chance that a peptide ion is selected for fragmentation also increases with the number of mass spectrometry measurements. The MaxQuant analysis software contains an option to compare the different mass spectrometry measurements to each other and it then checks whether the precursor mass that was selected for fragmentation in one of the samples was also present in other samples around the same retention time. Using this method, which is called 'match between runs', more peptides can be used for pairwise ratio determination. As a result, replicates having too low abundance of a protein to select its peptides for fragmentation may receive an LFQ value based on the identification of the same peptides in another sample [90].

In conclusion, LFQ-based quantification can be applied for comparison of many samples and likely results in identification of more proteins compared to stable isotope labeling methods, such as SILAC. However, each quantification approach has its advantages, especially for the characterization of protein-protein interactions. The combination of these different mass spectrometry techniques thus enables a detailed study of protein-DNA and protein-protein interactions, as is described in this thesis. In addition, several other commonly used biochemistry and cell biology techniques were used to complement the mass spectrometry experiments and to study the proteins *in vivo*.

ACKNOWLEDGEMENTS

We thank Susan L. Kloet and H. Irem Baymaz for helpful suggestions and discussions, and Nina C. Hubner for critical reading of the manuscript.

OVERVIEW OF THIS THESIS

Chapter 1 contains a general introduction of epigenetics and mass spectrometry. It explains the basics of histone marks and DNA methylation, and describes the NuRD complex, which is in the focus of this thesis. Furthermore, the basic mass spectrometry techniques used in the thesis are described.

In **chapter 2** we describe step-by-step how to determine readers (and their interactors) for a DNA modification. This chapter describes how to perform SILAC labeling of cells, how to prepare nuclear extract and how to do DNA pull-downs. It includes all tips and tricks to perform the experiment in an optimal way.

The set up as described in chapter 2 was used for the screenings in **chapter 3**, where we identified novel readers for mC and its oxidized derivatives from mouse embryonic stem cells using SILAC labeling. In addition we applied label-free quantification to identify the readers of mC and hmC that are expressed in neuronal progenitor cells and adult mouse brain. We observed that the readers are distinct for many of the DNA modifications and that they are dynamic through development. In this chapter we describe some of the novel proteins and we determine the absolute protein abundance in our nuclear extracts using the iBAQ algorithm. The protein abundance could explain about 30% of the dynamic binding that we observed.

In **chapter 4**, the LFQ set up was used to identify readers for H3K4me3 and H3K9me3 in mouse liver, brain and kidney. Readers for histone modifications overlap much more between tissues than readers for DNA modifications. However, some tissue-specific readers were observed in brain and testis. Furthermore, the binding pattern of a number of zinc finger proteins showed high similarity to known NuRD subunits. Affinity purifications of these proteins confirmed that they are interactors of the NuRD complex.

The next chapter describes the novel NuRD subunit CDK2AP1. This very small protein was previously observed in multiple NuRD affinity purifications and **chapter 5** confirms that this protein is a *bona fide* subunit of the complex. The protein co-localizes with MBD2 in immunofluorescence microscopy experiments and is able to recruit the transcriptional repression activity of NuRD in transactivation assays.

In **chapter 6** we characterize the interaction between the NuRD complex and its transient interactor ZMYND8, one of the proteins identified in chapter 4. Using both LFQ- and SILAC-based interaction proteomics, we show that the MYND domain is required and sufficient for the interaction of ZMYND8 with NuRD. Furthermore, ChIP-sequencing experiments show that ZMYND8 and MBD3 occupy partly the same loci. The fact that ZMYND8 closely interacts with a number of zinc finger-containing proteins suggests that ZMYND8 may be a recruiter for the NuRD complex at non-methylated target sites.

Chapter 7 gives a perspective on the role of DNA methylation and the novel readers that we identified in chapter 3. Since many of these proteins are sequence-specific transcription factors and DNA methylation is also found in low CpG dense enhancers of actively transcribed genes, the general repressive function of DNA methylation can be questioned.

Finally, in **chapter 8**, I will give an overall conclusion of the work described in this thesis. I will discuss the remaining open questions and suggest possible experiments to answer these questions.

REFERENCES

1. Li, B., M. Carey, and J.L. Workman, *The role of chromatin during transcription.* Cell, 2007. **128**(4): p. 707-19.
2. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.
3. Vermeulen, M., et al., *Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4.* Cell, 2007. **131**(1): p. 58-69.
4. Vermeulen, M., et al., *Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers.* Cell, 2010. **142**(6): p. 967-80.
5. Bartke, T., et al., *Nucleosome-interacting proteins regulated by DNA and histone methylation.* Cell, 2010. **143**(3): p. 470-84.
6. van Nuland, R., et al., *Multivalent engagement of TFIID to nucleosomes.* PLoS One, 2013. **8**(9): p. e73495.
7. Yang, Y., et al., *TDRD3 is an effector molecule for arginine-methylated histone marks.* Mol Cell, 2010. **40**(6): p. 1016-23.
8. Matthews, A.G., et al., *RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination.* Nature, 2007. **450**(7172): p. 1106-10.
9. Bird, A.P., *CpG-rich islands and the function of DNA methylation.* Nature, 1986. **321**(6067): p. 209-13.
10. Jurkowska, R.Z., T.P. Jurkowski, and A. Jeltsch, *Structure and function of mammalian DNA methyltransferases.* Chembiochem, 2011. **12**(2): p. 206-22.
11. Kriaucionis, S. and N. Heintz, *The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain.* Science, 2009. **324**(5929): p. 929-30.
12. Tahiliani, M., et al., *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, 2009. **324**(5929): p. 930-5.
13. Ito, S., et al., *Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine.* Science, 2011. **333**(6047): p. 1300-3.
14. Munzel, M., et al., *Quantification of the sixth DNA base hydroxymethylcytosine in the brain.* Angew Chem Int Ed Engl, 2010. **49**(31): p. 5375-7.
15. Pfaffeneder, T., et al., *Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA.* Nat Chem Biol, 2014. **10**(7): p. 574-81.
16. Maiti, A., et al., *TDG excision of fC may be a predominant element of pathways for active DNA demethylation.* Faseb Journal, 2013. **27**.
17. Maiti, A. and A.C. Drohat, *Thymine DNA glycosylase can rapidly excise 5-formylcytosine and 5-carboxylcytosine: potential implications for active demethylation of CpG sites.* J Biol Chem, 2011. **286**(41): p. 35334-8.
18. Mellen, M., et al., *MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system.* Cell, 2012. **151**(7): p. 1417-30.
19. Yildirim, O., et al., *Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells.* Cell, 2011. **147**(7): p. 1498-510.
20. Kaslow, D.C. and B.R. Migeon, *DNA methylation stabilizes X chromosome inactivation in eutherians but not in marsupials: evidence for multistep maintenance of mammalian X dosage compensation.* Proc Natl Acad Sci U S A, 1987. **84**(17): p. 6210-4.
21. Shiota, K., *DNA methylation profiles of CpG islands for cellular differentiation and*

development in mammals. Cytogenet Genome Res, 2004. **105**(2-4): p. 325-34.

22. Jones, P.L., et al., *Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription.* Nat Genet, 1998. **19**(2): p. 187-91.
23. Thomson, J.P., et al., *CpG islands influence chromatin structure via the CpG-binding protein Cfp1.* Nature, 2010. **464**(7291): p. 1082-6.
24. Prendergast, G.C. and E.B. Ziff, *Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region.* Science, 1991. **251**(4990): p. 186-9.
25. Hendrich, B. and A. Bird, *Identification and characterization of a family of mammalian methyl-CpG binding proteins.* Mol Cell Biol, 1998. **18**(11): p. 6538-47.
26. Filion, G.J., et al., *A family of human zinc finger proteins that bind methylated DNA and repress transcription.* Mol Cell Biol, 2006. **26**(1): p. 169-81.
27. Bogdanovic, O. and G.J. Veenstra, *DNA methylation and methyl-CpG binding proteins: developmental requirements and function.* Chromosoma, 2009. **118**(5): p. 549-65.
28. Defossez, P.A. and I. Stancheva, *Biological functions of methyl-CpG-binding proteins.* Prog Mol Biol Transl Sci, 2011. **101**: p. 377-98.
29. Ng, H.H., et al., *MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex.* Nat Genet, 1999. **23**(1): p. 58-61.
30. Smits, A.H., et al., *Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics.* Nucleic Acids Res, 2013. **41**(1): p. e28.
31. Mansfield, R.E., et al., *Plant homeodomain (PHD) fingers of CHD4 are histone H3-binding modules with preference for unmodified H3K4 and methylated H3K9.* J Biol Chem, 2011. **286**(13): p. 11779-91.
32. Musselman, C.A., et al., *Binding of the CHD4 PHD2 finger to histone H3 is modulated by covalent modifications.* Biochem J, 2009. **423**(2): p. 179-87.
33. Potts, R.C., et al., *CHD5, a brain-specific paralog of Mi2 chromatin remodeling enzymes, regulates expression of neuronal genes.* PLoS One, 2011. **6**(9): p. e24515.
34. Alqarni, S.S., et al., *Insight into the architecture of the NuRD complex: structure of the RbAp48-MTA1 subcomplex.* J Biol Chem, 2014. **289**(32): p. 21844-55.
35. Kloet, S.L., et al., *Towards elucidating the stability, dynamics and architecture of the nucleosome remodeling and deacetylase complex by using quantitative interaction proteomics.* FEBS J, 2014.
36. Wu, M., et al., *The MTA family proteins as novel histone H3 binding proteins.* Cell Biosci, 2013. **3**(1): p. 1.
37. Bowen, N.J., et al., *Mi-2/NuRD: multiple complexes for many purposes.* Biochim Biophys Acta, 2004. **1677**(1-3): p. 52-7.
38. Fujita, N., et al., *MTA3, a Mi-2/NuRD complex subunit, regulates an invasive growth pathway in breast cancer.* Cell, 2003. **113**(2): p. 207-19.
39. Le Guezennec, X., et al., *MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties.* Mol Cell Biol, 2006. **26**(3): p. 843-51.
40. Hendrich, B. and S. Tweedie, *The methyl-CpG binding domain and the evolving role of DNA methylation in animals.* Trends Genet, 2003. **19**(5): p. 269-77.
41. Bogdanovic, O. and G.J. Veenstra, *Affinity-based enrichment strategies to assay methyl-CpG binding activity and DNA methylation in early Xenopus embryos.* BMC Res Notes, 2011. **4**: p. 300.

42. Roder, K., et al., *Transcriptional repression by Drosophila methyl-CpG-binding proteins.* Mol Cell Biol, 2000. **20**(19): p. 7401-9.
43. Baubec, T., et al., *Methylation-dependent and -independent genomic targeting principles of the MBD protein family.* Cell, 2013. **153**(2): p. 480-92.
44. Gunther, K., et al., *Differential roles for MBD2 and MBD3 at methylated CpG islands, active promoters and binding to exon sequences.* Nucleic Acids Res, 2013. **41**(5): p. 3010-21.
45. Menafra, R., et al., *Genome-wide binding of MBD2 reveals strong preference for highly methylated loci.* PLoS One, 2014. **9**(6): p. e99603.
46. Walavalkar, N.M., N. Gordon, and D.C. Williams, Jr., *Unique features of the anti-parallel, heterodimeric coiled-coil interaction between methyl-cytosine binding domain 2 (MBD2) homologues and GATA zinc finger domain containing 2A (GATAD2A/p66alpha).* J Biol Chem, 2013. **288**(5): p. 3419-27.
47. Gnanapragasam, M.N., et al., *p66Alpha-MBD2 coiled-coil interaction and recruitment of Mi-2 are critical for globin gene silencing by the MBD2-NuRD complex.* Proc Natl Acad Sci U S A, 2011. **108**(18): p. 7487-92.
48. Brackertz, M., et al., *p66alpha and p66beta of the Mi-2/NuRD complex mediate MBD2 and histone interaction.* Nucleic Acids Res, 2006. **34**(2): p. 397-406.
49. Reddy, B.A., et al., *Drosophila transcription factor Tramtrack69 binds MEP1 to recruit the chromatin remodeler NuRD.* Mol Cell Biol, 2010. **30**(21): p. 5234-44.
50. Miccio, A. and G.A. Blobel, *Role of the GATA-1/FOG-1/NuRD pathway in the expression of human beta-like globin genes.* Mol Cell Biol, 2010. **30**(14): p. 3460-70.
51. Hong, W., et al., *FOG-1 recruits the NuRD repressor complex to mediate transcriptional repression by GATA-1.* EMBO J, 2005. **24**(13): p. 2367-78.
52. Wang, Y., et al., *LSD1 is a subunit of the NuRD complex and targets the metastasis programs in breast cancer.* Cell, 2009. **138**(4): p. 660-72.
53. Lejon, S., et al., *Insights into association of the NuRD complex with FOG-1 from the crystal structure of an RbAp48.FOG-1 complex.* J Biol Chem, 2011. **286**(2): p. 1196-203.
54. Lu, Y., et al., *Alternative splicing of MBD2 supports self-renewal in human pluripotent stem cells.* Cell Stem Cell, 2014. **15**(1): p. 92-101.
55. Kaji, K., et al., *The NuRD component Mbd3 is required for pluripotency of embryonic stem cells.* Nat Cell Biol, 2006. **8**(3): p. 285-92.
56. Luo, M., et al., *NuRD blocks reprogramming of mouse somatic cells into pluripotent stem cells.* Stem Cells, 2013. **31**(7): p. 1278-86.
57. Rais, Y., et al., *Deterministic direct reprogramming of somatic cells to pluripotency.* Nature, 2013. **502**(7469): p. 65-70.
58. Hendrich, B., et al., *Closely related proteins MBD2 and MBD3 play distinctive but interacting roles in mouse development.* Genes Dev, 2001. **15**(6): p. 710-23.
59. Jaber-Hijazi, F., et al., *Planarian MBD2/3 is required for adult stem cell pluripotency independently of DNA methylation.* Dev Biol, 2013. **384**(1): p. 141-53.
60. Marhold, J., et al., *The Drosophila MBD2/3 protein mediates interactions between the MI-2 chromatin complex and CpT/A-methylated DNA.* Development, 2004. **131**(24): p. 6033-9.
61. Kon, C., et al., *Developmental roles of the Mi-2/NURD-associated protein p66 in Drosophila.* Genetics, 2005. **169**(4): p. 2087-100.

62. Willemsen, M.H., et al., *GATAD2B loss-of-function mutations cause a recognisable syndrome with intellectual disability and are associated with learning deficits and synaptic undergrowth in Drosophila.* J Med Genet, 2013. **50**(8): p. 507-14.

63. Marino, S. and R. Nusse, *Mutants in the mouse NuRD/Mi2 component P66alpha are embryonic lethal.* PLoS One, 2007. **2**(6): p. e519.

64. Pfefferli, C., et al., *Specific NuRD components are required for fin regeneration in zebrafish.* BMC Biol, 2014. **12**: p. 30.

65. Smeenk, G., et al., *The NuRD chromatin-remodeling complex regulates signaling and repair of DNA damage.* J Cell Biol, 2010. **190**(5): p. 741-9.

66. Chou, D.M., et al., *A chromatin localization screen reveals poly (ADP ribose)-regulated recruitment of the repressive polycomb and NuRD complexes to sites of DNA damage.* Proc Natl Acad Sci U S A, 2010. **107**(43): p. 18475-80.

67. Polo, S.E., et al., *Regulation of DNA-damage responses and cell-cycle progression by the chromatin remodelling factor CHD4.* EMBO J, 2010. **29**(18): p. 3130-9.

68. O'Shaughnessy, A. and B. Hendrich, *CHD4 in the DNA-damage response and cell cycle progression: not so NuRDy now.* Biochem Soc Trans, 2013. **41**(3): p. 777-82.

69. Pegoraro, G., et al., *Ageing-related chromatin defects through loss of the NURD complex.* Nat Cell Biol, 2009. **11**(10): p. 1261-7.

70. Lai, A.Y. and P.A. Wade, *Cancer biology and NuRD: a multifaceted chromatin remodelling complex.* Nat Rev Cancer, 2011. **11**(8): p. 588-96.

71. Li, D.Q., et al., *Metastasis-associated protein 1/nucleosome remodeling and histone deacetylase complex in cancer.* Cancer Res, 2012. **72**(2): p. 387-94.

72. Toh, Y., et al., *Overexpression of metastasis-associated MTA1 mRNA in invasive oesophageal carcinomas.* Br J Cancer, 1999. **79**(11-12): p. 1723-6.

73. Toh, Y., et al., *Overexpression of the MTA1 gene in gastrointestinal carcinomas: correlation with invasion and metastasis.* Int J Cancer, 1997. **74**(4): p. 459-63.

74. Covington, K.R., et al., *Metastasis tumor-associated protein 2 enhances metastatic behavior and is associated with poor outcomes in estrogen receptor-negative breast cancer.* Breast Cancer Res Treat, 2013.

75. Liu, S.L., et al., *Expression of metastasis-associated protein 2 (MTA2) might predict proliferation in non-small cell lung cancer.* Target Oncol, 2012. **7**(2): p. 135-43.

76. Fujita, N., et al., *Hormonal regulation of metastasis-associated protein 3 transcription in breast cancer cells.* Mol Endocrinol, 2004. **18**(12): p. 2937-49.

77. Pacifico, F., et al., *RbAp48 is a target of nuclear factor-kappaB activity in thyroid cancer.* J Clin Endocrinol Metab, 2007. **92**(4): p. 1458-66.

78. Hu, S.Y., et al., *High expression of RbAp46 gene in patients with acute leukemia or chronic myelogenous leukemia in blast crisis.* Chin Med J (Engl), 2005. **118**(15): p. 1295-8.

79. Hiyoshi, Y., et al., *p12CDK2-AP1 is associated with tumor progression and a poor prognosis in esophageal squamous cell carcinoma.* Oncol Rep, 2009. **22**(1): p. 35-9.

80. Sun, M., et al., *Cyclin-dependent kinase 2-associated protein 1 suppresses growth and tumorigenesis of lung cancer.* Int J Oncol, 2013. **42**(4): p. 1376-82.

81. Zolochevska, O. and M.L. Figueiredo, *Cell cycle regulator cdk2ap1 inhibits prostate cancer cell growth and modifies androgen-responsive pathway function.* Prostate, 2009. **69**(14): p. 1586-97.

82. Kolla, V., et al., *Role of CHD5 in human cancers: 10 years later.* Cancer Res, 2014.

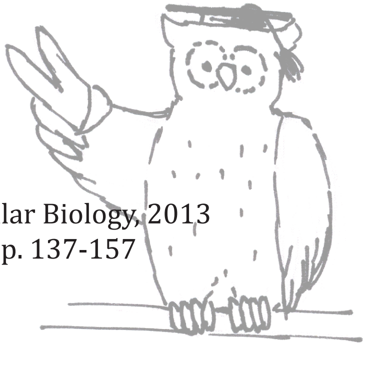
74(3): p. 652-8.

83. Sansom, O.J., et al., *Deficiency of Mbd2 suppresses intestinal tumorigenesis.* Nat Genet, 2003. **34**(2): p. 145-7.
84. Hebert, A.S., et al., *The one hour yeast proteome.* Mol Cell Proteomics, 2014. **13**(1): p. 339-47.
85. Rappsilber, J., Y. Ishihama, and M. Mann, *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics.* Anal Chem, 2003. **75**(3): p. 663-70.
86. Loo, J.A., H.R. Udseth, and R.D. Smith, *Peptide and protein analysis by electrospray ionization-mass spectrometry and capillary electrophoresis-mass spectrometry.* Anal Biochem, 1989. **179**(2): p. 404-12.
87. Biemann, K., *Sequencing of peptides by tandem mass spectrometry and high-energy collision-induced dissociation.* Methods Enzymol, 1990. **193**: p. 455-79.
88. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
89. Elias, J.E. and S.P. Gygi, *Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry.* Nat Methods, 2007. **4**(3): p. 207-14.
90. Cox, J., et al., *Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ.* Mol Cell Proteomics, 2014. **13**(9): p. 2513-26.
91. Neilson, K.A., et al., *Less label, more free: approaches in label-free quantitative mass spectrometry.* Proteomics, 2011. **11**(4): p. 535-53.
92. Ross, P.L., et al., *Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents.* Mol Cell Proteomics, 2004. **3**(12): p. 1154-69.
93. Gygi, S.P., et al., *Quantitative analysis of complex protein mixtures using isotope-coded affinity tags.* Nat Biotechnol, 1999. **17**(10): p. 994-9.
94. Ong, S.E., et al., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-86.
95. Eberl, H.C., M. Mann, and M. Vermeulen, *Quantitative proteomics for epigenetics.* Chembiochem, 2011. **12**(2): p. 224-34.
96. Vermeulen, M., N.C. Hubner, and M. Mann, *High confidence determination of specific protein-protein interactions using quantitative mass spectrometry.* Curr Opin Biotechnol, 2008. **19**(4): p. 331-7.
97. Kruger, M., et al., *SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function.* Cell, 2008. **134**(2): p. 353-64.
98. Larance, M., et al., *Stable-isotope labeling with amino acids in nematodes.* Nat Methods, 2011. **8**(10): p. 849-51.
99. Geiger, T., et al., *Super-SILAC mix for quantitative proteomics of human tumor tissue.* Nat Methods, 2010. **7**(5): p. 383-5.

Chapter 2

Identifying specific protein-DNA interactions using SILAC-based quantitative proteomics

Cornelia G. Spruijt, H. Irem Baymaz & Michiel Vermeulen



Methods in Molecular Biology, 2013
Volume 977, p. 137-157

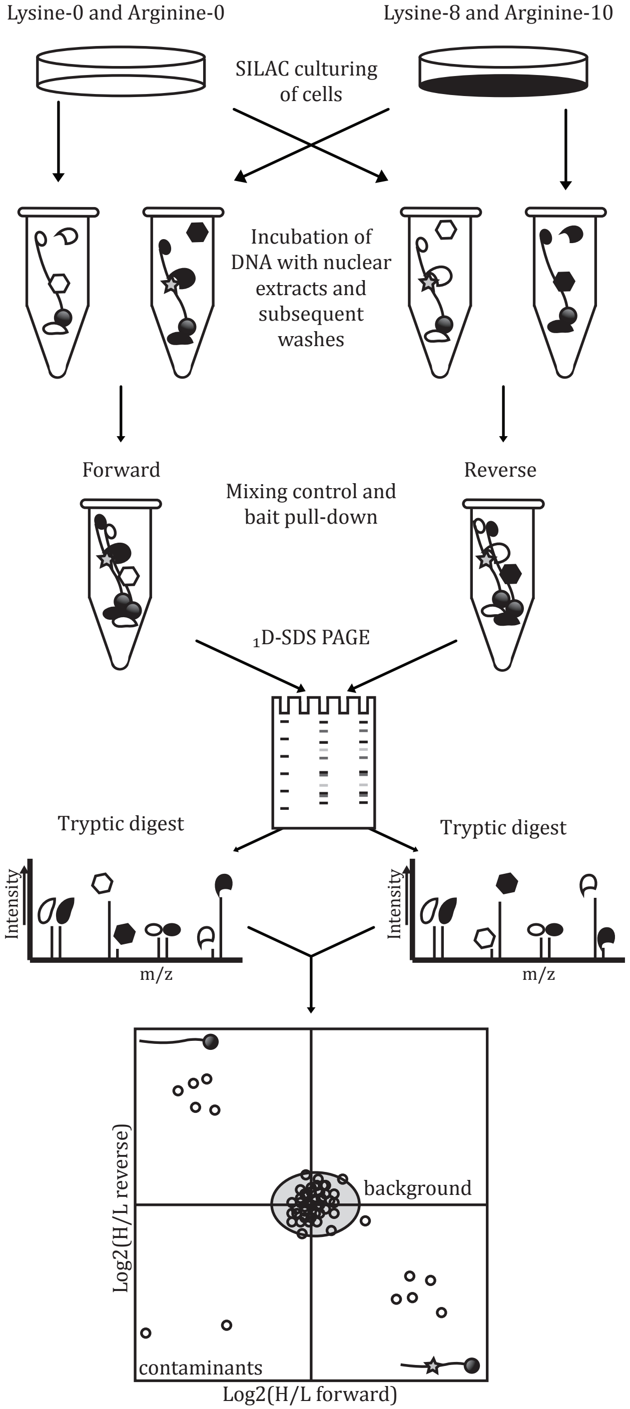
ABSTRACT

A comprehensive identification of protein-DNA interactions that drive processes such as transcription and replication, both in pro- and eukaryotic organisms, remains a major technical challenge. In this chapter, we present a SILAC-based DNA affinity purification method that can be used to identify specific interactions between proteins and functional DNA elements in an unbiased manner.

1. INTRODUCTION

The human genome consists of three billion basepairs, but only a small percentage encodes for genes. Apart from well-characterized regulatory sequences such as promoters and enhancers, the rest of the genome used to be considered 'junk DNA'. However, during the last decade it has become clear that a much larger percentage of the human genome is transcribed in the form of long and short non-coding RNAs. In addition, intergenic DNA sequences contain far more regulatory regions than previously thought [1]. Proteins and non-coding RNAs interact with these DNA sequences in a spatio-temporal manner to regulate transcription and replication. A comprehensive characterization of DNA-protein interactions is therefore essential to increase our understanding of the aforementioned processes in the nucleus. To identify sequence specific protein-DNA interactions, researchers have traditionally made use of methods such as the electromobility shift assay (EMSA) and footprinting. These assays are used to characterize a putative interaction between a candidate protein and a DNA sequence of interest. However, an unbiased identification of interactors for a specific DNA sequence requires other methods. In this regard, mass spectrometry-based proteomics has recently emerged as a powerful tool. Modern instrumentation and software enable the identification of hundreds of proteins in a sample in a few hours [2, 3]. Similar amounts of proteins can be identified in DNA affinity purifications from crude nuclear extracts. However, the majority of these proteins are highly abundant background proteins that bind non-specifically to the beads or DNA and only a small fraction represents sequence-specific interactors. This implies a need for a quantitative filter that can be used to discriminate specific interactors from non-specific background proteins. In recent years numerous methods have been developed that add a quantitative dimension to mass spectrometry measurements. In most of these methods, proteins or peptides of two conditions are labeled with different, 'light' or 'heavy', stable isotopes on specific amino acids. The two samples are then combined prior to mass spectrometry analysis. Each peptide that is identified in the mass spectrometer will have a 'light' and a 'heavy' peak and the ratio of these two signals corresponds to the relative abundance of that peptide (and the corresponding protein) in the two functional states. When applying this technology to protein-DNA interaction studies, by incubating two different DNA sequences with 'light' and 'heavy' nuclear extracts, the measured peptide ratio indicates the relative affinity of a protein for each of the two DNA probes (Figure 1).

Recently, we and others have established a DNA affinity purification protocol that makes use of an *in vivo* stable isotope labeling approach called SILAC (Stable Isotope Labeling by Amino acids in Cell culture) [4]. This generic method can be used to identify proteins binding to DNA sequences of interest, including transcription factor binding sites [5], single nucleotide polymorphisms [6] and methylated CpG islands [5, 7-9] (**see note 1**). In this chapter we describe the workflow behind this method in detail.


Lysine-0 and Arginine-0
Lysine-8 and Arginine-10
SILAC culturing of cells
Incubation of DNA with nuclear extracts and subsequent washes
Forward
Reverse
Mixing control and bait pull-down
1D-SDS PAGE
Tryptic digest
Tryptic digest
Intensity
m/z
Intensity
m/z
background
contaminants
Log2(H/L reverse)
Log2(H/L forward)

2. MATERIALS

All buffers are prepared with ultrapure water of 18.2 MΩ cm resistance (MilliQ, Millipore). To prevent the accumulation of polymers in the samples, avoid the use of autoclaved pipette tips during the experiment. Furthermore, solvents and buffers are best kept in high quality glass bottles (Schott).

Table top centrifuges with cooling capacity for Eppendorfs and 50 ml tubes are required throughout the protocol.

2.1 SILAC culture (see note 2, 3 and 4)

1. SILAC Dulbecco's Modified Eagle Medium without arginine, lysine and glutamine (PAA, E15-086).
2. Dialyzed serum (Gibco, 26400-044).
3. Glutamine (Lonza, BE17-605E).
4. Penicillin/Streptomycin (Lonza, DE 17-602E).
5. L-Lysine ('light' or 'K0' (Sigma, L8662)), dissolved in MilliQ.
6. L-Lysine 4,4,5,5-D_4-L-lysine ('medium' or 'K4' (Sigma, 616192 or Silantes, 211103912)), dissolved in MilliQ. Only in case of triple labeling **(see note 2).**
7. L-Lysine ($^{13}C_6{}^{15}N_2$) ('heavy' or 'K8' (Sigma, 608041 or Silantes, 211603902)), dissolved in MilliQ.
8. L-Arginine ('light' or 'R0'(Sigma, A6969)), dissolved in MilliQ.
9. L-Arginine $^{13}C_6$-monohydrochloride ('medium' or 'R6' (Sigma, 643440 or Silantes, 201203902)), dissolved in MilliQ. Only in case of triple labeling **(see note 2)**.
10. L-Arginine $^{13}C_6{}^{15}N_4$-monohydrochloride ('heavy'or 'R10' (Sigma, 608033 or Silantes, 201603902)), dissolved in MilliQ.
11. Dulbecco's Phosphate buffered saline (DPBS) (Lonza, BE17-512F).
12. Trypsin-EDTA (Lonza, BE17-161E) or, depending on the cell line, Accutase (Sigma-Aldrich, A6964-100).
13. 90 ml mouse embryonic stem cell serum substitute (Thermo scientific, 88213) (only for specific cell types, **see note 3**).
14. 100x Non essential amino acids (which contains proline, but no lysine or arginine) (Lonza, BE13-114E) (only for specific cell types, **see note 3**).
15. 100 mM sodium pyruvate (Lonza, BE13-115E) (only for specific cell types, **see note 3**).
16. Leukemia inhibitory factor (LIF), β-mercaptoethanol and '2i' inhibitors (CHIR99021 and PD0325901) (**see note 3**).
17. RPMI without arginine, lysine and glutamine (PAA, E15-087) (only needed for cells growing in suspension, **see note 3**).
18. 50 ml syringes (BD plastikpak, 300865).
19. 0.22 µm filters (Corning, 431219).

Figure 1: Schematic representation of the workflow described in this chapter. Bait and control DNA are incubated (separately) with light and heavy nuclear extracts (NE) from cells grown in light or heavy SILAC media. Bait DNA incubated with heavy NE is combined with control DNA incubated with light NE (forward experiment) and bait DNA incubated with light NE is combined with control DNA incubated with heavy NE (reverse experiment). The two experiments are fractionated using $_1$D SDS-PAGE, followed by in-gel digestion and mass spectrometry. The results can be visualized in a scatterplot. Specific interactors of the bait DNA are located in the lower right quadrant (high forward ratio, low reverse ratio) whereas proteins that are repelled by the bait DNA end up in the upper left quadrant (low forward ratio, high reverse ratio). High-abundant background proteins and non-specific DNA binders cluster together around the origin of the graph.

2.2 Nuclear extract (NE) preparation

1. Dulbecco's Phosphate buffered saline (DPBS) (Lonza BE17-512F).
2. Buffer A: 10 mM Hepes KOH pH 7.9, 1.5 mM $MgCl_{2,}$ 10 mM KCl.
3. Buffer C: 420 mM NaCl, 20 mM Hepes KOH pH 7.9, 20% glycerol (v/v), 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% Igepal CA-630 (v/v)/NP40 (Sigma-Aldrich, I8896-100ML). Add fresh before use: Complete protease inhibitors EDTA-free (Roche, 05056489001, 1 tablet for 50 mL buffer) and 0.5 mM DTT.
4. Glass douncer with type B pestle (tight), available in different sizes: 500 µl (Kimble Kontes, 885300-000), 2 ml (Kimble Kontes, 885303-0002 or 885301-0002) and 7 ml (Wheaton, 357542).

2.3 Bradford protein concentration

1. Bio-Rad Protein assay 5x solution (Biorad, 500-0006)
2. Bovine serum albumin (BSA), (1mg/ml solution in MilliQ) (Sigma-Aldrich, A9647-50G)
3. UV/Vis spectrophotometer
4. Cuvettes (1 ml).

2.4 DNA preparation

1. Oligonucleotides (HPLC-purified from any company).
2. TE: 10 mM Tris pH 8.0, 1 mM EDTA.
3. 2x Annealing buffer: 20 mM Tris pH 8.0, 100 mM NaCl, 2 mM EDTA.
4. T4 Polynucleotide kinase, (T4 PNK (10.000 U/ml)) (New England Biolabs (NEB), M0201S).
5. T4 DNA ligase (400 U/µl) (NEB, M0202S).
6. 100 mM ATP in MilliQ, pH 7.5 adjusted using NaOH.
7. Phenol/Chloroform (Sigma, P4557).
8. Ice-cold 100% ethanol.
9. Ice-cold 70% ethanol (v/v).
10. 3 M sodium acetate, pH 5.2.
11. Klenow fragment 5'exo- (NEB, M0212S), NEB buffer 2.
12. Biotin-14-dATP (Invitrogen, 19524-016) (make aliquots and store them at –20^{o}C).
13. Sephadex G-50, 50% slurry in 20% ethanol (v/v) (VWR, 17-0043-01)
14. 1 ml syringes (BD plastikpak, 300013) without needle

2.5 DNA affinity purification

1. Magnetic microtube rack.
2. Dynabeads MyOne C1 (Invitrogen, 650.01) **(see note 5).**
3. DNA binding buffer: 1 M NaCl, 10 mM Tris pH 8.0, 1 mM EDTA pH 8.0, 0.05% Igepal CA-630 (NP40, Sigma-Aldrich, I8896-100ML).
4. Poly-dIdC (Sigma-Aldrich, P4929-10UN) or poly-dAdT (Sigma-Aldrich, P0883-10UN) **(see note 6).**
5. Protein binding buffer: 150 mM NaCl, 50 mM Tris pH 8.0, 1 mM DTT, 0.25% Igepal CA-630 (NP40, Sigma-Aldrich, I8896-100ML) and complete protease inhibitors EDTA-free (Roche, 05056489001, 1 tablet for 50 ml).

2.6 In gel digestion

1. Gel running system (Invitrogen).
2. NuPage sample buffer (Invitrogen, NP0007).
3. MOPs buffer (Invitrogen, NP0001).
4. NuPAGE Novex 4-12% gradient gels (Invitrogen, NP0321BOX).
5. Colloidal blue stain kit (Invitrogen, LC6025).
6. Methanol (Merck, 1.06009.2500).
7. Acetic acid (Merck, 1.00063.2500).
8. ABC: 50 mM Ammonium bicarbonate (Fluka, 09830).
9. Destain solution : 25 mM ABC/50% ethanol (v/v).
10. Acetonitrile (Biosolve, 01200702).
11. Fixing solution: 50% methanol (v/v), 10% acetic acid (v/v) in MilliQ.
12. Staining solution: 55 ml MilliQ, 20 ml methanol, 20 ml Colloidal Blue Solution A.
13. 1 M 1,4-Dithiothreitol.
14. 0.55 M Iodoacetamide (Sigma, I1149).
15. Sequencing grade modified Trypsin (Promega, V5111).
16. 10% Trifluoric acid (TFA) (v/v) (Sigma, 302031).
17. Vacuum centrifuge.
18. Thermoshaker.

2.7 Peptide desalting and purification (Stage tipping)

1. C18 disks (Empore, 22125-C18).
2. 200 µl pipette tips (Rainin).
3. Hollow needle with a 1.2 mm diameter (BD Microlance 3, 304622). Make the end blunt and use a piece of nano tubing as a plunger.
4. Methanol (Merck, 1.06009.2500).
5. Buffer A: 0.5% acetic acid (v/v) (Merck, 1.00063.2500) in ultrapure water (Biosolve, 232141B1).
6. Buffer B: 0.5% acetic acid (v/v) (Merck, 1.00063.2500), 80% acetonitrile (v/v) (Biosolve, 01200702) in ultrapure water (Biosolve, 232141B1).

2.8 Mass spectrometry

1. Buffer A: 5% acetic acid (v/v) (Merck, 1.00063.2500) in ultrapure water (Biosolve, 232141B1).
2. Buffer B: 5% acetic acid (v/v) (Merck, 1.00063.2500), 80% acetonitrile (v/v) (Biosolve, 01200702) in ultrapure water (Biosolve, 232141B1).
3. 96 well thermofast robotic PCR plate (Thermo, 96 AB-1300) .
4. Nanoflow HPLC system.
5. Column oven from Sonation (PRSO-V1).
6. Fused silica based emitters (30 cm length, 360 mm OD, 75 um ID) (New Objective, FS360-75-8-N-5-C30) packed in-house with Reprosil-Pur 120 C18-AQ, 3 µm (Dr. Maisch GMBH, Germany).
7. High performance mass spectrometer such as an LTQ-Orbitrap-Velos or Q- Exactive instrument from Thermo fisher.

3. METHODS

3.1 SILAC labeling

Cells are SILAC-labeled by culturing them for at least 8 cell doublings in medium containing 'light' or 'heavy' amino acids. Note that the proliferation rate for some cell types is decreased in SILAC medium compared to normal medium due to the use of dialyzed serum during cell culture. Dialysis is necessary to get rid of non-labeled amino acids in the serum, but this also removes growth factors and other small molecules which may be important for proliferation. For some cell types, such as mouse ES cells, a SILAC compatible serum substitute is available (see **note 3**).

1. Prepare a bottle of 'light' and a bottle of 'heavy' SILAC medium **(see note 3&4).** For each:
   a. Take a bottle of 500 ml SILAC Dulbecco's Modified Eagle Medium without arginine, lysine and glutamine.
   b. Transfer 30-40 ml of medium from the bottle into a 50 ml falcon tube and add the appropriate amounts of arginine (light or heavy) and lysine (light or heavy) to this aliquot of DMEM. Add 29.4 µg/ml of arginine and 73 µg/ml of lysine. Filter this medium containing the amino acids using a syringe and a 0.22 micron filter back into the bottle.
   c. Add 50 ml dialyzed serum.
   d. Add 2 mM glutamine.
   e. Add 100 units/ml Penicillin/Streptomycin. Medium can be kept at 4^{o}C for up to six weeks.
2. Trypsinize a 10 cm dish of cells grown to ~80-100% confluence in regular medium (not light or heavy).
3. Neutralize the trypsin with regular medium and divide the suspension equally over two tubes.
4. Spin cells for 5 minutes at 400x*g*.
5. Resuspend the cell pellet of one tube in 4 ml of light medium and the other cell pellet in 4 ml of heavy medium. Seed 1 ml of this suspension in a 10 cm dish and add 9 ml of light or heavy medium.
6. Grow the cells at 37^{o}C in 5% CO_2 until they reach 80-100% confluency. Split the cells once more in a ratio of 1:8. Make sure to spin down the cells and resuspend them in fresh light or heavy medium after trypsinization since trypsin can be a source of non-labeled amino acids. In some cases, the splitting should be done differently depending on the cell type. Mouse ES cells, for example, are to be split 1:4 only. In this case, cells need to be split more often to ensure the minimal amount of 8 cell doublings required for efficient labeling.
7. Depending on the growth rate of the cells, labeling usually takes between 1 and 2 weeks. During the labeling it is recommended to perform an incorporation check on the heavy cells to make sure that the proteins are completely labeled **(see note 7)**.
8. When incorporation is complete, cells can be expanded to the desired amount. Typically, around 2 mg of nuclear extract is obtained from five 15 cm dishes, but this may vary depending on the cell line that is used.

3.2 Nuclear extract (NE) preparation

It is critical to be as consistent as possible when preparing nuclear extracts. Small differences in sample handling, especially during the douncing, can cause proteins to be differentially extracted between different samples. This makes it more difficult to discriminate true outliers from background proteins. This nuclear extraction protocol is based on Dignam *et al.* [10].

1. Wash cells with 10 ml of PBS and trypsinize them with 2 ml of trypsin per 15 cm dish. Neutralize trypsin by adding 10 ml of SILAC medium to the cells. Collect the cells in a 50 ml tube and rinse the plates once more with PBS to collect the remaining cells. Perform all subsequent steps at 4^{o}C.
2. Centrifuge the cells for 5 minutes at 400x*g* and aspirate the supernatant.
3. Wash cells with 50 ml of PBS and centrifuge for 5 minutes at 400x*g*, aspirate the supernatant.
4. Resuspend cells in 8 ml of PBS and transfer the cells to a 15 ml tube. Rinse the 50 ml tube with 5 ml of PBS and transfer this to the 15 ml tube containing the cell suspension.
5. Centrifuge for 5 minutes at 400x*g* **(see note 8)** and aspirate the supernatant.
6. Determine the volume of the cell pellet and add 5 volumes of cold buffer A. Resuspend the cells and incubate for 10 minutes on ice.
7. Centrifuge the cells for 5 minutes at 400x*g*, remove supernatant.
8. Determine the volume of the cell pellet (cell volume should increase due to osmotic uptake of buffer A by the cells, **see note 9**) and add 2 volumes of buffer A containing complete protease inhibitors and 0.15% Igepal NP40 (v/v). Resuspend cells and transfer the suspension to a dounce homogenizer **(see note 10)**.
9. Apply 30-40 strokes up and down with a type B pestle (tight) **(see note 11)**.
10. Transfer the suspension back to a 15 ml tube and centrifuge for 15 minutes at 3200x*g*. The supernatant is the cytoplasmic extract. Collect or discard the supernatant. When keeping the supernatant, add glycerol (10% final concentration) and NaCl (150 mM final concentration).
11. Wash the pellet once with 10 volumes of PBS. Gently pipette up and down once.
12. Centrifuge for 5 minutes at 3200x*g* and discard the supernatant.
13. The pellet consists of crude nuclei. Determine the volume and add 2 volumes of buffer C.
14. Resuspend and transfer crude nuclei to an Eppendorf tube. Homogenize the pellet by pipetting up and down (10x). For some cell lines the pellet may be difficult to resuspend.
15. Incubate the suspension for one hour at 4^{o}C on a rotating wheel. Due to the lysis of the nuclei and the release of chromatin the suspension will become viscous and white clouds of chromatin should appear.
16. Centrifuge the suspension for 45 minutes at 20800x*g* in a table top centrifuge at 4^{o}C.
17. Transfer the supernatant to a new tube. This is the nuclear extract (NE) that will be used for DNA pull-downs and it contains soluble nuclear proteins. The pellet contains the insoluble chromatin fraction and consists of DNA and proteins tightly bound to chromatin.
18. Aliquot (approximately 150 µl per Eppendorf tube) and snap-freeze the extracts in liquid N_2. The nuclear pellet can be snap-frozen too. Store at -80^{o}C.

3.3 Protein concentration determination

1. Prepare a 1 mg/ml stock solution of BSA in MilliQ.
2. Dilute 2 µl of nuclear extract with 18 µl of MilliQ.
3. Transfer 4 and 10 µl of the diluted nuclear extracts to separate Eppendorf tubes.
4. For the standard curve, pipette 0, 1, 2, 5, 7 and 10 µl of the BSA solution in Eppendorf tubes.
5. Prepare a 1x Biorad protein assay solution by diluting the reagent 5 times with MilliQ.
6. Add 1 ml of 1x Biorad protein assay solution to the Eppendorf tubes containing the standard curve and the nuclear extract samples.
7. Transfer the samples to cuvettes and measure absorbance at 595 nm at the spectrophotometer.
8. Fit a linear curve through the absorbance values of the BSA standard and extract protein concentrations of the nuclear extracts by matching the absorbance of the samples to this curve.

3.4 DNA preparation

1. Design complementary pairs of oligonucleotides of about 30 bases that contain your sequence of interest **(see note 1)**. Include two thymidines on the 5'end of one oligonucleotide and two adenines on the 5'end of the reverse complementary oligonucleotide. For each bait a control pair of oligonucleotides should be designed. For example, a bait containing a methylated CpG dinucleotide should be combined with a control bait that is not methylated.
2. Dissolve the oligonucleotides to a concentration of 0.3 mM in TE buffer by shaking at room temperature (RT) for 1 hour. Store DNA at -20°C until use.
3. Combine 12.5 µl of the forward and reverse oligonucleotides and add 25 µl of 2x Annealing buffer in an Eppendorf tube.
4. Incubate the sample at 95°C for 5 minutes in a water bath or heat block.
5. Spin down the sample and put it back to 95°C.
6. Switch of the heating and let the sample cool down slowly to RT. The oligonucleotides will be annealed at this point.
7. Phosphorylate the annealed oligonucleotides by adding 10 µl of 10x ligase buffer, 5 µl of T4 Polynucleotide kinase (10.000 U/ml) and 35 µl of MilliQ. Incubate for 2 hours at 37°C **(see note 12)**.
8. Ligate the oligonucleotides by adding 10 µl of 100 mM ATP pH 7.5 and 2 µl of T4 DNA ligase (400 U/µl) **(see note 12)**.
9. Incubate for 4 hours at RT and subsequently overnight at 4°C. The ligation efficiency can be investigated by loading 2 µl on a 1.5% agarose gel. The ligation products should form a ladder, as shown in Figure 2A.

10. Perform a Phenol/Chloroform extraction:
    a. Adjust the volume of the sample to 200 µl with MilliQ.
    b. Add 200 µl of Phenol/Chloroform, vortex for 1 minute, centrifuge for 2 minutes at 18400x*g* and transfer the upper phase to a new tube.
    c. Precipitate DNA by adding 500 µl of 100% ice-cold ethanol and 10 µl of 3 M NaAc pH 5.2. Incubate for at least 30 minutes at -20^{o}C.
    d. Centrifuge for 10 minutes at 18400x*g* at 4^{o}C.
    e. Aspirate the supernatant carefully and wash the DNA pellet with 500 µl of ice-cold 70% ethanol (v/v).
    f. Centrifuge for 5 minutes at 18400x*g* and aspirate the supernatant.
    g. Air-dry the pellet.
    h. Dissolve the DNA in 37 µl of MilliQ.
11. Add 5 µl of 10x NEB buffer 2, 3 µl of Klenow exo- (50 U/µl) and 5 µl of Biotin-14-dATP (0.1 mM) to the DNA and incubate for 3 hours at RT.
12. Prepare Sephadex G-50 columns (see Figure 2B and **note 13)**. Add 100 µl of TE to the DNA strands and load them onto the column. Centrifuge at 490x*g* for 1 minute at 4^{o}C. This step is performed to separate the DNA strands from the free biotin-ATP.
13. Measure the DNA concentration of the eluent.

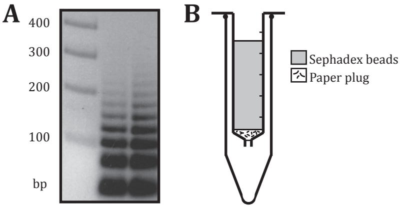


Figure 2: DNA preparation.
A. Shown are one bait and one control DNA sample with similar ligation efficiencies. **B.** A Sephadex G-50 column prepared by inserting a syringe without a plunger into a 15 ml tube. The syringe contains a paper plug at the bottom and is packed with Sephadex G-50 resin.

3.5 DNA affinity purification

The protocol below describes a so-called 'forward' and 'reverse' experiment. In the forward experiment, the control DNA is incubated with light extract, while the bait of interest is incubated with heavy extract. In the reverse experiment a label-swap is performed in which the control DNA is incubated with heavy extract and the bait DNA is incubated with light extract. This set-up constitutes a biological replicate.

1. Take four Eppendorf tubes and pipette 75 µl of Dynabeads MyOne C1 in each of them **(see note 6)**.
2. Add 0.5 ml of DNA binding buffer and place the tubes in a magnetic Eppendorf holder.
3. Aspirate the supernatant once the solution has cleared.
4. Take the tubes out of the holder, add 0.5 ml of DNA binding buffer and invert the tubes until the beads are completely resuspended.
5. Centrifuge briefly and place the tubes back into the magnetic holder. Aspirate the supernatant.
6. Take two times 10 µg of bait DNA and two times 10 µg of control DNA (obtained in section 3.4) and adjust the salt concentration to 1 M NaCl. Add the DNA in a total volume of 350 µl of DNA binding buffer to each of the tubes. Two tubes should contain bait DNA and the other two should contain control DNA.
7. Incubate for 1 hour at RT on a rotation wheel.
8. Briefly centrifuge and place the samples in the magnetic rack. Check the coupling

of the DNA to the beads by assessing the depletion of the DNA from the solution **(see note 14)**.

9. Wash the beads two times with 0.5 ml of DNA binding buffer as described in points 4 and 5.
10. Wash the beads two times with 0.5 ml of Protein binding buffer.
11. Add 400 µg of nuclear extract to the beads in a total volume of 600 µl protein binding buffer, including 10 µg of poly-dIdC or poly-dAdT **(see note 7)**. **NB:** Add 'light' NE to one tube with control DNA and to one tube containing bait DNA. Add 'heavy' NE to the other two tubes.
12. Incubate for 90 minutes at 4°C on a rotation wheel.
13. Wash the beads three times with 0.5 ml protein binding buffer. After the last wash, remove supernatant completely, also from the lid.
14. Resuspend the beads of the control DNA pull-down in 30 µl 2X NuPAGE loading buffer containing 20 mM DTT. Add the control DNA pull-down suspension to the beads containing the bait DNA, as such that the heavy bait pull-down is mixed with the light control pull-down (forward experiment) and vice versa (reverse experiment).
15. Incubate the samples for 5 minutes at 95°C.

3.6 In gel digestion

The DNA pull-down procedure described in section 3.5 results in two samples to be processed for mass spectrometry using in-gel trypsin digestion [11] (one forward and one reverse experiment). Wear gloves at all times and work as cleanly as possible. Keratin contamination of the samples can compromise the identification of proteins in the experiment.

1. Load the samples on a precast 4-12% gradient gel. Keep a blank lane between all the samples, including the lane between the molecular weight marker and the first sample.
2. Run the gel at 200 Volt.
3. Fix the gel for 10 minutes in fixing solution in a clean plastic box (50% methanol (v/v), 10% acetic acid (v/v) in MilliQ) on a shaker.
4. Incubate the gel for 5 minutes in 55 ml MilliQ, 20 ml methanol and 20 ml Colloidal Blue Solution A.
5. Add 5 ml of Colloidal Blue solution B and incubate for one hour.
6. Destain the gel in MilliQ for at least 2 hours and refresh MilliQ a couple of times. It is also possible to store the gel at 4°C in MilliQ for up to a week.
7. Clean a glass plate with MilliQ and absolute ethanol and air-dry the plate.
8. Put the gel on the glass plate and cut out one lane at a time. Divide each lane into 6 - 10 slices depending on the protein amount in the sample. Make sure to

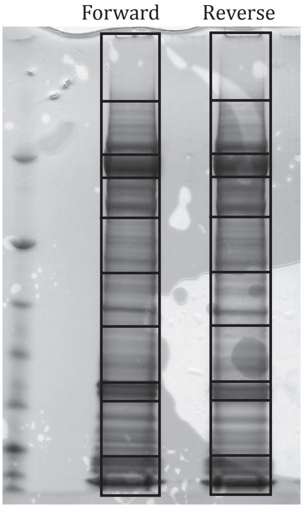


Figure 3: $_1$D SDS-PAGE fractionation of proteins obtained in a forward and a reverse DNA pull-down. The lines around the proteins indicate how these lanes should be sliced into 6 - 10 pieces (10 slices in this case), isolating abundant proteins in a single slice.

cut the lanes of the forward and the reverse experiment in a similar pattern and try to isolate very abundant proteins in a single gel slice (see Figure 3).

9. Cut each gel slice up into smaller pieces of about 1 mm^3 and transfer them to an eppendorf tube.
10. Incubate the gel pieces two times for 1 hour in 1 ml of destain solution (50% ethanol, 25 mM ABC) in a thermoshaker at RT at 1200 rpm. Aspirate all the liquid after each incubation step.
11. Dehydrate the gel pieces in 1 ml of acetonitrile for 5 - 10 minutes in a thermoshaker. The gel pieces will shrink and become white opaque.
12. Swell the gel pieces in 1 ml 50 mM ABC for 5 - 10 minutes.
13. Dehydrate the gel pieces two times with 1 ml of acetonitrile for 5 - 10 minutes in a thermoshaker. Aspirate the supernatant after each incubation.
14. Vacuum centrifuge the gel pieces (with lids of Eppendorf tubes open) for 5 - 10 minutes until all the liquid has evaporated. At this point it is possible to store the samples at 4°C up to a week.
15. Add 200 µl of reducing buffer (10 mM DTT in 50 mM ABC) and incubate for 45 minutes at 55°C without shaking.
16. Carefully remove all the liquid.
17. Add 300 µl of 55 mM iodoacetamide in 50 mM ABC and incubate for 30 minutes at RT in the dark (iodoacetamide is light-sensitive).
18. Carefully remove all the liquid.
19. Wash the gel pieces for 15 minutes in 1 ml of 50 mM ABC.
20. Wash the gel pieces twice with 1 ml of acetonitrile in a thermoshaker, remove the supernatant after each incubation.
21. Vacuum centrifuge the gel pieces.
22. Add 30 µl of sequence grade trypsin at 10 ng/µl in 50 mM ABC.
23. Incubate for 10 minutes at RT until the gel pieces have absorbed the trypsin solution.
24. Add 50 mM ABC until the gel pieces are completely covered and incubate overnight at 37°C.
25. Add 100 µl of 30% acetonitrile (v/v) and 3% TFA (v/v) in MilliQ to the gel pieces and shake for 10-15 minutes in a thermoshaker.
26. Transfer all the liquid to new tubes and repeat steps 25 and 26.
27. Add 100 µl of 100% acetonitrile to the gel pieces and shake for 10 - 15 minutes.
28. Centrifuge briefly and transfer the liquid to the collection tubes. Repeat steps 27 and 28.
29. Vacuum centrifuge (45 - 90 minutes, depending on the number of samples) until ~100 µl of liquid is left (the acetonitrile in the sample should be completely evaporated).

3.7 Peptide desalting and purification

Following vacuum centrifugation of the tryptic peptides, it is common practice to desalt and purify the peptides using self-made or commercial C18 columns **(see note 15)** called ‘stop and go extraction’ or ‘stage tips’ (Figure 4) [12]. During this procedure residual salt and small contaminants are removed and the peptides are captured on a small 200 µl tip containing a small plug of C18 material. Peptides bound to stage tips can be stored for months at 4°C.

1. Prepare stage tips by stamping out small disks from a double layer of C18 Empore filter using a blunt ended syringe needle. Eject the C18 disks from the needle into a 200 µl pipette tip and fix the material at the narrow end of the tip. Do not apply too much force since this will hinder buffer flow through the column. Prepare one stage tip per gel slice.
2. Punch a hole into the lid of 2 ml Eppendorf tubes and place the stage tips into the holes. Activate the stage tips by applying 50 µl of methanol and centrifuge at 1500x*g* for about 2 - 5 minutes. Make sure that all the liquid has passed through the column.
3. Wash the stage tips by applying 50 µl of buffer B and centrifugation at 1500x*g*.
4. Wash the stage tips twice with 50 µl of buffer A and centrifuge at 1500x*g*.
5. Load the samples on the stage tips and centrifuge at 380x*g* until all the liquid has passed through the column. This takes about 10 - 20 minutes.
6. Wash the stage tips with 50 µl of buffer A and centrifuge at 1500x*g*.
7. Store stage tips at 4°C or proceed with elution and mass spectrometry as described below.

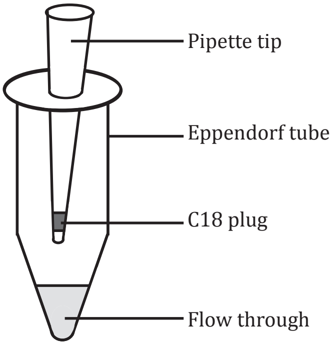


Figure 4: Stage tips. Schematic representation of a stage tip (200 µl pipette tip with C18 plug) inserted into a 2 ml Eppendorf to collect the flow through.

3.8 Mass spectrometry

Modern mass spectrometers are very sensitive, fast and are able to sequence thousands of peptides in a short period of time. However, their operation and raw data analysis require extensive expertise and training. Therefore, the protocol below only provides a rough guideline for the liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) following in gel digestion.

1. Wash the stage tips once with 30 µl of buffer A and centrifuge at 1500x*g*.
2. To elute the peptides, load 30 µl of buffer B onto the stage tips and elute into a new Eppendorf tube by centrifugation at 380x*g*.
3. Dry down the samples in a vacuum centrifuge to about 4 µl. Do not dry down the sample completely since this will result in a loss of peptides.
4. Add 4 µl of buffer A to the samples and transfer it to a 96-well plate that is compatible with the nano-HPLC.
5. Program the autosampler of the nano-HPLC to inject 4 µl onto the nano-HPLC column.
6. The peptides are eluted from the nano-HPLC column using a ~120 minute, 5-30% acetonitrile (v/v) gradient followed by a sharp increase to 60% acetonitrile (v/v) in 10 minutes. Setting up these gradients requires extensive expertise, supervision by an experienced mass spectrometrist is highly recommended.
7. When using an LTQ-Orbitrap-Velos mass spectrometer, the following basic data acquisition settings are recommended: Acquire precursor MS spectra at an m/z range of 300 - 1750 at a resolution of 60.000 and a target value of 1 million ions per full scan. MS/MS spectra can be acquired in HCD or CID mode. For protein

identification experiments we usually obtain the highest number of protein identifications in CID mode. When using CID, select the 15 most intense precurser ions of every full scan for fragmentation at a minimal ion count target value of 500. Fragment and record peptides in the dual pressure linear ion trap using a normalized collision energy of 35% and acquire these spectra in centroid mode. Enable dynamic exclusion (repeat duration 30 seconds, list size 500, exclusion duration 30 seconds, early expiration enabled (count 2, S/N threshold 2)).

3.9 Raw data processing and data analysis

We make use of the MaxQuant software to process and analyze the raw data generated by the LTQ-Orbitrap-Velos mass spectrometer [13, 14]. This software is freely available and can be downloaded at www.maxquant.org. Installation instructions and guidelines regarding the basic recommended settings during data processing can also be obtained through this website. Also available at www.maxquant.org is a suite of downstream data analysis tools embedded in the Maxquant software called Perseus. In addition, there is an online MaxQuant google group where practical questions regarding usage of the software are posted and answered.

When analyzing the raw data from the forward and reverse DNA pull-downs, it is important to specify the forward and reverse mass spectrometry runs in the ‘experimentalDesignTemplate.txt’ file that is generated by MaxQuant. In the ‘experiment’ column in the experimental design file simply name all the forward runs ‘forward’ and all the reverse runs ‘reverse’. Alternative names may also be used, but avoid numbers. Maxquant will now report separate protein ratios for the forward and the reverse pull-down. The Proteingroups.txt output table that MaxQuant generates contains all the basic information regarding identified proteins and their ratio in the forward and the reverse pull-down. This table should be filtered for contaminants and reverse hits. Furthermore, we recommend a minimal ratio count of 3 for each protein, both in the forward and the reverse pull-down. The ratios are then log2 transformed and eventually the ratios of all the proteins in the forward and reverse pull-down are plotted against each other in a two dimensional graph (see Figure 1). In this graph the x-axis and y-axis represent the H/L ratio in the forward and the reverse experiment, respectively. Background proteins will cluster together at the origin of the graph, showing roughly a one to one ratio in both experiments. Proteins that specifically bind to the bait of interest cluster in the bottom right quadrant, whereas proteins that are repelled appear in the upper left quadrant. Proteins that are significant outliers from the background population can be deduced using boxplot statistics or by making use of the ‘significance B’ value that can be calculated using the Perseus software.

4. NOTES

1. The described protocol is optimized to identify proteins binding to a transcription factor binding site, a single nucleotide polymorphism (SNP) or to an epigenetic DNA modification such as cytosine methylation. Since the method makes use of synthetic oligonucleotides, the bait length is restricted to about 60 basepairs. In principle, the method can be adapted to identify specific interactions for any given DNA sequence such as enhancers or locus control regions. However, these sequences are generally longer and a single point mutation and/or a modification may not be sufficient to abolish all interactions with these elements. Therefore,

designing the control DNA sequences is not straightforward. As a general guideline, the control sequence should be of the same length as the bait and should have roughly the same nucleotide composition. Palindromic sequences should be avoided. Note that sequences longer than 60 basepairs have to be cloned into a plasmid for amplification and digestion prior to biotinylation.

2. To study interactions with two DNA sequences (bait and control DNA), culturing of cells in light (K0R0) and heavy (K8R10) medium is required. Interactions with three different DNA sequences can also be studied in a single experiment using a so-called triple pull-down. This requires growing cells in light (K0R0), medium (K4R6) and heavy (K8R10) medium. Nuclear extracts from these cells are incubated with the three different stretches of DNA and combined prior to mass spectrometry analysis. Each peptide identified in the mass spectrometer will now appear as a triplet and the ratio between the three peptide peaks will indicate the relative affinity of a protein for each of the three DNA sequences.
3. This SILAC medium can be used for most commonly used cell lines such as HeLa, MCF7 and HEK293T cells. However, certain cell lines, like embryonic stem cells (ESC), are difficult to grow in medium containing dialyzed serum. For mouse ESCs a serum substitute is available that is SILAC compatible. A bottle of mouse ESC SILAC medium consists of the following components: 500 ml SILAC Dulbecco's Modified Eagle Medium without arginine, lysine and glutamine, 90 ml mouse embryonic stem cell serum substitute, 3.3 mM Glutamine, 100 units/ml Penicillin/ Streptomycin , 1x Non-essential amino acids (which contains proline, but no lysine or arginine), 1 mM sodium pyruvate, 73 µg/ml L-Lysine (light or heavy) and 29.4 µg/ml arginine (light or heavy), LIF (1000 U/ml), 4.2 µl β-mercaptoethanol and 2i inhibitors (CHIR99021 and PD0325901, 3 and 1 µM respectively). For cells growing in suspension, SILAC medium can be made on the basis of RPMI medium.
4. In some cell types, arginine to proline conversion may cause problems during SILAC labeling. Arginine10, when converted, results in a proline that is 6 Da heavier compared to normal proline. In addition to the 'normal' heavy peak containing a labeled arginine or lysine on the C-terminus, a third peak can now be observed for proline containing peptides. This peak contains a heavy proline and a heavy arginine or lysine. During quantification, this results in an underestimation of the peptide ratio. The extent of conversion can be minimized by titrating the amount of arginine and proline in the medium. Adding too much proline, however, may reduce arginine labeling, since conversion can take place in the other direction as well. Another solution is to make use of lysine-only labeling combined with Lys-C digestion instead of trypsin.
5. Different types of streptavidin-conjugated beads are available, for example Dynabeads M280; each of them may require optimization of the protocol (amount of DNA and nuclear extract, etc).
6. Different types of competitor DNA may influence the proteins that will be identified. When using CG-rich strands for the pull-down, it is recommended to use poly-dAdT as competitor.
7. A label check can be done by running a small amount of cell lysate on a gel and performing an in gel trypsin digestion and mass spectrometry analysis (on one gel slice) as described in section 3.6. The incorporation of the heavy amino acids can be deduced from the observed peptide ratios in the mass spectrometer (for example

a peptide ratio of 50 indicates 98% incorporation). When the incorporation of the heavy amino acids is 90%, the maximum observable ratio in a pull-down is 10, at 80% incorporation 5, etc. It is therefore recommended to strive for at least 95% incorporation of the heavy amino acids. Note that labeling never reaches 100% due to impurities in the 'heavy' amino acids and small amounts of non-labeled amino acids in the culture medium.

8. After this step it is possible to leave the cells on ice for 30-60 minutes. Centrifuge the cells again before continuing with the protocol. When continuing without a break, the first next step that can be prolonged is the incubation of the nuclei in buffer C (step 16 of section 3.2).
9. The increase of cell volume after incubation with buffer A is cell type-specific. HeLa cells increase their volume about 2 fold, while HEK293T cells hardly swell.
10. Pre-cool the dounce homogenizers on ice and rinse them with buffer A before usage. Clean and rinse the douncer with cold buffer A between different samples. Depending on the cell volume that is obtained after harvesting the cells, different douncer sizes should be used. For swollen cell pellet volumes up to 50 µl, use a 100 µl douncer. For volumes between 50 and 600 µl, use a 2 ml douncer and for volumes between 600 and 2.5 ml, use a 7 ml douncer. Douncers for even larger volumes are also available.
11. Keep the douncer on ice while douncing and do so in a slow steady rhythm. Wait for 45 seconds after every 10 strokes. Friction during the douncing results in an increase in temperature which may affect protein stability.
12. Both the T4 Polynucleotide kinase and the T4 DNA ligase exhibit 100% activity in 1x T4 DNA Ligase Buffer which contains 1 mM ATP. However, because the kinase uses most of the ATP in the buffer (step 7 in section 3.4), when adding the T4 DNA ligase (step 8 in section 3.4) make sure to add fresh ATP to the solution.
13. Commercial Sephadex G-50 columns are available: Illustra ProbeQuant G-50 micro columns (GE Healthcare, 28-9034-08) or Illustra NAP-10 columns (GE Healthcare, 17-0854-02). Use these columns according to the manufacturer's protocol. Alternatively, prepare your own: Take a 1 ml syringe without a needle, put a paper plug (tissue) in the bottom of the syringe and put it into a 15 ml tube. Fill the syringe with Sephadex G-50 slurry, centrifuge for 1 minute at 490x*g* and add more slurry. Repeat this step a couple of times until the column is filled with ~1 ml of beads. Wash the column twice with 0.5 ml of TE buffer and centrifuge for 1 minute at 490x*g*. Put the column into a new 15 ml tube, add 100 µl of TE buffer to the DNA and load it onto the column. Centrifuge for 1 minute at 490x*g* and measure the DNA concentration of the eluent (See figure 2).
14. Load 0.5 µg of input DNA and 17.5 µl of the supernatant on an agarose gel. Adjust the NaCl concentration of the input DNA to 1 M. High salt concentrations affect DNA migration and equalizing the salt concentration makes it easier to see the extent of DNA depletion from the solution.
15. Companies offering commercial stage tips include Proxeon and Millipore.

ACKNOWLEDGEMENTS

We thank members of the Vermeulen lab for critical reading of the manuscript. The Vermeulen lab is supported by grants from the Netherlands organization for scientific Research (NWO-VIDI), the Dutch Cancer Society (KWF) and the Cancer Genomics Centre (CGC-NGI).

REFERENCES

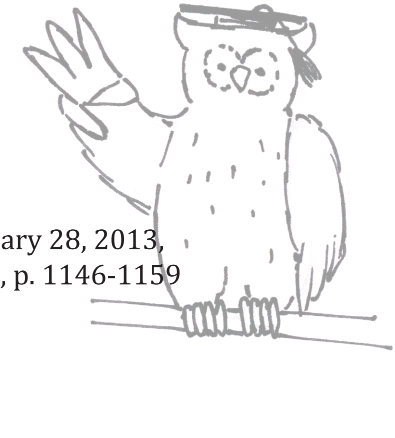
1. Alexander, R.P., et al., *Annotating non-coding regions of the genome.* Nat Rev Genet, 2010. **11**(8): p. 559-71.
2. Cox, J. and M. Mann, *Quantitative, high-resolution proteomics for data-driven systems biology.* Annu Rev Biochem, 2011. **80**: p. 273-99.
3. Beck, M., M. Claassen, and R. Aebersold, *Comprehensive proteomics.* Curr Opin Biotechnol, 2011. **22**(1): p. 3-8.
4. Ong, S.E., et al., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-86.
5. Mittler, G., F. Butter, and M. Mann, *A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements.* Genome Res, 2009. **19**(2): p. 284-93.
6. Butter, F., et al., *A domesticated transposon mediates the effects of a single-nucleotide polymorphism responsible for enhanced muscle growth.* EMBO Rep, 2010. **11**(4): p. 305-11.
7. Bartels, S.J., et al., *A SILAC-based screen for Methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein.* PLoS One, 2011. **6**(10): p. e25884.
8. Spruijt, C.G., et al., *CDK2AP1/DOC-1 is a bona fide subunit of the Mi-2/NuRD complex.* Mol Biosyst, 2010. **6**(9): p. 1700-6.
9. Bartke, T., et al., *Nucleosome-interacting proteins regulated by DNA and histone methylation.* Cell, 2010. **143**(3): p. 470-84.
10. Dignam, J.D., R.M. Lebovitz, and R.G. Roeder, *Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei.* Nucleic Acids Res, 1983. **11**(5): p. 1475-89.
11. Shevchenko, A., et al., *In-gel digestion for mass spectrometric characterization of proteins and proteomes.* Nat Protoc, 2006. **1**(6): p. 2856-60.
12. Rappsilber, J., M. Mann, and Y. Ishihama, *Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips.* Nat Protoc, 2007. **2**(8): p. 1896-906.
13. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
14. Cox, J., et al., *A practical guide to the MaxQuant computational platform for SILAC-based quantitative proteomics.* Nat Protoc, 2009. **4**(5): p. 698-705.

Chapter 3

Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives

Cornelia G. Spruijt[1#], Felix Gnerlich[2#], Arne H. Smits[1], Toni Pfaffeneder[2], Pascal W.T.C. Jansen[1], Christina Bauer[3], Martin Münzel[2], Mirko Wagner[2], Markus Müller[2], Fariha Khan[4,5], H. Christian Eberl[6], Anneloes Mensinga[1], Arie B. Brinkman[7], Konstantin Lephikov[8], Udo Müller[3], Jörn Walter[8], Rolf Boelens[5], Hugo van Ingen[5], Heinrich Leonhardt[3], Thomas Carell[2*] and Michiel

1. Dep. Molecular Cancer Research, Proteomics and Chromatin Biology, UMC Utrecht, 3584 CG, Utrecht, The Netherlands. 2. Center for Integrated Protein Science at the Fakultät für Chemie und Pharmazie, Ludwig-Maximilians-Universität München, 81377, Munich, Germany. 3. Center for Integrated Protein Science at the Fakultät für Biologie, Ludwig-Maximilians-Universität München, 82152, Planegg-Martinsried, Germany. 4. University Institute of Biochemistry and Biotechnology, Pir Mehr Ali Shah Arid Agriculture University Rawalpindi, Rawalpindi, Pakistan. 5. NMR Spectroscopy Research Group, Bijvoet Center for Biomolecular Research, Utrecht University Utrecht, Padualaan 8, 3584 CH, Utrecht, The Netherlands. 6. Proteomics and Signal Transduction, Max-Planck-Institut für Biochemie, D-82152 Martinsried, Germany. 7. Department of Molecular Biology; Nijmegen Centre for Molecular Life Sciences; Radboud University Nijmegen, 6525 GA, Nijmegen, The Netherlands. 8. Genetik / Epigenetik, Universität des Saarlandes, 66123, Saarbrücken, Germany



Cell, February 28, 2013,
Volume 152, p. 1146-1159

ABSTRACT

Tet proteins oxidize 5-methylcytosine (mC) to generate 5-hydroxymethyl (hmC), 5-formyl (fC) and 5-carboxylcytosine (caC). The exact function of these oxidative cytosine bases remains elusive. We applied quantitative mass spectrometry-based proteomics to identify readers for mC and hmC in mouse embryonic stem cells (mESC), neuronal progenitor cells (NPC) and adult mouse brain tissue. Readers for these modifications are only partially overlapping and some readers, such as Rfx proteins, display strong specificity. Interactions are dynamic during differentiation, as for example evidenced by the mESC-specific binding of Klf4 to mC and the NPC-specific binding of Uhrf2 to hmC, suggesting specific biological roles for mC and hmC. Oxidized derivatives of mC recruit distinct transcription regulators as well as a large number of DNA repair proteins in mouse ES cells, implicating the DNA damage response as a major player in active DNA demethylation.

INTRODUCTION

Methylation of cytosine residues at carbon atom 5 of the base (mC) represents a major mechanism via which cells can silence genes. Cytosine methylation mostly occurs in a CpG dinucleotide context. However, CpG islands (CGIs), which are characterized by a very high CpG density and are often found in promoter regions of genes, are typically hypomethylated. Methylation of these CGIs results in transcriptional silencing. The molecular mechanisms underlying the association between DNA methylation and repression of transcription have proven difficult to decipher. The classic view is that methylation of DNA results in the recruitment of methyl-CpG binding proteins (MBPs) that possess transcriptionally repressive enzymatic activities [1]. However, *in vivo* validation for this model on a genome wide level is still lacking. In contrast, recent *in vivo* data has revealed that CXXC-domain containing proteins specifically bind to non-methylated cytosines. In this case, hypomethylated CGIs serve as a recruitment signal for CXXC-domain containing activators that establish a transcriptionally active chromatin state [2].

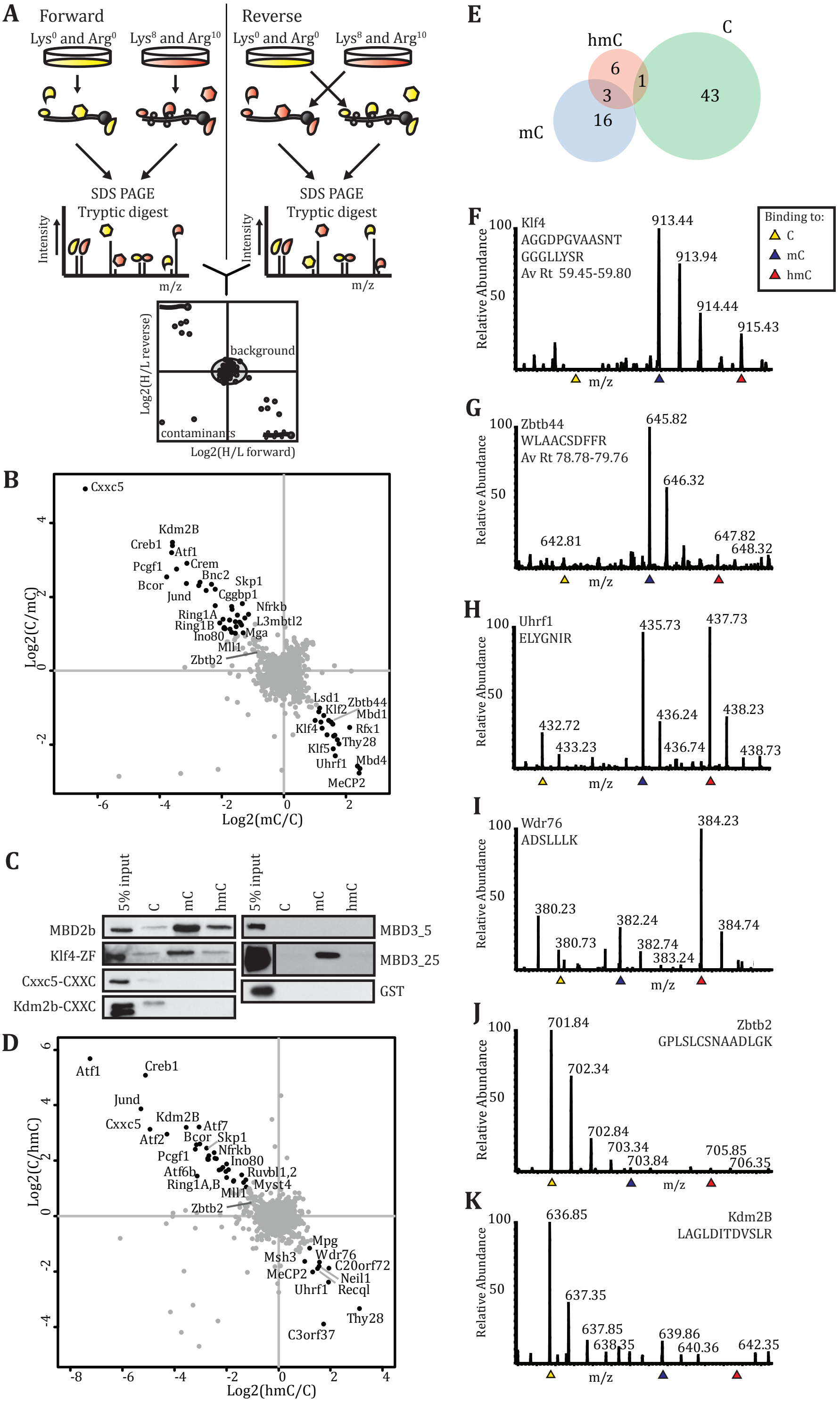
Four years ago it was discovered that Tet enzymes convert mC to 5-hydroxymethylcytosine (hmC) [3, 4]. This modification is particularly abundant in the brain and in embryonic stem cells but is detectable in all tissues tested [5, 6]. Tet enzymes can catalyze further oxidation of hmC to 5-formylcytosine (fC) and 5-carboxylcytosine (caC) [7-9]. Formylcytosine and caC can subsequently serve as substrates for Thymine-DNA glycosylase (Tdg) which eventually results in the generation of a non-methylated cytosine [8, 10]. Therefore, this Tet-Tdg pathway represents an active DNA demethylation pathway. It is not clear if hmC, fC and caC have additional DNA demethylation-independent functions as very few specific binders or 'readers' for these oxidized versions of mC have been described so far.

We applied quantitative mass spectrometry-based proteomics to identify a large number of readers for mC and its oxidized derivatives in mouse embryonic stem cells (mESCs). Furthermore, we also identified readers for mC and hmC in neuronal progenitor cells (NPCs) and adult mouse brain. Our data reveal that each cytosine modification recruits a distinct and dynamic set of proteins. The known biology of these interacting proteins suggests a role for hmC, fC and caC in active DNA demethylation pathways via base excision repair (BER), as well as an epigenetic recruitment function in certain cell types.

RESULTS

Identification of mC and hmC readers in mESCs

To identify readers for methylcytosine and its oxidized derivatives, we made use of a DNA pull-down approach combined with quantitative mass spectrometry. In brief, nuclear extracts from mESCs grown in 'light' or 'heavy' SILAC medium were incubated with a non-modified or a modified double stranded DNA sequence (5'-AAG.ATG.ATG.AXG.AXG.AXG.AXG.ATG.ATG-3'), with X representing C, mC or hmC ('forward' pull-down; Figure 1A). As a control, a label-swap or 'reverse' experiment was performed. Following incubation and washes, beads were combined and bound proteins were in-gel digested with trypsin and analyzed by LC-MS/MS. Raw mass spectrometry data were analyzed using MaxQuant [11]. Specific interactors are distinguishable from background proteins by their heavy/light ratio. Proteins binding selectively to the modified DNA have a high ratio in the forward pull-down and a low ratio in the reverse pull-down, whereas


A
Forward
Lys^0 and Arg^0
Lys^8 and Arg^{10}
Reverse
Lys^0 and Arg^0
Lys^8 and Arg^{10}
SDS PAGE
Tryptic digest
SDS PAGE
Tryptic digest
Intensity
m/z
Intensity
m/z
background
contaminants
Log2(H/L reverse)
Log2(H/L forward)
B
Cxxc5
Kdm2B
Creb1
Atf1
Crem
Pcgf1
Bnc2
Skp1
Bcor
Jund
Cggbp1
Nfrkb
Ring1A
Ring1B
L3mbtl2
Ino80
Mga
Mll1
Zbtb2
Lsd1
Klf2
Zbtb44
Mbd1
Klf4
Rfx1
Klf5
Thy28
Uhrf1
Mbd4
MeCP2
Log2(C/mC)
Log2(mC/C)
4
2
0
-2
-6
-4
-2
0
2
C
5% input
C
mC
hmC
5% input
C
mC
hmC
MBD2b
Klf4-ZF
Cxxc5-CXXC
Kdm2b-CXXC
MBD3_5
MBD3_25
GST
D
Atf1
Creb1
Jund
Cxxc5
Kdm2B
Atf7
Atf2
Bcor
Skp1
Nfrkb
Pcgf1
Ino80
Atf6b
Ruvbl1,2
Ring1A,B
Myst4
Mll1
Zbtb2
Mpg
Msh3
Wdr76
MeCP2
C20orf72
Neil1
Uhrf1
Recql
Thy28
C3orf37
Log2(C/hmC)
Log2(hmC/C)
6
4
2
0
-2
-4
-8
-6
-4
-2
0
2
4
E
hmC
C
mC
6
1
3
43
16
F
Relative Abundance
100
50
Klf4
AGGDPGVAASNT
GGGLLYSR
Av Rt 59.45-59.80
913.44
913.94
914.44
915.43
m/z
Binding to:
C
mC
hmC
G
Relative Abundance
100
50
Zbtb44
WLAACSDFFR
Av Rt 78.78-79.76
645.82
646.32
642.81
647.82
648.32
m/z
H
Relative Abundance
100
50
Uhrf1
ELYGNIR
435.73
437.73
432.72
433.23
436.24
436.74
438.23
438.73
m/z
I
Relative Abundance
100
50
Wdr76
ADSLLLK
384.23
380.23
380.73
382.24
382.74
383.24
384.74
m/z
J
Relative Abundance
100
50
701.84
Zbtb2
GPLSLCSNAADLGK
702.34
702.84
703.34
703.84
705.85
706.35
m/z
K
Relative Abundance
100
50
636.85
Kdm2B
LAGLDITDVSLR
637.35
637.85
638.35
639.86
640.36
642.35
m/z

readers for the non-modified DNA show opposite binding (low forward ratio, high reverse ratio). Background proteins will have a ~1:1 ratio in both pull-downs (Figure 1A).

As shown in Figure 1B and Table S1, we identified 19 proteins enriched for mC compared to C in mESC nuclear extracts (P <0.05 and ratio > 2 in both pull-downs). Among these are the methyl-CpG binding proteins MeCP2, Mbd1, Mbd4 and Uhrf1 [1]. Other interactors include Rfx1 and Zfhx3, which were previously identified as mC readers [12, 13]. Interestingly, three Klf proteins were identified as mC readers: Klf2, 4 and 5. These proteins carry three Krüppel-like zinc fingers, just like the Kaiso family of mC binding proteins. Klf4 is one of the four Yamanaka reprogramming factors and has not been previously identified as a mC binding protein in HeLa or U937 cells [13, 14]. This may be due to the low expression of Klf4 in differentiated cells relative to mESCs. We confirmed the direct binding of the Klf4 Krüppel-like zinc fingers to mC using recombinant protein and two different DNA sequences (Figure 1C and S1A). Using a motif bearing similarities to a recently published consensus binding site for Klf4, as determined by ChIP-seq (GGGXGTG) [15], revealed that Klf4 binds this motif with the highest affinity when 'x' is mC (Figure S1A). These results establish Klf4 as a novel sequence-specific mC binding protein.

Mining published bisulfite sequencing data of mESCs and NPCs [16] and overlapping this data with the Klf4 ChIP-seq profile in mESCs [15] revealed a substantial amount of methylated Klf4 binding sites in mESCs (Figure S1B), which are mainly intronic and intergenic (Figure S1C). Out of the 7321 Klf4 binding sites in mESCs that were covered in the bisulfite sequencing dataset, 1356 show high levels of DNA methylation in mESCs (18.5%). Many of these Klf4 binding sites contain a methylated Klf4 binding motif, such as GGCGTG (Figure S1D and S1E). Interestingly, many Klf4 binding sites that are non-methylated in ES cells become hypermethylated in NPC cells [16] (Figure S1B and S1D). This finding may be highly relevant in the context of Klf4-mediated cellular reprogramming. During reprogramming, Klf4 may be able to bind these methylated loci in differentiated cells to initiate stem cell-specific gene expression patterns. Enrichment analyses for functional domains among the mC interactors revealed DNA-binding zinc fingers to be significantly enriched (Benj.Hoch.FDR=$10^{-2.45}$, Fig S3A). These zinc fingers may also interact with the methylated DNA in a sequence-specific manner.

In addition to the cluster of mC binding proteins, a large number of proteins displayed preferential binding to non-methylated DNA (Figure 1B, upper left quadrant). Consistent with previous observations, this cluster of proteins contains a number of CXXC-domain containing proteins which are known to preferentially bind to non-methylated CpGs [2, 17]. Examples include Cxxc5, Kdm2b and Mll1 (also see Figure 1C). We also identified other subunits of the Mll1 and PRC1.1 (Bcor) complexes, which most likely bind to the non-methylated DNA indirectly via Mll1 and Kdm2b, respectively. Other

Figure 1: Identification of mC and hmC specific readers in mouse embryonic stem cells. **A.** Schematic overview of the workflow. **B.** Scatterplot of a SILAC-based mC DNA pull-down in mESCs nuclear extracts. **C.** Validation of the mC specific binding of Klf4 and non-methyl C specific binding of Cxxc5 and Kdm2b. DNA pull-downs were performed with recombinant GST-fusion proteins followed by western blotting. For MBD3_25µl an empty lane was cut out. **D.** Scatterplot of a SILAC-based hmC DNA pull-down in mESC nuclear extract. **E.** Venn diagram showing overlap of readers for C, mC and hmC. **F-K.** Representative mass spectra obtained in the triple SILAC DNA pull-down in mESCs. Each spectrum shows the relative affinity of the indicated peptides and proteins for non-methylated (yellow), methylated (blue) and hydroxymethylated (red) DNA. See also Figure S1 and TableS1.

interactors include the Ino80 chromatin remodeling complex, zinc finger-containing transcription factors such as Zbtb2 as well as basic-leucine zipper-containing proteins (enriched Benj.Hoch.FDR=$10^{-5.57}$, Figure S3A) such as JunD, Creb1 and Atf7, for which sequence-specific DNA binding is most likely abolished by DNA methylation.

Readers for hmC showed partial overlap with proteins observed to interact with mC (Figure 1D, lower right quadrant and Figure 1E) as only three proteins interacted with both modified baits: MeCP2, Uhrf1 and Thy28. Uhrf1 and MeCP2 are known to bind both mC and hmC, although MeCP2 clearly binds with a higher affinity to mC compared to hmC [18-20]. Thy28 is an uncharacterized protein that is associated with apoptosis [21] and contains an EVE domain, which is possibly involved in (ds) RNA binding [22]. Interestingly, two DNA glycosylases (Mpg and Neil3) and a helicase (Recql) were identified as hmC readers in mESCs. These proteins might be involved in active DNA demethylation pathways to convert hmC back to cytosine via base excision repair mechanisms, as has been suggested previously [23, 24]. In addition, a number of previously uncharacterized proteins, Wdr76 and C3orf37, preferentially bound to hmC compared to C. We purified WDR76 as a GFP fusion protein from HeLa cells and found interactions with OCR, HELLS and GAN (Figure S1F). Hells, or Lsh, is a DNA helicase which has previously been implicated in regulating DNA methylation levels in cells [25]. Interestingly, OCR or Spindlin-1 is a protein known to bind trimethylated H3 lysine 4 (H3K4me3) [13]. A large number of proteins preferentially bound to the non-modified DNA, as was observed for the mC pull-down (Figure S1G). We validated some of these findings using western blotting for endogenous proteins (Figure S1H).

To further investigate the relative affinity of proteins for C versus mC versus hmC in a single experiment, we made use of a triple pull-down approach [26], in which mESCs are grown in three different SILAC media. ‘Light’, ‘medium’ and ‘heavy’ nuclear extracts derived from these cells are incubated with C, mC and hmC-containing DNA, respectively (Table S1). Quantitative mass spectrometry is used to visualize the relative abundance of a protein in each of the three different pull-downs. This experiment confirmed most of the observations made in Figure 1B and 1D, although for some proteins the ratios in the triple pull-down are lower. As shown in Figure 1F and G, Klf4 and Zbtb44 preferentially bind to the methylated DNA. Other proteins bind to both modified baits, such as Uhrf1 (Figure 1H). Kdm2b preferentially binds to the non-modified DNA (Figure 1K). Contrary to a previous report [27], we did not observe a specific interaction between MBD3 and hmC (forward ratio 0.448 and reverse ratio 1.823). We validated these observations using recombinant protein (Figure 1C). At higher concentrations of recombinant MBD3 protein, we observed a specific interaction with mC (Figure 1C), which is in agreement with a recent study that revealed that MBD3 has the highest affinity for mC compared to hmC and C [18].

Taken together, these experiments reveal that mC and hmC both recruit distinct proteins in mESCs with little overlap. Furthermore, a large number of proteins preferentially binds to non-modified DNA. The amount of observed interactions with hmC is moderate and some of these suggest that hmC acts as an intermediate in active DNA demethylation pathways in mESCs.

fC and caC recruit a large number of proteins in mouse embryonic stem cells, including DNA glycosylases and transcription regulators

We also applied our SILAC-based DNA pull-down approach to identify readers for fC and caC in mESCs. Colloidal blue analysis revealed that the total amount of protein binding to each bait is similar (Figure S2A). Ratios of the forward and reverse pull-downs with hmC, fC or caC were individually averaged and these average ratios were then plotted against each other in two-dimensional graphs (Figure 2A-C, Table S1). From these plots, it is clear that both fC (blue, purple and green) and caC (yellow and green) recruit many more proteins than hmC does (red and purple). Strikingly, there is only limited overlap between fC and caC binders (green) (Figure 2D). One of the proteins that binds to fC and caC, but not hmC, is Tdg, which is consistent with its reported substrate specificity [10]. We validated this binding behavior using recombinant protein in electromobility shift assays (EMSA) (Figure 2E and 2F). We also purified GFP-Tdg from mESCs to identify Tdg interaction partners (Figure S2B and TableS1). None of the Tdg interactors were identified as specific readers in the fC and caC pull-down, indicating that these fC and caC interactions are Tdg independent. Another fC specific reader is the p53 protein, which plays an important role in the DNA damage response [28]. Interestingly, Dnmt1 specifically interacted with caC. This interaction was confirmed by EMSA as well as western blotting using an antibody against endogenous protein (Figure 2F and S2C). We also identified subunits of the Swi/Snf chromatin remodeling complex, such as Baf170, as readers for caC. Three proteins bind to all oxidized derivatives of mC: Thy28, C3orf37 and Neil1. GO term enrichment for biological processes shows that fC significantly enriches for proteins that are related to DNA repair (Benj.Hoch.FDR=10-$^{2.71}$) (Figure S3A), whereas caC interactors are not enriched for any biological process. RNA binding proteins, mitochondrial proteins and other proteins which are less likely to be associated with regulation of gene expression or DNA repair binding were identified as binders for fC and caC (Table S3). Some of these may have a basic affinity for the formyl and carboxyl groups on the DNA strands, which are more reactive than methyl or hydroxymethyl. To exclude the possibility that many fC and caC interactors are binding to damaged or abasic DNA, we validated the homogeneity of the DNA strands using HPLC (Figure S2D). Furthermore, we analyzed the DNA before (blue) and after incubation (red) with mESC nuclear extract by MALDI-TOF-MS (Figure S2E). Quantification of the modified residues by LC-MS/MS shows that there is no significant loss of the modified bases after incubation with nuclear extract (Figure S2F). Figure 2A-C also show that the group of proteins that bind preferentially to non-modified cytosine (black, lower left quadrant) shows a large overlap between the three pull-downs and contains the PRC1.1, Mll1 and Ino80 complexes. To compare the relative affinity of proteins for these three modifications in a single experiment, we performed a triple pull-down. Analyses of the triple pull-down ratios for the identified fC and caC readers show similar trends, although some of the observed ratios are less prominent. As shown in Figure 2 G-L (and Table S1), the representative spectra of the indicated peptides of Tdg, Neil3, Mpg, Dnmt1, MeCP2 and Uhrf1 show relative ratios that are in agreement with ratios obtained in the independent experiments shown in Figure 2A-C.

In summary, our data suggest that oxidized cytosine bases may induce a DNA damage response and trigger base excision repair pathways, which may finally result in DNA demethylation. In addition, each of these modifications recruits transcription

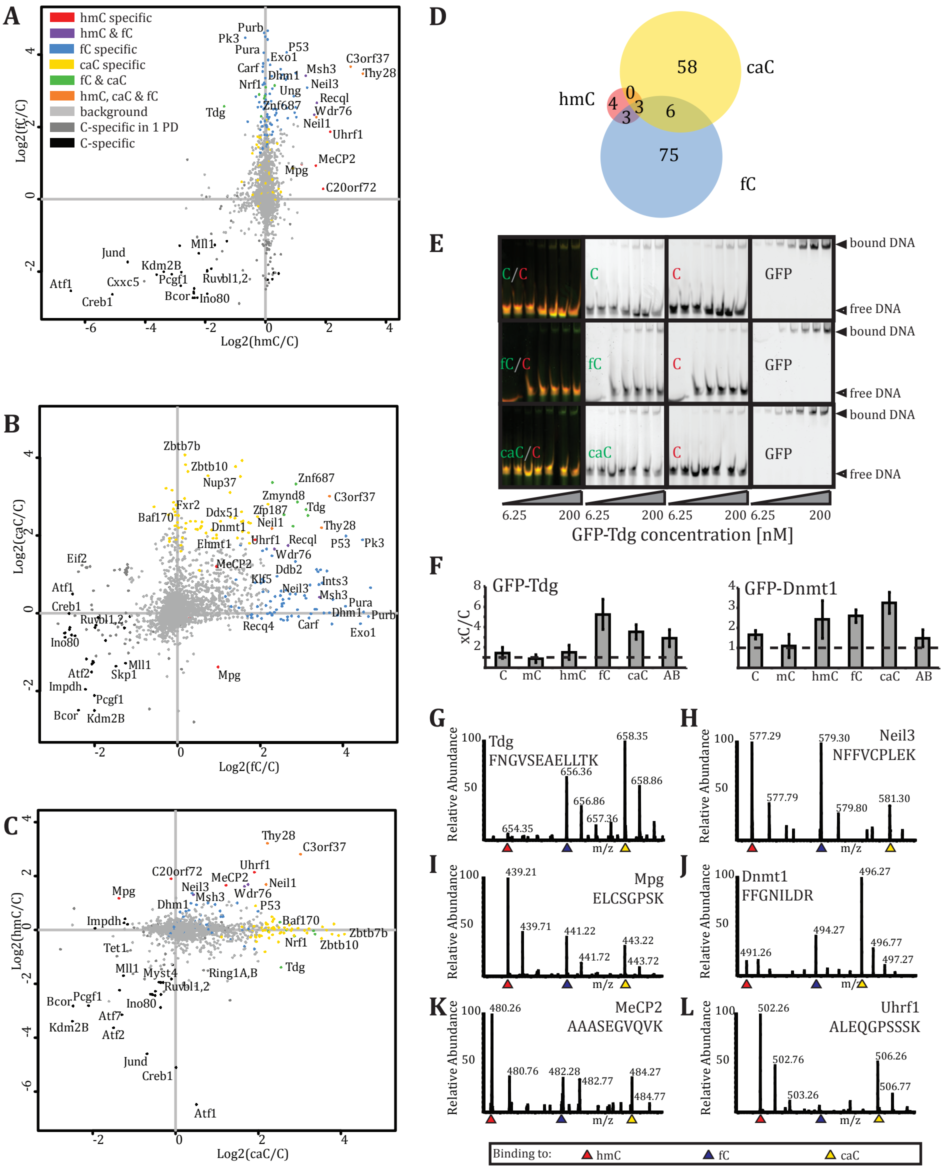


Figure 2: fC and caC recruit a large number of non-overlapping proteins in mouse embryonic stem cells. A-C. Scatterplots of SILAC-based hmC, fC and caC DNA pull-downs in mESC nuclear extract. The average ratio of all the identified and quantified proteins in the forward and reverse experiment for each of the three modifcations is plotted on the X, Y and Z-axes of a three dimensional cube. Shown in A-C are different side views of the cube. Colors indicate in which of the three pull-downs a protein was significantly enriched. **D.** Venn diagram showing the number of significantly enriched proteins for each of the baits. **E.** Electrophoretic mobility shift assay with GFP-Tdg at increasing protein concentrations (6.25 nM to 200 nM) incubated with dsDNA (250 nM of differentially labeled xC- and C-containing oligonucleotide, each). **F.** EMSAs as shown in (E) performed with GFP-Tdg and GFP-Dnmt1 for all 6 residue variants (C, mC, hmC, fC, caC, abasic site (AB)) in direct comparison to unmodified DNA. The binding preference was determined as the ratio of fluorescence signals of the different DNA substrates in the shifted bands. Shown are the means of three experiments and error bars represent standard

regulators and other proteins which are not likely to be related to active DNA demethylation.

NPCs contain a distinct set of mC and hmC readers, including Uhrf2, which has the highest affinity for hmC

To investigate whether interactions with mC and hmC are dynamic during differentiation, we differentiated mESCs to NPCs. Nuclear extracts were generated from these cells followed by DNA pull-downs. Since no SILAC-compatible neurobasal medium is available, these experiments were performed using label-free quantification (LFQ) [29, 30]. Each DNA pull-down is analyzed separately and in triplicate. For all the identified proteins (Table S1), we used ANOVA statistics (P=0.025 and S_0=2) to compare the relative enrichment of proteins for each of the three baits. All significant outliers (192) were hierarchically clustered based on correlation after normalization by row mean subtraction (Figure 3A). Protein enrichment is indicated in red whereas lack of enrichment is shown in blue. A large number of proteins bind to C or mC, whereas fewer proteins are specifically enriched in the pull-downs with hmC. Three smaller groups of proteins bind specifically to two of the baits (C/hmC, C/mC or mC/hmC). As was observed in the DNA pull-downs from mESC nuclear extracts, CXXC-domain containing proteins (Kdm2b and Mll, indicated in black) and their associated factors Bcor/Ring1a/b (blue) and Rbbp5/Ash2l (black) are enriched in the DNA pull-downs with non-modified DNA relative to mC and hmC-containing DNA. We identified Mbd2 and associated Mi-2/NuRD complex subunits as mC readers (indicated in yellow). Other identified MBD proteins include Mbd4, MeCP2 and Mbd1. Furthermore, a number of winged-helix (WH) domain-containing proteins bound specifically to mC, including Rfx5 and its associated factors Rfxap and Rfxank (orange), which have previously been identified as methyl-CpG interactors [13].

Strikingly, these proteins bind more strongly to C compared to hmC. We further substantiated these observations using recombinant protein (Figure 3B). This result indicates that for some readers, oxidation of mC not only weakens the interaction, but repels the mC interactor. The homeobox domain is significantly enriched in the cluster of mC specific readers (Benj.Hoch.FDR=$10^{-1.8}$, Figure S3A), which is consistent with a previous study [13]. In addition, several known mC readers bind both modified forms of cytosine, such as Kaiso, Uhrf1 and Mbd4. A number of DNA glycosylases bind specifically to hmC (Neil1, Neil3), as well as some helicases (Hells, Harp, Recql and its homolog Bloom), which again suggests a DNA repair-involved DNA demethylation pathway (GO DNA repair, Benj.Hoch.FDR=$10^{-3.91}$, Figure S3A). Although homeobox proteins are known to bind specifically to mC, a number of homeobox proteins show preferential binding to hmC in NPC extracts (examples include Zhx1 and 2). Finally, Uhrf2 was identified as a specific hmC binding protein in NPCs, which we confirmed using recombinant protein (Figure 3B). Uhrf2 is not expressed in mESCs and its levels increase upon differentiation [31]. This explains why Uhrf2 was not identified as an hmC specific reader in mESC DNA pull-downs.

Figure 2 (continued). deviation. **G-L.** Representative spectra of the indicated peptides obtained in the triple labeled DNA pull-down in mESCs. Each spectrum shows the relative affinity of the indicated peptides and proteins for hmC (red), fC (blue) and caC (yellow) containing DNA. Spectra are shown for Tdg (**G**), Neil3 (**H**), Mpg (**I**), Dnmt1 (**J**), MeCP2 (**K**) and Uhrf1 (**L**). See also Figure S2 and Table S1.

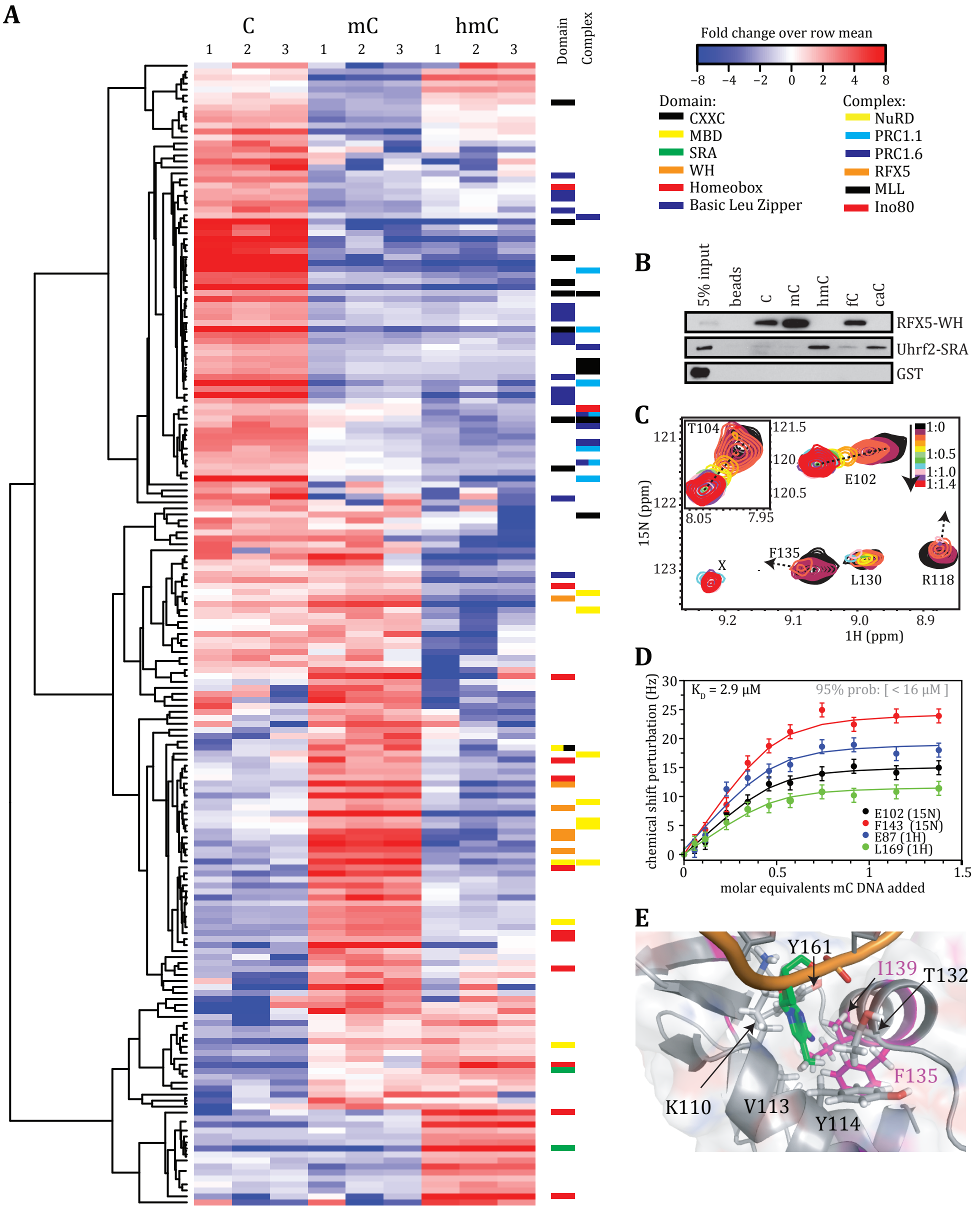


Figure 3: Hierarchical clustering of NPC-specific C, mC and hmC readers. A. Correlation-based clustering of the LFQ intensities after log2 transformation and normalization by row mean subtraction. Included in the clustering are proteins that are significantly binding to at least one of the baits as determined by an ANOVA test. Blue indicates lack of enrichment whereas enrichment is indicated in red. Domain and Complex columns indicate the DNA binding domain(s) that may be responsible for direct binding to the bait and the complexes that readers are part of, respectively. **B.** Biochemical validation experiments using DNA pull-downs with recombinant DNA binding domains. **C.** Overlay of Rfx5-WH HSQC spectra with increasing amounts of mC DNA added, color-coded on the indicated scale listing the WH domain:DNA ratio. Some residues, such as F135 and R118, cannot be unambiguously tracked to their bound states, because their chemical shift changes are very large. Peaks corresponding to their bound state, such as "X" appear only after addition of a full molar equivalent of DNA. **D.** Selected

Taken together, these experiments reveal that interactions with mC and hmC are highly dynamic during differentiation. Furthermore, the observations made in NPCs strengthen our hypothesis that oxidation of mC serves as a trigger for active DNA demethylation. Nevertheless, some hmC specific readers in NPCs do not appear to be linked to DNA repair mechanisms, indicating that in these cells hmC may also serve a role as a 'classical' epigenetic mark that recruits transcriptional regulators.

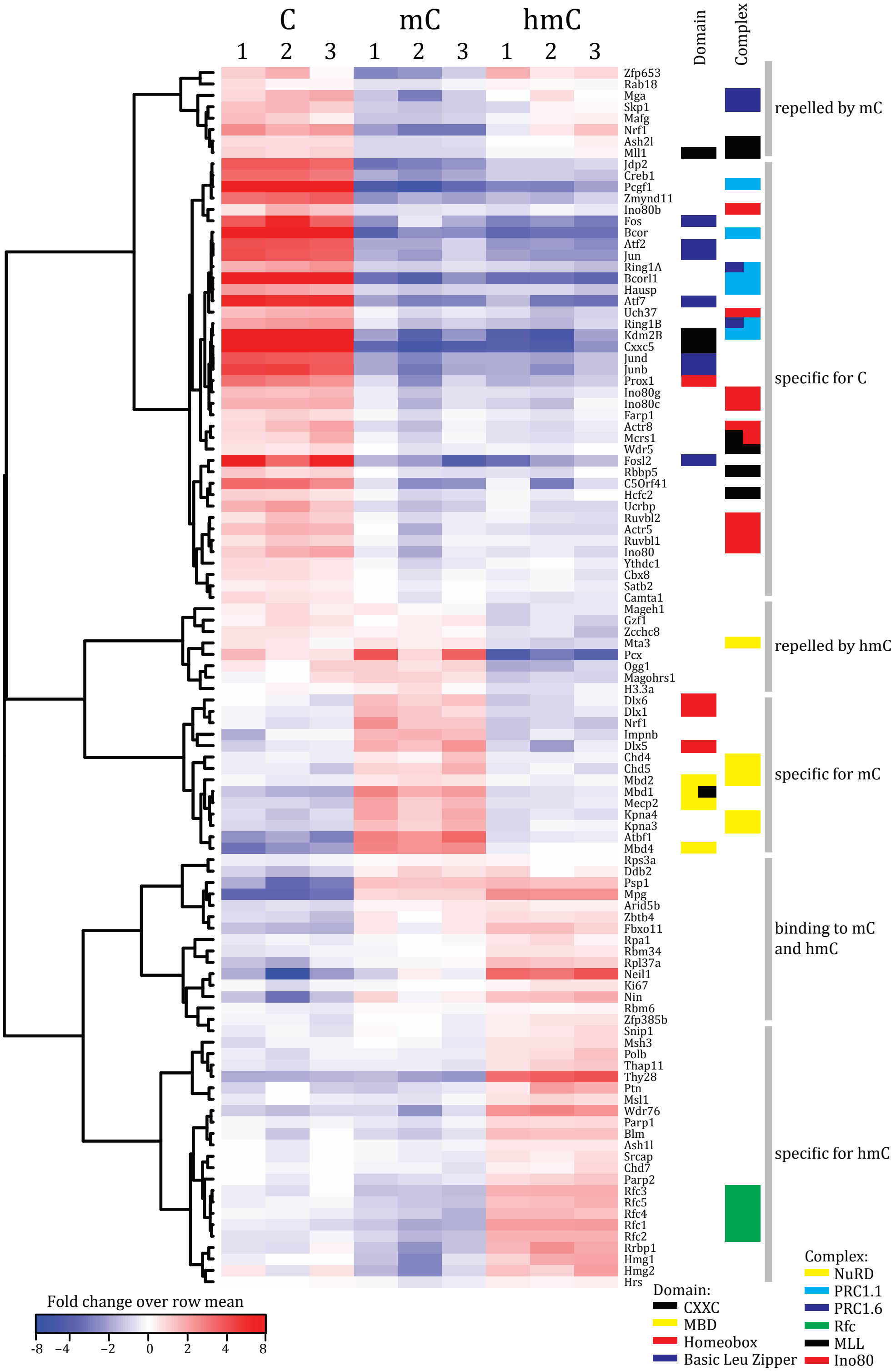
NMR based analysis of the Rfx5 winged helix domain bound to mC DNA

The specific interaction between the Rfx5 winged helix (WH) domain and mC DNA was studied in detail using solution NMR spectroscopy in order to derive binding affinity and identify the mC binding site. Addition of a singly methylated 18bp DNA fragment to the Rfx5-WH domain results in large changes in the ^{1}H-^{15}N HSQC 'fingerprint' spectrum (Figure 3C). After addition of a slight molar excess of DNA, the spectrum does not show any further changes, indicating that Rfx5-WH strongly binds mC DNA, and preferentially at only one of the two mC sites (Figure 3C). The affinity of Rfx5 for mC DNA was derived from the observed peak displacement for residues in the fast exchange regime, such as T104 and E102, assuming the two mCs are independent and equivalent, which resulted in an apparent dissociation constant $K_{D,app}$ of ~3 μM (with 95% probability limits 10 nM < K_D < 16 μM) (Figure 3D and Suppl. information). Based on DNA pull-downs done with recombinant protein, which revealed a quantitative depletion of the WH-domain from the lysate, we anticipate the K_D to be in the nM range (Figure 3B). To identify the residues responsible for specific mC binding, we used the DNA-bound Rfx1 winged helix domain crystal structure (PDB-id 1DP7; sequence identity 35%; [32, 33] to construct a homology model structure of Rfx5-WH and validated it against the experimental chemical shifts (data not shown). The homology model contains a hydrophobic pocket that includes residues with the largest chemical shift changes and is well aligned with an extended basic surface which is responsible for DNA binding in Rfx1. This binding pocket, formed by the side chains of K110, V113, Y114, T132, F135, L139 and Y169, is appropriately shaped to capture the mC base via a flip-out mechanism as seen in the case of UHRF1 (Figure 3E). Steric clashes introduced by the presence of an additional hydroxyl group could cause the observed specificity for mC. Given the apparent high affinity and DNA sequence-independent binding to mC, we propose that the WH-domain present in Rfx proteins is a *bona fide* mCpG binding domain.

Brain-specific readers for mC and hmC include Dlx proteins

The adult brain is the organ with the highest levels of hmC [5]. Tet enzymes and hmC have been shown to play a role in active DNA demethylation of certain genes in this organ [34]. To identify readers for C, mC and hmC in the adult brain, nuclear extracts were prepared from this tissue and these extracts were used for DNA pull-downs. LFQ was used to determine differential binders (Table S1). In brain extracts, we identified fewer specific readers compared to NPCs (108, P=0.025 and S_0=0) (Figure 4), most likely due to the presence of highly abundant structural proteins derived from

Figure 3 (continued). binding curves and fits for resonances that are in the fast-exchange regime throughout the titration.The error bars (standard deviations) for the peak positions are set to 1.2 Hz. **E.** Close up of the putative mC binding pocket in the RFX5 WH domain. The methylated cytosine is indicated in green. See also Table S1.


C
mC
hmC
1
2
3
1
2
3
1
2
3
Domain
Complex
Zfp653
Rab18
Mga
Skp1
Mafg
Nrf1
Ash2l
Mll1
Jdp2
Creb1
Pcgf1
Zmynd11
Ino80b
Fos
Bcor
Atf2
Jun
Ring1A
Bcorl1
Hausp
Atf7
Uch37
Ring1B
Kdm2B
Cxxc5
Jund
Junb
Prox1
Ino80g
Ino80c
Farp1
Actr8
Mcrs1
Wdr5
Fosl2
Rbbp5
C5Orf41
Hcfc2
Ucrbp
Ruvbl2
Actr5
Ruvbl1
Ino80
Ythdc1
Cbx8
Satb2
Camta1
Mageh1
Gzf1
Zcchc8
Mta3
Pcx
Ogg1
Magohrs1
H3.3a
Dlx6
Dlx1
Nrf1
Impnb
Dlx5
Chd4
Chd5
Mbd2
Mbd1
Mecp2
Kpna4
Kpna3
Atbf1
Mbd4
Rps3a
Ddb2
Psp1
Mpg
Arid5b
Zbtb4
Fbxo11
Rpa1
Rbm34
Rpl37a
Neil1
Ki67
Nin
Rbm6
Zfp385b
Snip1
Msh3
Polb
Thap11
Thy28
Ptn
Msl1
Wdr76
Parp1
Blm
Ash1l
Srcap
Chd7
Parp2
Rfc3
Rfc5
Rfc4
Rfc1
Rfc2
Rrbp1
Hmg1
Hmg2
Hrs
repelled by mC
specific for C
repelled by hmC
specific for mC
binding to mC
and hmC
specific for hmC
Fold change over row mean
-8
−4
−2
0
2
4
8
Domain:
CXXC
MBD
Homeobox
Basic Leu Zipper
Complex:
NuRD
PRC1.1
PRC1.6
Rfc
MLL
Ino80

connective tissue and extracellular matrix in these nuclear extracts. Interestingly, more proteins specifically bind to hmC compared to mC in brain extracts. This is in contrast to NPCs and mESCs, in which more interactions with mC relative to hmC are observed, which may imply a specific role for hmC in brain tissue.

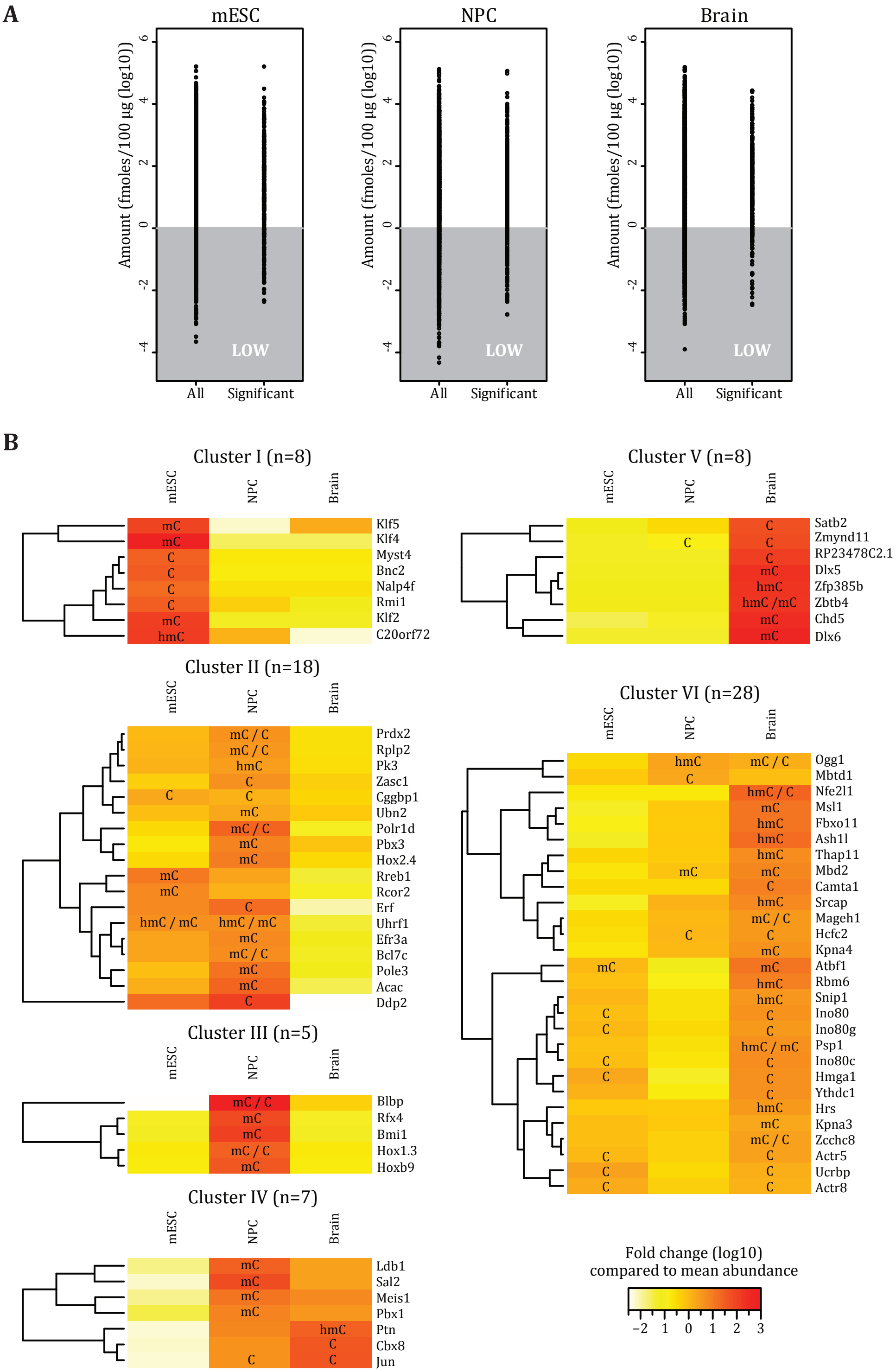
The non-modified DNA pull-down enriched for the same factors as those observed in mESCs and NPCs, including Cxxc5, Kdm2b, Bcor (CXXC-domains indicated in black, PRC1 complex in blue and Ino80 in red). In this case, mC DNA was bound by the Mbd2/NuRD complex which contains the brain-specific ATPase Chd5 [30, 35] (indicated in yellow). Interestingly, we identified 3 distal-less homeobox proteins (Dlx1, 5 and 6) as specific mC interactors. Dlx proteins play a role in the development of the brain and are also expressed in specific regions of the adult brain [36, 37]. Wdr76 and Thy28 are hmC specific, as was also observed in NPCs. Thap11 (or Ronin) is identified as a brain-specific hmC reader. Interestingly, this protein is highly expressed in certain regions of the brain, including Purkinje cells [38]. Finally, we identified all four subunits of replication factor C (Rfc2-5) and the associated factor Rfc1 as hmC specific readers (indicated in green).

Altogether, these experiments further emphasize the dynamic nature of the mC and hmC interactomes during development.

Global absolute quantification of protein levels in mESCs, NPCs and adult mouse brain extracts reveals expression level-dependent and independent interaction dynamics

Our screening for mC and hmC specific readers in mESCs, NPCs and adult mouse brain revealed a large number of cell-type or organ specific interactors (Figure S3B). The most obvious explanation for these observed differential interactions is regulation of reader abundance at the protein level. Alternatively, the interaction between a reader and (modified) DNA may be affected by post-translational modifications (PTMs). To investigate global absolute protein levels in the different nuclear extracts that were used for the pull-downs, we made use of a method called intensity-Based Absolute Quantification (iBAQ) [39]. Approximately 8000 proteins were quantified in at least one of the extracts (Table S2). All proteins with at least a 10-fold change in concentration were clustered based on their expression pattern (Figure S4B). The cluster of mESC specific proteins is enriched for anchoring junction (Benj.Hoch.FDR=$10^{-2.96}$) and cell adhesion (Benj.Hoch.FDR=$10^{-2.14}$), whereas proteins in brain enriched GO terms such as synaptic transmission (Benj.Hoch.FDR=$10^{-3.77}$) and cognition (Benj.Hoch.FDR=$10^{-2.75}$), as expected (Figure S4C). The molar concentrations of proteins that are significantly enriched in one of the DNA pull-downs are spread over several orders of magnitude, indicating that our screening is not biased towards high-abundant proteins (Figure 5A). Of the 259 proteins that showed dynamic interactions through development (Table S3), 20 proteins were not quantified in the iBAQ measurements. The 74 proteins (~31%) that do show a correlation between interaction pattern and protein abundance in the different extracts can be divided into 6 clusters (Figure 5B). A correlation was defined as gaining or losing an interaction accompanied by at least a two-fold change in protein abundance. An example of a protein that was identified as a specific (mC) reader only

Figure 4: Hierarchical clustering of brain-specific C, mC and hmC readers. Correlation-based clustering of the LFQ intensities of proteins in C, mC and hmC DNA pull-downs in adult mouse brain nuclear extracts. See also Table S1.


A
mESC
NPC
Brain
Amount (fmoles/100 µg (log10))
6
4
2
0
-2
-4
LOW
All
Significant
B
Cluster I (n=8)
mESC
NPC
Brain
mC
mC
C
C
C
C
mC
hmC
Klf5
Klf4
Myst4
Bnc2
Nalp4f
Rmi1
Klf2
C20orf72
Cluster V (n=8)
C
C
C
mC
hmC
hmC /mC
mC
mC
Satb2
Zmynd11
RP23478C2.1
Dlx5
Zfp385b
Zbtb4
Chd5
Dlx6
Cluster II (n=18)
mC / C
mC / C
hmC
C
C
C
mC
mC / C
mC
mC
mC
mC
C
hmC / mC
hmC / mC
mC
mC / C
mC
mC
C
Prdx2
Rplp2
Pk3
Zasc1
Cggbp1
Ubn2
Polr1d
Pbx3
Hox2.4
Rreb1
Rcor2
Erf
Uhrf1
Efr3a
Bcl7c
Pole3
Acac
Ddp2
Cluster VI (n=28)
hmC
mC / C
C
hmC / C
mC
hmC
hmC
hmC
mC
mC
C
hmC
mC / C
C
C
mC
mC
mC
hmC
hmC
C
C
C
C
hmC / mC
C
C
C
C
C
hmC
mC
mC / C
C
C
C
C
C
C
Ogg1
Mbtd1
Nfe2l1
Msl1
Fbxo11
Ash1l
Thap11
Mbd2
Camta1
Srcap
Mageh1
Hcfc2
Kpna4
Atbf1
Rbm6
Snip1
Ino80
Ino80g
Psp1
Ino80c
Hmga1
Ythdc1
Hrs
Kpna3
Zcchc8
Actr5
Ucrbp
Actr8
Cluster III (n=5)
mC / C
mC
mC
mC / C
mC
Blbp
Rfx4
Bmi1
Hox1.3
Hoxb9
Cluster IV (n=7)
mC
mC
mC
mC
hmC
C
C
C
Ldb1
Sal2
Meis1
Pbx1
Ptn
Cbx8
Jun
Fold change (log10)
compared to mean abundance
−2
1
0
1
2
3

in mESCs was Klf4. As shown in figure 5B, this protein is highly expressed in mESCs but is less abundant in NPCs or in the adult mouse brain. Another example is represented by the Dlx5 and Dlx6 proteins, which are high abundant in brain nuclear extract and exclusively bind to mC in pull-downs from these extracts. For about 185 proteins, no correlation is observed between expression levels (at least 2-fold change) and binding behavior. For these proteins, the cause of differential binding may be explained through PTMs that affect the interaction between a reader and DNA or a differentially expressed co-factor. A good example of the latter is the Mi-2/NuRD complex. Although most of its subunits display equal expression levels in mESCs, NPCs and brain, mC-specific interactions are not observed in mESCs. This can be explained by the fact that Mbd2, which is the direct reader of mC within the NuRD complex, is low abundant in mESCs and is upregulated during differentiation (Figure 5B). Thereby, it controls the mC specific binding of the entire complex. In mESCs, the majority of the Mi-2/NuRD complex contains Mbd3, which is the MBD-containing protein that has lost its high affinity mC binding ability. Furthermore, technical reasons for not identifying an interactor could be the presence of high-abundant structural proteins in the brain lysate or binding competition amongst different readers in the extracts.

Altogether, the absolute quantification of protein abundance in the different nuclear extracts revealed large differences in protein levels between mESC, NPCs and adult mouse brain. This dataset serves as a rich resource on its own, but also enables us to explain many of the differential interactions that we identified using quantitative mass spectrometry-based interactomics.

Uhrf2 stimulates the sequential activity of the Tet1 enzyme

The first protein that was identified as a hmC binder was Uhrf1 [19], a protein involved in maintenance of DNA methylation [40]. Our data revealed that Uhrf1 binds with a similar affinity to mC and hmC, which is consistent with previously published data [19]. This is in contrast to Uhrf2, which we identified as a high affinity hmC binding protein in NPCs that shows a lower affinity for mC. The function of Uhrf2 is not well understood. It is clear, however, that Uhrf2 cannot rescue the phenotype of Uhrf1 knock-out cells, which lose DNA methylation [31, 41]. Uhrf1 is highly expressed in mESCs, while Uhrf2 levels increase during differentiation (Table S3 and [31]). Altogether, this prompted us to investigate whether Uhrf2 expression affects the levels of mC and its oxidized derivatives. The Tet1-catalytic domain was transfected into HEK-293T cells with and without co-expression of Uhrf2. Total genomic DNA and modification levels were determined using LC-MS/MS (Figure 6 and Supplementary Information). As shown in Figure 6D, Uhrf2 over-expression increases the level of hmC. More striking is the increase of fC and caC levels upon Uhrf2 co-expression together with the Tet1 catalytic domain. Because fC and caC serve as substrates for Tdg and BER, the detected increase in the levels of fC and caC following Uhrf2 expression may be an underestimation of the actual production of these bases. It therefore seems that Uhrf2

Figure 5: Global absolute protein quantification in mESCs, NPCs and adult mouse brain. A. Graphs indicating the concentration of all proteins identified in the nuclear extract (all) and the identified readers (significant) in each of the cell types. The grey area indicates the concentration at which protein quantification is inaccurate. **B.** Readers for which protein expression levels correlate with DNA binding patterns were clustered into six groups based on their expression in the three different nuclear extracts. The color indicates protein levels (white = low and red = high), while binding preference is indicated by C, mC, hmC or combinations thereof. See also Figure S4 and Table S2 and Table S3.

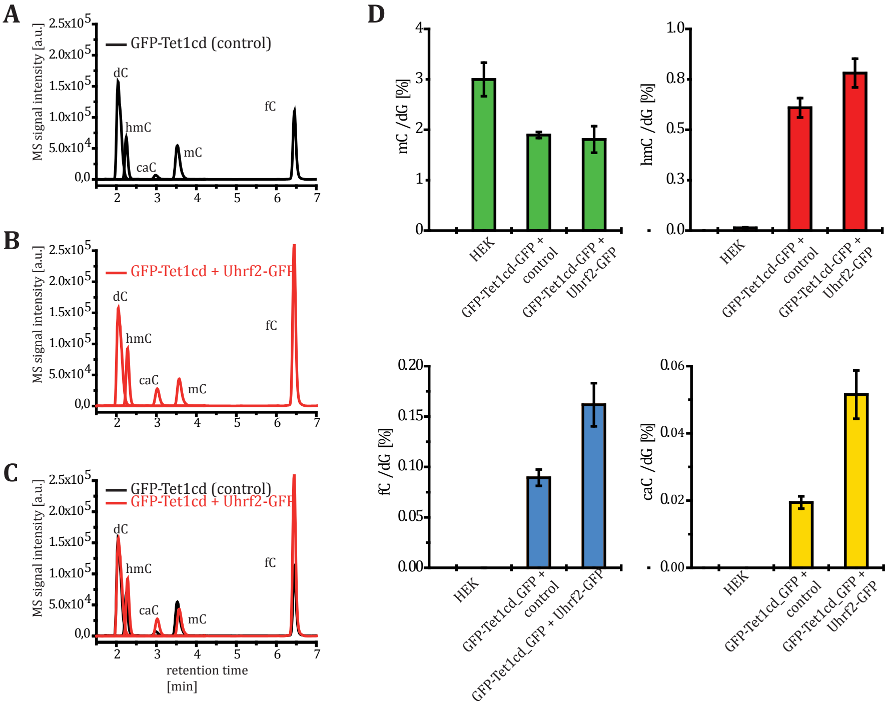


Figure 6: Stable isotope-dilution based LC-ESI-MS/MS quantification of cytosine derivatives in HEK-293T cells. A. Non-quantitative LC-MS/MS-chromatogram of digested genomic DNA from HEK-293T cells cotransfected with Tet1-catalytic domain-GFP (GFP-Tet1cd) and an unrelated expression construct (control). Depicted are the overlaid ion-chromatograms of the MS/MS-transitions for dC and the cytosine derivatives (black curves). dC, mC and hmC were measured by a factor of approx. 10^2-10^3 less sensitive in comparison to caC and fC. **B.** Same as A., except that Uhrf2-GFP was co-expressed together with GFP-Tet1cd. The MS signal intensities were normilized to the dC content of A. **C.** Superposition of A and B. **D.** Levels of cytosine derivatives relative to the total cytosine content (dG) as determined by quantitative LC-MS/MS mass spectrometry. Shown are the means of technical triplicates and error bars reflect standard deviation.

promotes repetitive oxidation of mC by the Tet proteins. We hypothesize that flipping the modified cytosine base out of the DNA double helix, as has been described for Uhrf1 binding to methylated and hydroxymethylated DNA [19, 42], may enhance accessibility of the hydroxymethylated
base to the Tet enzymes, thereby promoting further oxidation.

DISCUSSION

In this study we have used quantitative mass spectrometry-based proteomics to identify readers for mC and its oxidized derivatives in mESCs as well as readers for mC and hmC in NPCs and adult mouse brain. Readers for individual modifications were found to be highly dynamic throughout the three cell types and tissues that we investigated (Figure 7). This is in contrast to interactions with histone modifications, such as trimethylated lysines on histone H3. For these modifications, the majority of interactors are constant between different cell types or developmental stages ([30] and M.V., unpublished observations). Readers for distinct cytosine modifications show

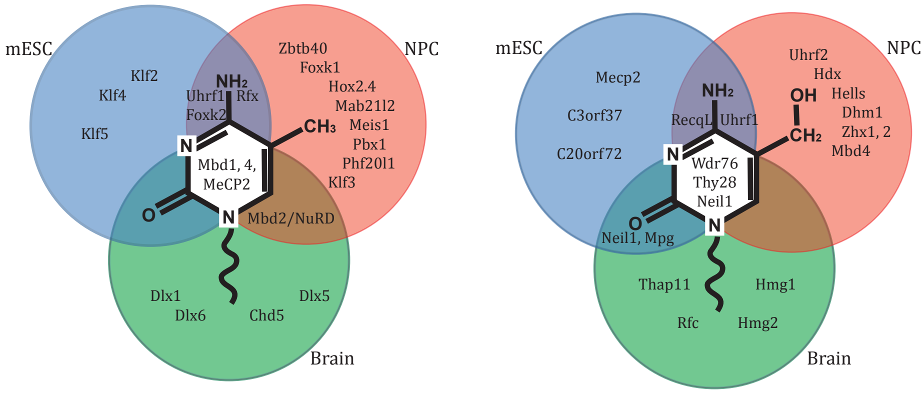


Figure 7. Venn diagram showing examples of mC (**A**) and hmC (**B**) readers that were identified in mESCs (blue), NPCs (orange) and adult mouse brain (green). See also Figure S3 and Table S3.

limited overlap. This indicates that, at least from a biochemical perspective, mC, hmC, fC and caC behave quite differently. Although little overlap was observed with regard to proteins that interact with each of the epigenetic marks, they all repelled a common set of proteins, such as several CXXC-domain containing proteins and their interactors. It remains to be determined which of the consequences of DNA (hydroxy)methylation is functionally most relevant: recruitment of transcriptionally repressive complexes or preventing the binding of certain (activating) proteins to unmodified DNA. A detailed biochemical characterization of the interactions and their dissociation constants will be important to answer this question.

Our experiments revealed a number of DNA glycosylases and DNA repair proteins that bind to hmC, fC and caC, whereas we identified few such proteins binding to mC. The enriched binding of DNA repair-associated proteins was most pronounced for fC. From this observation, one can conclude that the conversion of hmC to fC is a signal that is likely to result in repair-associated removal of the modified base by proteins which are rather ubiquitously expressed. It is therefore surprising that in different cell types and tissues, rather constant levels of hmC, fC and caC are found. The maintenance of such constant levels of these bases in mESCs may indicate a high turnover of DNA methylation, probably involving a constant "correction" by *de novo* methylation. Regardless, it will be important to investigate which mechanisms control Tet enzyme conversion of mC to hmC and further oxidation to fC and caC. Our data reveal that co-expression of Uhrf2 with the catalytic domain of Tet1 results in a (transient) upregulation of hmC, fC and caC, indicating that Uhrf2 promotes the sequential oxidation of mC by Tet1. One of the other factors influencing the catalytic activity of the Tet enzymes is the concentration of cellular metabolites. It has been shown that oncometabolites such as 2-hydroxyglutarate can competitively inhibit the activity of 2-oxo-glutarate dependent enzymes, such as the Tet proteins [43, 44]. Furthermore, mutations in IDH1 and 2, which generate 2-oxo-glutarate, are phenocopied by mutations in the TET enzymes and result in cancer [45]. Mutations in the IDH2 and TET2 genes were also linked to lower genomic hmC levels and altered gene expression patterns in myeloid cancers [46, 47]. In support of these observations, which clearly link hmC to cancer, we noticed that many hmC, fC and caC readers are implicated in cancer, including UHRF2, CARF, p53 and Hells [48]. Interestingly, mutations in the Hells helicase, which we identified as a hmC reader in NPCs, result in a decrease of DNA methylation levels in cells [49]. It seems clear that regulating the levels of mC and its oxidized derivatives is essential for normal cell homeostasis and that deregulation of the readers, writers and erasers of these marks results in a disturbance of the balance between cell proliferation and differentiation during development.

MATERIALS AND METHODS

Cell culture

IB10 mESCs were cultured in 'light' (R^0K^0) or 'heavy' ($R^{10}K^8$) SILAC medium in the presence of 2i compounds. For triple labeling, a third type of medium was used containing medium-labeled L-lysine (K^4) and L-arginine (R^6). mESCs were differentiated to NPCs in N2B27 medium and cultured in NSA medium, consisting of NSA MEM, 1% glutamine, 1x N2 supplement, 10 ng/mL bFGF and 10 ng/mL EGF.

DNA pull-downs

Nuclear extracts were generated as described previously [26, 30]. DNA (See Table S4) immobilized on Dynabeads MyOne C1 was incubated with nuclear extract in 50mM Tris-HCl pH8, 150 mM NaCl, 1mM DTT, 0.25% NP40 and complete protease inhibitors (Roche, EDTA-free) in the presence of poly-dAdT. After extensive washes (using incubation buffer w/o poly-dAdT), bound proteins were in-gel digested using trypsin. After sample preparation, peptides were desalted on Stage-tips [50].

Mass spectrometry

Peptides were separated using an EASY-nLC (Proxeon) connected online to an LTQ-Orbitrap Velos mass spectrometer (Thermo) as described [51]. Raw data was analysed using MaxQuant version 1.2.2.5 and searched against protein database ipi.MOUSE.v3.68.fasta. Using Perseus data were filtered and scatter plots were made using R.

Recombinant protein expression and DNA pull-downs

DNA-binding domains were cloned into the GST-containing PRP256NB vector. The Uhrf2(aa416-626) GST–fusion construct was kindly provided by Dr. Jiemin Wong. Protein expression was performed in E. coli BL21 codon$^+$ cells. Bacterial lysate was cleared by ultracentrifugation. DNA pull-downs were performed as described above with the addition of 10 μM $ZnCl_2$ to the incubation buffer.

iBAQ

iBAQ was performed essentially as described in [39] and the supplementary information.

LC-MS/MS analysis of genomic DNA

Co-transfections were performed in HEK293 cells and genomic DNA was purified according to [52]. Quantification of DNA-nucleosides from genomic DNA is based on a further development of our isotope dilution method ([9] and manuscript in preparation). LC-MS/MS analysis was performed on an Agilent 6490 triple quadrupole mass spectrometer coupled to an Agilent 1290 UHPLC system. For general source- and compound-dependent parameters see Supplementary Methods and Tables S5 and S6. The transitions of the nucleosides were analyzed in the positive ion selected reaction monitoring mode (SRM) operating MS1 and MS2 under unit mass resolution conditions.

ACKNOWLEDGEMENTS

We would like to thank Dr. Jiemin Wong for providing constructs. We are very thankful to Dr. Nikolay Outchkourov for providing NPCs and we thank Tony Hyman and Ina Poser for providing the BAC-GFP WDR76 cell line. We would also like to thank Deepani Poramba Liyanage for support. Furthermore we would like to thank all group members that contributed to fruitful discussions and Susan Kloet for critical reading of the manuscript. Work in the Vermeulen lab is supported by a grant from the Netherlands Organization for Scientific Research (NWO-VIDI) and an ERC Starting Grant. The Carell group thanks the DFG Normalverfahren CA275/8-4, SFB 749, SFB 646 and the Volkswagen Foundation for financial support. Work in the Leonhardt group is supported by the SFB 646 and SPP 1463 grants from The Deutsche Forschungsgemeinschaft (DFG). Martin M. and T. P. thank the Fonds der Chemischen Industrie for pre-doctoral fellowships. F.K. is supported through the International Research Support Initiative Program from the Higher Education Commission, Pakistan. H.v.I. is supported by a VENI fellowship from the Netherlands Organization for Scientific Research, NWO-CW. K. Lepikhov and J. Walter thank the DFG (WA1029/6) for support through SPP1356.

REFERENCES

1. Defossez, P.A. and I. Stancheva, Biological Functions of Methyl-CpG-Binding Proteins. Modifications of Nuclear DNA and Its Regulatory Proteins, 2011. 101: p. 377-398.
2. Thomson, J.P., et al., CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature, 2010. 464(7291): p. 1082-U162.
3. Kriaucionis, S. and N. Heintz, The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. Science, 2009. 324(5929): p. 929-30.
4. Tahiliani, M., et al., Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. Science, 2009. 324(5929): p. 930-935.
5. Globisch, D., et al., Tissue Distribution of 5-Hydroxymethylcytosine and Search for Active Demethylation Intermediates. Plos One, 2010. 5(12).
6. Szwagierczak, A., et al., Sensitive enzymatic quantification of 5-hydroxymethylcytosine in genomic DNA. Nucleic Acids Research, 2010. 38(19).
7. Ito, S., et al., Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. Science, 2011. 333(6047): p. 1300-3.
8. He, Y.F., et al., Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. Science, 2011. 333(6047): p. 1303-1307.
9. Pfaffeneder, T., et al., The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. Angewandte Chemie-International Edition, 2011. 50(31): p. 7008-7012.
10. Maiti, A. and A.C. Drohat, Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES. Journal of Biological Chemistry, 2011. 286(41): p. 35334-35338.
11. Cox, J. and M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nature Biotechnology, 2008. 26(12): p. 1367-1372.
12. Sengupta, P.K., M. Ehrlich, and B.D. Smith, A methylation-responsive MDBP/RFX site is in the first exon of the collagen alpha 2(I) promoter. Journal of Biological Chemistry, 1999. 274(51): p. 36649-36655.

13. Bartke, T., et al., Nucleosome-interacting proteins regulated by DNA and histone methylation. Cell, 2010. 143(3): p. 470-84.
14. Bartels, S.J.J., et al., A SILAC-Based Screen for Methyl-CpG Binding Proteins Identifies RBP-J as a DNA Methylation and Sequence-Specific Binding Protein. Plos One, 2011. 6(10).
15. Chen, X., et al., Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell, 2008. 133(6): p. 1106-1117.
16. Stadler, M.B., et al., DNA-binding factors shape the mouse methylome at distal regulatory regions (vol 480, pg 490, 2011). Nature, 2012. 484(7395): p. 550-550.
17. Blackledge, N.P., et al., CpG Islands Recruit a Histone H3 Lysine 36 Demethylase. Molecular Cell, 2010. 38(2): p. 179-190.
18. Hashimoto, H., et al., Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. Nucleic Acids Research, 2012. 40(11): p. 4841-4849.
19. Frauer, C., et al., Recognition of 5-Hydroxymethylcytosine by the Uhrf1 SRA Domain. Plos One, 2011. 6(6).
20. Mellen, M., et al., MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system. Cell, 2012. 151(7): p. 1417-30.
21. Toyota, H., et al., Thy28 partially prevents apoptosis induction following engagement of membrane immunoglobulin in WEHI-231 B lymphoma cells. Cellular & Molecular Biology Letters, 2012. 17(1): p. 36-48.
22. Bertonati, C., et al., Structural genomics reveals EVE as a new ASCH/PUA-related domain. Proteins-Structure Function and Bioinformatics, 2009. 75(3): p. 760-773.
23. Hajkova, P., et al., Genome-Wide Reprogramming in the Mouse Germ Line Entails the Base Excision Repair Pathway. Science, 2010. 329(5987): p. 78-82.
24. Wossidlo, M., et al., Dynamic link of DNA demethylation, DNA strand breaks and repair in mouse zygotes. Embo Journal, 2010. 29(11): p. 1877-1888.
25. Dennis, K., et al., Lsh, a member of the SNF2 family, is required for genome-wide methylation. Genes & Development, 2001. 15(22): p. 2940-2944.
26. Vermeulen, M., et al., Quantitative Interaction Proteomics and Genome-wide Profiling of Epigenetic Histone Marks and Their Readers. Cell, 2010. 142(6): p. 967-980.
27. Yildirim, O., et al., Mbd3/NURD Complex Regulates Expression of 5-Hydroxymethylcytosine Marked Genes in Embryonic Stem Cells. Cell, 2011. 147(7): p. 1498-1510.
28. Kastan, M.B., et al., Participation of P53 Protein in the Cellular-Response to DNA Damage. Cancer Research, 1991. 51(23): p. 6304-6311.
29. Hubner, N.C. and M. Mann, Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC). Methods, 2011. 53(4): p. 453-9.
30. Eberl, H.C., et al., A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-free Interaction Proteomics. Molecular Cell, 2012.
31. Pichler, G., et al., Cooperative DNA and Histone Binding by Uhrf2 Links the Two Major Repressive Epigenetic Pathways. Journal of Cellular Biochemistry, 2011. 112(9): p. 2585-2593.

32. Avvakumov, G.V., et al., Structural basis for recognition of hemi-methylated DNA by the SRA domain of human UHRF1. Nature, 2008. 455(7214): p. 822-U13.
33. Gajiwala, K.S., et al., Structure of the winged-helix protein hRFX1 reveals a new mode of DNA binding. Nature, 2000. 403(6772): p. 916-921.
34. Guo, J.U., et al., Hydroxylation of 5-Methylcytosine by TET1 Promotes Active DNA Demethylation in the Adult Brain. Cell, 2011. 145(3): p. 423-434.
35. Potts, R.C., et al., CHD5, a Brain-Specific Paralog of Mi2 Chromatin Remodeling Enzymes, Regulates Expression of Neuronal Genes. Plos One, 2011. 6(9).
36. Jones, D.L., et al., Deletion of Dlx1 results in reduced glutamatergic input to hippocampal interneurons. Journal of Neurophysiology, 2011. 105(5): p. 1984-1991.
37. Wang, B., T. Lufkin, and J.L.R. Rubenstein, Dlx6 Regulates Molecular Properties of the Striatum and Central Nucleus of the Amygdala. Journal of Comparative Neurology, 2011. 519(12): p. 2320-2334.
38. Dejosez, M., et al., Ronin is essential for embryogenesis and the pluripotency of mouse embryonic stem cells (vol 133, pg 1162, 2008). Cell, 2008. 134(4): p. 692-692.
39. Schwanhausser, B., et al., Global quantification of mammalian gene expression control. Nature, 2011. 473(7347): p. 337-342.
40. Bostick, M., et al., UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science, 2007. 317(5845): p. 1760-1764.
41. Zhang, J.Q., et al., S phase-dependent interaction with DNMT1 dictates the role of UHRF1 but not UHRF2 in DNA methylation maintenance. Cell Research, 2011. 21(12): p. 1723-1739.
42. Arita, K., et al., Recognition of hemi-methylated DNA by the SRA protein UHRF1 by a base-flipping mechanism. Nature, 2008. 455(7214): p. 818-U12.
43. Xu, W., et al., Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of alpha-ketoglutarate-dependent dioxygenases. Cancer Cell, 2011. 19(1): p. 17-30.
44. Chowdhury, R., et al., The oncometabolite 2-hydroxyglutarate inhibits histone lysine demethylases. Embo Reports, 2011. 12(5): p. 463-469.
45. Figueroa, M.E., et al., Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell, 2010. 18(6): p. 553-67.
46. Konstandin, N., et al., Genomic 5-hydroxymethylcytosine levels correlate with TET2 mutations and a distinct global gene expression pattern in secondary acute myeloid leukemia. Leukemia, 2011. 25(10): p. 1649-1652.
47. Ko, M., et al., Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. Nature, 2010. 468(7325): p. 839-843.
48. Lee, D.W., et al., Proliferation-associated SNF2-like gene (PASG): A SNF2 family member altered in leukemia. Cancer Research, 2000. 60(13): p. 3612-3622.
49. Myant, K., et al., LSH and G9a/GLP complex are required for developmentally programmed DNA methylation. Genome Research, 2011. 21(1): p. 83-94.
50. Rappsilber, J., Y. Ishihama, and M. Mann, Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Analytical Chemistry, 2003. 75(3): p. 663-670.
51. Smits, A.H., et al., Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. Nucleic

Acids Res, 2013. 41(1): p. e28.

52. Munzel, M., et al., Quantification of the sixth DNA base hydroxymethylcytosine in the brain. Angew Chem Int Ed Engl, 2010. 49(31): p. 5375-7.

Figure S1: Genome-wide localization of Klf4 partially correlates with DNA methylation. Related to Figure 1. A. DNA pull-downs with recombinant GST-fusion proteins of DNA binding domains and western blotting analysis. **B**. DNA methylation of Klf4 sites in mESCs and NPCs. Whole-genome bisulfite sequencing was used to determine DNA methylation within a window of +/- 50 bp around Klf4 peak centers. Darker coloring indicates high density of datapoints. **C.** Pie charts showing the genomic distribution of Klf4 sites as presented in the different quadrants of (B). **D.** Distribution of DNA methylation specifically within the GGCGTG sequence present underneath Klf4 sites. **E.** Example of DNA methylation profiles and Klf4 binding (ChIP-seq), showing binding of Klf4 to both methylated and unmethylated sites. Yellow squares indicate the presence of the GGCGTG sequence underneath Klf4 sites. **F.** SILAC-based GFP-purification from HeLa cells stably expressing WDR76-GFP. Significant interactors are indicated in black (high forward WDR76-GFP/control ratio, low reverse control/ WDR76-GFP ratio). **G.** Venn diagram showing the overlap of C-specific readers in the mC and hmC DNA pull-downs from mESC nuclear extracts. **H.** Validation of C and mC specific binders by DNA pull-downs in HeLa nuclear extract and western blotting for the endogenous proteins.

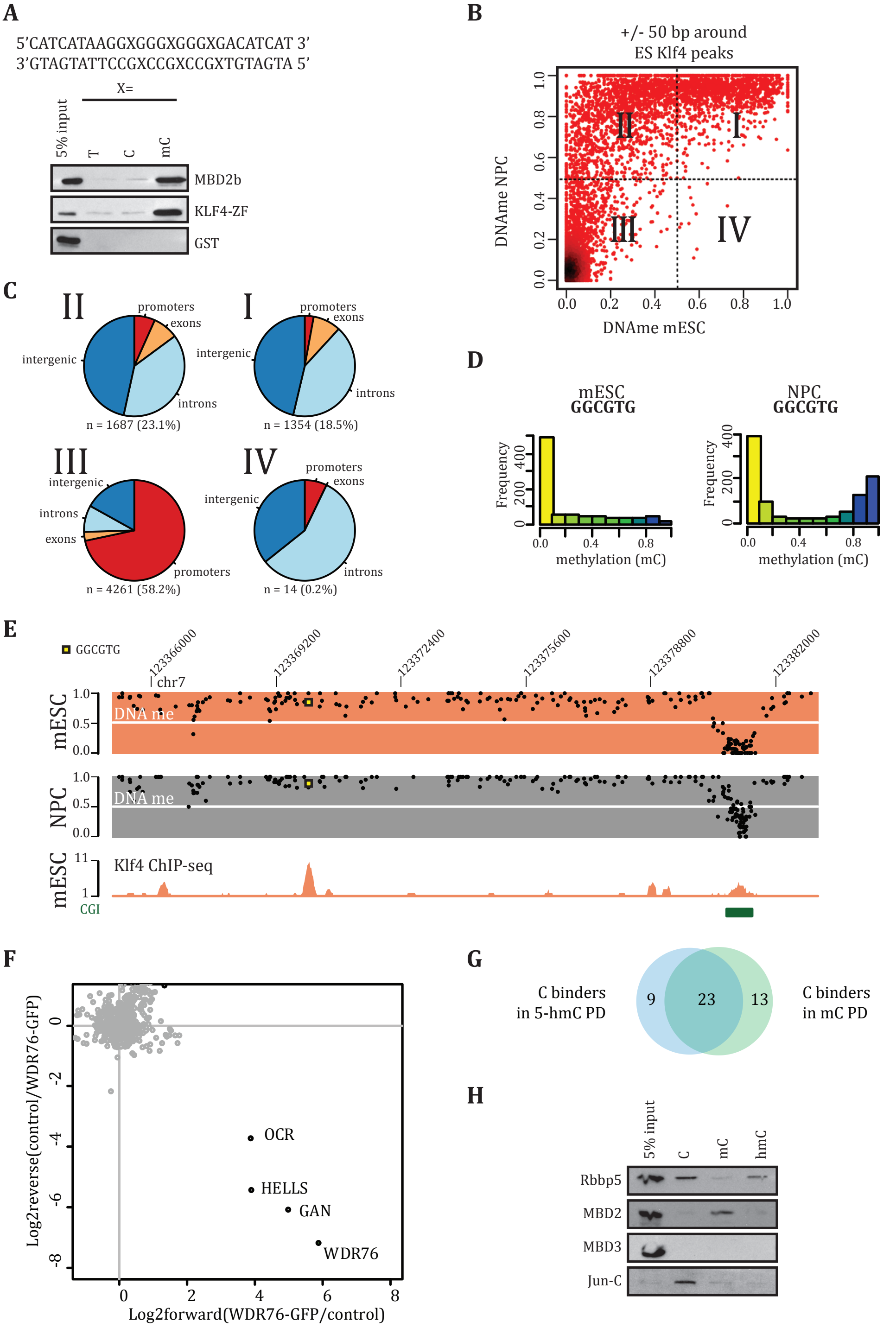
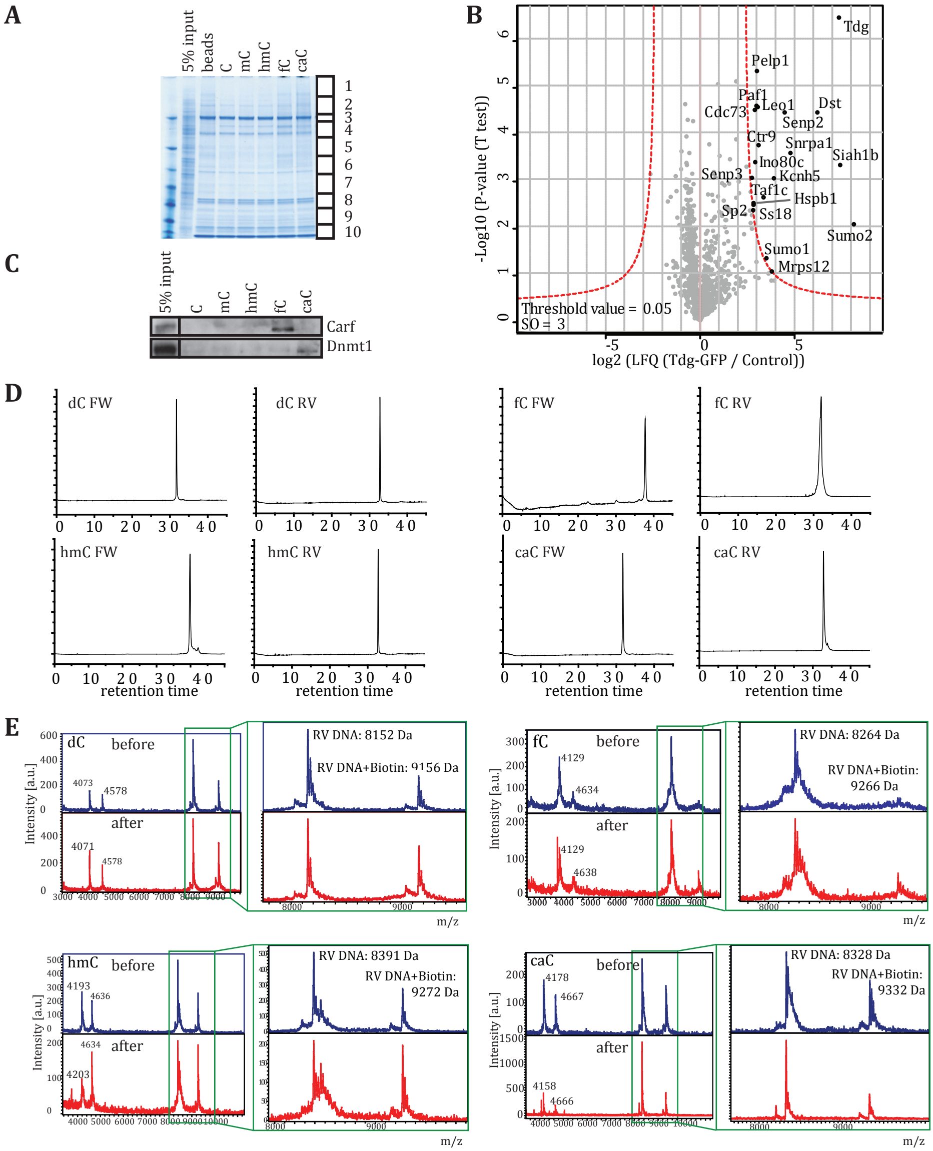

A
5'CATCATAAGGXGGGXGGGXGACATCAT 3'
3'GTAGTATTCCGXCCGXCCGXTGTAGTA 5'
X=
5% input
T
C
mC
MBD2b
KLF4-ZF
GST
B
+/- 50 bp around
ES Klf4 peaks
DNAme NPC
DNAme mESC
I
II
III
IV
0.0
0.2
0.4
0.6
0.8
1.0
C
II
I
III
IV
promoters
exons
intergenic
introns
n = 1687 (23.1%)
n = 1354 (18.5%)
n = 4261 (58.2%)
n = 14 (0.2%)
D
mESC
GGCGTG
NPC
GGCGTG
Frequency
0
200
400
methylation (mC)
0.0
0.4
0.8
E
GGCGTG
123366000
123369200
123372400
123375600
123378800
123382000
chr7
mESC
DNA me
NPC
DNA me
1.0
0.5
0.0
mESC
Klf4 ChIP-seq
11
1
CGI
F
Log2reverse(control/WDR76-GFP)
Log2forward(WDR76-GFP/control)
0
-2
-4
-6
-8
0
2
4
6
8
OCR
HELLS
GAN
WDR76
G
C binders
in 5-hmC PD
C binders
in mC PD
9
23
13
H
5% input
C
mC
hmC
Rbbp5
MBD2
MBD3
Jun-C

Figure S2: **Identification of Tdg interactors, western blot verification of fC and caC interactors and validation of bait DNA quality. A**. The indicated immobilized DNA baits were incubated with mouse nuclear extract. Following washes, bound proteins were analyzed by colloidal blue staining. Note that the elution profile of all these baits looks similar, indicating that specific interactors are masked by a large number of high abundant background binders. **B.** Volcano plot of a label-free GFP-Tdg pull-down in mESC nuclear extract. Significant interactors of GFP-Tdg are identified by permutation-based *t*-test (FDR=0.05 & S0=3). The LFQ intensity of the GFP pull-down over the control is plotted against the –Log10(p-value). The red line indicates the permutation-based FDR. Also see Table S1. **C.** Western blot validation of the fC-specific binding of Carf and caC-specific binding of Dnmt1 in mESC nuclear extract. A single empty lane was removed from the blot. **D.** HPL-Chromatograms of the purified FW and RV DNA obtained from solid phase DNA synthesis showing the purity of the employed strands. **E.** The mass spectra of the DNA before (blue) and after (red) NE incubation as determined by MALDI MS showing the expected m/z before and after NE incubation. Major alterations of the DNA, like degradation or strand breaks, can be excluded. **F.** Synthetic DNA-strands that were used for DNA pull-downs were compared without (w/o) and with nuclear extract (NE) treatment (2 h, 4 °C) to proof the stability of the indicated modifications. The quantification of the nucleoside content was carried out by LC-MS/MS. For this, the DNA was digested to the nucleoside level and spiked with a specific amount of the following internal standards for precise quantification: [$^{15}N_2$]-dC, [D_2,$^{15}N_2$]-hmC, [$^{15}N_2$]-fC, [$^{15}N_2$]-caC and [D_3]-dT. The absolute amount (pmol) of each nucleoside was calculated by calibration curves (not shown). Depicted are ratios of the modified nucleoside (pmol) to deoxy-thymidine (dT; pmol), which were obtained from three independent measurements. The relative standard deviation was between 0.3-6.2%. No or only marginal loss of the modified nucleosides was observed.



F

	NE incubation	dC/dT	hmC/dT	β-fC/dT	α-fC/dT	caC/dT
dC	w/o	0.9340				
dC	2 hr	0.9172				
hmC	w/o	0.0738	0.7258	0.0009		
hmC	2 hr	0.0745	0.7102	0.0009		
fC	w/o	0.4251		0.2818	0.1794	0.0035
fC	2 hr	0.4172		0.2950	0.1811	0.0034
caC	w/o	0.3780				0.5729
caC	2 hr	0.3755				0.5414

Figure S3: Modification and cell-type specific GO term enrichment analysis. Related to Figure 1, 2, 3 and 4. A. Shows GO term enrichment and enriched domains for the different baits (C, mC, hmC, fC and caC) in mESC, NPCs and adult mouse brain. **B.** Venn diagrams showing the overlap between C, mC and hmC readers within each cell type and the overlap between C, mC and hmC readers between mESCs, NPCs and adult mouse brain.

A

mESC

	GO biological process	-log(Benj.Hoch.FDR)
C	RNA biosynthetic process	7.61
	cellular biosynthetic process	6.65
	biosynthetic process	6.48
	DNA recombination	5.14
	DNA metabolic process	5.06
	nucleobase-containing compound metabolic process	4.49
	transcription, DNA-dependent	4.48
	nucleic acid metabolic process	4.42
	cellular macromolecule biosynthetic process	4.30
	macromolecule biosynthetic process	4.29
	cellular nitrogen compound metabolic process	4.28
	nitrogen compound metabolic process	4.21
	regulation of macromolecule biosynthetic process	3.85
	regulation of cellular macromolecule biosynthetic process	3.85
	transcription initiation, DNA-dependent	3.82
	regulation of transcription, DNA-dependent	3.80
	regulation of cellular biosynthetic process	3.77
	regulation of biosynthetic process	3.75
	DNA repair	3.71
	regulation of gene expression	3.69
	regulation of nucleobase-containing compound metabolic process	3.53
	regulation of nitrogen compound metabolic process	3.53
	regulation of RNA metabolic process	3.33
	regulation of primary metabolic process	3.32
	regulation of macromolecule metabolic process	3.31
	response to DNA damage stimulus	3.26
	regulation of cellular metabolic process	3.12
	regulation of metabolic process	3.05
	cellular response to stress	2.97
	RNA metabolic process	2.72
	cellular macromolecule metabolic process	2.58
	response to stress	2.57
	macromolecule metabolic process	2.37
	eye morphogenesis	2.05
	cellular metabolic process	2.04
	primary metabolic process	1.86
	metabolic process	1.72
mC	transcription, DNA-dependent	2.84
	RNA biosynthetic process	2.81
	regulation of cellular biosynthetic process	1.76
	macromolecule biosynthetic process	1.76
	one-carbon metabolic process	1.74
	regulation of biosynthetic process	1.73
	regulation of macromolecule biosynthetic process	1.73
	regulation of nitrogen compound metabolic process	1.72
	regulation of RNA metabolic process	1.71
	nucleic acid metabolic process	1.70
	regulation of macromolecule metabolic process	1.70
fC	DNA repair	2.71
	response to DNA damage stimulus	2.60
	DNA metabolic process	2.59
	cellular response to stress	1.79

	Cellular Compartment	-log(Benj.Hoch.FDR)
C	transcription factor TFIID complex	2.17
	TFIID complex	7.65
	MLL1 complex	6.06
	histone methyltransferase complex	4.20
	methyltransferase complex	4.37
	transcription factor complex	3.24
	nucleoplasm part	1.73

	Domains	-log(Benj.Hoch.FDR)
C	BRLZ	7.40
	bZIP_1	5.57
mC	ZnF_C2H2	2.45
	MBD	2.24

NPC

	GO biological process	-log(Benj.Hoch.FDR)
C	regulation of nitrogen compound metabolic process	5.04
	regulation of biosynthetic process	5.03
	regulation of nucleobase-containing compound metabolic process	5.01
	regulation of RNA metabolic process	5.01
	transcription, DNA-dependent	5.00
	regulation of cellular biosynthetic process	4.94
	RNA biosynthetic process	4.93
	regulation of macromolecule biosynthetic process	4.90
	regulation of cellular macromolecule biosynthetic process	4.88
	regulation of transcription, DNA-dependent	4.85
	regulation of gene expression	4.81
	cell differentiation	2.60
	positive regulation of transcription from RNA polymerase II promoter	2.51
	cellular response to metal ion	2.49
	regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	2.48
	cellular response to inorganic substance	2.47
	positive regulation of biosynthetic process	2.46
	cellular response to chemical stimulus	2.45
	cellular response to calcium ion	2.45
	regulation of transcription from RNA polymerase II promoter	2.44
	gland development	2.39
	cellular developmental process	2.37
	positive regulation of RNA metabolic process	2.35
	adipose tissue development	2.33
	positive regulation of metabolic process	2.26
	positive regulation of nucleobase-containing compound metabolic process	2.21
	positive regulation of cellular biosynthetic process	2.16
	positive regulation of nitrogen compound metabolic process	2.14
	positive regulation of transcription, DNA-dependent	1.99
	positive regulation of macromolecule metabolic process	1.98
	positive regulation of gene expression	1.92
	response to calcium ion	1.88
	positive regulation of macromolecule biosynthetic process	1.85
	regulation of developmental process	1.83
	response to inorganic substance	1.79
	anatomical structure development	1.78
	regulation of cell differentiation	1.74
	response to metal ion	1.73
mC	positive regulation of macromolecule biosynthetic process	2.25
	positive regulation of cellular biosynthetic process	2.20
	positive regulation of biosynthetic process	2.17
	positive regulation of RNA metabolic process	2.14
	regulation of transcription from RNA polymerase II promoter	2.14
	positive regulation of gene expression	2.11
	positive regulation of cellular metabolic process	2.10
	dorsal spinal cord development	2.07
	positive regulation of metabolic process	2.04
	positive regulation of nucleobase-containing compound metabolic process	2.00
	positive regulation of cellular process	1.95
	positive regulation of nitrogen compound metabolic process	1.94
	positive regulation of biological process	1.86
hmC	DNA metabolic process	4.97
	DNA repair	3.91
	response to DNA damage stimulus	3.61
	base-excision repair	3.32
	cellular response to stress	2.96

	Cellular Compartment	-log(Benj.Hoch.FDR)
C	histone methyltransferase complex	2.02
	intracellular membrane-bounded organelle	1.86

	Domains	-log(Benj.Hoch.FDR)
C	bZIP_1	14.42
	zf-CXXC	7.91
	MBT	1.97
	Jun	1.77
mC	RFX_DNA_binding	1.98
	MBD	1.81
	Homeobox	1.80

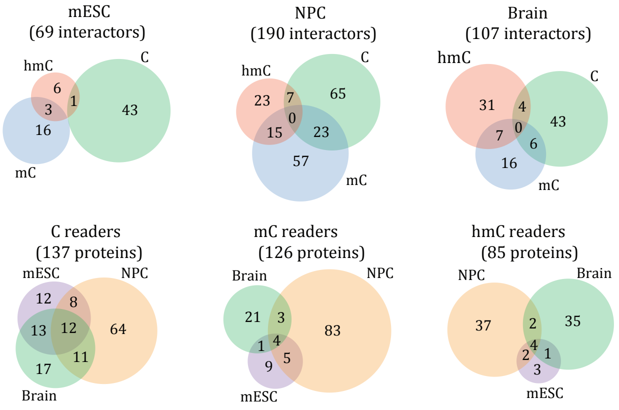
Brain

	GO biological process	-log(Benj.Hoch.FDR)
C	transcription, DNA-dependent	5.32
	RNA biosynthetic process	5.31
	regulation of biosynthetic process	5.24
	regulation of transcription, DNA-dependent	5.18
	regulation of macromolecule biosynthetic process	5.02
	regulation of nucleobase-containing compound metabolic process	5.00
	regulation of nitrogen compound metabolic process	4.99
	regulation of cellular macromolecule biosynthetic process	4.99
	regulation of gene expression	4.95
	regulation of cellular biosynthetic process	4.95
	regulation of RNA metabolic process	4.94
	regulation of primary metabolic process	4.70
	regulation of macromolecule metabolic process	4.65
	regulation of metabolic process	4.33
	DNA recombination	4.18
	regulation of cellular metabolic process	4.15
	cellular macromolecule biosynthetic process	4.08
	biosynthetic process	3.59
	response to DNA damage stimulus	2.75
	cellular response to stimulus	2.61
	cellular response to stress	2.58
	regulation of transcription from RNA polymerase II promoter	2.58
	DNA repair	2.47
	anatomical structure morphogenesis	2.46
	cellular response to calcium ion	2.15
	histone H4 acetylation	2.13
	eye morphogenesis	2.06
	response to calcium ion	2.03
	cellular response to metal ion	2.01
	cellular response to inorganic substance	2.00
	response to metal ion	1.71
	DNA metabolic process	1.70
hmC	DNA metabolic process	4.57
	DNA replication	3.54
	base-excision repair	1.78

	Cellular Compartment	-log(Benj.Hoch.FDR)
C	histone methyltransferase complex	5.42
hmC	Brd4-Rfc complex	6.11
	DNA replication factor C complex	3.93

	Domains	-log(Benj.Hoch.FDR)
C	bZIP_1	8.60
	Jun	1.93
mC	MBD	2.70
hmC	Rep_fac_C	3.61
	AAA	2.11

B



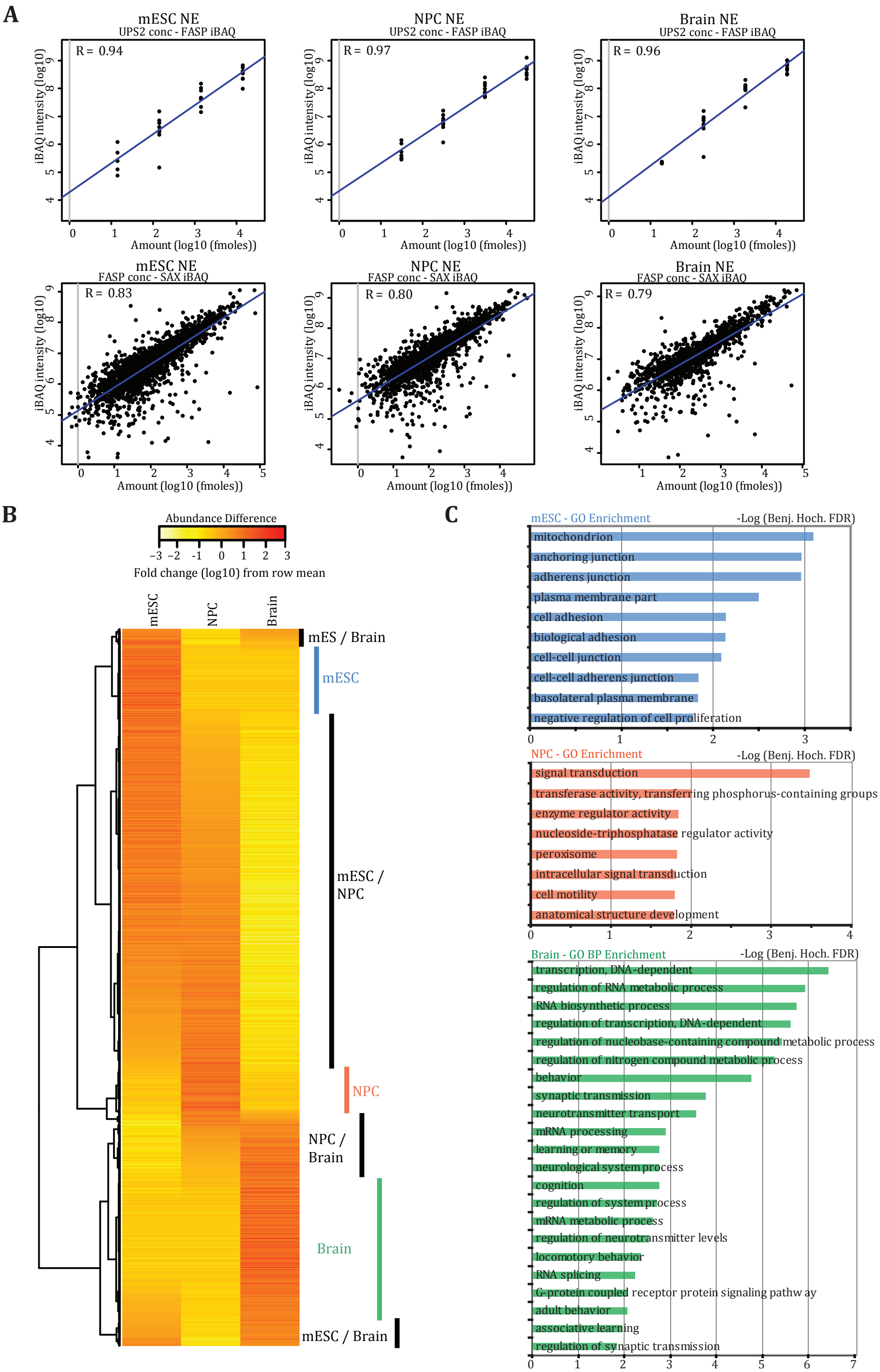

A
mESC NE
UPS2 conc - FASP iBAQ
R = 0.94
NPC NE
UPS2 conc - FASP iBAQ
R = 0.97
Brain NE
UPS2 conc - FASP iBAQ
R = 0.96
iBAQ intensity (log10)
4 5 6 7 8 9
0 1 2 3 4
Amount (log10 (fmoles))
mESC NE
FASP conc - SAX iBAQ
R = 0.83
NPC NE
FASP conc - SAX iBAQ
R = 0.80
Brain NE
FASP conc - SAX iBAQ
R = 0.79
0 1 2 3 4 5
Amount (log10 (fmoles))
B
Abundance Difference
−3 −2 −1 0 1 2 3
Fold change (log10) from row mean
mESC
NPC
Brain
mES / Brain
mESC
mESC / NPC
NPC
NPC / Brain
Brain
mESC / Brain
C
mESC - GO Enrichment
-Log (Benj. Hoch. FDR)
mitochondrion
anchoring junction
adherens junction
plasma membrane part
cell adhesion
biological adhesion
cell-cell junction
cell-cell adherens junction
basolateral plasma membrane
negative regulation of cell proliferation
0 1 2 3
NPC - GO Enrichment
-Log (Benj. Hoch. FDR)
signal transduction
transferase activity, transferring phosphorus-containing groups
enzyme regulator activity
nucleoside-triphosphatase regulator activity
peroxisome
intracellular signal transduction
cell motility
anatomical structure development
0 1 2 3 4
Brain - GO BP Enrichment
-Log (Benj. Hoch. FDR)
transcription, DNA-dependent
regulation of RNA metabolic process
RNA biosynthetic process
regulation of transcription, DNA-dependent
regulation of nucleobase-containing compound metabolic process
regulation of nitrogen compound metabolic process
behavior
synaptic transmission
neurotransmitter transport
mRNA processing
learning or memory
neurological system process
cognition
regulation of system process
mRNA metabolic process
regulation of neurotransmitter levels
locomotory behavior
RNA splicing
G-protein coupled receptor protein signaling pathw ay
adult behavior
associative learning
regulation of synaptic transmission
0 1 2 3 4 5 6 7

Figure S4: iBAQ analyses of mESC, NPC and adult mouse brain nuclear extracts. Related to Figure 5. A. Standard and linear regression curves for the iBAQ of protein abundance in the different nuclear extracts that were used for the DNA pull-downs. **B.** Correlation based clustering of proteins that show at least a 10 fold change in protein levels. Yellow is low abundance, red is high. **C.** GO term enrichment for mESC specific proteins (indicated in blue in Fig. S4B), NPC (indicated in red in Fig. S4B) and adult mouse brain (indicated in green in Fig. S4).

SUPPLEMENTARY MATERIALS AND METHODS

SILAC labeling of ES cells.

IB10 murine Embryonic stem cells were cultured feeder-free on gelatin coated dishes in medium consisting of 500 ml SILAC Dulbecco's Modified Eagle Medium without arginine, lysine and glutamine (PAA, E15-086), supplemented with 15% MESC serum substitute (Thermo Scientific), Glutamine, Penicillin/Streptomycin, 1x Non-essential amino acids, sodium pyruvate, 73 μg/ml L-Lysine (light/K^0 (Sigma, A6969), medium/K^4 (Sigma, 616192 or Silantes, 211103912) or heavy/K^8 (Sigma, 608041 or Silantes, 211603902)) and 29.4 μg/ml arginine (light/R^0 (Sigma, A6969), medium/R^6 (Sigma, 643440 or Silantes, 201203902) or heavy/R^{10} (Sigma, 608033 or Silantes, 201603902)), LIF (1000 U/ml), β-mercaptoethanol and 2i compounds (CHIR99021 and PD0325901, 3 and 1 μM respectively). Cells were cultured in SILAC medium until labeling efficiency exceeded 95% after which cells were expanded and harvested to generate nuclear extracts.

NPC culturing

Neuronal progenitor cells were kindly provided by Dr. N.S. Outchkourov. They were cultured in medium consisting of NSA MEM (Euromed EVM0883LD), 1% glutamine, 1x N2 supplement, 10 ng/mL bFGF (RD systems 233-F3) and 10 ng/mL EGF (235-E9) on gelatin-coated dishes. Cells were detached from culture plates using accutase. Nuclear extracts were made as described below.

Mice brain nuclear extracts

Nuclei from adult mouse brain were purified by centrifugation through a sucrose cushion following homogenization, modified from [1]. Then nuclei were lysed as described below.

Nuclear extract preparation

This protocol is based essentially on Dignam *et al.* [2]. Briefly, cells were trypsinized and washed two times with PBS. Using a hypotonic buffer, the cells were swollen, after which the cells were lysed by dounce homogenizing in the presence of 0.15% NP40 and complete protease inhibitors. After centrifugation, the pellet consisting of nuclei was lysed by 90 minutes incubation in 2 volumes of nuclear lysis buffer (420 mM NaCl, 20 mM Hepes pH 7.9, 20 % v/v glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1 % NP40, complete protease inhibitor w/o EDTA (Roche) and 0.5 mM DTT). After centrifugation, the supernatant containing the soluble nuclear extract was aliquoted and snap frozen until further usage. Protein concentrations of the nuclear extracts were determined using the Biorad Protein assay.

DNA Synthesis

The synthesis of the oligonucleotides for DNA pull-downs for analysis by mass spectrometry or western blot (see Table S4), was performed on an ABI 394 DNA/RNA Synthesizer (Applied Biosystems) using typical reagent concentrations (activator: 0.25 M benzylthiotetrazole in MeCN (10 ppm H_2O), detritylation: 3% dichloroacetic acid in CH_2Cl_2, oxidation: 25 mM I_2 in MeCN/H_2O/2,6-lutidine (11/5/1), capping: Ac_2O/2,6-lutidine/MeCN (30 ppm H_2O) (20/30/50) and 20% N-methylmidazole in MeCN (10 ppm H_2O). The oligonucleotide syntheses were performed on 200 nmol low-volume polystyrene carriers using 0.1 M DNA CE-phosphoramidites: A (Bz-dA), C (Bz-dC), G (iBu-dG), T, mC (Bz-mC) obtained from Glen Research or Link Technologies. hmC, fC and caC phosphoramidites were synthesized according to literature [3] and incorporated into DNA using the standard protocol. Benzylthiotetrazole was prepared according to literature [4]. The coupling times for the modified bases were increased to 3 min to ensure maximum coupling efficiency.

The mC and the unmodified strands were treated with ethanolic ammonia for cleavage of the carrier and removal of the permanent protecting groups. hmC, fC and caC containing DNA was cleaved and deprotected using 0.4 M NaOH in MeOH/H_2O 4:1 for 18 h at room temperature. After addition of 600 µL triethylammonium acetate (1 M) and centrifugation, the supernatant was concentrated to 30% of the original volume in a speedvac. Analysis and purification was performed on a Waters HPLC system (Waters Alliance 2695 with PDA 2996, preparative HPLC: 1525EF with 2482 UV detector) with VP 250/10 Nucleosil 100-7 C 18 columns from Macherey Nagel using a gradient of 0.1 M triethylamine/acetic acid in water and 80% acetonitrile. The quality of the strands was determined by MALDI-MS. The forward and reverse oligo's were combined and annealed in 10 mM Tris pH 8; 50 mM NaCl and 1 mM EDTA. Biotin-14-ATP was used to fill in the TT-overhang using Klenow exo-, followed by purification of the DNA on sephadex-G50 columns.

DNA pull-downs

For each DNA pull-down, 10 µg of DNA (See Table S4) was immobilized on 75 µL of Dynabeads MyOne C1 (Invitrogen) by incubating for 1 hour at room temperature in a total volume of 350 µl of DNA binding buffer (1 M NaCl, 10 mM Tris-HCl pH 8, 1 mM EDTA pH 8 and 0.05% NP40). Coupling of the DNA to the beads was always verified by agarose gel electrophoresis. Beads containing immobilized DNA were then incubated with 400 µg of nuclear extract in a total volume of 600 µL of protein binding buffer (50 mM Tris-HCl pH 8, 150 mM NaCl, 1 mM DTT, 0.25% NP40 and complete protease inhibitors (Roche, EDTA-free)) in the presence of 10 µg poly-dAdT for 2 hours at 4^oC. Baits were then washed three times with 0.5 ml of protein binding buffer after which beads containing different DNA modifications and different SILAC labels were combined and loaded on 4-12% NuPage gradient gels (Invitrogen) (for example, C-beads with light extract were combined with mC beads that were incubated with heavy extract; forward pull-down). For the label-free analysis, three separate DNA pull-downs with every bait were performed and each of these was loaded on gel separately. For western blot validation using endogenous antibodies, protein amounts were scaled down by a factor of four.

In gel digestion

Samples were analyzed on 4-12% precast NuPage gels (Invitrogen) and subsequently stained using colloidal blue staining (Invitrogen). Each lane was cut into 8-12 gel slices and each of these slices was subjected to in-gel trypsin digestion overnight. Tryptic peptides were desalted on Stage-tips [5].

Mass spectrometry

Peptides were separated on an EASY-nLC (Proxeon) connected online to an LTQ-Orbitrap-Velos mass spectrometer. Spectra were recorded in CID mode. A gradient of organic solvent (5-30% acetonitrile) was applied (120 minutes) and the top 15 most abundant peptides were fragmented for MS/MS, using an exclusion list of 500 proteins for 45 seconds.

Data analysis

Raw data were analyzed using Maxquant version 1.2.2.5 and the integrated Andromeda search engine against protein database ipi.MOUSE.v3.68. Using Perseus, data was filtered for contaminants, reverse hits, number of peptides (>1) and unique peptides (>0). Ratios were logarithmitized (log2) and groups (consisting of forward and reverse) were defined. Proteins were filtered to have at least 2 valid values in one of the groups and missing values were imputed based on a normal distribution (width=0.2 and shift=0), after which Significance B was calculated (Benj.Hoch.FDR=0.05). Scatterplots were made using R. Proteins were defined to be significant when both forward and reverse significance $p<0.05$ and minimal ratios were >2 in both experiments. The H/L ratios shown in Figure2A-C were calculated using the formula (log(forward ratio) – log(reverse ratio))/2.

Label-free quantification

LFQ values, based on the summed measured intensities of all tryptic peptides of a single protein, allow for comparing the relative abundance of a protein in different pull-downs. Changes in the LFQ intensity of a protein between pull-downs with different DNA modifications indicate preferential binding of that protein to one modification over another. Raw data were analyzed using Maxquant version 1.2.2.5 and protein database ipi.MOUSE.v3.68.fasta. Settings that were different from SILAC analyses were: multiplicity set at 1 and the options for ‘label-free quantification’ and ‘match between runs’ were selected. Using Perseus, data were filtered for contaminants, reverse hits, number of peptides (>1) and unique peptides (>0). LFQ intensities were logarithmitized (log2). After defining each triplicate as a group, proteins were filtered to have at least 3 values in a single group, assuming that when a protein binds specifically to one modification, it may only be identified in the three pull-downs with that modification. The missing values were imputed using a normal distribution (with=0.3, shift=1.8). Groups were defined and the significant outliers were calculated using ANOVA (FDR=0.025, S_0=2 for NPC and S_0=0 for brain). Correlation based clustering was done in R for the ANOVA-outliers only, using LFQ-values which had been normalized by row-mean-subtraction.

Purification of GFP-fusion proteins for EMSA

HEK 293T cells were transfected with expression constructs encoding for GFP-Tdg or GFP-Dnmt1. 48 hours after transfection, cells were lysed 30 min on ice in Lysis-Buffer (50 mM NaH_2PO_4, 150 mM NaCl, 10 mM Imidazole, 0.5 mM EDTA, 0.5% Tween, 1 g/l DNaseI, 2 mM $MgCl_2$, 0.5 mM $CaCl_2$, 1 mM PMSF, 1x Protease-Inhibitor-Mix M (SERVA Electrophoresis GmbH)). The lysate was cleared by centrifugation (14 000 rpm, 10 min, 4°C) and followed by incubation of the supernatant with equilibrated Ni-NTA beads (Qiagen) in IP-buffer (50 mM NaH_2PO_4, 150 mM NaCl, 10 mM Imidazole, 0.5 mM EDTA, 0.05% Tween). After centrifugation (2200 rpm, 2 min) the supernatant was added to equilibrated GBP-Ni-NTA beads (Chromotek) in IP-buffer and rotated for 2 hours at 4°C. After washing three times with Washing-Buffer (50 mM NaH_2PO_4, 300 mM NaCl, 10 mM Imidazole, 0.1% Tween), the GFP-fusion proteins were eluted with 50 mM NaH_2PO_4, 150 mM NaCl, 250 mM Imidazole, 0.05% Tween. The elution buffer was exchanged to 20 mM TrisHCl, pH 7.5, 150 mM NaCl, 0.5 mM EDTA, 1 mM DTT for EMSA reactions. The glycosylase activity of the purified Tdg was tested on T/G mismatch containing DNA (data not shown).

Electrophoretic mobility shift assays of fluorescent DNA oligonucleotides with GFP-fusion proteins

GFP-Tdg and GFP-Dnmt1 at decreasing concentrations (200 nM, 150 nM, 100 nM, 50 nM, 25 nM 12.5 nM and 6.25 nM) were incubated for 30 min on ice with a 1:1 mixture of two distinctly labelled fluorescent 42mers (See Table S4, MWG-Eurofins, 250 nM each) containing a central CG site. The ATTO647N-labelled oligonucleotide contains only canonical bases whereas the ATTO550-labelled DNA bears different cytosine modifications (C, mC, hmC, fC and caC) or an abasic site at the CG position on both strands. Samples were run on a 6% non-denaturating polyacrylamide gel (pre-run 1 hour with 0.5x TBE) at 4°C. Oligonucleotide- and GFP-fluorescence was detected by the Typhoon Scanner (GE Healthcare). Quantifications were done with ImageJ.

DNA purification and analysis after NE incubation

DNA pull-downs were performed as described above, but all amounts were scaled up 3 times. As a control, all baits were also incubated in buffer plus poly-dAdT without nuclear extract for 2 hrs at 4°C. The beads were washed 3x using 1 ml of incubation buffer and 1x using 1M NaCl, 10mM Tris-HCl pH8, 1mM EDTA and 0.05% NP40, to reduce contamination with DNA from the nuclear extracts. Beads were then resuspended in 200 uL incubation buffer and DNA was purified using phenol/chloroform extraction from the beads. The DNA-strands were finally dissolved in milliQ, enzymatically hydrolyzed to nucleosides and analysed in triplicate (15 pmol each) by MALDI-MS or LC-MS/MS.

GFP pull-downs

HeLa wild-type cells and a BAC-GFP transgenic cell line (WDR76) were cultured in SILAC medium for eight cell doublings, after which cells were expanded and nuclear extracts were made. For each pull-down 20 µL of GFP-trap slurry (50% v/v; Chromotek) was washed and incubated for 90 minutes at 4°C with 1 mg of nuclear extract (WT L, WT H, GFP L and GFP H) in a total volume of 400 µL incubation buffer (300 mM NaCl, 20 mM Hepes KOH pH 7.9, 20% v/v glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% NP40, complete protease inhibitor w/o EDTA (Roche) and 0.5 mM DTT) in the presence of 2 µL ethidium bromide (10mg/ml, final concentration 50 µg/ml). Beads were then washed two times with this incubation buffer, twice with PBS + 0.5% NP40 and two times with PBS only. During the last wash, beads of light control and heavy GFP pull-down were mixed and vice versa. Bound proteins were then subjected to on-bead trypsin digestion [6] and significant proteins were determined as described for the SILAC DNA pull-downs. For the GFP-Tdg pull-down, mESC were cultured in normal mESC medium and a transient transfection with the GFP-Tdg plasmid (15 µg/15cm dish) using PEI (ratio DNA:PEI = 1:3) was performed. GFP-Tdg was purified in a label-free method, thus 3 pull-downs were performed using GFP-trap beads and as a control the same extract was incubated in triplicate with control blocked agarose beads (Chromotek). For each pull-down 20 µL of bead slurry (50% v/v) was washed and incubated for 90 minutes at 4^{o}C with 1 mg of the nuclear extract in a total volume of 400 µL incubation buffer (150 mM NaCl, 50 mM Tris-HCl pH 8, 1 mM DTT, 0.25% NP40 and complete protease inhibitor w/o EDTA (Roche)) to mimic the conditions of the DNA pull-downs as close as possible in the presence of 50 µg/ml of ethidium bromide. Beads were then washed two times with 0.5 ml of incubation buffer, twice with PBS + 0.5% NP40 and two times with PBS only, after which bound proteins were on-bead digested. The Tdg-GFP purification was analyzed using a permutation-based *t*-test (FDR=0.05 & S0=3) to determine significant interactors.

Recombinant protein expression/ DNA pull-downs

Klf4(aa396-483), KDM2B(aa606-647), Cxxc5(aa234-293), MBD3(aa1-77) and Rfx5(aa85-173) were cloned into PRP256NB vector, containing a GST with a C-terminal multiple cloning site. Uhrf2 (aa416-626) GST fusion was kindly provided by Dr. Jiemin Wong. hMBD2b-GST was provided by Stefanie Bartels.

Protein expression was performed in E. coli BL21-DE3 Codon+ by growing them at 37^{o}C until OD_{600} of 0.5, after which expression was induced using 1 mM IPTG and culturing for 3 additional hours at 25^{o}C. Cells were lysed in 50 mM Tris-HCl pH 8.0, 20% sucrose, 1 mM EDTA, 0.5 mM PMSF, 1 mM DTT, 1 µg/ml aprotinin using lysozyme and Triton-X100 and repeated freeze-thawing. Bacterial debris was removed by ultracentrifugation.

DNA pull-downs were performed using 2.5 µg DNA coupled to 16.75 µL MyOne beads and 5 µl of bacterial lysate/ nuclear extract in 250 µl total volume (50 mM Tris-HCl pH 8, 150 mM NaCl, 1 mM DTT, 0.25% NP40 and complete protease inhibitors (Roche, EDTA-free)) in the presence of 2.5 µg poly-dAdT. After 3 times of washing with 0.5 ml of this buffer, beads were boiled in sample buffer. 5% of the input material and 100% of the bound material was loaded on gel for western blot analyses.

Western blot

Gels were blotted onto nitrocellulose membranes. Blots were blocked using 5% skimmed milk in TBST. Used antibodies are: MouseαMBD3 (IBL, 3A3), GoatαMBD2 (Everest Biotech, EB07538), RabbitαRBBP5 (Bethyl, BL766), GoatαJun-C (SantaCruz), RabbitαDNMT1 (Abcam, ab13537), RabbitαCarf (Abcam, ab140519), RabbitαGST (Santa Cruz, SC-138), RabbitαGFP (home made), DonkeyαmouseHRP and DonkeyαRabbitHRP.

NMR spectroscopy based interaction study of Rfx5 and mC DNA

The winged-helix (WH) domain of human Rfx5 (residues 85-173, plus 18 additional residues at the N-terminus) was expressed as a GST-fusion in BL21-DE3 Codon$^+$ bacterial strains at 25°C in M9 minimal medium with $^{15}NH_4Cl$ and/or ^{13}C-glucose. The protein was purified by binding to a Glutathione agarose (GA) column (Sigma) and eluted with 50 mM reduced glutathione (Sigma). After thrombin digestion, Rfx5-WH was purified over a Sephadex-75 (HiLoad 16/60) column in buffer A (50 mM KPi pH 7, 100 mM KCl, 5 mM DTT, 0.5 mM PMSF and protease inhibitors). NMR samples used for backbone assignment contained ca. 0.3 mM WH domain in 90/10% H_2O/D_2O in buffer A. NMR spectra (HNCACB, CBCACONH, HNCA, and HNCO) were recorded at 298K on a 600 or 750 MHz Bruker Avance II spectrometer, processed using the NMRPipe package [7], and analyzed using CcpNmr Analysis [8]. Backbone assignments were obtained for 90 out of 106 residues in the Rfx5-WH construct.

Interaction study with mC DNA was done using an 18bp DNA fragment (see Table S4; (Biolegio)) carrying a single mC on each strand. Annealed DNA oligos were lyophilized and dissolved in buffer A to a stock concentration of 620 µM. The Rfx5-WH domain (103 µM) was titrated with mC DNA, and after each addition (11 points in total) the ^1H-^{15}N HSQC spectrum of Rfx5-WH as recorded (298K / 600 MHz Bruker Avance II). Since the DNA sequence used is not palindromic, the two mC may be inequivalent in their capability to bind Rfx5. At high DNA:Rfx ratios, several peaks appear split in two in a roughly 1:1 ratio, suggesting that although the Rfx5-WH domain senses the distinct DNA sequence context of the two mC sites, it recognizes both with similar affinities (data not shown). Although a few residues showed non-linear titration profiles, most peak displacements were linear. For further analysis, the binding sites were treated as being independent, resulting an apparent dissociation constant for the Rfx5-WH – mC interaction.

Titration data were fitted using using MatLAB scripts (MATLAB version 7.13.0, The MathWorks Inc., 2011) using the fast-exchange assumption for residues with observed chemical shift perturbations between 10 and 30 Hz (fast-exchange regime; 15 residues) in a global fit. The error bars for the observed peak position was set to 1.2 Hz. The overall reduced chi-squared for the fit was 2.17. The error in the fitted K_d was estimated using 1000 MonteCarlo simulations resulting in an average of 3.2 ± 0.9 µM. The range of acceptable fits was examined using F-statistics from a grid search, resulting in 95% probability limits of $10\ \text{nM} < K_d < 16\ \mu\text{M}$.

A homology model of Rfx5-WH domain was constructed on the basis of the DNA-bound crystal structure of the Rfx1 winged helix domain (PDB-id: 1DP7; 35% sequence identity) using the SwissModel server [9]. The model was validated against the predicted backbone dihedral angles from the observed backbone chemical shifts using TALOS+ [10]. The model of mC bound to the putative binding pocket was constructed in PyMol by superimposing the mC DNA from the UHRF1-mDNA crystal structure (PDB-id 3CLZ) onto the Rfx1-bound DNA, such that the binding pocket and mC are aligned. To achieve a proper fit, the mC base was set to a syn-conformation. The side chain orientations of K110 and Y161 were adjusted manually to minimize clashes.

In silico analysis of Klf4 ChIP-seq profile and bisulfite sequencing data in mESCs cells and NPCs

Klf4 binding data (ChIP-seq) was taken from [11] (GSM288354), and DNA methylation data (whole-genome bisulfite sequencing) was taken from [12] (GSE30202). Annotated Klf4 peak centers (mESC) were extended with 50 bp on both sides to obtain 100-bp Klf4 binding regions. The mean CpG methylation of each 100-bp region was calculated for mESCs and NPCs and plotted as a scatterplot (Fig. S1B). For each quadrant of this scatterplot, the genomic distribution of the 100-bp Klf4 binding regions was calculated and plotted as a Venn diagram (Fig. S1C). Promoters were defined as -/+ 1 kb upstream and downstream from transcription start sites of the RefSeq mm9 annotation. The DNA sequences of the 100-bp Klf4 binding regions were used to search for the GGCGTG motif, and the CpG methylation within these motifs was calculated. The obtained distribution was plotted as a histogram (Suppl. Fig. S1D). Analyses were done using Python, Perl and R.

iBAQ

iBAQ was performed essentially as described in [13]. 3.3 μg of UPS2 standard (Sigma) was added to 10 μg of nuclear extract, which was digested using the FASP protocol [14]. In addition, 100 μg of NE was digested using FASP after which the peptides were separated into 8 fractions using SAX. Each of these samples was measured during a 4 hour gradient of LC-MSMS. A linear fit was made for the known amounts of the UPS2 standard and the measured iBAQ intensities in the 10 μg sample. Using this curve, iBAQ values of all other identified proteins in the 10μg sample were converted to amounts. A linear fit was again made using these amounts and the iBAQ values in the eight SAX fractions, which were used to extrapolate absolute protein amounts of all identified proteins in these samples (Fig S4A).

Cell culture and transfection experiments

The mammalian GFP-Tet1cd expression vector was generated by PCR amplification of mouse (E14) cDNA encoding the catalytic domain of Tet1 (amino acids 1365 to 2057) and N-terminal GFP fusion. HEK-293 cells were grown at 37°C and 5% CO_2 in Dulbecco's Modified Eagle Medium (DMEM, Invitrogen 41966-029) supplemented with 10% fetal bovine serum and 1% Penicillin-Streptomycin. Cells were passaged at 80% confluency. All transfections were performed using the jetPRIME system (PEQLAB Biotechnologie GmbH) according to the manufacturer's instructions. HEK-293 cells were seeded 24 h prior to transfection at a density of $3x10^6$ cells per 75 cm^2 flask and incubated in 10 mL of medium at 37°C and 5% CO_2 for 24 hrs. Cotransfection of GFP-Tet1cd plasmid (6 μg) either with mouse Uhrf2-GFP plasmid DNA (6 μg) [15] or 6 μg of pCMV6-Cdk5Rap1-v2 (Origene RG216600) as an unrelated control was carried out in a 75 cm^2 flask containing 10 mL of fresh medium. The transfection solution (500 μL of jetPRIME buffer, 12 μg of plasmid DNA and 24 μL of jetPRIME reagent) was added to the medium and the cells were incubated at 37°C and 5% CO_2 for 48 hrs. After removal of the medium the cells were washed once with PBS and then lysed for DNA extraction according to [16]. The DNA was enzymatically digested to the nucleosides and subsequently analyzed by LC-ESI-MS/MS.

LC-MS/MS analysis of genomic DNA and synthetic DNA

The following LC-MS/MS method for the quantification of DNA-nucleosides is based on a further development of our precise and sensitive isotope dilution method ([17] and manuscript in preparation). In the following we shortly summarize the parameters of the method. Genomic or synthetic DNA was enzymatically digested to the nucleoside level. A specific amount of internal standards with a stable isotope label were spiked to the digestion mixture for precise quantification. The following labeled nucleosides were used as internal standards: [$^{15}N_2$]-dC, [D_3]-mC, [D_2,$^{15}N_2$]-hmC, [$^{15}N_2$]-fC, [$^{15}N_2$]-caC and [D_3]-dT. In case of genomic DNA the dC- or dG-content was determined by LC-UV-Detection.

LC-MS/MS analysis was performed on an Agilent 6490 triple quadrupol mass spectrometer coupled to an Agilent 1290 UHPLC system. The general source-dependent parameters were as follows: Gas Temp 50°C, Gas Flow 15 L/min, Nebulizer 30 psi, Sheath Gas Heater 300°C, Sheath Gas Flow 11 L/min, Capillary Voltage 2500 V and Nozzle Voltage 500 V. For compound-dependent parameters used for genomic DNA see Table S5, for compound-dependent parameters used for synthetic DNA see Table S6. The transitions of the nucleosides were analyzed in the positive ion selected reaction monitoring mode (SRM) operating MS1 and MS2 under unit mass resolution conditions.

For the analysis a C8 column from Agilent was used (1.8 µm, 2.1 mm x 150 mm). The compounds were separated by a gradient using water and acetonitril with 0.0075% formic acid. The column temperature was maintained at 30 °C. The flow rate was 400 µL min^{-1}, and the injection volume amounted to 29 µL. The effluent up to 1.5 min (total run time of 12 min) was diverted to waste by a Valco valve in order to protect the mass spectrometer.

Validation of quantification method for genomic DNA modifications

In accordance with the FDA guidance for bioanalytical method validation, linearity, precision, and accuracy (i.e., recovery determined from spiked matrix samples) of the established method were investigated. Validation for the established LC-MS/MS quantification method was based on five different series (i.e., calibration functions and quality control samples) accomplished on different days. Calibration standards were analyzed at least in triplicates. Quality control samples to evaluate accuracy, intra- and inter-batch (see intra- and inter-assay) precision were determined using a biological sample with internal standards. Furthermore, each validation experiment was complemented by matrix blanks (analyzed in triplicates) to ensure selectivity and specificity of the method. Additionally, acceptable accuracy (80–120%) as well as precision (<20% RSD) was required. Linear regression was applied to obtain calibration curves. Therefore, the peak area ratio (y) of the unlabeled nucleoside to the internal standard vs. the concentration ratio of the unlabeled nucleoside to the internal standard (x) was plotted. Calibration functions were calculated without weighting. Long-term stability of aqueous solutions of the labeled and unlabeled nucleosides at a storage temperature of −20 °C was investigated over two months including several freeze and thaw cycles by analyzing the MS/MS-responses with each batch. Short-term stability at room temperature was studied in overnight experiments. In this process, the results of quantification by LC-ESI-MS/MS directly after preparing the samples were compared with those obtained from samples kept overnight at room temperature.

LEGENDS FOR SUPPLEMENTARY TABLES
Supplementary tables are available at:
http://www.sciencedirect.com/science/article/pii/S0092867413001529

Table S1: Table that summarizes all the quantitative mass spec data. Related to Figure 1, 2, 3 and 4.
Tab 1 contains the data obtained in the mC, hmC, fC and caC pull-down in mESCs. For each protein, the protein name, gene name and Uniprot identifier are indicated. Furthermore, GO term, PFAM and Corum annotations are listed. The Log2 transformed normalized ratios in the forward and reverse pull-downs for mC (K, L), hmC (M,N), fC (O,P) and caC (Q,R) are then listed followed by the significance B value in each pull-down. Proteins were considered significant if both Significance B values are <0.05 and the ratio is at least twofold. Proteins identified as a significant interactor in a particular pull-down are indicated with a '+' in columns AA-AI. Note that many proteins show preferential binding to fC and caC but are not significantly enriched due to the large amount of outliers in these pull-downs. Examples include subunits of the Sin3/HDAC complex, which show preferential binding to fC and caC. The ratios and normalized ratio's for all the proteins in the different triple pull-downs are listed in columns AJ-AU. The rest of the table contains information on the number of identified peptides, molecular weight, non-normalized protein ratios, mass spec intensities and so forth. **Tab 2** contains the mass spec data obtained in the C/mC/hmC triple pull-down in mESC nuclear extract. **Tab 3** contains the mass spec data obtained in the hmC/fC/caC pull-down in mESC nuclear extract. **Tab 4** contains a list of the significant interactors of Tdg-GFP. Shown are the LFQ values for the three GFP pull-downs and three control pull-downs, as well as the t-test difference and P-value. **Tab 5** contains the LFQ data of the C, mC and hmC pull-downs in nuclear extracts from NPC cells. The LFQ intensities for the three C pull-downs are listed in column K-M, the LFQ intensities for the three mC pull-downs are listed in columns N-P and the LFQ intensities for the three hmC pull-downs are listed in columns Q-S. These are then followed by the log transformed values in columns T-AB. In column AD-AF significant interactors as determined by ANOVA are indicated with a '+'. **Tab 6** contains the LFQ data of the C, mC and hmC pull-downs in nuclear extracts from adult mouse brains.

Table S2: iBAQ to quantify protein levels in different nuclear extracts. Related to Figure 5.
The iBAQ method was used to quantify absolute protein amounts of mESCs, NPCs and adult mouse brain nuclear extracts that were used for the DNA pull-downs. This table summarizes the result of this analysis with the molar amount for each protein in fmol/100 microgram in the mESCs (column AH), NPCs (column AK) and adult mouse brain (column AN)

Table S3: Summarizing table of the interactors that were identified in each of the cell types, including their absolute protein abundance. Related to Figure 7.
Table that lists all the proteins that were significantly enriched in one or more of the quantitative DNA pull-downs that were done in this study. Significant binding to the different baits in the different cell types is indicated with a '+' in columns F-S. Columns F-I indicate preferential binding to C vs each of the indicated modified baits in mESCs.

Table S4: DNA sequences used in the experiments

Experiment	Forward and Reverse	X=
Mass spec DNA pull-downs, western blot validations	AAG.ATG.ATG.AXG.AXG.AXG.AXG.ATG.ATG TTC.ATC.ATX.GTX.GTX.GTX.GTC.ATC.ATC	C, mC, hmC, fC or caC
Klf4 validation	TTCATCATAAGGXGGGXGGGXGACATCAT ATGATGTXGCCXGCCXGCCTTATGATG	T, C or mC
EMSA	GGATGATGACTCTTCTGGTCXGGATGGTAGTTAAGTGTTGAG CCTACTACTGAGAAGACCAGGXCTACCATCAATTCACAACTC	C, mC, hmC, fC, caC or Abasic
RFX5 NMR	CCTGATGAXGACGTACCG CGGTACGTXGTCATCAGG	mC

Table S5: Compound-dependent parameters for LC-MS/MS used in the analysis of genomic DNA.

Compound	Precursor Ion (MS1)	Product Ion (MS2)	Dwell time (ms)	CE (V)	Cell Acc (V)
[$^{15}N_2$]-caC	274.08	158.03	90	5	4
caC	272.09	156.04	90	5	4
[$^{15}N_2$]-fC	258.09	142.04	30	5	4
fC	256.09	140.05	30	5	4
[$^{15}N_2$,D_2]-hmC	262.12	146.07	40	27	1
hmC	258.11	142.06	40	27	1
[D_3]-mC	245.13	129.09	30	60	1
mC	242.11	126.07	30	60	1
dC	228.1	112.05	1	1	0
dG	268.1	152.06	1	1	0
dT	243.1	127.05	1	3	1
dA	252.11	136.06	1	50	0

Table S6: Compound-dependent parameters for LC-MS/MS used in the analysis of synthetic DNA.

Compound	Precursor Ion (MS1)	Product Ion (MS2)	Dwell time (ms)	CE (V)	Cell Acc (V)
[$^{15}N_2$]-caC	274.08	158.03	50	2	5
caC	272.09	156.04	50	2	5
[$^{15}N_2$]-fC	258.09	142.04	20	2	5
fC	256.09	140.05	20	2	5
[$^{15}N_2$,D_2]-hmC	262.12	146.07	50	1	1
hmC	258.11	142.06	50	1	1
[D_3]-mC	245.13	129.09	50	60	1
mC	242.11	126.07	50	60	1
[$^{15}N_2$]-dC	230.1	114.1	80	2	5
dC	228.1	112.05	80	2	5
[D_3]-dT	246.12	130.07	20	3	3
dT	243.1	127.05	20	3	3

SUPPLEMENTARY REFERENCES

1. Lavery, D.J. and U. Schibler, Circadian Transcription of the Cholesterol 7-Alpha Hydroxylase Gene May Involve the Liver-Enriched Bzip Protein Dbp. Genes & Development, 1993. 7(10): p. 1871-1884.
2. Dignam, J.D., R.M. Lebovitz, and R.G. Roeder, Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Res, 1983. 11(5): p. 1475-89.
3. Munzel, M., et al., Efficient Synthesis of 5-Hydroxymethylcytosine Containing DNA. Organic Letters, 2010. 12(24): p. 5671-5673.
4. Welz, R. and S. Muller, 5-(benzylmercapto)-1H-tetrazole as activator for 2 ‘-O-TBDMS phosphoramidite building blocks in RNA synthesis. Tetrahedron Letters, 2002. 43(5): p. 795-797.
5. Rappsilber, J., Y. Ishihama, and M. Mann, Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. Analytical Chemistry, 2003. 75(3): p. 663-670.
6. Hubner, N.C. and M. Mann, Extracting gene function from protein-protein interactions using Quantitative BAC InteraCtomics (QUBIC). Methods, 2011. 53(4): p. 453-9.
7. Delaglio, F., et al., NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR, 1995. 6(3): p. 277-93.
8. Vranken, W.F., et al., The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins, 2005. 59(4): p. 687-96.
9. Schwede, T., et al., SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res, 2003. 31(13): p. 3381-5.
10. Shen, Y., et al., TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR, 2009. 44(4): p. 213-23.
11. Chen, X., et al., Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell, 2008. 133(6): p. 1106-17.
12. Stadler, M.B., et al., DNA-binding factors shape the mouse methylome at distal regulatory regions (vol 480, pg 490, 2011). Nature, 2012. 484(7395): p. 550-550.
13. Schwanhausser, B., et al., Global quantification of mammalian gene expression control. Nature, 2011. 473(7347): p. 337-342.
14. Wisniewski, J.R., et al., Universal sample preparation method for proteome analysis. Nat Methods, 2009. 6(5): p. 359-62.
15. Pichler, G., et al., Cooperative DNA and Histone Binding by Uhrf2 Links the Two Major Repressive Epigenetic Pathways. Journal of Cellular Biochemistry, 2011. 112(9): p. 2585-2593.
16. Munzel, M., et al., Quantification of the sixth DNA base hydroxymethylcytosine in the brain. Angew Chem Int Ed Engl, 2010. 49(31): p. 5375-7.
17. Pfaffeneder, T., et al., The Discovery of 5-Formylcytosine in Embryonic Stem Cell DNA. Angewandte Chemie-International Edition, 2011. 50(31): p. 7008-7012.

Chapter 4

A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics

H. Christian Eberl[1], Cornelia G. Spruijt[2], Christian D. Kelstrup[3], Michiel Vermeulen[2,*] and Matthias Mann[1,*]

1. Department of Proteomics and Signal Transduction, Max-Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany.
2. Department of Molecular Cancer Research, University Medical Center Utrecht, Utrecht, The Netherlands.
3. Department for Proteomics, NNF Center for Protein Research, Faculty of Health Sciences, University of Copenhagen, Blegdamsvej 3b, DK-2200 Copenhagen, Denmark.
**Correspondence: m.vermeulen-3@umcutrecht.nl (M.V.), mmann@biochem.mpg.de (M.M.)*

Molecular Cell, January 24, 2013
Volume 49, p. 368–378

ABSTRACT

Post-translational modifications on core histones can serve as binding scaffolds for chromatin associated proteins. Proteins that specifically bind to or 'read' these modifications were previously identified in mass spectrometry-based proteomics screens based on stable isotope-labeling in cell lines. Here we describe a sensitive, label-free histone peptide pull-down technology with extracts of different mouse tissues. Applying this workflow to the classical activating and repressive epigenetic marks on histone H3, H3K4me3 and H3K9me3, we identified known and novel potential readers in extracts from brain, liver, kidney and testis. A large class of proteins were specifically repelled by H3K4me3. Our screen reached near saturation of direct interactors, most of which are ubiquitously expressed. In addition, it revealed a number of specialized readers in tissues such as testis. Apart from defining the chromatin interaction landscape in mouse tissues, our workflow can be used for peptides with different modifications and cell types of any organism.

INTRODUCTION

The genetic information of eukaryotes is stored in the nucleus by wrapping the DNA around octamers of histone proteins, forming the basic building blocks of chromatin, the nucleosomes [1]. Besides compacting and storing DNA, nucleosomes play an active role in regulated processes such as transcription and DNA repair. Post-translational modifications (PTMs) of the N-terminal tails of the core histones often serve as docking sites for 'chromatin readers', which can subsequently modify chromatin in *cis* or directly activate or repress transcription [2]. Prominent examples include the binding of HP1 proteins to H3K9me3 (K9me3) or the wide variety of H3K4me3 (K4me3) binding modules like e.g. BPTF [3], ING proteins [4], SGF29 [5] or PHF8 [6]. A number of reader domains have evolved that recognize specific PTMs in a protein sequence. These domains form special binding pockets, which probe the surrounding amino acid sequence in addition to containing a very selective interaction surface discriminating the unmodified from the modified state of a specific amino acid [7].

Histone modifications and their readers play important roles during cellular differentiation, development and in tumorigenesis [8, 9]. They contribute to maintaining gene expression differences between tissues. Even at the bulk histone levels, differences in the modification pattern between tissues can be observed [10]. Clearly the repertoire of chromatin readers and associated proteins varies between cell types and developmental stages. A classical example is the PHD finger containing protein RAG2, which is expressed in B cells during VDJ recombination. Its binding to K4me3 is crucial for the recombination event that these cells undergo during maturation [11]. Currently it is not known if RAG2 is an example for a larger group of specific chromatin readers or a specialized exception.

Mass spectrometry (MS)-based proteomics has played a crucial role in defining the global histone modification landscape in cells and in characterizing the subunit composition of chromatin related protein complexes (reviewed in [12]). A principal strength of MS-based methods is that they are hypothesis-free, making them well suited to discovering new interactors [13]. The combination of histone peptide pull-downs from crude nuclear extracts with quantitative MS is a particularly powerful approach to identify novel chromatin readers. Pull-downs are performed with modified and unmodified peptides and a quantitative filter distinguishes specific PTM readers from the vast amount of background binders that are typically present. We first applied this approach in HeLa cells that were metabolically labeled 'heavy' or 'light' using SILAC [14] to identify TFIID as a reader for K4me3 [15] and later characterized readers for five major lysine trimethylation sites on histone H3 and H4 [5]. Similar workflows identified proteins that specifically recognize combinations of histone modifications and DNA methylation [16], and enabled the study of interactions with reconstituted modified nucleosomal arrays [17].

All of the abovementioned studies were performed in a single cancer cell line, which restricted the identifiable interactors to proteins and protein complexes expressed in that system. Because reader complexes could differ by cell type and tissue or developmental stage, we wished to remove this limitation and develop a label-free technology that would be applicable to any sample and organism. Investigation of the binding to the activating K4me3 and the repressive K9me3 mark across tissues resulted in a very high coverage of known reader complexes, most of which are ubiquitously expressed in all the tissues we screened. We also observe a large group of proteins that

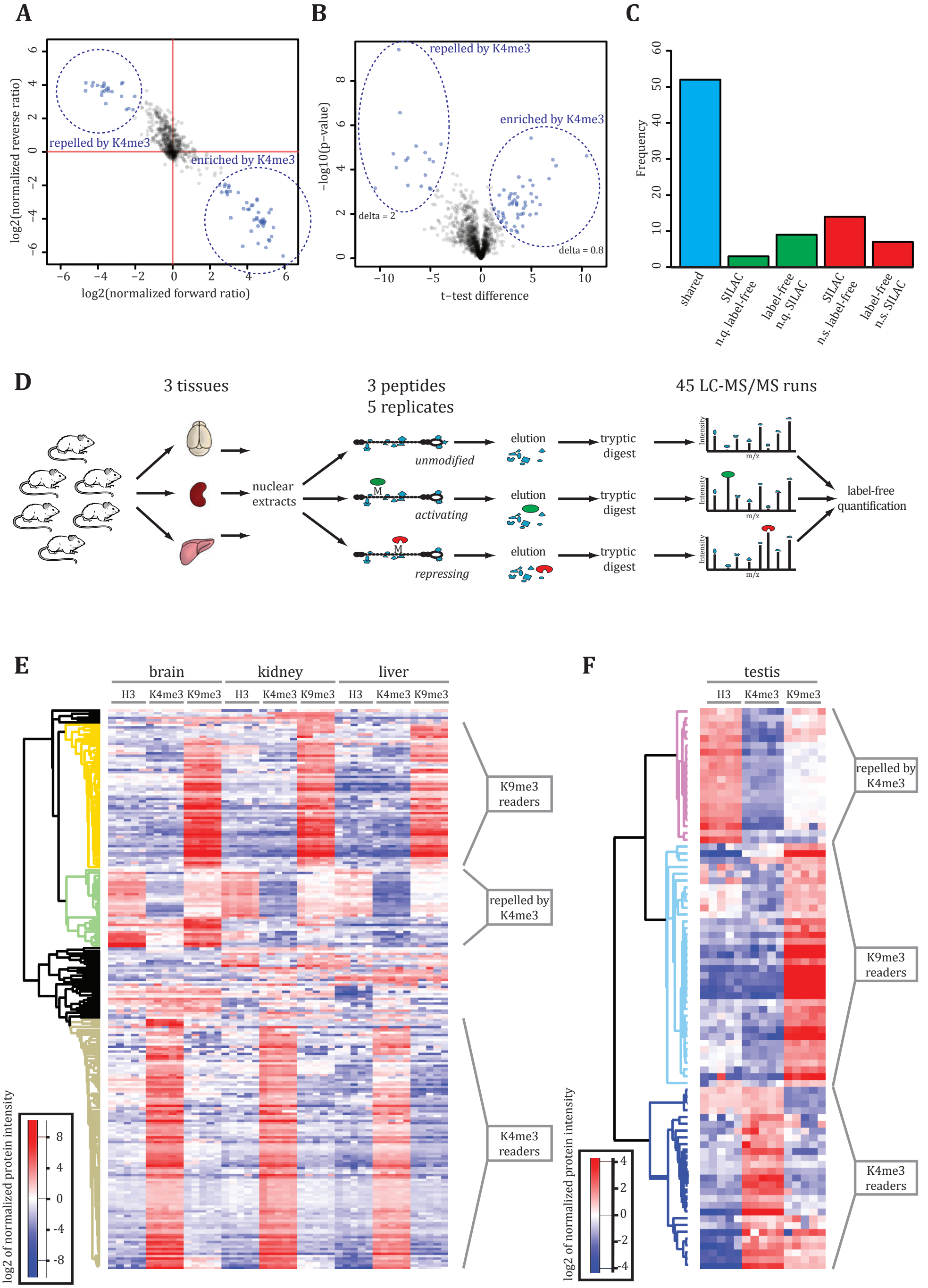


Figure 1: Label-free quantification is as powerful as SILAC-based quantification. A. SILAC-based peptide pull-down of H3K4me3 versus H3 unmodified. Significant outliers are marked in blue. **B.** Same pull-down using label-free quantification. Outliers that show significance in a modified *t*-test-based analysis are marked in blue. **C.** Overlap of outliers between SILAC and parallel label-free experiment:

are repelled by the K4 trimethyl mark as well as tissue specific subunits of chromatin reader complexes. Whereas the majority of chromatin reader complexes is conserved between tissues, some of the ubiquitously expressed chromatin reader complexes have evolved to contain tissue specific subunits, which could enable regulation of tissue specific target genes, or fine tune enzymatic activities. Some of these tissue specific subunits of chromatin reading complexes are DNA binding transcription factors which may serve to recruit reader complexes to tissue specific target genes in the genome.

RESULTS

A label-free interaction pipeline allows rapid screening for chromatin readers

Our previous workflow required individual analysis of each pull-down including separation by 1D gel electrophoresis followed by LC-MS/MS analysis of eight fractions [5, 15]. Here we placed sepharose beads in wells with a coarsely meshed bottom, which are impenetrable for aqueous solutions under normal conditions but enable liquid removal by slow centrifugation. This allowed switching to a 96 well format, increasing throughput and reproducibility. Furthermore, we made use of the increased sequencing speed of a linear ion trap – Orbitrap mass spectrometer [18] as well as longer gradients, to reduce the measurement of pull-downs to single LC-MS/MS runs. Finally, we replaced isotope based quantification by a sophisticated label-free quantification algorithm within the MaxQuant software suite [19].

To test this workflow, we performed SILAC-based and label-free peptide pull-downs in parallel for K4me3 readers from a mouse liver cell line (Table S1). The SILAC experiment was done in forward (i.e. incubating the modified peptide with the heavy and the unmodified peptide with the light extracts) and reverse (swapping of the labels). We found 46 proteins to be enriched and 23 proteins to be repelled by K4me3; these outliers encompassed many of the known K4me3 interactors (Figure 1A). Label-free pull-downs were performed in triplicate and analyzed by a modified *t*-test [20] (Figure 1B). The K4me3 mark enriched 49 proteins and specifically repelled 18. The large majority of the outliers were found in both experiments (blue in Figure 1C). Several proteins were only identified or quantified in one of them (green in Figure 1C). In accordance with a previous comparison [21], the larger dynamic range of the label-free experiment led to proteins only identified in this set of experiments (red in Figure 1C) whereas the higher quantitative accuracy of SILAC ensured statistical significance for borderline cases. For instance, the K4me3 interactor MORC3 or the K4me3 associated EMSY were significant in the SILAC experiment but close to threshold in the label-free experiment. The fact that some proteins are only outliers in one experiment but not the other is expected based on the different statistical behavior of binders in label-free and SILAC analysis. Overall, we concluded that label-free quantification is a viable alternative to SILAC for

Figure 1 (continued). blue: outliers that were identified and significant in both; green: outliers that were only identified in one experiment; red: outliers significant in one experiment but not in the other, n.q.: not quantified, n.s.: not significant. **D.** Workflow for screening chromatin readers from mouse tissue extracts: nuclear extract pools were prepared from mouse brain, liver and kidney. Pull-downs were performed with each extract with three different peptides (H3 unmodified, K4me3 and K9me3 modified) resulting in a total of 45 samples. Samples were measured separately and a label-free quantification algorithm was applied. **E.** Heat map of significant outliers from peptide pull-downs for H3K9me3 and H3k4me3 from brain, kidney and liver nuclear extracts. Readers with the same pattern are clustered together and are indicated on the right (see also Table S2). **F**. Similar heat map as in E for testis. In contrast to E, whole cell extracts were used (see also Table S3).

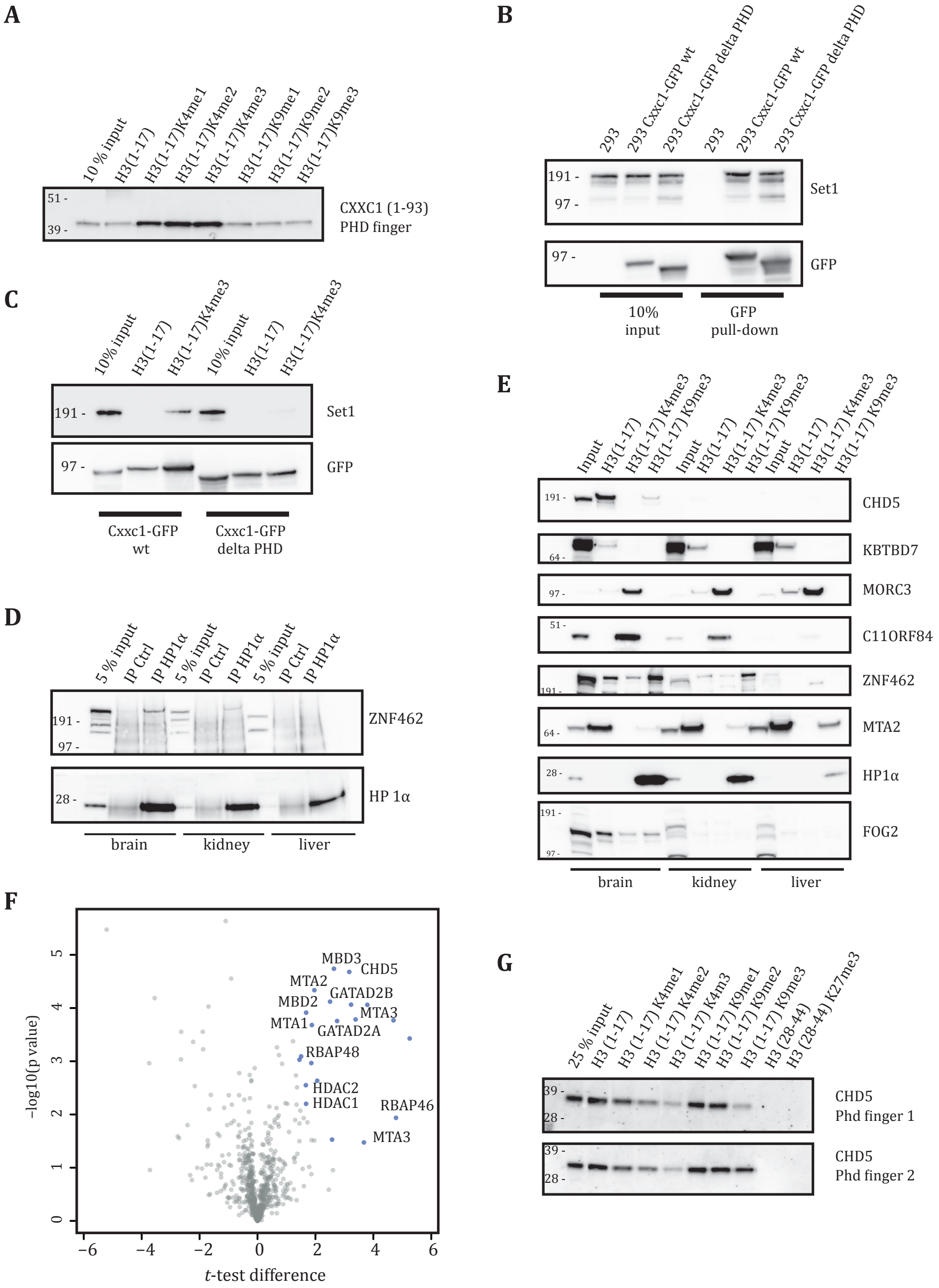
Figure 2: Verification of general and tissue specific chromatin readers and associated proteins. **A**. Peptide pull-down using purified CXXC1 PHD finger 1: specific binding of the SET1 complex subunit CXXC1 to H3K4me3. **B**. Overexpression of GFP-tagged mouse Cxxc1 full length (wt) and delta-PHD in HEK293 cells. Set1 co-precipitates with both constructs. **C**. Peptide pull-down with HEK293 nuclear extracts overexpressing Cxxc1-GFP wt and delta PHD. Cxxc1 wt is enriched on the H3K4me3 peptide compared to the unmodified peptide. The delta PHD mutant only shows background binding. Set1 binding to H3K4me3 is seen in the Cxxc1 wt extracts, but not when Cxxc1 delta PHD is overexpressed, demonstrating that Cxxc1 recruits Set1 to H3K4me3. **D**. HP1α Co-IP: ZNF462 - which is enriched on H3K9me3 - is enriched from brain and kidney but not from liver extracts. **E**. Western blot verification of selected readers. **F**. CHD5 Co-IP from brain nuclear extracts, followed by label-free quantitative proteomics: CHD5 enriches members of the NuRD complex. **G**. Peptide pull-down using purified CHD5 PHD fingers reveals specific repulsion by H3K4me3.

discovering chromatin readers, especially if quantitative accuracy is further boosted by increasing the number of replicates.

Having established a label-free high-throughput histone peptide pull-down interaction screening platform, we decided to use it to screen for tissue-specific chromatin readers of the key activating and repressive histone modifications K4me3 and K9me3, respectively. Nuclear extracts were prepared from pooled mouse brain, liver and kidney and these were separately incubated with unmodified, K4me3 and K9me3 modified peptides (Figure 1D). Every pull-down was analyzed in quintuplicate to maximize statistical significance.

We tested significant binding between the three possible pairs of bait peptides for each organ (nine *t*-test comparisons). Hierarchical clustering of all outliers generated in this way showed distinct groups (Figure 1E): enriched on K4me3 (115 proteins), enriched on K9me3 (64 proteins) and de-enriched on K4me3 (41 proteins) (Table S1). Inspecting the group of proteins significantly binding to these chromatin marks, we found almost only proteins annotated to be nuclear and very few apparent interactors from unexpected cellular compartments. Of the 31 K4me3 binders found by both Vermeulen *et al.* [5] and Nikolov *et al.* [17], our tissue based screen included 28. For the repressive K9me3 mark these studies had only 14 interactors in common, of which 11 are statistically significant in our data set. Thus our tissue-based screen appears to have reached very high coverage of previously established chromatin readers.

As an example of a tissue that cannot easily be mimicked in cell culture, we chose testis. This is a particularly interesting system to study chromatin readers, as sperm maturation and concomitant massive chromatin remodeling take place in this organ. Although nucleosomes are replaced to a large extent by protamines during sperm maturation, conventional histones, histone variants and modifications such as K4me3 can still be detected in mature sperm cells in developmentally important loci [22]. Because of the relatively low tissue mass, we performed pull downs from total tissue extract. Although the different extraction procedure precludes a direct comparison to the pull downs with the other organs, many of the same interactors were found, showing that chromatin readers can efficiently be retrieved even from total tissue extracts available in small amounts. In total we found 21 proteins associated with K4me3, 29 proteins associated with K9me3 and 19 proteins being repelled by K4me3 in testis (Figure 1F; Table S1).


A
10 % input
H3(1-17)
H3(1-17)K4me1
H3(1-17)K4me2
H3(1-17)K4me3
H3(1-17)K9me1
H3(1-17)K9me2
H3(1-17)K9me3
51 -
39 -
CXXC1 (1-93)
PHD finger
B
293
293 Cxxc1-GFP wt
293 Cxxc1-GFP delta PHD
293
293 Cxxc1-GFP wt
293 Cxxc1-GFP delta PHD
191 -
97 -
Set1
97 -
GFP
10%
input
GFP
pull-down
C
10% input
H3(1-17)
H3(1-17)K4me3
10% input
H3(1-17)
H3(1-17)K4me3
191 -
Set1
97 -
GFP
Cxxc1-GFP
wt
Cxxc1-GFP
delta PHD
D
5 % input
IP Ctrl
IP HP1α
5 % input
IP Ctrl
IP HP1α
5 % input
IP Ctrl
IP HP1α
191 -
97 -
ZNF462
28 -
HP 1α
brain
kidney
liver
E
Input
H3(1-17)
H3(1-17) K4me3
H3(1-17) K9me3
Input
H3(1-17)
H3(1-17) K4me3
H3(1-17) K9me3
Input
H3(1-17)
H3(1-17) K4me3
H3(1-17) K9me3
191 -
CHD5
64 -
KBTBD7
97 -
MORC3
51 -
C11ORF84
191 -
ZNF462
64 -
MTA2
28 -
HP1α
191 -
97 -
FOG2
brain
kidney
liver
F
−log10(p value)
0
1
2
3
4
5
−6
−4
−2
0
2
4
6
t-test difference
MBD3
CHD5
MTA2
GATAD2B
MBD2
MTA3
MTA1
GATAD2A
RBAP48
HDAC2
HDAC1
RBAP46
MTA3
G
25 % input
H3 (1-17)
H3 (1-17) K4me1
H3 (1-17) K4me2
H3 (1-17) K4m3
H3 (1-17) K9me1
H3 (1-17) K9me2
H3 (1-17) K9me3
H3 (28-44)
H3 (28-44) K27me3
39 -
28 -
CHD5
Phd finger 1
39 -
28 -
CHD5
Phd finger 2

General and organ specific chromatin associated complexes

The large majority of reader proteins were found as specific binders in all three organs studied. Table 1 lists these proteins, grouped into known chromatin reader complexes, where possible. We found 9 such complexes for the K4me3 mark and in most of these cases the entire set of established complex members was found as significant interactors. This indicates that our screen reached unprecedented coverage. Interestingly, the SET1 complex, which itself methylates H3K4, was one of the complexes bound to K4me3. In yeast, direct binding of SET1 complex member SPP1 to H3K4me3, which recruits yeast SET1, has been described [23]. However, in mammals none such interaction has been described yet. We therefore tested the PHD finger of the complex member CXXC1 for binding to K4me3 and indeed observed a specific interaction with H3K4me3 (Figure 2A). Moreover, overexpressed CXXC1 devoid of the PHD finger still interacts with Set1 (Figure 2B). Furthermore, it shows a dominant negative effect on Set1 binding to the H3K4me3 peptide (Figure 2C). Thus we conclude that CXXC1 recruits SET1 to H3K4me3.

The proteins associated with K9me3 encompass most of the known direct readers of this modification, including several that were only described very recently (Table 1). As expected among the specific binders to this repressive mark were many Polycomb group members as well as many HP1 interactors reported in a recent HP1 interactome study [24]. It is noteworthy that both among the already known and the newly described K9me3 associated proteins were many with zinc finger motifs. These proteins could couple a DNA sequence specific read-out to the detection of the repressive mark in a similar manner as already described for the HP1 interactor POGZ [24].

We tested several of the outliers of specific interest as well as some completely uncharacterized proteins by Western blotting. In each of the cases, the Western blot verified the result of our global analysis (Figure 2E).

Next, we inspected our quantitative data for tissue specific chromatin readers and associated proteins. Mass spectrometry and western blotting found ZNF462 as a specific binder to K9me3 in brain and kidney but not in liver, where this protein appears not to be expressed (Figure 2D and E). ZNF462 is a zinc finger protein with a role in development [25], and its knockdown leads to mislocalization of HP1α [26]. In conjunction with the enrichment of ZNF462 on K9me3, this suggested that it is an HP1α interactor. Indeed ZNF462 is present in HP1α immuno-precipitations from brain and kidney but not from liver extracts (Figure 2D). Thus we conclude that ZNF462 is a tissue specific and restricted HP1 interactor.

In brain extracts but none of the other extracts, CHD5 was enriched with the unmodified and K9me3 modified peptide as compared to K4me3. This was also confirmed by western blotting, which furthermore indicated absence of the protein in the input material in kidney and liver extracts (Figure 2E). To obtain insights into the function of CHD5, we performed interaction proteomics with the above described platform but coupling an antibody against CHD5 to the beads. Members of the NuRD complex (MBD2/3, MTA1/2/3, GATAD2A/B, HDAC1/2 and RBBP7) were significantly enriched, except for CHD3 and CHD4 (Figure 2F). Together with a very recent report

Table 1: Chromatin readers and associated proteins. Summary of all specific interaction partners for the investigated chromatin marks (for details see Table S2 and S3). Proteins are grouped into complexes or interaction networks according to their description in literature. [a] only found in brain, [b]only found in brain and kidney, [c]only found in testis.

Table 1. Chromatin Readers and Associated Proteins

Reader group	Complex	Direct binder	Complex members
K4me3	TFIID	TAF3	TAF1, 2, 3, 4a, 4b, 5, 6, 7, 8, 9b, 10, 11, 12, 13, TBP
	SAGA	SGF29	ATXN7, ATXN7L1, ATXN7L2, ATXN7L3, CHD1, FAM48A, USP22, TAF5L, TAF6L, SUPT3H, SUPT7L, TADA1L, SGF29
	SET1	CXXC1	ASH2L, SETD1A, SETD1B, CXXC1
	NuA4 HAT	ING3	BRD8, DMAP1, EP400, EPC1, TIP60, ING3, MORF4L1, MORF4L2, RUVBL1, RUVBL2, YL1, YEATS4, MRGBP, TRRAP
	ATAC		TADA3L, CSRP2BP, GCN5L2, PCAF, SGF29, YEATS2, MBIP, TADA2L, ZZZ3
		JARID1A	EMSY, GATAD1, JARID1A, SIN3B, PHF12, MORF4L1
	HBO1 (ING5 complex)	ING4/5	HBO1, ING4/5, PHF15, PHF16, PHF17, MEAF6, BRD1, BRPF3
	SIN3A	ING2	ING2, SIN3A, SAP130, SAP30L, SUDS3, SAP180, ARID4A, BRMS1L
	MLL		DPY30, HCFC1, HCFC2, JMJD3, MLL2, MLL5, CHD8, RBBP5, MEN1
	NURF	BPTF	C17ORF49(BAP18), HMGB2L1, SMARCA1
	Not yet assigned to complexes	DIDO1, ING1, JHDM1D (KDM7), JHDM1B, JMJD2A, PHF8, MORC3, PHF13, PHF2, PHF23, SPIN1	BOD1L, BAF53B, EPC2, GTF2A1, H2AFV, JARID1B, JAZF1, PCYOX, MBTD1, SMARCA5, TADA2B, C11ORF84 homolog, SMC1A, SMC3, UBXD7[c], EHMT1[c], EHMT2[c], BRWD1[c], CRCP[c], SSTY1[c], SSTY2[c], SLY[c], SLXL1[c], KLHL36[c]
K9me3		HP1α	ADNP, AHDC1, FBXL11, ZNF828, POGZ, SENP7, RLF, NIPBL, PRR14, C1ORF103 homolog, ZNF462[b], TRIM66[c], CHAF1A[c]
		HP1β	
		HP1γ	
	ORC		LRWD1, ORC2
	Polycomb		SUZ12, RING1A, RING1B, EED, EZH1, EZH2, MGA, L3MBTL2, MAZ, PCGF6, PHF1, CBX4
	Not yet assigned to complexes	CDYL, CDYL2, ATRX, MPHOSPH8, UHRF1, UHRF2	Hypothetical protein LOC72123, ADNP2, PRDM10, HDGFRP2, HOMEZ, ZMYM2, ZMYM3, ZMYM4, ZMYM5, ZMYM6, SMCHD1, TRIM33, MIER1, MIER2, ZFP280C, ZFP280D, ZNF518B, PAP20, TRIM28, PPHLN1, NSD3, P91A, TRIM24, ZFP15, ZFP524, ZFP297, C19ORF68 homolog, FAM208A, SFRS2, SCAI, UBR7[c], PHF10[c], KPNA3[c], KPNA4[c]
Repelled by K4me3	NuRD	CHD3, CHD4, CHD5[a]	RBAP48, RBAP46, HDAC1, HDAC2, MBD2, MBD3, MTA1, MTA2, MTA3, CHD3, CHD4, CHD5a, FOG2[a], GATAD2A, GATAD2B, DOC1, MBD3L[c]
	NuRD associated	CHD4	ZNF687, ZMYND8, ZNF592, ZNF532
			RAI1, PHF14, TCF20
		BHC80 (PHF21A)	
	Not yet assigned to complexes	DNMT3A[c], DNMT3B[c]	BCL7A, CFL1, DGKE, DHX30, FLYWCH1, PRMT5, PWWP2A, PPIG, KBTBD7, MYT1L[a], PABP1, ZBTB43, ZNF428, GABRG1[c], H1FX[c], HAT1[c], RPS10

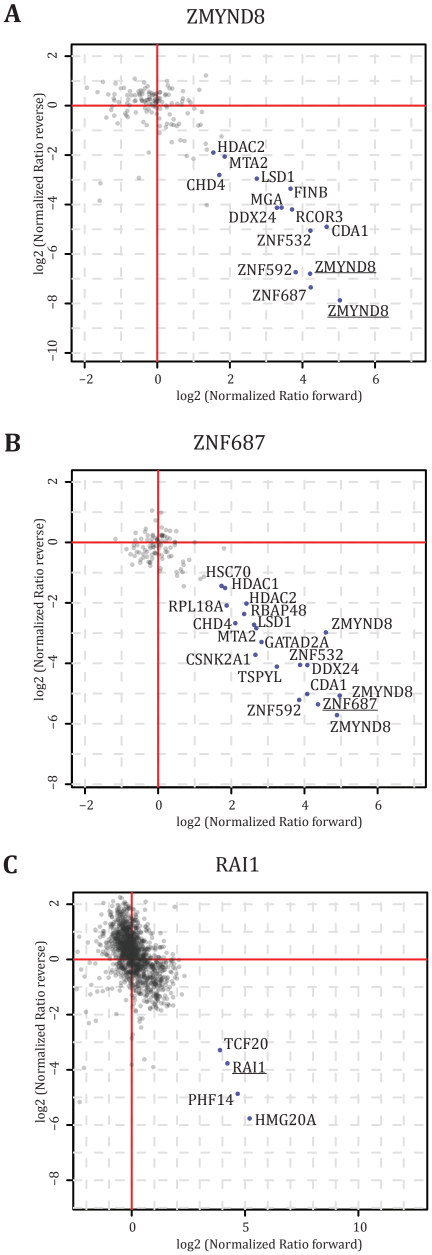


Figure 3: Interaction proteomics for proteins repelled by H3K4me3. SILAC GFP pull-downs from HeLa nuclear extracts for ZMYND8 (**A**), ZNF687 (**B**) and RAI1 (**C**); proteins are expressed at near endogenous levels in HeLa cells. Interaction partners can be found on the right lower quadrant and are marked with their names.

[27], this demonstrates that CHD5 is a member of a NuRD like complex. The NuRD complex represses transcription by nucleosome remodeling and deacetylation [28, 29]. As its interaction with the H3 tail is mediated by the two PHD fingers of CHD3 or CHD4 [30], neither of which interacted with CHD5, we tested if CHD5 could take over this function. We expressed the PHD fingers of CHD5 and found that both bind to the unmodified peptide and are repelled by K4me3 (Figure 2G). The binding patterns of the CHD5 PHD fingers mirror CHD4, whose two PHD fingers bivalently recognize both H3 tails on a single nucleosome [31]. We hypothesize that CHD5 takes the position of CHD3 or CHD4 in a neuronal NuRD complex and that it is responsible for binding to the H3 tail.

Several readers were exclusively found in testis, reflecting the unique chromatin remodeling events in spermatogenesis. Among the known testis specific readers and associated proteins, we detected MBD3L, a testis specific NuRD subunit [32] that clusters with other NuRD complex members in the typical repulsion pattern from K4me3. TRIM66 (TIF1δ) is an HP1 interactor predominantly expressed in testis [33] and was enriched on the K9me3 modification. DNMT3A is a DNA methyltransferase preferentially expressed in cells undergoing *de novo* methylation such as testis, and was enriched on unmodified H3 as described before [34]. In addition, the testis specific proteins SSTY1 and SSTY2 were specifically enriched on K4me3. Both proteins are encoded in many copies on the Y chromosome of mice and are expressed during sperm development [35]. Deletions of these genes lead to severe sperm head defects and sterility [36]. Interestingly, SPIN1, a known K4me3 reader [37] has 55 and 52 % sequence identity towards SSTY1 and SSTY2, respectively. These proteins share the same domain and the amino acids suggested to mediate the interaction with the modified lysine residue in SPIN1 (F141, Y170 and Y177) [37] are conserved. We therefore speculate that SSTY is a direct binder of K4me3 in testis. Additional testis specific proteins that specifically bound to K4me3 include SLX, SLXL1 and SLY.

Complexes specifically repelled by K4 trimethylation

Apart from readers for K4me3 and K9me3, our screen also identified a group of proteins that specifically showed reduced binding to the K4me3 modification (Table 1). Among these is the already mentioned NuRD complex with its known subunits and BHC80, the first PHD finger-containing protein described to bind preferentially to unmodified H3K4 via its PHD finger [38]. In proteomic datasets published so far the focus has been on readers of modified amino acids, rather than proteins that are specifically repelled by a modification. We found 41 such proteins, all of which were repelled by K4me3, whereas no readers specifically repelled by K9me3 were apparent, in accordance with an absence of literature reports of proteins specifically recognizing unmodified H3K9. As all of these repelled proteins – with the exception of CHD5 – showed nearly equal binding in all three tissues they appear to perform general and non-tissue specific functions.

To further elucidate these functions, we used cell line based methods to assign them into complexes. Specifically, we employed the recently developed BAC technology [39] to perform SILAC-based GFP pull-downs of proteins expressed at endogenous levels [21]. We analyzed protein-protein interactions for three proteins not described in the context of reading unmodified histone H3 (Table S2). Of particular interest was a series of zinc finger proteins, including ZMYND8, a zinc finger protein that also contains a PWWP domain, a bromodomain and a PHD type zinc finger. It interacts with CHD4, the NuRD complex member that is responsible for binding of the complex to unmodified and K9me3 [40] thereby explaining the observed binding pattern (Figure 3A). The zinc finger proteins ZNF592, ZNF687 and ZNF532, which we also found to be enriched in our peptide pull-down, likewise specifically interacted with ZMYND8. Moreover, when pulling-down ZNF687, we reciprocally enriched ZMYND8, as well as ZNF592 and ZNF532 (Figure 3B). CHD4 and further NuRD complex members specifically interacted with ZNF687 as well. The zinc finger proteins ZMYND8, ZNF592 and ZNF687 have been shown to form a subcomplex [41] and our data now links them to the NuRD complex as auxiliary members. Given the large number of zinc fingers in these proteins, we hypothesize that some of them serve to recruit the NuRD complex to specific target genes in the genome.

Another protein associated with unmodified histone H3 was retinoic acid induced protein 1 (RAI1), which is implicated in Smith-Magenis syndrome, a developmental disorder characterized by mental retardation and craniofacial and skeletal abnormalities [42]. In the GFP pull-down we found PHF14, TCF20 (Kiaa0292) and HMG20A specifically associated with RAI1 (Figure 3C); these four proteins may form a novel chromatin associated complex whose members possess several PHD fingers.

Chromatin readers of the H3K4me1 mark

To demonstrate extensibility of our pull-down methodology not only for specialized tissues (Figure 1F) but also for different baits, we performed pull-downs with brain and liver nuclear extracts for monomethylated H3K4 (Figure 4A, Table S3), a histone modification generally associated with enhancers [43]. We enriched for the known

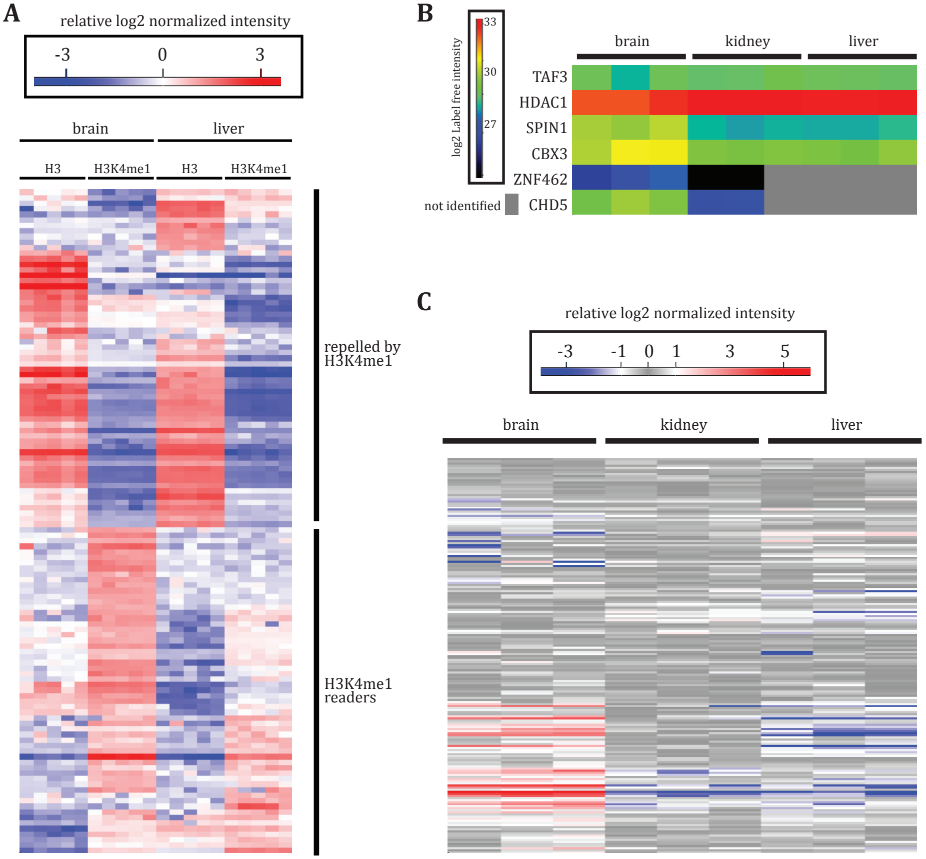


Figure 4: Extension of the proteomic screen. A. Label free interaction screen for readers of H3K4me1 from mouse brain and liver nuclear extracts. **B.** Protein expression profiles of selected chromatin readers. General chromatin readers show nearly equal expression levels over the analyzed tissues, whereas organ specific chromatin readers show organ specific expression profiles. **C.** Proteomic expression profiles of chromatin readers identified in this study.

H3K4me1 readers CHD1 [44] and the TIP60 complex [45] with its members EP400, EPC1, BRD8, YL1 and ING3. Interestingly, the H3K4me3 readers Morc3, Spindlin1, PHF2 and PHF23 were also significantly enriched compared to the unmodified peptide. In contrast, the large group of direct H3K4me3 interactors described above (Table 1) was not clear and significantly binding to H3K4me1. Finally, we observed tissue specific interactions, like the already observed FOG2 and CHD5, which are brain specific and repelled by H3K4me1, as well as ZHX2 and ZHX3, which are repelled by H3K4me1 in liver.

Deep proteomic quantification supports tissue binding patterns of chromatin readers

Next we complemented our interaction studies by a deep proteomic profile of nuclear extracts across the tissues (biological triplicates; more than 5000 proteins identified). This demonstrated that organ specific chromatin readers in our interaction screen also show organ specific expression patterns. This is exemplified by the brain specific CHD5 (Figure 4B). The testis specific readers SSTY1 and 2, as well as SLY or SLX were not identified in brain, kidney or liver. The HP1 interactor ZNF462, which was absent in the interaction screen in liver, also was not detected in the nuclear liver proteome. In line with the pull-down results, the large majority of chromatin readers observed in our screen showed approximately equal expression levels in all three tissue nuclear extracts (Figure 4C).

DISCUSSION

Here we have developed and demonstrated a high resolution and high accuracy workflow to detect interactions with modified peptides. It uses label-free quantification and is completely generic as it can be used for any synthesizable peptide modification as well as any suitable protein extract. The technology is highly sensitive, streamlined and scalable. The absence of any protein or peptide fractionation steps, with concomitant reduction in measurement time, enabled us to perform a relatively large number of replicates in different tissues, increasing statistical confidence. Compared to previous proteomics efforts on identifying chromatin readers, we obtained much improved coverage. This was evident, for instance, by the fact that subunits of chromatin reader complexes were in most cases completely recovered.

We applied our workflow to generate a reader map of interactors of the activating K4me3 and the repressive K9me3 chromatin mark from mouse tissue, which not only covers the large majority of known interactors, but also describes many associations for the first time. The increased depth and completeness of the measured interactome should make it a useful resource to the community. It also highlights the diversity and complexity of chromatin associated proteins for these marks. This is especially apparent for the activating K4me3 mark, for which we recover 16 known and 1 novel direct binder and most of their associated complex members as well as novel factors. These proteins represent a strikingly broad variety of different functions that they can perform on the surrounding chromatin, even including writing and erasing the K4me3 mark itself. Furthermore some readers play a general role for gene expression, such as TFIID, whereas others are only important for expression of a specialized subset of genes. One important question that remains is how all these different chromatin readers are recruited to their specific target genes in the genome, since it is clear that different K4me3 reading complexes bind to distinct and only partially overlapping clusters of K4me3 marked genes in human cells [5]. Part of this specificity may be brought about by additional chromatin marks that serve to differentially enhance or reduce the binding of readers to genes. We have previously shown how such fine-tuning modifications including H3R2me2a and H3S10P can selectively enhance or repress the binding of readers to K4me3 and K9me3, respectively [5]. But beyond these auxiliary modifications, many of the chromatin reading complexes described here most likely gain binding specificity for their target genes by DNA sequence driven recruitment events.

The combination of DNA sequence specific and histone modification mediated recruitment of chromatin associated complexes can best be seen on the repressive K9me3 mark, for which we describe new associated proteins. Among them, many harbor DNA binding modules like zinc finger domains. Furthermore even a tissue specific function can be connected to a general machinery by auxiliary factors like ZNF462 in brain and kidney, or TRIM66 in testis.

In addition to the interaction screen, we also used proteomics to correlate our results to organ specific expression patterns. The large majority of chromatin readers showed similar expression patterns across the tissues. However, all tissue specific binders also had tissue specific expression patterns. This restricted expression suggests unique functions necessary in the respective tissue.

The combination of interaction and deep expression proteomics can also be used in an inverse approach: the tissue, cell type or developmental stage specific expression of a putative chromatin reader could guide subsequent targeted experiments to determine if this protein binds to a specific mark in those contexts.

In conclusion, advances in proteomics technology increasingly make it possible to move from *in vitro* cell culture to *in vivo* derived tissues extracts. This allows surveying the binding of proteins expressed in diverse tissues including ones not expressed in standard cell lines. In particular, it allows surveying the interactome in specialized tissues that cannot easily be mimicked in cell culture, such as testis. Scalable and accurate mapping of the binders to single and combined chromatin marks should contribute to increased understanding of protein-chromatin interactions and their role in regulating tissue and cell type specific gene expression programs and cell fate decisions.

EXPERIMENTAL PROCEDURES

Extract preparation

Nuclear extracts from cell lines were prepared as described before [15].

Nuclei from brain, liver and kidney were purified by homogenization followed by pelleting through a sucrose cushion, modified from [46]. Nuclei were lysed in 2 volumes 420 mM NaCl, 20 mM Hepes pH 7.9, 20% v/v glycerol, 2 mM $MgCl_2$, 0.2 mM EDTA, 0.1% NP40, complete protease inhibitor w/o EDTA (Roche), 0.5 mM DTT.

Testis were snap frozen in liquid nitrogen, grinded in a beadmill (2x 3 minutes, 300 Hz) and 4 volumes lysis buffer (50 mM Tris pH 8.0, 20 mM NaCl, 0.25% NP40, 1 mM $MgCl_2$, complete Protease inhibitor , 0.5 mM DTT) were added followed by sonication. Samples were incubated with Benzonase until no pellet was visible anymore and subsequently precleared at 15000 g for 10 minutes. Extract were pooled for pull-downs to minimize variability introduced by extract preparation.

Peptide pull-downs

Peptide pull-downs were performed on 96 well plates modified from [5]. In brief: Histone peptides containing the N-terminal 17 amino acids of the histone H3 tail followed by two glycines and a biotinylated lysine were synthesized using the Fmoc strategy as described [47]. An excess of peptide was coupled to sepharose streptavidin beads (GE Healthcare). Beads were transferred to 96 well Multi screen filter plates (Millipore, MSBVN1210). Nuclear extracts (400 µg total protein) in 200 µl incubation buffer (150 mM NaCl, 50 mM Tris pH 8.0, 1% NP40, 0.5 mM DTT) were added and incubated for 3 h at 4 °C while gently shaking. Beads were washed 3 times (30 seconds, 60 g) with 200 µl wash buffer 1 (320 mM NaCl, 50 mM Tris pH 8.0, 0.5% NP40), followed by 5 washes with wash buffer 2 (150 mM NaCl, 50 mM Tris pH 8.0) to minimize residual detergent. 25 µl 2 M urea, 1mM DTT supplemented with 120 ng trypsin (Promega) was added and incubated for 30 min at room temperature and eluted, followed by two additional elution steps (50 µl 2 M urea, 5 mM iodoacetamide, 10 minutes incubation each). Proteins were digested over night at room temperature.

LC MS/MS analysis

Samples were measured using the LTQ-Orbitrap Velos or Q Exactive proteomic pipeline. Raw mass spectrometric data was analyzed using the MaxQuant pipeline [48].

GFP pull-downs

GFP pull-downs were performed as described before [5] followed by FASP and measurement as single run (ZMYND8 and ZNF687) or in gel digest and fractionation into 8 slices (RAI1). All samples were measured on a LTQ Orbitrap Velos using 120 min segmented gradients.

Protein Co-IPs

CHD5 antibody and rabbit control antibody, or HP1α antibody and goat control antibody were cross-linked to Protein G sepharose (GE Healthcare) using dimethyl pimilidate. CHD5 Co-IPs were performed on 96 well Mulit screen plates as described above using brain nuclear extracts (350 µg total protein). HP1α Co-IPs were performed in tube (600 µg total protein) and eluted by boiling in loading buffer.

ACKNOWLEDGEMENTS

We thank T. Viturawong and J. Cox for helpful discussions; S. Hake for critical reading of the manuscript; D. Walther, M. Robles and S. Dewitz for help with mouse work. We would like to acknowledge S. Uebel and S. Suppmann from the MPI Microchemistry Core facility for peptide synthesis and providing material, respectively. We are indebted to T. Hyman and I. Poser for providing BAC cell lines for ZMYND8, ZNF687 and RAI1.

Author contributions: HCE, MV and MM designed research. HCE performed most experiments and analyzed data. CDK and HCE developed 96-well plate based peptide pull-downs. CGS performed GFP pull-downs from HeLa cells. HCE, MV and MM wrote the paper.

This work was supported by the German National Genome Research Network (From Disease Genes to Protein Pathways, DiGtoP) and by the HEROIC grant (LSHG-CT-2005-018883) from the 6th EU Framework program. Research in the Vermeulen lab is supported by a grant from the Netherlands Organization for Scientific Research (NWO-VIDI).

REFERENCES

1. Luger, K., et al., *Crystal structure of the nucleosome core particle at 2.8 A resolution.* Nature, 1997. **389**(6648): p. 251-60.
2. Kouzarides, T., *Chromatin modifications and their function.* Cell, 2007. **128**(4): p. 693-705.
3. Li, H., et al., *Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF.* Nature, 2006. **442**(7098): p. 91-5.
4. Pena, P.V., et al., *Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2.* Nature, 2006. **442**(7098): p. 100-3.
5. Vermeulen, M., et al., *Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers.* Cell, 2010. **142**(6): p. 967-80.
6. Feng, W., et al., *PHF8 activates transcription of rRNA genes through H3K4me3 binding and H3K9me1/2 demethylation.* Nat Struct Mol Biol, 2010. **17**(4): p. 445-50.
7. Taverna, S.D., et al., *How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers.* Nat Struct Mol Biol, 2007. **14**(11): p. 1025-40.
8. Wang, G.G., et al., *Haematopoietic malignancies caused by dysregulation of a chromatin-binding PHD finger.* Nature, 2009. **459**(7248): p. 847-51.
9. Berdasco, M. and M. Esteller, *Aberrant epigenetic landscape in cancer: how cellular identity goes awry.* Dev Cell, 2010. **19**(5): p. 698-711.
10. Garcia, B.A., et al., *Tissue-specific expression and post-translational modification of histone H3 variants.* J Proteome Res, 2008. **7**(10): p. 4225-36.
11. Matthews, A.G., et al., *RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination.* Nature, 2007. **450**(7172): p. 1106-10.
12. Eberl, H.C., M. Mann, and M. Vermeulen, *Quantitative proteomics for epigenetics.* Chembiochem, 2011. **12**(2): p. 224-34.
13. Vermeulen, M., N.C. Hubner, and M. Mann, *High confidence determination of specific protein-protein interactions using quantitative mass spectrometry.* Curr Opin Biotechnol, 2008. **19**(4): p. 331-7.
14. Ong, S.E., et al., *Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.* Mol Cell Proteomics, 2002. **1**(5): p. 376-86.
15. Vermeulen, M., et al., *Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4.* Cell, 2007. **131**(1): p. 58-69.
16. Bartke, T., et al., *Nucleosome-interacting proteins regulated by DNA and histone methylation.* Cell, 2010. **143**(3): p. 470-84.
17. Nikolov, M., et al., *Chromatin affinity purification and quantitative mass spectrometry defining the interactome of histone modification patterns.* Mol Cell Proteomics, 2011.
18. Olsen, J.V., et al., *A dual pressure linear ion trap orbitrap instrument with very high sequencing speed.* Mol Cell Proteomics, 2009. **8**(12): p. 2759-69.
19. Luber, C.A., et al., *Quantitative proteomics reveals subset-specific viral recognition in dendritic cells.* Immunity, 2010. **32**(2): p. 279-89.
20. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-

21.
21. Hubner, N.C., et al., *Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions.* J Cell Biol, 2010. **189**(4): p. 739-54.
22. Hammoud, S.S., et al., *Distinctive chromatin in human sperm packages genes for embryo development.* Nature, 2009. **460**(7254): p. 473-8.
23. Shi, X., et al., *Proteome-wide analysis in Saccharomyces cerevisiae identifies several PHD fingers as novel direct and selective binding modules of histone H3 methylated at either lysine 4 or lysine 36.* J Biol Chem, 2007. **282**(4): p. 2450-5.
24. Nozawa, R.S., et al., *Human POGZ modulates dissociation of HP1alpha from mitotic chromosome arms through Aurora B activation.* Nat Cell Biol, 2010. **12**(7): p. 719-27.
25. Masse, J., et al., *ZFPIP/Zfp462 is involved in P19 cell pluripotency and in their neuronal fate.* Exp Cell Res, 2011. **317**(13): p. 1922-34.
26. Masse, J., et al., *Involvement of ZFPIP/Zfp462 in chromatin integrity and survival of P19 pluripotent cells.* Exp Cell Res, 2010. **316**(7): p. 1190-201.
27. Potts, R.C., et al., *CHD5, a brain-specific paralog of Mi2 chromatin remodeling enzymes, regulates expression of neuronal genes.* PLoS One, 2011. **6**(9): p. e24515.
28. Tong, J.K., et al., *Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex.* Nature, 1998. **395**(6705): p. 917-21.
29. Xue, Y., et al., *NURD, a novel complex with both ATP-dependent chromatin-remodeling and histone deacetylase activities.* Mol Cell, 1998. **2**(6): p. 851-61.
30. Mansfield, R.E., et al., *Plant homeodomain (PHD) fingers of CHD4 are histone H3-binding modules with preference for unmodified H3K4 and methylated H3K9.* J Biol Chem, 2011. **286**(13): p. 11779-91.
31. Musselman, C.A., et al., *Bivalent recognition of nucleosomes by the tandem PHD fingers of the CHD4 ATPase is required for CHD4-mediated repression.* Proc Natl Acad Sci U S A, 2012. **109**(3): p. 787-92.
32. Jiang, C.L., S.G. Jin, and G.P. Pfeifer, *MBD3L1 is a transcriptional repressor that interacts with methyl-CpG-binding protein 2 (MBD2) and components of the NuRD complex.* J Biol Chem, 2004. **279**(50): p. 52456-64.
33. Khetchoumian, K., et al., *TIF1delta, a novel HP1-interacting member of the transcriptional intermediary factor 1 (TIF1) family expressed by elongating spermatids.* J Biol Chem, 2004. **279**(46): p. 48329-41.
34. Otani, J., et al., *Structural basis for recognition of H3K4 methylation status by the DNA methyltransferase 3A ATRX-DNMT3-DNMT3L domain.* EMBO Rep, 2009. **10**(11): p. 1235-41.
35. Toure, A., et al., *A protein encoded by a member of the multicopy Ssty gene family located on the long arm of the mouse Y chromosome is expressed during sperm development.* Genomics, 2004. **83**(1): p. 140-7.
36. Toure, A., et al., *A new deletion of the mouse Y chromosome long arm associated with the loss of Ssty expression, abnormal sperm development and sterility.* Genetics, 2004. **166**(2): p. 901-12.
37. Wang, W., et al., *Nucleolar protein Spindlin1 recognizes H3K4 methylation and stimulates the expression of rRNA genes.* EMBO Rep, 2011. **12**(11): p. 1160-6.
38. Lan, F., et al., *Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression.* Nature, 2007. **448**(7154): p. 718-22.
39. Poser, I., et al., *BAC TransgeneOmics: a high-throughput method for exploration of*

protein function in mammals. Nat Methods, 2008. **5**(5): p. 409-15.

40. Musselman, C.A., et al., *Binding of the CHD4 PHD2 finger to histone H3 is modulated by covalent modifications.* Biochem J, 2009. **423**(2): p. 179-87.
41. Malovannaya, A., et al., *Analysis of the human endogenous coregulator complexome.* Cell, 2011. **145**(5): p. 787-99.
42. Slager, R.E., et al., *Mutations in RAI1 associated with Smith-Magenis syndrome.* Nat Genet, 2003. **33**(4): p. 466-8.
43. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-8.
44. Flanagan, J.F., et al., *Double chromodomains cooperate to recognize the methylated histone H3 tail.* Nature, 2005. **438**(7071): p. 1181-5.
45. Jeong, K.W., et al., *Recognition of enhancer element-specific histone methylation by TIP60 in transcriptional activation.* Nat Struct Mol Biol, 2011. **18**(12): p. 1358-65.
46. Lavery, D.J. and U. Schibler, *Circadian transcription of the cholesterol 7 alpha hydroxylase gene may involve the liver-enriched bZIP protein DBP.* Genes Dev, 1993. **7**(10): p. 1871-84.
47. Schulze, W.X. and M. Mann, *A novel proteomic screen for peptide-protein interactions.* J Biol Chem, 2004. **279**(11): p. 10756-64.
48. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.

SUPPLEMENTAL MATERIALS AND METHODS

Cell culture

Hepa 1-6 and HeLa cells were cultured in SILAC medium (PAA) containing heavy isotope–labeled $^{13}C_6{}^{14}N_4$-L-arginine and $^{13}C_6{}^{14}N_2$-L-lysine (Cambridge Isotope Laboratories) supplemented with 10% dialyzed fetal bovine serum (PAA). BAC cell lines were supplemented with 0.4 mg/ml G418 (Gibco).

Nuclear Extract preparation from mice tissue (detailed)

Mice were sacrificed by cervical dislocation. 25 ml homogenization buffer (2.2 M sucrose, 10 mM Hepes/ KOH pH 7.6, 15 mM KCl, 2 mM EDTA, 0.15 mM Spermine, 0.5 mM Spermidine, 1mM DTT, complete protease inhibitor w/o EDTA (Roche), 0.5 mM PMSF) was prepared. PBS was added to 1 liver, 3 brains or 6 kidneys to a total volume of 4 ml. 4 ml cushion buffer (2.05 M sucrose, 10 mM Hepes/ KOH pH 7.6, 15 mM KCl, 2 mM EDTA, 0.15 mM Spermine, 0.5 mM Spermidine, 1 mM DTT, complete protease inhibitor w/o EDTA (Roche), 0.5nmM PMSF) was added and solution was dounced until homogeneity and mixed with the residual homogenization buffer. Solution was stacked over 10 ml cushion buffer (2.05 M sucrose, 10 mM Hepes/ KOH pH 7.6, 15 mM KCl, 2 mM EDTA, 0.15 mM Spermine, 0.5 mM Spermidine, 1mM DTT, complete protease inhibitor w/o EDTA (Roche), 0.5mM PMSF) in a ultracentrifuge tube. Samples were centrifuged for 1 h at 4 °C at 24000 rpm in a SW28 rotor (Beckmann Coulter). Pellet containing nuclei was washed twice with 1 ml PBS (3000x*g*, 5 min at 4 °C), subsequently resuspended in 250 µl Buffer C (420 mM NaCl, 20 mM Hepes pH 7.9, 20% v/v glycerol, 2 mM MgCl2, 0.2 mM EDTA, 0.1% NP40, complete protease inhibitor w/o EDTA (Roche), 0.5 mM DTT) and incubated for 1 h on a rotating wheel at 4 °C. After centrifugation (1h, 16000x*g*, 4°C), supernatant was snap frozen. Several extract preparation for each organ were pooled to minimize variability.

Peptide pull-downs on 96-well plates (detailed)

Histone peptides containing the N-terminal 17 amino acids of the histone H3 tail followed by two glycines and a biotinylated lysine were synthesized using the Fmoc strategy as described [1]. An excess of peptide was coupled to sepharose streptavidin beads (GE Healthcare) in incubation buffer (150 mM NaCl, 50 mM Tris pH 8.0, 0.1% NP40) at room temperature for 1h. After extensive washing (150 mM NaCl, 50 mM tris pH 8.0, 1% NP40) beads were transferred 96 well Multi screen filter plates (Millipore, MSBVN1210). Per well 10 µl beads were used. Liquid was removed by slow centrifugation (30 seconds, 60x*g*). Nuclear extracts (400 µg total protein, pooled extracts) in 200 µl incubation buffer (150 mM NaCl, 50 mM Tris pH 8.0, 1% NP40, 0.5 mM DTT) were added and incubated for 3 h at 4 °C while gently shaking. Beads were washed 3 times (30 seconds, 60x*g*) with 200 µl wash buffer 1 (320 mM NaCl, 50 mM Tris pH 8.0, 0.5% NP40), followed by 5 washes with wash buffer 2 (150 mM NaCl, 50 mM Tris pH 8.0) to minimize residual detergent. We applied a modified in solution digest protocol on the column: 25 µl 2 M urea, 1mM DTT supplemented with 120 ng trypsin (Promega) was added and incubated for 30 min at room temperature and eluted, followed by two additional elution steps (50 µl 2 M urea, 5 mM iodoacetamide, 10 minutes incubation each). Proteins were digested over night at room temperature. Peptides were desalted by stage tipping with C18 material [2].

Sample preparation for nuclear extract proteome

20 µg total protein from each nuclear extract was precipitated with 4 volumes acetone, and resuspended in 20 µl 6 M Urea/2 M Thiourae in 50 mM ABC. Cysteines were reduced with 1 mM DTT for 30 minutes and alkylated with 5 mM iodoacetamide for 30 minutes. 500 ng LysC (Wako) was added and proteins were digested at room temperature for 3 hours. Samples were diluted 1:4 with 50 mM ABC and 500 ng trpysin (Promega) was added. Proteins were digested overnight. 20% per digest was desalted by stage tipping with C18 material [2].

LC-MS/MS analysis

Eluted peptides were analyzed by a nanoflow HPLC (Thermo Scientific, Odense) coupled on-line via a nano-electrospray ion source (Thermo Scientific, Odense) to a linear ion trap mass spectrometer (LTQ-Orbitrap Velos, Thermo Fisher Scientific, Germany) or a quadrupole-Orbitrap mass spectrometer (Q-Exactive, Thermo Fisher Scientific, Germany)[3]. Peptide mixtures were loaded with IntelliFlow at maximal 500 nl/min onto a C18-reversed phase column (18 cm long, 75 µm inner diameter, packed in-house with ReproSil-Pur C18-AQ 1.8 µm resin (Dr. Maisch GmbH)) in buffer A (0.5% acetic acid). Peptides were eluted with a multi-segment linear gradient of 5–60% buffer B (80% ACN and 0.5% acetic acid) at a constant flow rate of 250 nl/min over 180 min. The LTQ Orbitrap Velos was operated in the positive ion mode applying a data-dependent automatic switch between survey scan and tandem mass spectra (MS/MS) acquisition. A 'top 10' method was applied that acquires one Orbitrap survey scan in the range of m/z 300-1750 followed by MS/MS of the ten most intense ions in the LTQ. The target value in the LTQ-Orbitrap was 1,000,000 for survey scans at a resolution of 60,000 at m/z 400. Fragmentation in the LTQ was performed by collision-induced dissociation with a target value of 5,000 ions. The ion selection threshold was 500 counts. Selected sequenced ions were dynamically excluded for 90 seconds. The Q Exactive was also operated in the positive ion mode but using a data dependent top 5 method. Survey scans were acquired at a resolution of 70,000 at m/z 400. Up to the top 5 most abundant isotope patterns with charge ≥2 from the survey scan were selected with an isolation window of 2 Th and fragmented by HCD with normalized collision energies of 25. The maximum ion injection times for the survey scan and the MS/MS scans were 20 ms and 120 ms respectively and the ion target values were set to 3E6 and 1E6, respectively. Selected sequenced ions were dynamically excluded for 25 seconds.

MS Raw files and unfiltered proteingroups.txt, peptides.txt and evidence.txt can be downloaded from the TRANCHE repository (www.proteomecommons.com), using the following HASH keys:

YfIWcYmKH5GtbPP1VrHLtlS+e3ELsKgyZA3pXnvw1wkBc5uRmbTRpJkxLQjcNg7VFpJ2nlhSwZ78S3mPwIOmjJIg+B0AAAAAAABMng==

and

6pJlcJ77tmxxGGSf2yfhPExRZiZD1TnzLJVRiJPiSAeZTCS0RjFhisAYtcytzTgJrt6MEa33i4quLZNNh92wCLX

Data analysis

Raw mass spectrometric data were analyzed with the MaxQuant software (version 1.2.0.31 and 1.2.0.33) [4]. A false discovery rate (FDR) of 0.01 for proteins and peptides and a minimum peptide length of 6 amino acids were required. A time-dependent mass recalibration algorithm was used instead of lock masses for recalibration to improve the mass accuracy of precursor ions [5]. MS/MS spectra were searched by the Andromeda search engine which is incorporated into the MaxQuant software suite [6] against the IPI mouse data base (version 3.68, containing 56,729 entries) or IPI human data base (version 3.68, containing 87,061 entries) combined with 248 common contaminants and concatenated with the reversed versions of all sequences. For the search trypsin allowing for cleavage N-terminal to proline was chosen as enzyme specificity. Cysteine carbamidomethylation was selected as a fixed modification, while protein N-terminal acetylation and methionine oxidation were selected as variable modifications. Maximally two missed cleavages were allowed. Initial mass deviation for the precursor ion was up to 7 ppm, and maximum allowed mass deviation for fragment ions was 0.5 Da. Protein identification required two peptides one of which had to be unique to the protein group. Quantification in MaxQuant was performed using the built in label-free quantification algorithm [7], enabling the 'Match between runs' option (time window 2 minutes). SILAC pull-downs were analyzed using Significance B statistics as indicated in the figures.

Label-free pull-down experiments were analyzed with the freely available Perseus software, which includes all necessary functionalities (download available at http://www.**perseus**-framework.org/Perseus_1.3.0.4.zip). Proteingroups were filtered to require in at least one experimental group (all five replicates of a peptide – extract combination) five valid values and in addition at least 2 peptides (unique and razor) were required. Label free intensities were logarithmized and empty values were imputed with random numbers from a normal distribution, whose mean and standard deviation were chosen to best simulate low abundance values close to noise level. A modified *t*-test with permutation based FDR statistics [8, 9] was applied. We performed 250 permutations and required an FDR of 0.001. The parameter s_0 was empirically optimized to separate outliers from the background distribution. All peptide combinations within one tissue were tested separately for enrichment and repulsion and all entries that were significant in at least one test were kept for further analysis. Intensities were normalized by subtracting the mean in every row and hierarchically clustered by a correlation matrix. Proteins clustered into the different reader groups. Proteins either not clustering with a defined reader group or showing contradicting behavior in different tissues were removed In addition, proteins with a low number of peptides (generally below 5 in total) were manually inspected and in cases the additional evidence was not convincing were rejected (all proteins removed after initial testing can still be found in Table S1 in the panels "manually filtered").

Antibodies used in this study

For Western blotting the following antibodies were used: Chd5: M-182 (sc-68389, Santa Cruz), Kbtd7: N-15 (sc-84328, Santa Cruz), C11orf84 (HPA040128, Sigma), Mta2 (ab50209, abcam), HP1α (ab77256, abcam), Znf462 (HPA022283, Sigma), Set1 (ab70378, abcam), GFP (11 814 460 001, Roche), FOG2 (sc-10755, Santa Cruz). For Co-IP experiments the following control antibodies were used: Rabbit control antibody (sc2027, Santa Cruz), goat control antibody (sc2028, Santa Cruz)

CHD5 and CXXC1 Phd fingers

Human CHD5 Phd finger 1 (amino acids 339 - 393) and Phd finger 2 (amino acids 413 - 468) were cloned from OCABo5050B1010D (Imagenes), mouse CXXC1 Phd finger 1 (1-93) was cloned from IRAVp968H0261D (Imagenes), into pCoofy3 (a modified pETM33 vector). Proteins were expressed in BL21 in LB. Bacteria were lysed in lysis buffer (50 mM Tris pH 8.0, 150 mM NaCl, 0.1% NP40, 0.5 mM DTT, 0.1 mM $ZnSO_4$, 10% glycerol, 1x complete protease inhibitor without EDTA (Roche)) in the Fastprep machine (MP Biomedicals) using lysing matrix blue. Lysates were purified by incubation with glutathione magnetic beads (Pierce) for 1 h at 4°C washed with 150 mM NaCl, 50 mM Tris pH 8.0, 0.25% NP40 and eluted with 50 mM glutathione in lysis buffer. Eluates were dialyzed against lysis buffer over night at 4°C using Slide-A-Lyzer (Thermo Scientific) 3500 MWCO. Binding to histone peptides was tested by incubating 2 µl purified protein diluted in 50 µl lysis buffer with coupled beads for 2 h at 4°C. Beads were washed 3 times with 300 mM NaCl, Tris pH 8.0, 1% NP40 and bound protein was eluted by boiling in loading buffer. Western blots were probed for GST with goat polyclonal HRP conjugated antibody (abcam, ab 6649).

Cxxc1 full length and delta PHD (missing the first 93 amino acids) was TOPO cloned into pDEST47 (C-terminal GFP tag). Constructs were transfected into HEK293 cells, nuclear extracts were prepared as described above and binding to histone tail peptides was tested as described above.

Assignment of complexes and direct readers

To assign proteins specifically binding to histone tail peptides, we performed an extensive literature search. The following references were used: JARID1A [10], PHF23 [10], PHF2 [11], PHF13 (http://www.thesgc.org/structures/details?pdbid=3O7A/), PHF8 [12], TAF3 [13], KDM7 [14], ING proteins (here 1,3,4,5) [15], BPTF [16], DIDO1 [17], JMJD2A [18], SGF29 [19], SPIN1, MORC3 [20].

SUPPLEMENTARY REFERENCES

1. Schulze, W.X. and M. Mann, *A novel proteomic screen for peptide-protein interactions.* J Biol Chem, 2004. **279**(11): p. 10756-64.
2. Rappsilber, J., Y. Ishihama, and M. Mann, *Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics.* Anal Chem, 2003. **75**(3): p. 663-70.
3. Michalski, A., et al., *Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer.* Mol Cell Proteomics, 2011. **10**(9): p. M111 011015.
4. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
5. Cox, J., A. Michalski, and M. Mann, *Software lock mass by two-dimensional minimization of peptide mass errors.* J Am Soc Mass Spectrom, 2011. **22**(8): p. 1373-80.
6. Cox, J., et al., *Andromeda: a peptide search engine integrated into the MaxQuant environment.* J Proteome Res, 2011. **10**(4): p. 1794-805.
7. Luber, C.A., et al., *Quantitative proteomics reveals subset-specific viral recognition in dendritic cells.* Immunity, 2010. **32**(2): p. 279-89.
8. Hubner, N.C., et al., *Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions.* J Cell Biol, 2010. **189**(4): p. 739-54.
9. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 5116-21.
10. Wang, G.G., et al., *Haematopoietic malignancies caused by dysregulation of a chromatin-binding PHD finger.* Nature, 2009. **459**(7248): p. 847-51.
11. Wen, H., et al., *Recognition of histone H3K4 trimethylation by the plant homeodomain of PHF2 modulates histone demethylation.* J Biol Chem, 2010. **285**(13): p. 9322-6.
12. Feng, W., et al., *PHF8 activates transcription of rRNA genes through H3K4me3 binding and H3K9me1/2 demethylation.* Nat Struct Mol Biol, 2010. **17**(4): p. 445-50.
13. Vermeulen, M., et al., *Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4.* Cell, 2007. **131**(1): p. 58-69.
14. Horton, J.R., et al., *Enzymatic and structural insights for substrate specificity of a family of jumonji histone lysine demethylases.* Nat Struct Mol Biol, 2010. **17**(1): p. 38-43.
15. Shi, X., et al., *ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression.* Nature, 2006. **442**(7098): p. 96-9.
16. Li, H., et al., *Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF.* Nature, 2006. **442**(7098): p. 91-5.
17. Prieto, I., et al., *Synaptonemal complex assembly and H3K4Me3 demethylation determine DIDO3 localization in meiosis.* Chromosoma, 2009. **118**(5): p. 617-32.
18. Huang, Y., et al., *Recognition of histone H3 lysine-4 methylation by the double tudor domain of JMJD2A.* Science, 2006. **312**(5774): p. 748-51.
19. Vermeulen, M., et al., *Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers.* Cell, 2010. **142**(6): p. 967-80.

20. Li, X., et al., *Quantitative chemical proteomics approach to identify post-translational modification-mediated protein-protein interactions.* J Am Chem Soc, 2012. **134**(4): p. 1982-5.

SUPPLEMENTARY TABLES:
Supplementary tables are available at:
http://www.sciencedirect.com/science/article/pii/S1097276512009070

Table S1: related to Figure 1
Panel 1: Comparison of SILAC based and label free interaction proteomics
(related to Figure 1A-C)
All proteins that were significantly enriched in either the SILAC or the label-free peptide pull-downs from a mouse cell line. SILAC and label-free experiments were analyzed separately and merged using the IPI identifier. The columns Protein IDs (IPI identifier), Protein Names, Gene Names, Uniprot (Uniprot identifier) were retained for both experiments. NA indicates that this protein was not identified in the respective experiment. The column Proteins provides the number of IPI entries that were merged into the respective proteingroup. The columns Peptides, Razor and unique peptides and unique peptides lists the number of total peptides, razor and unique as well as only unique, respectively. Ratio.H.L.Normalized forward and reverse gives the SILAC ratios as computed and normalized by MaxQuant. Ratio.H.L.Normalized.reverse. Significance.B.SILAC and Ratio.H.L.Normalized.reverse.Significance.B.SILAC are the probabilities that this protein is an outlier in the distribution weighted by the intensity (see also [4] for details). The column t.test.labelfree gives the value for the standard t-test which is used for plotting. T.test.difference.labelfree is the difference between logarithmized labelfree intensities (corresponds to fold change). The column comment defines the overlap between labelfree and SILAC experiment and was used to generate Figure 1C.

Panel 2-5: Label-free interaction screen from mouse brain, kidney and liver nuclear extracts (related to Figure 1E)
All proteins that are significantly enriched in the label free screen for readers of H3K4me3 and H3K9me3 using mouse brain, kidney and liver nuclear extracts. Proteins are sorted into reader groups according to the hierarchical clustering as seen in Figure 1E.

Panels 2-4 contain the indictaed reader groups. Panel 5 contains proteins that were statistically significant, however were rejected in the further analysis.

This table contains Gene and Protein Names, IPI and Uniprot identifiers as well as number of peptides per proteingroup as already described for Table S1. In addition sequence coverage and posterior error probability (PEP) are added to judge confidence in the protein identification. The columns significant indicate in which tissue this protein was found to be significant (please note, that a protein can still be enriched from a tissue but not significantly thus making it not organ specific). The further columns contain the logarithmized and normalized labelfree intensities as used for generating the heatmap in Figure 1E. The name always consists of the tissue, the peptide and the replicate number separated by dashes.

Panel 6-9: Label-free interaction screen from mouse testis whole tissue extracts (related to Figure 1F)
All proteins that are significantly enriched in the label-free screen for readers of H3K4me3 and H3K9me3 using mouse testis whole tissue extracts. Proteins are sorted into reader groups according to the hierarchical clustering as seen in Figure 1F. This table contains the same columns as already described for Table S2.

Table S2: Interaction proteomics for ZMYND8, ZNF687 and RAI1
Related to Figure 3. All proteins identified and quantified in the forward and reverse GFP pull-downs for the tagged proteins. Proteins that were found to be enriched are indicated in yellow. Table contains identifiers and names for all proteins, numbers of peptides, sequence coverage and PEP as well as Normalized ratios for the forward and reverse experiment. The columns Ratio H/L count indicates the number of quantification events for the respective experiment.

Table S3: related to Figure 4A
All proteins that were found significant in the screen H3 unmodified versus H3K4me1 from mouse brain and liver nuclear extracts. Table contains the same columns as already described for Table S1.

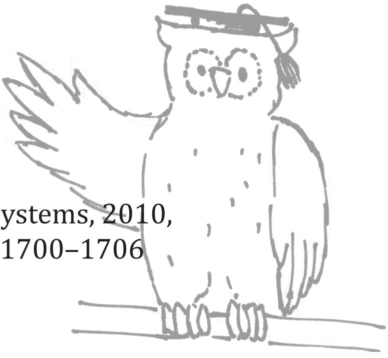
Table S4: related to Figure 4B
All proteins identified in the nuclear extract proteomes from mouse brain, liver and kidney. Proteins were filtered for at least 2 peptides (razor and unique). LFQ intensity is not logarithmized and normalized.

Chapter 5

CDK2AP1/DOC-1 is a *bona fide* subunit of the Mi-2/NuRD complex

Cornelia G. Spruijt[#a], Stefanie J. J. Bartels[#b], Arie B. Brinkman[b], Jorrit V. Tjeertes[$b], Ina Poser[c], Hendrik G. Stunnenberg[b] and Michiel Vermeulen[*a]

[a] Department of Physiological Chemistry and Cancer Genomics Centre, University Medical Center Utrecht, Utrecht, The Netherlands. E-mail: M.Vermeulen-3@umcutrecht.nl.
[b] Radboud University, Department of Molecular Biology, Nijmegen Centre for Molecular Life Sciences, Nijmegen, The Netherlands.
[c] Max Planck Institute for Molecular Cell Biology and Genetics, Dresden, Germany.
[#] These authors contributed equally.
[$] Present address: The Gurdon Institute, University of Cambridge, Cambridge, UK.



Molecular BioSystems, 2010,
Volume 6, p. 1700–1706

ABSTRACT

The Mi-2/NuRD (NUcleosome Remodeling and histone Deacetylase) chromatin remodeling complex is a large heterogeneous multiprotein complex associated with transcriptional repression. Here we apply a SILAC-based quantitative proteomics approach to show that all known Mi-2/NuRD complex subunits co-purify with Cyclin Dependent Kinase 2 Associated Protein1 (CDK2AP1), also known as Deleted in Oral Cancer 1 (DOC-1). DOC-1 displays *in vitro* binding affinity for methylated DNA as part of the meCpG binding MBD2/NuRD complex. In luciferase reporter assays, DOC-1 is a potent repressor of transcription. Finally, immunofluorescence experiments reveal co-localization between MBD2 and DOC-1 in mouse NIH-3T3 nuclei. Collectively, these results indicate that DOC-1 is a *bona fide* subunit of the Mi-2/NuRD chromatin remodeling complex.

INTRODUCTION

In eukaryotic cells, DNA is packed in a structural polymer called chromatin. Nucleosomes form the fundamental building blocks of chromatin and in general these nucleosomes are inhibitory to processes that require access to the DNA template, such as transcription and DNA repair. During the last two decades many protein complexes have been identified and characterized that use ATP hydrolysis to alter the position of nucleosomes on DNA. In doing so, these protein complexes can regulate the accessibility of transcription factors or repair proteins to DNA [1, 2]. One of these ATP dependent chromatin remodeling complexes is the Mi-2/NuRD complex (NUcleosome Remodeling and histone Deacetylase complex). This complex was biochemically purified by a number of labs more than a decade ago [3-5]. The two highly homologous proteins CHD3 and CHD4 (or Mi-2α and Mi-2β) represent the catalytic ATP hydrolyzing subunits in the complex. In addition, the complex contains two histone deacetylases, HDAC1 and HDAC2, RbAp48 and RbAp46, MTA1-3, p66α and β and MBD2 or MBD3. MBD2 and MBD3 were first described as common subunits within the NuRD complex [6] but our subsequent study revealed that MBD2 and MBD3 each assemble into a Mi-2/NuRD-like complex in a mutually exclusive manner [7]. MBD2, unlike MBD3, binds to methyl-CpG residues and it has been proposed that this protein forms the link between the MBD2/NuRD complex and transcriptionally silent CpG methylated promoters. In addition to the reported (core) subunits, a number of transcription factors have been shown to interact with the Mi-2/NuRD complex [8-14]. These transcription factors could serve to recruit the Mi-2/NuRD complex to specific loci in the genome.

Previously, we identified DOC-1 (Deleted in Oral Cancer-1) peptides in MBD2/NuRD and MBD3/NuRD complex purifications [7], indicating that this protein may be an interactor or a novel subunit of the Mi-2/NuRD complex. DOC-1 was first described as a protein that is commonly mutated or deleted in various malignancies [15, 16]. In addition, DOC-1 has been characterized as a Cyclin Dependent Kinase 2 Associated Protein (CDK2AP1) [17]. In this study it was shown that over-expression of DOC-1 in 293T cells results in a G1 arrest and significant growth retardation compared to wild-type cells consistent with loss of the protein in tumors. Recently, interactions between MBD3 and DOC-1 were shown by co-immunoprecipitation and western blot analyses [18]. However, convincing evidence that DOC-1 is a general Mi-2/NuRD interactor or a core subunit of the complex is still lacking [19]. Using a variety of biochemical and functional experiments, we here show that DOC-1 is indeed a *bona fide* subunit of the MBD2/NuRD and MBD3/NuRD complexes.

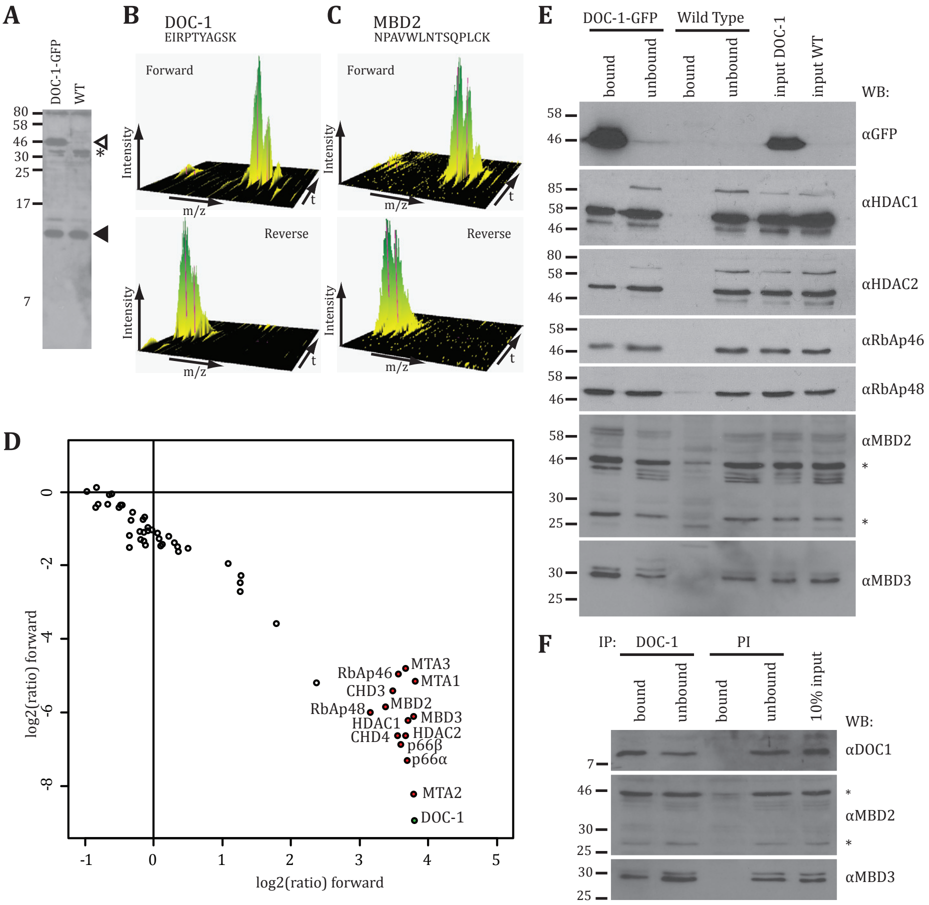


Figure 1: DOC-1 is a subunit of the Mi-2/NuRD complex. A. Nuclear extracts from DOC-1-GFP and wild-type HeLa cells were analyzed by western blotting using a DOC-1 antibody. Endogenous DOC-1 and DOC-1-GFP are indicated by ◀ and ◁, respectively. Note that the signal intensities for endogenous DOC-1 and DOC-1-GFP are about equal, indicating that DOC-1-GFP is expressed at roughly endogenous levels. The asterisk indicates antibody cross-reactivity. **B and C**. Three dimensional representations (m/z= x-axis, chromatographic retention time= y-axis and MS intensity= z-axis) of MS signals from DOC-1 (**B**) and MBD2 (**C**) peptides that were obtained in the forward (upper spectra) and reverse (lower spectra) DOC-1-GFP pull-downs. The indicated MBD2 peptide shows a high ratio in the forward pull-down and a low ratio in the reverse pull-down, indicating that MBD2 specifically interacts with DOC-1-GFP. **D**. Ratio versus ratio plot of all the proteins that were identified and quantified with at least two peptides in the DOC-1-GFP pull-downs. In this plot background proteins appear around the centre of the axes with ratios close to 1 in both the forward and the reverse pull-down. In contrast, DOC-1-GFP and associated proteins show a high ratio in the forward pull-down and a low ratio in the reverse pull-down and therefore cluster together in the bottom right quadrant of the graph. Note that all the identified DOC-1-GFP interacting proteins are known subunits of the Mi-2/NuRD complex. **E**. Nuclear extracts from DOC-1-GFP and wild-type HeLa cells were subjected to GFP pull-downs using GFP nanotrap beads and tested for the presence of the indicated proteins by western blotting. The eluates from the beads as well as 12.5% of the non-bound fraction and 10% of input extract was loaded on gel. Asterisks indicate MBD2a and b. **F**. Endogenous DOC-1 was immunoprecipitated from HeLa nuclear extracts using a DOC-1 antibody. Immunoprecipitates were tested for the presence of DOC-1, MBD2 and MBD3. PI= immunoprecipitation using pre-immune serum.

RESULTS

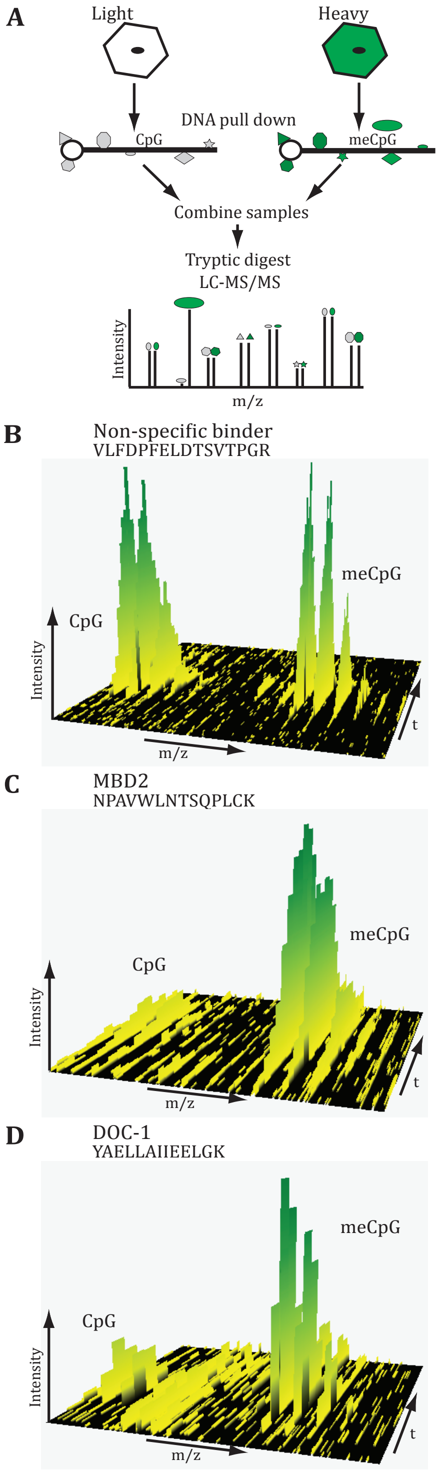
DOC-1 exclusively associates with Mi-2/NuRD complex subunits in the nucleus

To investigate a putative interaction between DOC-1 and the Mi-2/NuRD complex, we tagged and purified DOC-1 from human cells. We made use of the recently developed BAC-transgeneOmics approach [20] to obtain a HeLa cell line expressing DOC-1-GFP from its own promoter at endogenous levels (Figure 1A).

This cell line and wild-type HeLa cells were SILAC labeled 'heavy' and 'light', respectively, subjected to single step affinity purification on GFP-nanotrap beads [21] after which bound proteins were digested with LysC and measured in a single LC-MS run on an LTQ-Orbitrap mass spectrometer. Computational analysis of the data was done using the MaxQuant software [22]. In this approach, GFP-tagged proteins and proteins interacting with the bait are more abundant in the heavy compared to the light form and can therefore easily be distinguished from background binders that have a one to one ratio. As a control, a 'label swap' experiment is performed in which the GFP-tagged cell line is labeled light and the wild-type cells labeled heavy. In this case, the bait and associated proteins have a low heavy/light ratio. Plotting ratios of the 'forward' experiment against ratios of the 'reverse' experiment results in four quadrants in which the GFP-tagged protein and its interactors cluster together in a single quadrant. As expected, DOC-1-GFP derived peptides had a high ratio in the forward and a low ratio in the reverse pull-down, indicating that the bait protein was specifically enriched in both pull-downs consistent with the SILAC labeling scheme (Figure 1B). MBD2 derived peptides showed a similar pattern, indicating that MBD2 is a DOC-1-GFP interacting protein in this experiment (Figure 1C). A ratio vs ratio plot of all the proteins that were identified and quantified in the pull-downs revealed that DOC-1-GFP interacts specifically with essentially all Mi-2/NuRD complex subunits that have been described in the literature to date, including both MBD2 and MBD3 (Figure 1D and Supplementary Table 1). To further validate these findings we used nuclear extracts derived from DOC-1-GFP cells for pull-downs with GFP-nanotrap beads, which were then tested for the presence of Mi-2/NuRD complex subunits using western blotting (Figure 1E). Consistent with our mass spectrometry data, all the Mi-2/NuRD complex subunits we tested were specifically enriched on DOC-1-GFP containing beads, whereas no enrichment could be observed on beads that were incubated with wild-type HeLa nuclear extract. Finally, to study the interaction between endogenous DOC-1 and MBD2/MBD3 we used an antibody against DOC-1 to precipitate the protein from HeLa nuclear extract (Figure 1F, upper panel). MBD2 and MBD3 were specifically co-immunoprecipitated with endogenous DOC-1 (Figure 1F, middle and lower panel). Taken together, these experiments show that DOC-1 interacts with the MBD2/NuRD and MBD3/NuRD complexes. Furthermore, no additional protein-protein interactions could be detected for DOC-1 in HeLa nuclear extracts by mass spectrometry, indicating that, at least in mammalian nuclei, the protein is primarily and exclusively associated with the Mi-2/NuRD complex.

DOC-1 and MBD2 specifically interact with methylated CpGs *in vitro*

DOC-1 is a small (115 aa) protein that does not carry an obvious methyl-CpG binding motif. However, our biochemical data now clearly indicate that DOC-1 is part of the MBD2/NuRD complex. We therefore hypothesized that DOC-1 would indirectly bind to methylated DNA via an interaction with the MBD2/NuRD complex. To address



this question we applied a DNA pull-down approach in combination with SILAC-based quantitative proteomics (Figure 2A) [23, 24]. Methylated and non-methylated DNA bound to beads was incubated with heavy or light SILAC labeled U937 nuclear extracts, respectively. Following the pull-down and washes, beads from both pull-downs were combined and bound proteins were separated by one dimensional SDS PAGE. Proteins were subsequently digested with trypsin and peptide mixtures were measured by high-resolution LC-MS on an LTQ-Orbitrap hybrid mass spectrometer. Proteins that interact with DNA irrespective of DNA methylation or bind non-specifically to the beads are equally abundant in the light and heavy state and these proteins therefore show a one to one ratio in the mass spectrometer (Figure 2B). In contrast, proteins specifically interacting with the meCpGs are more abundant in the heavy form and have a heavy/light ratio higher than one. As a validation of the approach and consistent with previous observations, MBD2, one of the five "classic" proteins containing a meCpG binding domain (MBD) [25], was identified as a specific meCpG binding protein in our quantitative DNA pull-down experiment (Figure 2C and Supplementary table 2). In agreement with our hypothesis we also identified DOC-1 as a meCpG interactor in our pull-down (Figure 2D and Supplementary table 2). MBD3, which does not bind specifically to

Figure 2: DOC-1 and MBD2 are specifically recovered on methylated DNA *in vitro.* **A**. Schematic representation of the experimental approach. **B-D**. SILAC labeled nuclear extracts from U937 cells were incubated with non-methylated and methylated DNA immobilized on streptavidin conjugated dynabeads. Shown in the figures is the three dimensional representation of the MS signal for the indicated peptides and their relative binding to methylated versus non-methylated DNA. Note that for a background protein equal binding to methylated versus non-methylated DNA is observed (**B**), whereas for MBD2 (**C**) and DOC-1 (**D**), preferential methyl-DNA binding is observed.

methylated DNA [7, 25] was not identified in this experiment. These results indicate that DOC-1 binds to methylated DNA *in vitro* as part of the MBD2/NuRD complex.

DOC-1 is a repressor of transcription and co-localizes with MBD2 *in vivo*

To further functionally characterize the DOC-1 protein, we performed luciferase reporter gene assays using a DOC-1-Gal4 containing construct (Figure 3A). An expression construct containing the Gal4 DNA binding domain was used as a control and reveals the basal activity of the luciferase gene. Gal4-MBD2 and Gal4-Ash2L constructs were used as additional controls for repressive and activating activities, respectively. Consistent with previous observations and in line with its role in activation of transcription [26], Gal4-Ash2L potentiated reporter gene activity [27]. In contrast to this and in agreement with its known biological function, Gal4-MBD2 was a repressor of transcription in this experimental set-up (Figure 3B). DOC-1-Gal4 also conferred repression to the reporter gene in a dose dependent manner comparable to Gal4-MBD2, indicating that in this luciferase reporter assay, DOC-1 is a potent repressor of transcription.

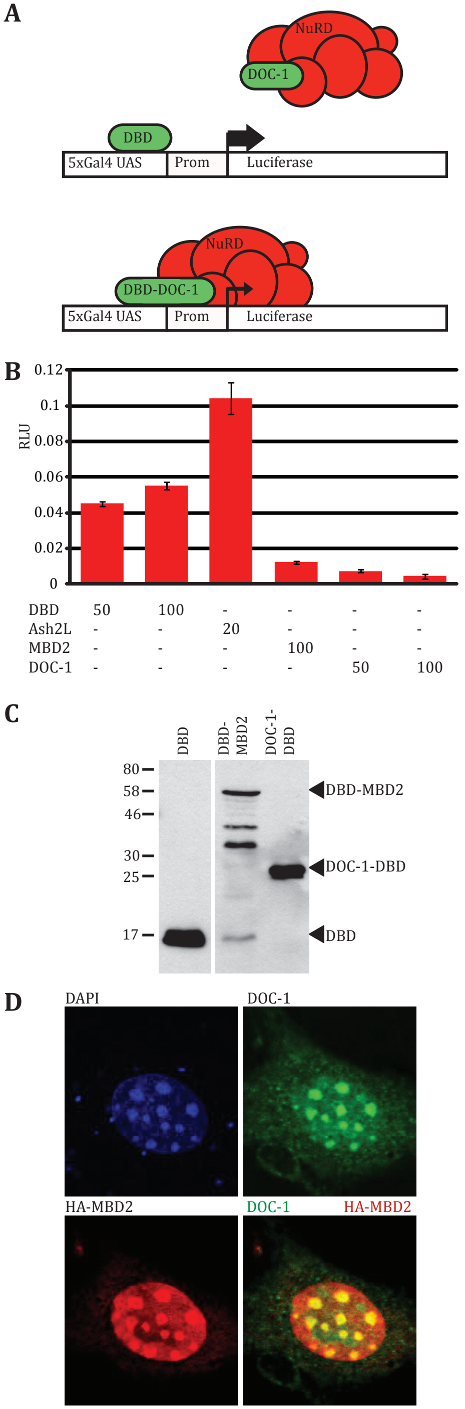


Figure 3: DOC-1 is a repressor of transcription and co-localizes with MBD2 *in vivo*. **A**. Schematic representation of the Gal4-luciferase assay. Gal4-DBD (DNA binding domain) binds to the Upstream Activating Sequence (UAS) in front of the reporter gene. Gal4-fusion proteins and their associated proteins can therefore be recruited to the TK promoter to exert their function. **B**. Luciferase reporter gene assays with the indicated constructs. Transfection amounts are given in nanograms. **C**. Anti-Gal4 western blot to confirm expression of the Gal4 fusion proteins that were used in figure 3B. 500 ng of the indicated constructs were transfected into 293T cells. **D**. Confocal microscopy of NIH-3T3 cells that were transiently transfected with HA-MBD2 reveals co-localization of endogenous DOC-1 and HA-MBD2 in the nucleus.

To further study the DOC-1/Mi-2-NuRD interaction *in vivo* we performed immunofluorescence experiments. Mouse NIH-3T3 cells were transfected with an HA-tagged MBD2 construct, and a combination of a mouse monoclonal HA antibody with a rabbit polyclonal antibody against endogenous DOC-1 was used to visualize the proteins in the cells. Both proteins were predominantly found in DNA dense regions in the nucleus and showed substantial overlap, indicating that MBD2 and DOC-1 co-localize in mammalian nuclei *in vivo*. These DNA dense regions in the mouse nuclei are known to be enriched for major and minor satellite repeats that are heterochromatic and transcriptionally silent [28]. It should be noted that DOC-1 interacts with both MBD2/NuRD and MBD3/NuRD, which are two distinct complexes. This may explain why MBD2 and DOC-1 do not co-localize completely.

Collectively, the biochemical and cell biological assays presented in this paper reveal that CDK2AP1/DOC-1 is a subunit of the Mi-2/NuRD complex and a repressor of transcription.

DISCUSSION

In this paper we have provided compelling evidence that CDK2AP1/DOC-1 is a *bona fide* subunit of the Mi-2/NuRD complex. DOC-1 was first reported as a protein that is deleted in oral cancer and was subsequently described as a cyclin dependent kinase 2 associated protein. Our quantitative mass spectrometry data did not reveal an association between CDK2 and DOC-1 in nuclear extracts. Although we cannot exclude that DOC-1 interacts with CDK2 in specific physiological conditions in the cytoplasm, western blotting as well as confocal microscopy revealed that DOC-1 is predominantly nuclear and our quantitative mass spectrometry data show that in the nucleus it exclusively associates with the Mi-2/NuRD complex. Interestingly, deletion of the Mi-2/NuRD subunit MBD2 in mice protects these mice from intestinal tumors [29]. This is in contrast to the pathology of DOC-1; reduced DOC-1 expression appears to be an inducer of malignant transformation [16, 30]. In agreement with this, over-expression of DOC-1 in 293T cells results in a partial G1/S arrest [17], whereas over-expression of MBD2 in 293T cells enhances cell proliferation (Xavier Le Guezennec and M.V, unpublished data). Whether these observations are indicative of antagonistic functions for DOC-1 and MBD2 in the Mi-2/NuRD complex or hinting towards a cytoplasmic DOC-1 function related to CDK2 remains unclear at this point. To further study the potential interplay between DOC-1 and MBD2 in tumorigenesis it would be informative to cross MBD2 deficient mice with a DOC-1 knock-out strain and look at survival rates in polyposis challenge experiments. Alternatively, immortalized MBD2 deficient MEFs could be subjected to DOC-1 siRNA in colony formation assays to look at their proliferation.

Given its apparent general presence in both the MBD2/NuRD and MBD3/NuRD complexes, it is surprising that DOC-1 has not been identified by mass spectrometry previously in Mi-2/NuRD complex purifications. However, given the small size of DOC-1, the protein was not visualized by silver or coomassie stainings prior to LC-MS/MS analyses in a number of studies [3-6] and therefore may have escaped identification. Although our study has clearly established DOC-1 as a Mi-2/NuRD subunit, future research is required to elucidate the molecular function of the protein within the complex, its putative association with methylated promoters as a component of the MBD2/NuRD complex and its link to carcinogenesis.

MATERIALS AND METHODS

Cell culture

HeLa Kyoto, NIH-3T3 and HEK 293 cells were cultured in DMEM containing 10% Fetal Calf Serum, 2mM Glutamine and 100 U/mL of Penicillin/Streptomycin (BioWhittaker), whereas U937 cells were cultured similarly in RPMI. The DOC-1-GFP BAC line was cultured in the presence of 400 µg/mL geneticin (G418) (Life Technologies/Gibco). For SILAC labeling experiments, DOC-1-GFP, wild-type HeLa and wild-type U937 cells were cultured in the presence of light and heavy lysine ($^{13}C_6{}^{15}N_2$, Isotec) (GFP pull-down) or light and heavy lysine and arginine ($^{13}C_6{}^{15}N_2$ and $^{13}C_6{}^{15}N_4$, Isotec) (DNA pull-down) for > 8 doublings to ensure full incorporation of the heavy isotope prior to preparation of nuclear extracts.

GFP-pull downs

Nuclear extracts (prepared essentially as described in [31]) derived from DOC-1-GFP and wild-type HeLa cells (200-300 µg for western blot analyses and 1 mg for mass spectrometric analysis) were incubated with 10 µl of GFP nanotrap beads (Chromotek) for 90 minutes at 4°C in binding buffer (PBS, 0.25% NP40, 0.5 mM DTT, 50 µg/ml ethidium bromide and complete protease inhibitors –EDTA (Roche)). Beads were washed extensively with binding buffer after which proteins were eluted using SDS PAGE loading buffer for western blot analyses or acidic glycine (0.1 M, pH 2.0) for subsequent mass spec analyses. The following antibodies were used for western blotting: MBD2 (Everest Biotech, EB07538); MBD3 (IBL, 3A3); RbAp46 (Abcam, 72457-100); RbAp48 (Abcam, 74188-100); HDAC1 (Santa Cruz Biotechnology, H51 sc-7872); HDAC2 (Santa Cruz Biotechnology, ACII sc-7899 -54); Gal4-DBD (Santa Cruz Biotechnology, RK5C1); GFP (Roche, 11814460001). A rabbit polyclonal antibody against recombinant full length DOC-1 was generated in-house.

Generation of anti-DOC-1 antibodies

A GST-DOC-1 fusion construct was created by ligating a DOC-1 cDNA-clone (IRATp970A0640D, RZPD-clone) into pGEX-2T. The DOC-1 cDNA was PCR-amplified using the following oligos: 5'CGCggatccATGTCTTACAAACCGAACTTGGC3' (forward) and 5'CCGgaattcCTAGGATCTGGCATTCCGTTC3' (reverse). The amplified product was ligated into pGEX-2T using BamHI/EcoRI restriction sites. GST-DOC-1 protein was produced in E.coli BL21 (DE3) and purified using Glutathione Sepharose 4B-beads (GE Healthcare) according to standard procedures. The GST-tag was removed by thrombin cleavage and DOC-1 was subsequently isolated from preparative SDS PAGE gel and used for immunization of rabbits.

Co-immunoprecipitation

2 µl of DOC-1 antiserum or 2 µl pre-immune serum was immobilized on 30 µl protein A Dynabeads slurry (Invitrogen). Beads were then incubated with 50 µl HeLa nuclear extract (~ 5 mg/ml) in 150 µl binding buffer for 2.5 hours at 4°C. Beads were washed extensively with binding buffer after which bound proteins were eluted in SDS PAGE loading buffer and analyzed by western blotting for the presence of DOC-1, MBD2 and MBD3.

DNA pull-down

The following oligos were used for preparation of pull-down DNA: 5’aagcagacactggcaggtttCGGCGGGAGTCCGCGGGACCCTCCAGAAGAGCGGCCGG-CGCCGTGACctaaggctaaggctcata3’ (forward) and 5’tttatgagccttagccttagGTCACGGCGCCGGCCGCTCTTCTGGAGGGTCCCGCGGACTCCCGCCGaaacctgccagtgtctgc3’ (reverse), containing a sequence derived from the *GSTP1* CpG-island (in capitals), sites for primer annealing, and a methylation-sensitive restriction site (underlined). PAGE purified oligos were annealed, phosphorylated and ligated, resulting in fragments with lengths ranging from 85 to 600 bp. Subsequently, biotinylation was performed by incorporation of biotin-14-dATP (Invitrogen) at the 3’end of the forward strand using Klenow Fragment (3’-5’exo-) (New England Biolabs). For the meCpG pull-down, DNA was methylated by M.SssI (New England Biolabs) and methylation was checked by methylation-sensitive digestion followed by quantitative PCR. 75 µl of Dynabeads MyOne Streptavidin C1 (Invitrogen) were incubated with 10 µg of DNA for 1h at RT in DNA binding buffer (150 mM NaCl, 50 mM Tris pH 8.0, 0.1% NP40). After washing, the beads with coupled DNA were incubated with 400 µg U937 nuclear extract and 10 µg poly(dI-dC) competitor DNA (Sigma) for 2h at 4°C in protein binding buffer (150 mM NaCl, 50 mM Tris pH 8.0, 0.25% NP40, 0.5 mM DTT, and complete protease inhibitors –EDTA (Roche)). Beads were washed extensively and bound proteins were eluted in SDS PAGE loading buffer and processed for mass spec analyses.

Mass spectrometry

Proteins eluted from the GFP-nanotrap beads were neutralized using Tris (pH 8.5) and subsequently digested with LysC (Wako) using the FASP protocol [32]. Proteins precipitated during the DNA pull-down were separated by SDS PAGE and subjected to in-gel trypsin digestion as described [27]. Collected peptides were desalted using StageTips [33] and measured on an LTQ-Orbitrap mass analyzer essentially as described [27]. Raw data were analyzed using the MaxQuant software package [22]. The DOC-1-GFP pull-down ratio vs ratio plot was generated using the open software package R.

Cloning

To generate a DOC-1-Gal4-DBD construct, the stop-codon between the HindIII cleavage site and the transcription start site in plasmid pCMV-DBD [34] was mutated into a glycine codon using primers 5’CCAAGCTTCCGGAAAGATGAAGC3’ (forward) and 5’AGGTGACACTATA3’ (reverse). The point mutation in the forward primer is underlined. The PCR product was ligated into the backbone vector. Full-length DOC-1 was amplified from a pCMX-DBD vector using primers (5’CCCAAGCTTATGTCTTACAAACCGAACTTG3’) and (5’ CCCAAGCTTGGGATCTGGCATTCCGTTCC3’). This fragment was then ligated into the mutated CMV-DBD vector, to obtain a C-terminal Gal4-DBD-fusion.

Confocal Immunofluorescence Microscopy

NIH-3T3 cells were seeded on coverslips in 12-well plates. At ~40% confluency, cells were transfected with 1 µg stII-3HA-MBD2 plasmid [7] using PEI (Polysciences). At ~80% confluency cells were fixed with 4% v/v paraformaldehyde. Permeabilization was performed by incubation with 0.2% Triton X-100 in PBS for 5 minutes at RT. Cells were then blocked with 1% Bovine Serum Albumin (Sigma) in PBS supplemented with 0.1% Triton X-100 for 30 minutes and subsequently incubated with the primary antibodies (DOC-1 and HA, (Santa Cruz Biotechnology, 12xA5)) in blocking buffer for at least 1 hour. This was followed by incubation with secondary antibodies (GαR Alexa 488 and GαM Alexa568 (Invitrogen)) for 1 hour in blocking buffer. DNA was stained using 10 µg/mL DAPI (4'-6-Diamidino-2-phenylindole). A Zeiss 510 Meta confocal microscope with a 63X/1.4 Oil DIC Plan-ApoChromat objective was used for microscopic analysis.

Luciferase assay

HEK293 cells were seeded in 12-well plates on day 1, transfected on day 2 (when confluency was ~30-40%). Transfection was done in triplicates, using 1.5 µl Fugene6 reagent (Roche), 15 ng pCMV-Renilla, 200 ng Gal4-TK-luciferase and 50/100 ng pCMV DBD; 50/100 ng pCMV DOC-1-DBD; 100 ng pCMX DBD-MBD2 or 20 ng pGal4-Ash2L per well. Cells were lysed by applying 150 µl 1x Passive lysis buffer (Dual-luciferase assay kit (Promega)) and incubation for 20 min at RT. 50 µl lysate was used for measurement in a 96-well Berthold LB96V MicroLumat Plus luminometer.

ACKNOWLEDGEMENTS

We thank Radhika Warrier for advice on luciferase reporter assays and Dr. Petra de Graaf and Radhika Warrier for help with microscopy. Furthermore, we thank Xavier Le Guezennec for critical reading of the manuscript, Luuk te Riet for assistance and Matthias Mann and Tony Hyman for their support. This research was supported by a grant from the Dutch Cancer Foundation (KWF, KUN 2008-4130). The Vermeulen lab is supported by grants from the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research (NWO) and NWO/VIDI. Work in the laboratory of Tony Hyman was funded by the Max Planck Society and the Bundesministerium für Bildung und Forschung grand DiGtoP [01GS0859].

SUPPLEMENTAL INFORMATION

Two supplemental tables are available with this chapter at:
http://pubs.rsc.org/en/Content/ArticleLanding/2010/MB/c004108d#!divAbstract

REFERENCES

1. Becker, P.B. and W. Horz, *ATP-dependent nucleosome remodeling.* Annu Rev Biochem, 2002. **71**: p. 247-73.
2. Kittler, R., et al., *Genome-scale RNAi profiling of cell division in human tissue culture cells.* Nat Cell Biol, 2007. **9**(12): p. 1401-12.
3. Wade, P.A., et al., *A multiple subunit Mi-2 histone deacetylase from Xenopus laevis cofractionates with an associated Snf2 superfamily ATPase.* Curr Biol, 1998. **8**(14): p. 843-6.
4. Zhang, Y., et al., *The dermatomyositis-specific autoantigen Mi2 is a component of a complex containing histone deacetylase and nucleosome remodeling activities.* Cell, 1998. **95**(2): p. 279-89.
5. Tong, J.K., et al., *Chromatin deacetylation by an ATP-dependent nucleosome remodelling complex.* Nature, 1998. **395**(6705): p. 917-21.
6. Feng, Q. and Y. Zhang, *The MeCP1 complex represses transcription through preferential binding, remodeling, and deacetylating methylated nucleosomes.* Genes Dev, 2001. **15**(7): p. 827-32.
7. Le Guezennec, X., et al., *MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties.* Mol Cell Biol, 2006. **26**(3): p. 843-51.
8. Denslow, S.A. and P.A. Wade, *The human Mi-2/NuRD complex and gene regulation.* Oncogene, 2007. **26**(37): p. 5433-8.
9. Wang, Y., et al., *LSD1 is a subunit of the NuRD complex and targets the metastasis programs in breast cancer.* Cell, 2009. **138**(4): p. 660-72.
10. Kim, J., et al., *Ikaros DNA-binding proteins direct formation of chromatin remodeling complexes in lymphocytes.* Immunity, 1999. **10**(3): p. 345-55.
11. Lauberth, S.M. and M. Rauchman, *A conserved 12-amino acid motif in Sall1 recruits the nucleosome remodeling and deacetylase corepressor complex.* J Biol Chem, 2006. **281**(33): p. 23922-31.
12. Hong, W., et al., *FOG-1 recruits the NuRD repressor complex to mediate transcriptional repression by GATA-1.* EMBO J, 2005. **24**(13): p. 2367-78.
13. Tan, C.P. and S. Nakielny, *Control of the DNA methylation system component MBD2 by protein arginine methylation.* Mol Cell Biol, 2006. **26**(19): p. 7224-35.
14. Kehle, J., et al., *dMi-2, a hunchback-interacting protein that functions in polycomb repression.* Science, 1998. **282**(5395): p. 1897-900.
15. Todd, R., et al., *Deleted in oral cancer-1 (doc-1), a novel oral tumor suppressor gene.* FASEB J, 1995. **9**(13): p. 1362-70.
16. Yuan, Z., T. Sotsky Kent, and T.K. Weber, *Differential expression of DOC-1 in microsatellite-unstable human colorectal cancer.* Oncogene, 2003. **22**(40): p. 6304-10.
17. Shintani, S., et al., *p12(DOC-1) is a novel cyclin-dependent kinase 2-associated protein.* Mol Cell Biol, 2000. **20**(17): p. 6300-7.
18. Deshpande, A.M., et al., *Cdk2ap1 is required for epigenetic silencing of Oct4 during murine embryonic stem cell differentiation.* J Biol Chem, 2009. **284**(10): p. 6043-7.
19. Bao, Y. and X. Shen, *SnapShot: chromatin remodeling complexes.* Cell, 2007. **129**(3): p. 632.
20. Poser, I., et al., *BAC TransgeneOmics: a high-throughput method for exploration of protein function in mammals.* Nat Methods, 2008. **5**(5): p. 409-15.

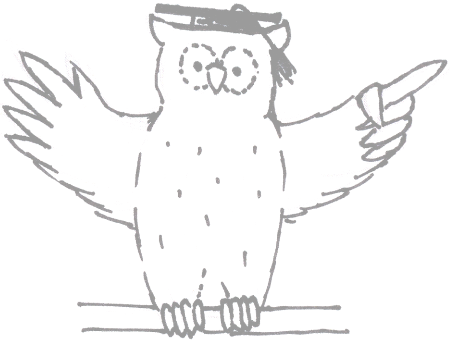
21. Rothbauer, U., et al., *A versatile nanotrap for biochemical and functional studies with fluorescent fusion proteins.* Mol Cell Proteomics, 2008. **7**(2): p. 282-9.
22. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
23. Mittler, G., F. Butter, and M. Mann, *A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements.* Genome Res, 2009. **19**(2): p. 284-93.
24. Butter, F., et al., *A domesticated transposon mediates the effects of a single-nucleotide polymorphism responsible for enhanced muscle growth.* EMBO Rep, 2010.
25. Hendrich, B. and A. Bird, *Identification and characterization of a family of mammalian methyl-CpG binding proteins.* Mol Cell Biol, 1998. **18**(11): p. 6538-47.
26. Shilatifard, A., *Molecular implementation and physiological roles for histone H3 lysine 4 (H3K4) methylation.* Curr Opin Cell Biol, 2008. **20**(3): p. 341-8.
27. Vermeulen, M., et al., *Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4.* Cell, 2007. **131**(1): p. 58-69.
28. Guenatri, M., et al., *Mouse centric and pericentric satellite repeats form distinct functional heterochromatin.* J Cell Biol, 2004. **166**(4): p. 493-505.
29. Sansom, O.J., et al., *Deficiency of Mbd2 suppresses intestinal tumorigenesis.* Nat Genet, 2003. **34**(2): p. 145-7.
30. Choi, M.G., et al., *Decreased expression of p12 is associated with more advanced tumor invasion in human gastric cancer tissues.* Eur Surg Res, 2009. **42**(4): p. 223-9.
31. Dignam, J.D., R.M. Lebovitz, and R.G. Roeder, *Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei.* Nucleic Acids Res, 1983. **11**(5): p. 1475-89.
32. Wisniewski, J.R., et al., *Universal sample preparation method for proteome analysis.* Nat Methods, 2009. **6**(5): p. 359-62.
33. Rappsilber, J., M. Mann, and Y. Ishihama, *Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips.* Nat Protoc, 2007. **2**(8): p. 1896-906.
34. Zwartjes, C.G., et al., *Repression of promoter activity by CNOT2, a subunit of the transcription regulatory Ccr4-not complex.* J Biol Chem, 2004. **279**(12): p. 10848-54.

Chapter 6

ZMYND8 interacts with NuRD through its MYND domain and co-localizes with MBD3 on a subset of its target genes

Cornelia G. Spruijt[1], Roberta Menafra[2], Moritz C. Voelker-Albert[1,3], Pascal W.T.C. Jansen[2], Raghu Edupuganti[2], Marijke Baltissen[2], Anneloes Mensinga[1], Hendrik G. Stunnenberg[2] and Michiel Vermeulen[1,2]

1. MCR, UMC Utrecht, Utrecht, The Netherlands.
2. RIMLS, Nijmegen, The Netherlands.
3. current address: Zentrallabor für Proteinanalytik, Ludwig-Maximilians-Universität München (LMU) , Munich, Germany



ABSTRACT

The MBD2/NuRD complex, which is generally known to be associated with repression of transcription, has been shown to occupy genomic loci containing high levels of DNA methylation. In addition, a small cluster of NuRD-bound loci have low levels of DNA methylation. The binding to methylated loci can be explained by the presence of the methyl-CpG binding MBD2 protein in the complex. How the NuRD complex is recruited to non-methylated loci, however, remains unclear. Here, we use quantitative mass spectrometry-based proteomics to show that the ZMYND8 protein bridges the NuRD complex to a number of putative DNA binding zinc finger proteins, ZNF687, ZNF592 and ZNF532. ZMYND8 most likely directly interacts with the NuRD complex subunit GATAD2A, which is mutually exclusive with GATAD2B, through its conserved MYND domain. Furthermore, ChIP-sequencing analyses reveal that ZMYND8 and MBD3 share a subset of genome wide binding sites, which mostly map to active promoters and enhancers. We thus hypothesize that the ZMYND8/ZNF module recruits the GATAD2A/NuRD complex to a subset of hypomethylated, transcriptionally active target sites in the genome.

INTRODUCTION

The Nucleosome Remodeling and histone Deacetylase complex (NuRD) is a highly conserved chromatin remodeling complex which is generally known to be associated with transcriptional repression. This multi-subunit complex contains two catalytic activities: the ATP dependent chromatin remodeling enzymes CHD3, CHD4, and CHD5 [1, 2] and the histone deacetylases HDAC1 and HDAC2. In addition to these catalytic activities, NuRD contains either MBD2 or MBD3, which are proteins containing a Methyl-CpG Binding Domain. However, multiple studies have shown that mammalian MBD3 lacks methyl-C binding affinity [3, 4]. Additional subunits of the NuRD complex are GATAD2A and GATAD2B; MTA1, MTA2 and MTA3; RBBP4 and RBBP7; and CDK2AP1 [1, 5]. Although Smits *et al.* recently determined the stoichiometry of the complex [6], questions remain regarding which subunits are mutually exclusive to each other and how many different functional compositions of the complex exist.

The domains and enzymatic activities in the NuRD complex are generally associated with repression of transcription, although until now this repression has mostly been shown on reporter genes in transactivation assays [7, 8]. For example, the CHROMO domains present in CHD3, 4 and 5 bind to H3K4me0 and H3K9me3 [9, 10] and the HDACs remove histone acetylation.

Recently, two groups have published Chromating immuno-precipitation followed by high-throughput sequencing (ChIP-seq) datasets for both MBD2 and MBD3 [11, 12]. The ChIP-seq profile of MBD2 in mouse embryonic stem cells shows two distinct MBD2 clusters. One cluster correlates with high levels of DNA methylation on CpG islands and in gene bodies, while the other, much smaller group localizes to non-methylated promoters and overlaps with the Mbd3 and Chd4 profiles [12]. The Renkawitz data also show a correlation between MBD2 binding and high DNA methylation levels, while the MBD3 peaks show an anti-correlation with DNA methylation [11]. Although these ChIP-seq datasets have revealed NuRD target genes in different cell types, the question remains how NuRD is recruited to promoters which lack DNA methylation.

In previous reports, we and others have identified a number of substoichiometric interactors of the NuRD complex, including ZMYND8 and ZNF687 [6, 9, 13]. The ZNF687 protein contains 10 C2H2-type zinc fingers. Zinc finger domains chelate zinc and often function as (sequence-specific) DNA binding domains or protein-protein interaction domains. ZMYND8 has a completely different domain architecture with a PHD finger, BROMO domain and PWWP domain at its N-terminus, and a MYND domain located closer to the C-terminus [14]. This protein is also called Protein Kinase C- Binding Protein (PKCBP1) or RACK7 [15].

Here, we show that Zmynd8 directly binds to the NuRD complex through its conserved MYND domain. Furthermore, we show that GATAD2A and GATAD2B assemble in mutually exclusive NuRD sub-complexes. ZMYND8 purifications only enrich for GATAD2A/NuRD, which is a strong indication that GATAD2A is the direct interaction partner of the MYND domain. Finally, we show that ZMYND8 and MBD3 share a subset of genome wide non-methylated loci, suggesting that ZMYND/ZNF687/ZNF592 recruits GATAD2A-containing NuRD complexes to a subset of its target genes. Interestingly, the enhancers and promoters bound by ZMYND8 and MBD3 are decorated with histone modifications which are commonly associated with active transcription.

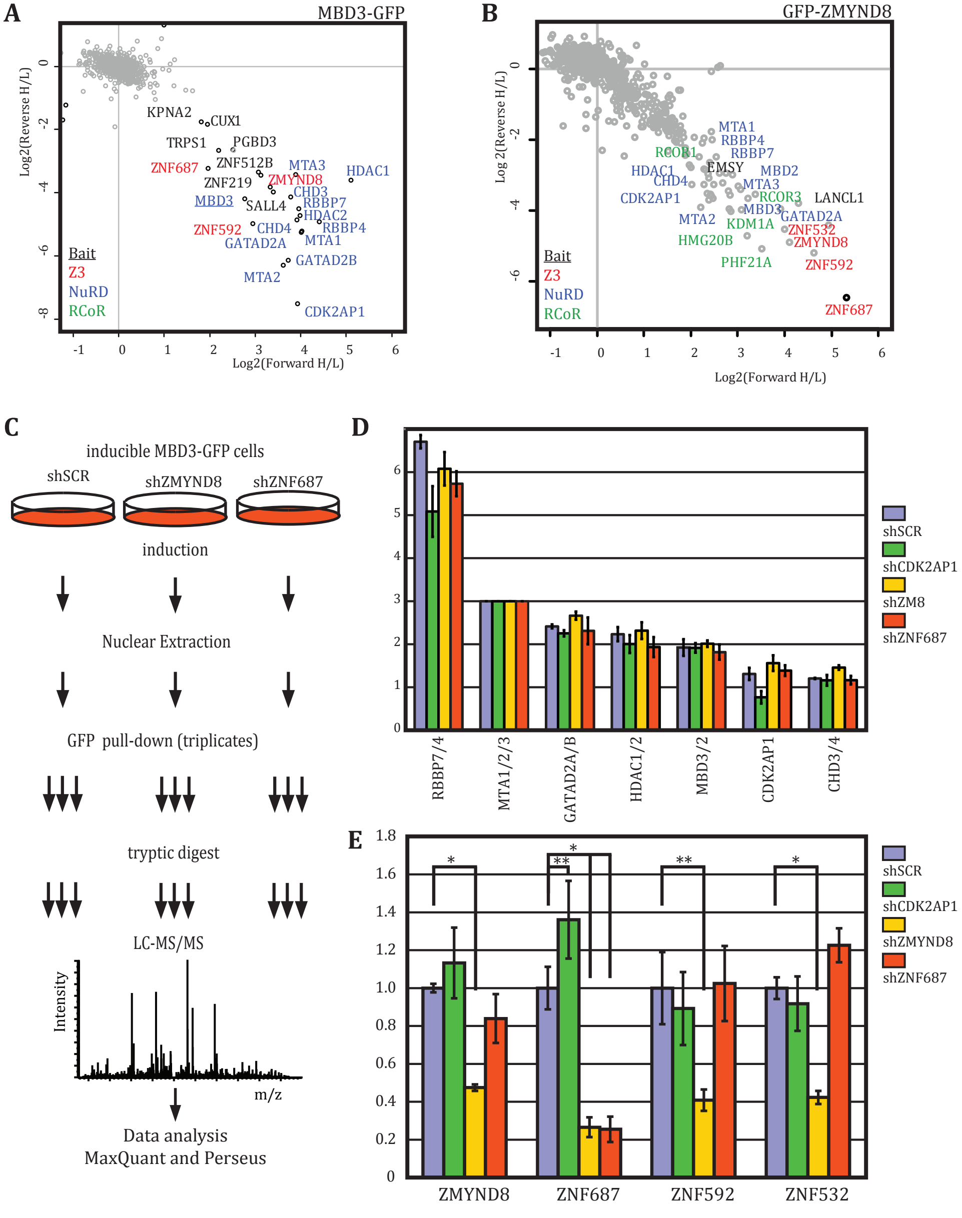


Figure1: ZMYND8 mediates the interaction between the Z3 module and NuRD. A. SILAC-based purification of GFP-MBD3 shows all NuRD subunits and a number of substoichiometric interactors. **B.** SILAC-based purification of GFP-ZMYND8 shows subunits of different co-repressor complexes. NuRD and BHC complex subunits are indicated by blue and green, respectively. ZNF proteins belonging to the Z3 module are indicated in orange. **C.** Schematic representation of the LFQ-based GFP-purifications from HeLa cells inducibly expressing MBD3-GFP and having a shRNA mediated knock-down of CDK2AP1, ZMYND8 or ZNF687. As a control, purifications are performed using a cell line expressing a scrambled shRNA. Purifications are performed in triplicates. **D.** Stoichiometry determination for NuRD subunits based on the LFQ purifications, according to [6]. The stoichiometry of the complex is stable in the various knock-down lines, except for CDK2AP1, which shows a lower stoichiometry in the CDK2AP1

RESULTS

ZMYND8 is a sub-stoichiometric interactor of the MBD3/NuRD complex.

Recently, we and others have identified ZMYND8 as a novel putative interactor of the NuRD compex [6, 9, 13]. To identify ZMYND8 and MBD3 interactors in HeLa cells, we performed SILAC-based GFP-affinity purifications [16, 17]. NuRD core subunits were convincingly enriched in the MBD3-GFP purification (Figure 1A), as well as some ZNF proteins and the known NuRD interactor SALL4. In contrast, purification of GFP-ZMYND8 resulted in the identification of multiple protein complexes. In addition to NuRD, BHC (consisting of LSD1, RCOR1-3, PHF21A and HMG20B) and EMSY were all significantly enriched. This is in agreement with the data obtained by Malovannaya *et al.* [13], who identified the ZMYND8 protein as a central hub in a large transcription regulation network (Figure 1B). Finally, in the GFP-ZMYND8 purification, three zinc finger proteins were enriched which have together been described as the Z3 module [13].

ZMYND8 links the Z3 module to the NuRD complex

To investigate which protein mediates the interaction between the NuRD complex and the Z3 hub, we performed label-free quantification (LFQ)-based purifications of GFP-MBD3 from stable cell lines containing either a scrambled shRNA or an shRNA targeting ZMYND8, ZNF687 or CDK2AP1 (Figure 1C) [18]. We obtained iBAQ values from these purifications, which we used to calculate the stoichiometry of the core subunits of the NuRD complex (Figure 1D) [6]. The three ZNFs share a number of tryptic peptides and thus cannot be distinguished in iBAQ-based stoichiometry determination. Therefore, we calculated the total intensity of the peptides unique to each of the ZNF proteins in each of the purifications (the unique intensity). Then, we divided these by the summed intensity of all identified peptides in the same sample to correct for concentration differences between the samples. Next, we compared these corrected unique intensities to those in the scrambled knock-down, which resembles the wild-type situation (Figure 1E).

Although the knock-downs in the respective lines were only partial (data not shown), we were able to show an effect of each knock-down on the fraction of the Z3 complex that is associated with NuRD. Knock-down of CDK2AP1, which often shows a very high ratio in ZMYND8 purifications suggestive of a direct interaction, did not affect the stoichiometry of ZMYND8 or any of the zinc fingers much. In contrast, knock-down of ZMYND8 reduced the levels of the zinc fingers in the purified sample by about 50-70% ($p < 0.05$, Figure 1E). Finally, knock-down of ZNF687 significantly reduced the levels of NuRD-associated ZNF687 protein itself, whereas the levels of the other ZNF proteins and ZMYND8 were not affected. In summary, these results indicate that ZMYND8 mediates the interaction between NuRD and ZNF687, -532 and -592.

The MYND domain is required and sufficient for interaction with NuRD

Since ZMYND8 mediates the interaction between the ZNF proteins and NuRD, we set out to identify the domain that is required for these interactions. ZMYND8 contains three domains that may be involved in histone-tail binding: a PHD finger, a

Figure1 (continued). knock-down line as expected. **E.** Comparison of the summed intensities of peptides unique for each of the ZNF proteins in the MBD3 purifications from each of the knock-down lines. Error bars in **D.** and **E.** indicate standard deviation. * p< 0.005, ** p< 0.01.

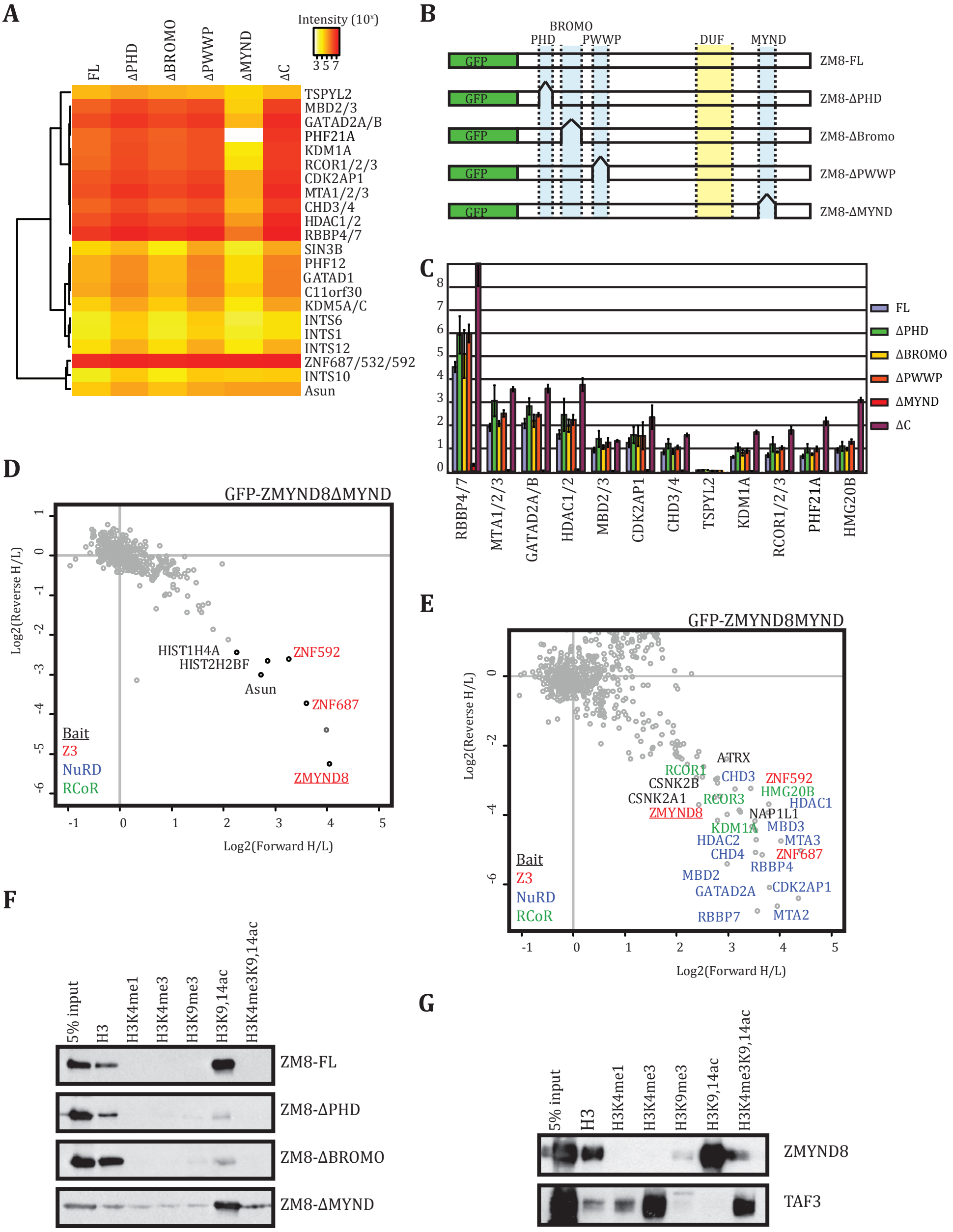


Figure 2: MYND domain is required and sufficient for the interaction with the transcription regulation complexes. A. Heatmap showing the average LFQ intensities of the ZMYND8 interactors (indicated on the right) in GFP purifications of different ZMYND8 deletion mutants (indicated on the top). The intensity of the Z3 module is equal in all purifications, but the intensities of subunits of both the NuRD and BHC complexes are lower in the ZMYND8ΔMYND mutant. **B.** Schematic representation of the GFP-fused domain-deletion mutants of ZMYND8. **C.** Stoichiometry of NuRD and BHC subunits derived from the LFQ-based GFP purifications of the ZMYND8 deletion mutants. Error bars indicate standard deviation. A clear loss of interaction can be observed for all subunits of the NuRD and BHC

BROMO domain, and a PWWP domain. In addition to these domains, ZMYND8 has a MYND (MYeloid, Nervy and Deaf) domain, which is a well-conserved protein-protein interaction domain [19].

Using a label-free quantification method to purify full-length ZMYND8 and deletion mutants lacking either one of these domains revealed that the MYND domain-deletion mutant does not interact with co-repressor complexes (Figure 2A and B). A heatmap of the LFQ values of interactors in purifications of the different deletion mutants shows a loss of NuRD and BHC intensity in the purification of the MYND deletion mutant only. In contrast, the intensity of the Z3 module is similar in the ZMYND8 purifications, indicating that this interaction is not mediated by any of the domains. Calculation of the stoichiometry of the NuRD and BHC core subunits shows remarkably stable values for a transient transfection-based purification (Figure 2C). The only mutant showing clearly deviating values is the ΔMYND mutant. These data thus show that the ZMYND8 MYND domain binds to co-repressor complexes whereas this domain is not required for the interaction between ZMYND8 and the ZNF proteins.

Having established that neither the PHD or BROMO domain are required for the interaction between ZMYND8 and NuRD, we wanted to further test the requirement of the MYND-domain. In a SILAC-based purification of ZMYND8-ΔMYND, lacking only the MYND domain, all interactions with NuRD and BHC were lost (Figure 2D), showing that the MYND domain is required for the interaction.

To test whether the MYND domain is also sufficient for the interaction with NuRD, we performed SILAC GFP affinity purifications using a GFP-MYND construct containing the C-terminus of ZMYND8. This purification resulted in interactions with both NuRD, BHC, as well as the Z3 proteins (Figure 2E). In conclusion, the MYND domain is required and sufficient for ZMYND8 to interact with the co-repressor complexes. The interaction surface for the zinc finger proteins within the ZMYND8 protein is less clear.

Since the PHD, BROMO and PWWP deletion mutants did not affect the interaction between NuRD and ZMYND8, we set out to test if they are indeed involved in binding to histone modifications. To this end, biotinylated synthetic peptides representing the first 17 amino acids of the H3-tail, with and without known histone modifications, were coupled to streptavidin beads and incubated with nuclear extract containing different deletion mutants. As shown in Figure 2F, full length ZMYND8 specifically recognizes H3K9,14Ac but not when H3 is methylated at K4. This repulsion by H3K4me3 is also observed in the context of non-acetylated peptides. The PHD finger of ZMYND8 is most likely responsible for the recognition of non-modified H3K4. However, deletion of this domain also diminishes binding to the H3K9,14Ac peptide, probably due to reduced stability or disturbed conformation of the PHD-BROMO module. Interestingly, binding to unmodified H3 is maintained in the BROMO-domain deletion mutant, whereas H3K9,14Ac binding is lost. Binding of the ΔMYND-mutant to the unmodified H3 and the

Figure 2 (continued). complexes in the ZMYND8ΔMYND mutant. **D.** SILAC-based GFP purification of ZMYND8ΔMYND, showing only ZNF687, ZNF592 and some histone proteins as interactors. **E.** SILAC-based GFP purification of ZMYND8MYNDonly. Specific enrichment of most NuRD and BHC complex subunits is observed. **F.** Affinity enrichments with immobilized histone-tail resembling peptides (indicated at the top) in nuclear extracts from cells expressing GFP-ZMYND8-FL or one of the deletion mutants. Full-length ZMYND8 strongly binds to H3K9,14Ac peptides and to a lesser extent to the unmodified H3 peptide. Deletion of the PHD or BROMO domain diminishes binding to the acetylated peptide, whereas deletion of the MYND domain does not affect binding to the peptides. **G.** As in F. but the western blot was performed with anti-ZMYND8 or anti-TAF3 antibody.

H3K9,14Ac peptides shows clearly that this binding occurs independently of the NuRD complex. Notably, unmodified H3 and H3K9,14Ac binding can also be observed for the endogenous ZMYND8 protein, whereas TAF3 clearly prefers binding to H3K4me3 and H3K4me3K9,14Ac peptides (Figure 2G). The PWWP domain is expected to bind H3K36me3 [20], but further experiments are needed to investigate this.

GATAD2A and B are mutually exclusive and ZMYND8 only binds GATAD2A/NuRD

Close inspection of the SILAC-based GFP-purifications of ZMYND8-FL and ZMYND8ΔMYND revealed that in most cases all paralogues of each NuRD subunit are specifically enriched. The exception to this are the GATAD2 paralogues of which only GATAD2A co-purifies with ZMYND8. This protein shows the highest enrichment ratios in most of the ZMYND8 purifications. Smits *et al.* have determined that the NuRD complex contains two GATAD2 molecules per complex [6], but whether GATAD2A and GATAD2B can form heterodimers is not clear. Since our purifications suggest that

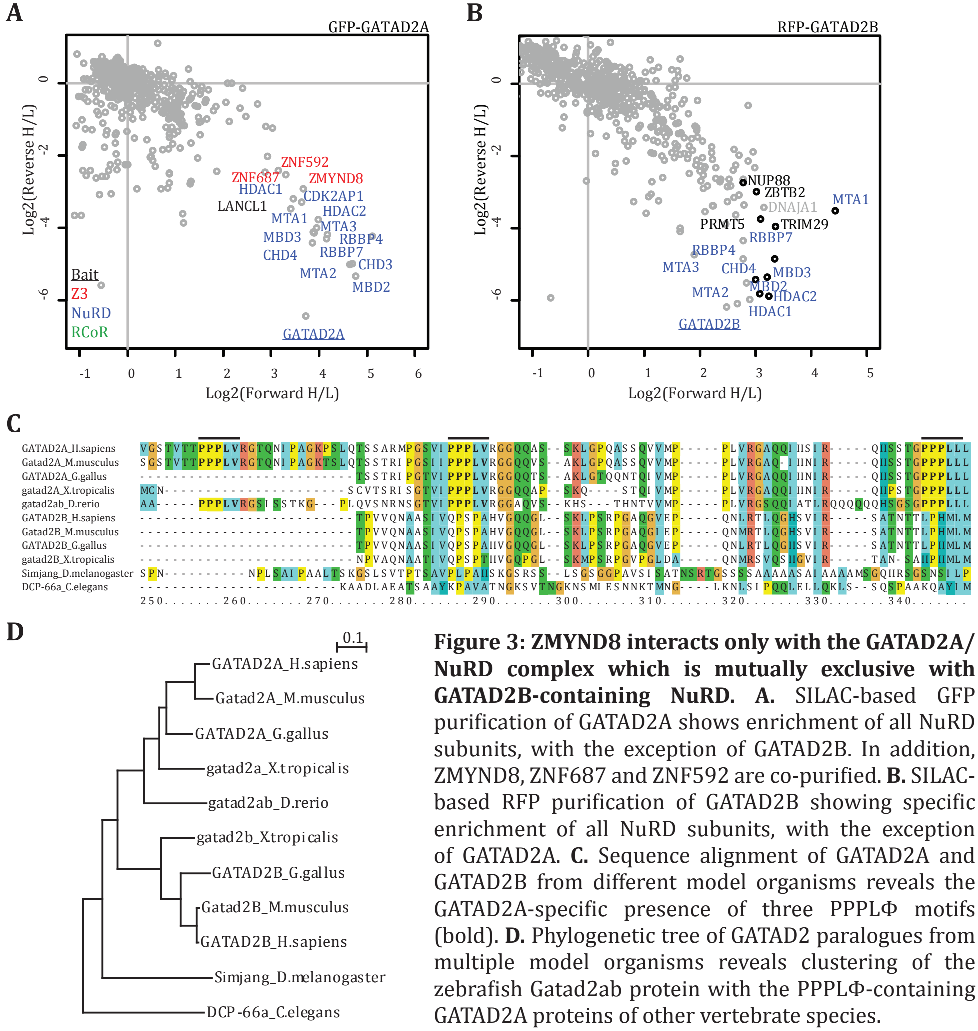


Figure 3: ZMYND8 interacts only with the GATAD2A/NuRD complex which is mutually exclusive with GATAD2B-containing NuRD. A. SILAC-based GFP purification of GATAD2A shows enrichment of all NuRD subunits, with the exception of GATAD2B. In addition, ZMYND8, ZNF687 and ZNF592 are co-purified. **B.** SILAC-based RFP purification of GATAD2B showing specific enrichment of all NuRD subunits, with the exception of GATAD2A. **C.** Sequence alignment of GATAD2A and GATAD2B from different model organisms reveals the GATAD2A-specific presence of three PPPLΦ motifs (bold). **D.** Phylogenetic tree of GATAD2 paralogues from multiple model organisms reveals clustering of the zebrafish Gatad2ab protein with the PPPLΦ-containing GATAD2A proteins of other vertebrate species.

ZMYND8 only interacts with GATAD2A, we purified GFP-GATAD2A and GATAD2B-RFP using SILAC-based quantitative proteomics.

The GATAD2A purification resulted in identification of most NuRD subunits, as well as ZMYND8, ZNF687 and ZNF592 (Figure 3A). However, whereas GATAD2A had a log2(H/L) ratio of ~8, GATAD2B was identified as a background binder. Comparing the ratios of the unique and shared peptides of these two proteins shows clearly that GATAD2A containing complexes do not contain GATAD2B (data not shown). The purification of GATAD2B-RFP again shows specific enrichment of all NuRD subunit paralogues, except for GATAD2A (Figure 3B). The ZMYND8 and ZNF proteins were probably so low abundant in this purification that they were not even identified. This is strong evidence that GATAD2A and GATAD2B are mutually exclusive and thus form homodimers only. Furthermore, ZMYND8 only interacts with the GATAD2A/NuRD complex, most likely via the GATAD2A protein itself. If ZMYND8 would directly bind to another NuRD subunit, which all can interact with both GATAD2A as well as GATAD2B, no GATAD2A/NuRD specific binding of ZMYND8 could be observed.

GATAD2A has conserved MYND interaction motifs

The fact that ZMYND8 exclusively interacts with GATAD2A/NuRD and not with GATAD2B/NuRD suggests that ZMYND8 binds directly to a motif present in GATAD2A which is lacking in GATAD2B. Multiple publications about the structure and interactions of MYND domains suggest that this domain may recognize different amino acid motifs [19, 21, 22]. Ansieau *et al.* described that the MYND domain of ZMYND11 binds PxLxP motifs, while the same study shows that ZMYND8 does not bind to this amino acid sequence [19]. However, the protein fragment used for recombinant expression and interaction studies may have been too small to achieve the required conformation. Liu *et al.* described how the MYND domain of ETO (also called ZMYND2) recognizes a PPPLΦ motif in N-CoR [21]. So, although the sequence of the MYND domain of ZMYND8 is more similar to the MYND domain of ZMYND11, we set out to find PPPLΦ motifs in NuRD core subunits.

Interestingly, GATAD2A contains three consensus PPPLΦ motifs, while GATAD2B lacks these motifs (Figure 3C, indicated in bold). We performed phylogenetic analysis on the GATAD2A and B proteins of the most commonly used model-organisms, representing vertebrates (mammal, fish, amfibia, bird) and invertebrates (insect and roundworm). Interestingly, whereas most vertebrates contain two GATAD2 paralogues, zebrafish has only one *Gatad2ab* gene, similar to invertebrates. However, this Gatad2ab protein has the putative motifs required for the interaction and therefore it clusters with all the GATAD2A genes (Figure 3D). In conclusion, this analysis might hint towards occurrence of the PPPLΦ motif during early vertebrate development, coinciding with the presence of DNA methylation and the potential changes in function of the NuRD complex [23]. However, inclusion of many more species in the analysis is required for a solid statement and more research is required to investigate whether this PPPLΦ motif is indeed required for the interaction with the ZMYND8- MYND domain.

ZMYND8 and MBD3 occupy active promoters and enhancers genome-wide

To investigate whether ZMYND8 and MBD3 also functionally interact with each other *in vivo*, we performed ChIP-seq experiments to determine genome-wide occupancy patterns. We used antibodies against endogenous proteins and chromatin derived from doubly cross-linked HeLa cells. ZMYND8 ChIP-seq resulted in the identification of roughly 10 000 peaks (Figure 4A). A heatmap centered on the ZMYND8 peaks shows a high correlation with the MBD3 genome-wide occupancy pattern (Figure 4B). The peaks are divided in three clusters. Based on co-occurring histone modifications, cluster 2 consists of active gene promoters. Clusters 1 and 3 overlap for about 53% with active enhancers, as defined by the presence H3K4me1 and H3K27Ac and DNAseI hypersensitive sites (Figure 4C). The Renkawitz and Schübeler groups have mainly shown MBD3/NuRD binding at non-methylated promoters [11, 12], so it is unclear what the role of ZMYND8 at the enhancers could be. ChIP-sequencing in a genetic

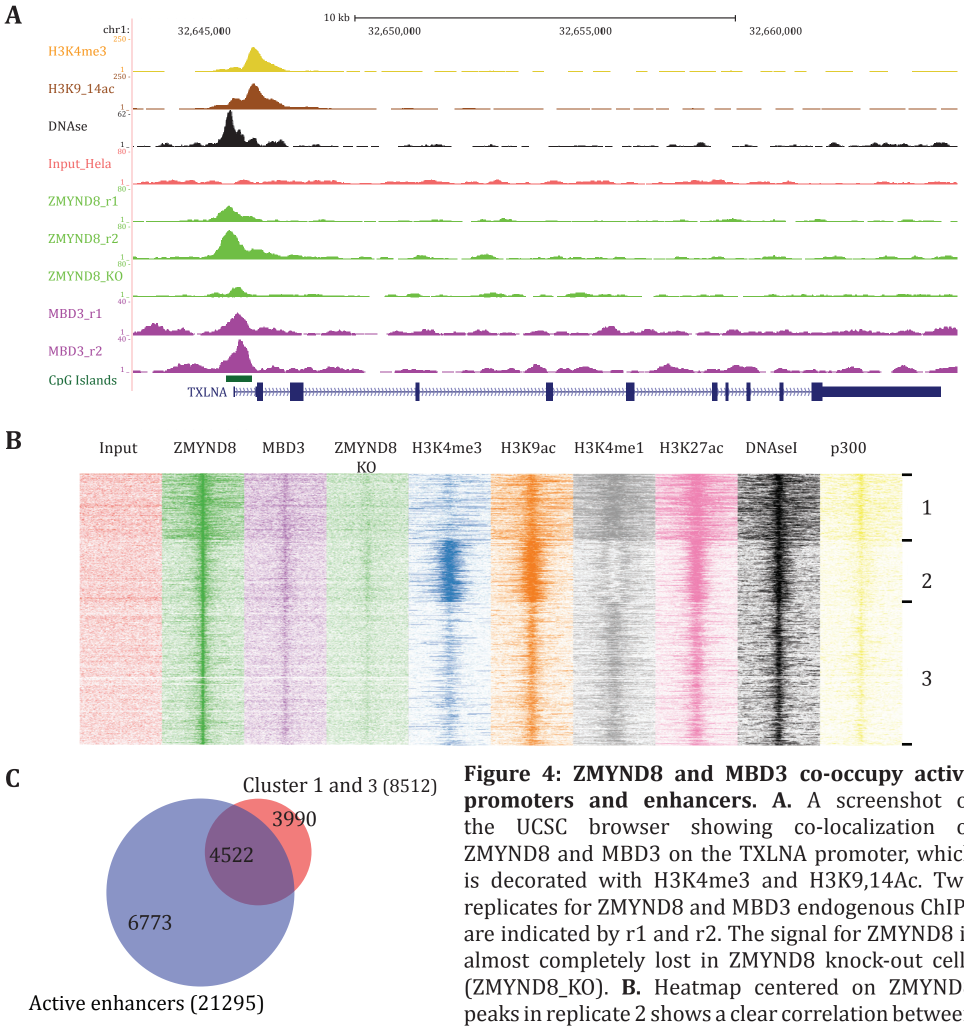


Figure 4: ZMYND8 and MBD3 co-occupy active promoters and enhancers. A. A screenshot of the UCSC browser showing co-localization of ZMYND8 and MBD3 on the TXLNA promoter, which is decorated with H3K4me3 and H3K9,14Ac. Two replicates for ZMYND8 and MBD3 endogenous ChIPs are indicated by r1 and r2. The signal for ZMYND8 is almost completely lost in ZMYND8 knock-out cells (ZMYND8_KO). **B.** Heatmap centered on ZMYND8 peaks in replicate 2 shows a clear correlation between ZMYND8 and MBD3 peaks. Cluster 2 consists mainly of H3K4me3 and H3K9Ac rich promoters, whereas ~53% of clusters 1 and 3 overlaps with active enhancers (**C.**), based on presence of DHS, H3K4me1 and H3K27ac and absence of H3K4me3.

CRISPR-based ZMYND8 knock-out HeLa cell line confirms that all peaks observed with the endogenous ZMYND8 antibody are specific (Figure 4B: ZMYND8 KO).

In conclusion, genome wide binding studies of ZMYND8 and MBD3 reveals extensive overlap between these proteins at active promoters and enhancers, suggesting a functional link *in vivo*.

DISCUSSION

In this study we have characterized the interaction between ZMYND8 and the NuRD complex, which was previously reported to be sub-stoichiometric [6, 9]. We show how ZMYND8 mediates an interaction between the putative DNA binding Z3 module and different corepressor complexes, such as NuRD and BHC. We have determined that the MYND domain is required and sufficient for the interaction between ZMYND8 and co-repressor complexes. This domain likely binds to the three PPPLΦ motifs present in the GATAD2A protein, since we found ZMYND8 to be associated exclusively with GATAD2A/NuRD. A schematic model of the confirmed and hypothetic interactions is shown in Figure 5.

Malovannaya *et al.* described multiple other complexes as interactors of ZMYND8, but we did not identify complexes other than NuRD, BHC and EMSY/Sin3 in HeLa cells [13]. A possible explanation for this discrepancy may be the use of nuclear versus whole cell extracts. One of the proteins described to be a structural/constitutive binder of ZMYND8 is TSPYL [13] which we did identify. However, in our analysis, this protein displayed a much lower stoichiometry than the NuRD core subunits.

Although MBD3 purifications identified only ZMYND8 and the Z3 module as NuRD interactors, we observed both the NuRD and the BHC complex in ZMYND8 purifications. This implies that ZMYND8 binds to either NuRD or BHC in a mutually exclusive manner, which is in agreement with the fact that both NuRD and BHC require the MYND of ZMYND8 for the interaction.

We have collected evidence to support our hypothesis that the ZMYND8-MYND domain binds to GATAD2A. Evolutionary analysis shows that a number of putative MYND-interaction motifs (PPPLΦ) in the GATAD2A protein are conserved from mammals to fish. In addition, our mass spectrometry experiments have shown that GATAD2A, but not GATAD2B, is specifically enriched in GFP-ZMYND8 purifications. Together with interaction data for GFP-GATAD2A and B showing that these proteins are mutually exclusive within NuRD, this suggests that ZMYND8 may only recruit GATAD2A/NuRD complexes.

Genome wide localization of ZMYND8 and MBD3

To confirm our hypothesis that the ZNF module recruits NuRD to certain target genes, we performed ChIP-seq for endogenous MBD3 and ZMYND8. Indeed, these proteins show a significant overlap in genome wide occupancy, suggesting that they interact *in vivo* as well. However, MBD3 ChIP-sequencing should be performed in wild-type and ZMYND8 knock-out cells to unambiguously establish MBD3 recruitment to these loci by ZMYND8.

The loci bound by ZMYND8 and MBD3 are mainly active promoters and enhancers that are also enriched for H3K4me3 and H3K9,14Ac or H3K4me1 and H3K27Ac, respectively. No enrichment of MBD3 on heterochromatic loci was observed. However, *in vitro* binding assays have shown that ZMYND8 preferentially binds

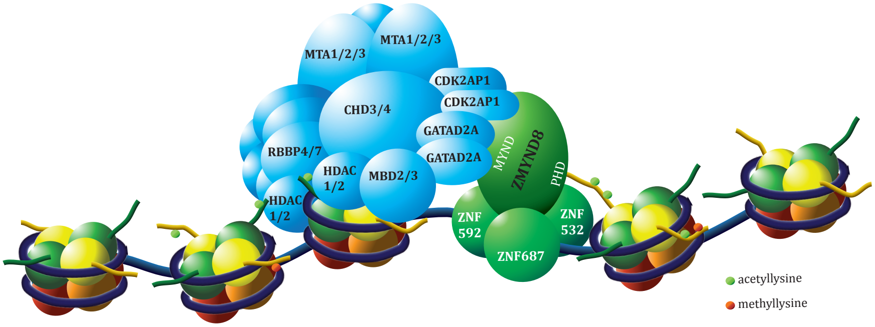


Figure 5: Hypothetic model of NuRD recruitment by ZMYND8. A schematic model showing sequence-specific DNA binding by the Z3 module and histone-tail binding by ZMYND8. Furthermore, ZMYND8 mediates the interaction between the Z3 module and the NuRD complex, by directly contacting GATAD2A with its MYND domain.

H3K9,14Ac when H3K4 is not modified. Even on non-acetylated peptides, ZMYND8 specifically binds non-methylated H3K4, and binds weakly to H3K4me1. The apparent co-occurrence of ZMYND8 and H3K4me3 *in vivo* can be explained in multiple ways. First of all, ChIP-seqencing only shows correlations. Observed signals are based on populations of cells, so the two enriched marks do not necessarily co-exist on one nucleosome or could even exist on the two different H3-tails protruding from one nucleosome. Second, the NuRD complex contains many additional histone binding domains that may guide it to promoters. Third, the specific localization to promoters could also be based on DNA sequence binding by the Z3 modules. Applying ChIP-reChIP may give an answer to some of these questions and could reveal which histone modifications coincide with ZMYND8 binding to nucleosomes. However, whether both K4 and K9 on a single histone tail are modified cannot be resolved using this technique. Furthermore, ZMYND8 ChIPs are very specific but have a very low efficiency, which makes ChIP-reChIP very difficult. Additionally, performing ZMYND8 ChIPs in ZNF knock-out cells may reveal whether the genome-wide ZMYND8 binding is driven by sequence-specific binding of the Z3 module or by specific recognition of histone modifications. Most likely, an interplay between these two determines genome-wide ZMYND8 binding. Whereas DNA binding by the Z3 module may be important to achieve binding specificity, the interaction between ZMYND8 and acetylated histone tails may stabilize the protein on chromatin.

MATERIALS AND METHODS

Cloning

ZMYND8-FL was PCR amplified from the Thermo Scientific cDNA clone 9052809/MHS1768-213246149 and ligated into pEGFP-C3 using HindIII and BamHI. This cDNA clone lacks some amino acid stretches, but no functional domains. Deletion mutants were made by a 3-point ligation of two PCR products. The vector was digested using HinDIII and BamHI, whereas the inserts had either HinDIII and SalI (for N-terminal part) or SalI and BamHI (for C-terminal part) restriction sites. GFP-GATAD2A was constructed by PCR from a pool of cDNA and ligation into the pEGFP-C3 vector. pcDNA5-FRT-TO-MBD3-GFP was created by Gateway cloning.

Cell culture and transfection

HeLa Kyoto or HeLa-FRT-TO MBD3-GFP cells were grown in Dulbecco's modified essential medium (DMEM) containing 4.5 g/ml glucose, 10% FBS and 1% penicillin/streptomycin. For all HeLa-FRT-TO lines, hygromycin and blasticidin were added to the medium.

For SILAC experiments, cells were cultured in SILAC DMEM plus 10% dialyzed serum, 1% glutamine, 1% penicillin/streptomycin, and 30 µg/ml of either light arginine (R0) and 73 µg/ml lysine (K0) or 73 µg/ml heavy lysine (K8) and 30 µg/ml arginine (R10) for at least 8 cell divisions. Cells were then expanded and transfected using PEI. After 20 hours, cells were harvested for nuclear extract preparation. The stable inducible MBD3-GFP HeLa-FRT-TO cell line was made by co-transfection of pcDNA5-FRT-TO-MBD3-GFP and pOG44 containing the flippase into HeLa-FRT-TO cells, after 2 days followed by hygromycin selection.

shRNA knock down

COS7 cells were grown in DMEM F12 containing 4.5 g/ml glucose, 10% FBS and 1% penicillin/streptomycin. The cells were transfected with lentiviral packaging vectors in combination with an shRNA construct using PEImax. After 24 and 48 hours, virus-containing media were collected, filtered and concentrated before transducing the target cells in the presence of polybrene. After another 24 hours, puromycin selection was started to select cells positive for shRNA integration. The knock-down efficiency was checked by RT-qPCR and, when possible, by western blot. For some knock-down lines monoclonals were grown and tested similarly.

Nuclear extract preparation

Nuclear extracts (NE) were prepared as described in Smits *et al.* [6].

GFP affinity purification

GFP purifications were performed essentially as described by Baymaz *et al.* [17]. When using transient transfection, all cells were transfected to avoid transcriptional side effects of PEI or ZMYND8 overexpression. In the HeLa-FRT-TO system, we likewise induced all cells with doxycycline. For the GFP-purifications performed with GFP-trap beads (Chromotek) (or RFP-trap beads (Chromotek) for GATAD2B-RFP), blocked agarose-beads (Chromotek) were used as a negative control. Incubation and wash buffers are as described before [17], with 0.25% NP40 during incubation and 0.5% NP40 in wash buffer C. On-bead digest was performed using Trypsin.

Mass spectrometry

Mass spectrometry measurements were performed on an easy nanoHPLC-1000 (Proxeon) operating a C18 column online with an LTQ-Orbitrap Velos in top15 CID mode with an exclusion list of 30 proteins for 30 seconds. An acetonitrile gradient of 5-80% was applied for 2.5 hours.

Data analysis

Raw data were analyzed using MaxQuant software package 1.3.0.5 using multiplicity 2 for SILAC experiments [24]. We filtered for contaminants and reverse hits using Perseus. The normalized forward and reverse ratios were logarithmized and significance B was calculated, after which scatterplots were made using R.

For label-free quantification (LFQ) [18], MaxQuant was applied using multiplicity 1 and boxes for 'match between runs' and 'iBAQ quantification' checked. For LFQ experiments, LFQ intensities were logarithmized and triplicates were assigned to the same group. We then filtered for 3 valid values in at least one group, assuming that very specific interactors may only be identified in the triplicates of the specific purification. Missing values were imputed using a normal distribution and default settings. A two-sample *t*-test was performed between the control and the experiment to obtain p-values for each protein, after which a volcano plot was made using R. A two-tailed ANOVA test was performed to calculate specific outliers when more than two conditions were compared. Stoichiometry determination was performed as described in Smits *et al.* [6] for the ANOVA significant proteins.

Western blot

For western blot analysis, protein samples were applied to SDS-PAGE and transferred to nitrocellulose membrane. The membrane was blocked using 5% skimmed milk in TBS-T, after which it was incubated with the primary antibody in milk solution for at least one hour and up to overnight incubation. After extensive washing, the membrane was incubated with HRP-fused secondary antibody in milk solution for one hour. The blot was washed and Enhanced ChemiLuminescence was used to visualize proteins on Kodak films. Antibodies used were: RabbitαGFP (Abcam, ab290), RabbitαZMYND8 (Sigma, HPA020949), RabbitαCDK2AP1 V41 (in-house [5]) and RabbitαTAF3 (Bethyl, A302-359A).

Peptide affinity purification

Peptide affinity purifications were performed essentially as described in [25]. Biotinylated peptides were coupled to streptavidin-agarose beads (GE Healthcare), after which unbound peptide was washed away using incubation buffer (150 mM NaCl, 50 mM Tris-HCl pH 8.0, 0.25% (v/v) NP40, 0.5 mM DTT, 10 μM $ZnCl_2$, 150 nM TSA and complete protease inhibitors (EDTA free)). The peptides were incubated with NE for 2 hours. Beads were washed extensively and bound proteins were eluted using sample buffer and boiling. Proteins were separated on SDS-PAGE and western blot was used to visualize the proteins.

ChIP-sequencing

HeLa Kyoto cells were seeded on 15 cm dishes. After 24 hours, cells were fixed using DSG for 45 minutes, followed by formaldehyde crosslinking for 10 minutes. Cells were collected and sonication was used to shear the chromatin into 100-300 bp fragments. Antibodies (RabbitαZMYND8 (Sigma, HPA020949) and RabbitαMBD3 (Bethyl, A302-528A)) were first coupled to the beads in the presence of BSA, after which we incubated this complex with chromatin overnight. Extensive washes were performed and DNA was de-crosslinked using a four hour incubation at 65^{o}C. DNA was purified using the Qiagen PCR-clean up kit. After this, bound DNA was analyzed by deep sequencing on a HiSeq.

ACKNOWLEDGEMENTS

We are grateful to Prof. Dr. Rainer Renkawitz for sharing the GATAD2B-RFP construct. The authors would like to thank the entire Timmers and Vermeulen groups for constructive discussions. Furthermore we would like to thank Holger Rehmann for advice concerning cloning of protein fragments.

REFERENCES

1. Allen, H.F., P.A. Wade, and T.G. Kutateladze, *The NuRD architecture.* Cell Mol Life Sci, 2013. **70**(19): p. 3513-24.
2. Potts, R.C., et al., *CHD5, a brain-specific paralog of Mi2 chromatin remodeling enzymes, regulates expression of neuronal genes.* PLoS One, 2011. **6**(9): p. e24515.
3. Hendrich, B. and A. Bird, *Identification and characterization of a family of mammalian methyl-CpG binding proteins.* Mol Cell Biol, 1998. **18**(11): p. 6538-47.
4. Spruijt, C.G., et al., *Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives.* Cell, 2013. **152**(5): p. 1146-59.
5. Spruijt, C.G., et al., *CDK2AP1/DOC-1 is a bona fide subunit of the Mi-2/NuRD complex.* Mol Biosyst, 2010. **6**(9): p. 1700-6.
6. Smits, A.H., et al., *Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics.* Nucleic Acids Res, 2013. **41**(1): p. e28.
7. Jiang, C.L., et al., *MBD3L1 and MBD3L2, two new proteins homologous to the methyl-CpG-binding proteins MBD2 and MBD3: characterization of MBD3L1 as a testis-specific transcriptional repressor.* Genomics, 2002. **80**(6): p. 621-9.
8. Bakker, J., X. Lin, and W.G. Nelson, *Methyl-CpG binding domain protein 2 represses transcription from hypermethylated pi-class glutathione S-transferase gene promoters in hepatocellular carcinoma cells.* J Biol Chem, 2002. **277**(25): p. 22573-80.
9. Eberl, H.C., et al., *A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics.* Mol Cell, 2013. **49**(2): p. 368-78.
10. Mansfield, R.E., et al., *Plant homeodomain (PHD) fingers of CHD4 are histone H3-binding modules with preference for unmodified H3K4 and methylated H3K9.* J Biol Chem, 2011. **286**(13): p. 11779-91.
11. Gunther, K., et al., *Differential roles for MBD2 and MBD3 at methylated CpG islands, active promoters and binding to exon sequences.* Nucleic Acids Res, 2013. **41**(5): p. 3010-21.
12. Baubec, T., et al., *Methylation-dependent and -independent genomic targeting principles of the MBD protein family.* Cell, 2013. **153**(2): p. 480-92.
13. Malovannaya, A., et al., *Analysis of the human endogenous coregulator complexome.* Cell, 2011. **145**(5): p. 787-99.
14. Fossey, S.C., et al., *Identification and characterization of PRKCBP1, a candidate RACK-like protein.* Mamm Genome, 2000. **11**(10): p. 919-25.
15. Ansieau, S. and A. Sergeant, *[BS69 and RACK7, a potential novel class of tumor suppressor genes].* Pathol Biol (Paris), 2003. **51**(7): p. 397-9.
16. Vermeulen, M., et al., *Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4.* Cell, 2007. **131**(1): p. 58-69.
17. Baymaz, H.I., C.G. Spruijt, and M. Vermeulen, *Identifying nuclear protein-protein*

interactions using GFP affinity purification and SILAC-based quantitative mass spectrometry. Methods Mol Biol, 2014. **1188**: p. 207-26.

18. Cox, J., et al., *Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ.* Mol Cell Proteomics, 2014. **13**(9): p. 2513-26.
19. Ansieau, S. and A. Leutz, *The conserved Mynd domain of BS69 binds cellular and oncoviral proteins through a common PXLXP motif.* J Biol Chem, 2002. **277**(7): p. 4906-10.
20. Wen, H., et al., *ZMYND11 links histone H3.3K36me3 to transcription elongation and tumour suppression.* Nature, 2014. **508**(7495): p. 263-8.
21. Liu, Y., et al., *Structural basis for recognition of SMRT/N-CoR by the MYND domain and its contribution to AML1/ETO's activity.* Cancer Cell, 2007. **11**(6): p. 483-97.
22. Kateb, F., et al., *Structural and functional analysis of the DEAF-1 and BS69 MYND domains.* PLoS One, 2013. **8**(1): p. e54715.
23. Hendrich, B. and S. Tweedie, *The methyl-CpG binding domain and the evolving role of DNA methylation in animals.* Trends Genet, 2003. **19**(5): p. 269-77.
24. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
25. Vermeulen, M., *Identifying chromatin readers using a SILAC-based histone peptide pull-down approach.* Methods Enzymol, 2012. **512**: p. 137-60.

Chapter 7

DNA methylation: old dog, new tricks?

Cornelia G. Spruijt[1] & Michiel Vermeulen[1-3]

[1]. Department of Molecular Cancer Research, University Medical Center Utrecht, Utrecht, the Netherlands.
[2]. Department of Molecular Biology, Radboud Institute for Molecular Life Sciences, Radboud University Nijmegen, Nijmegen, the Netherlands.
[3]. Cancer Genomics Netherlands, the Netherlands.

Nature Structural & Molecular Biology, November 2014
Volume 21, Number 11, p. 949-954

ABSTRACT

DNA methylation is an epigenetic modification that is generally associated with repression of transcription initiation at CpG island promoters. Here we argue, based on recent high-throughput genomic and proteomic screenings, that DNA methylation can also have different outcomes, including activation of transcription. This is evident from the fact that transcription factors can interact with methylated DNA sequences. Furthermore, in certain cellular contexts, genes containing methylated promoters are highly transcribed. Interestingly, this uncoupling between methylated DNA and repression of transcription seems to be particularly evident in early vertebrate development. Thus, contrary to previous assumptions, DNA methylation is not ubiquitously associated with repression of transcription initiation.

INTRODUCTION

Epigenetics refers to changes in gene expression and phenotype that occur without changes in DNA sequence. Epigenetic regulation of gene expression is at least partially achieved through post-translational modifications of histone proteins or by chemical modification of the DNA itself. DNA methylation was the first epigenetic modification to be described. In higher eukaryotes, DNA is mainly methylated on the position 5 carbon of the cytosine base (mC), and most frequently occurs on symmetrical CpG dinucleotides, although in ES cells and aging brain cells, non-CpG methylation is quite abundant [1, 2]. In vertebrates, DNA methylation in promoter regions is generally associated with gene silencing in *cis* and can thus affect the repertoire of expressed genes, thereby influencing cellular phenotype.

DNA methylation is catalysed by DNA methyltransferases (DNMTs). In mammals, three DNMTs have been identified: the 'maintenance' methyltransferase DNMT1, and the '*de novo*' DNMTs 3A and 3B [3]. Following DNA replication, both daughter duplexes contain one methylated and one non-methylated strand. DNMT1 methylates the newly synthesized strand of these hemi-methylated duplexes, thereby ensuring faithful transfer of DNA methylation patterns during cell division [3]. DNMT3A and DNMT3B generally catalyse methylation of DNA in a DNA replication-independent manner, although recent data indicate that this functional distinction between DNMT1 and DNMT3A/B is not as strict as previously thought [4, 5].

Dynamic DNA methylation patterns are very important during early development. During lineage commitment, differentiating cells are thought to methylate promoters of non-transcribed genes specific to other lineages to permanently silence them. In contrast, genes that are essential for lineage specification are kept non-methylated. DNA methylation-mediated gene silencing is thought to involve multiple mechanisms that are still not completely understood. Methylation can directly interfere with the binding of transcription factors to DNA [6-9]. Methylated DNA also recruits transcriptionally repressive methyl-CpG (mCpG) binding proteins [10-13]. Furthermore, DNA methylation can affect nucleosome positioning [14]. Non-methylated CpG islands, which are defined as DNA stretches containing a high density of CpG dinucleotides, on the other hand, are bound by CXXC domain-containing activating complexes [15-17]. Recently, a large number of putative novel mC-binding proteins have been identified, mainly using high throughput screenings. Many of these proteins are transcription factors that are generally not known to function as repressors of transcription. Furthermore, large scale DNA methylation profiling has revealed that in certain cellular contexts, genes with methylated promoters are highly transcribed. Here we review these recent studies and summarize the novel insights they have generated (Figure 1).

DNA methylation as a repressive epigenetic mark

The link between DNA methylation and repression of transcription originates from experiments in which methylated and non-methylated reporter constructs were transfected into mammalian cells and Xenopus oocytes [18, 19]. Because the non-methylated reporter gene is expressed at much higher levels compared to the methylated reporter in these assays [10, 11], the biological function of mC was believed to be exerted through proteins that differentially interact with cytosine in its methylated and non-methylated form.

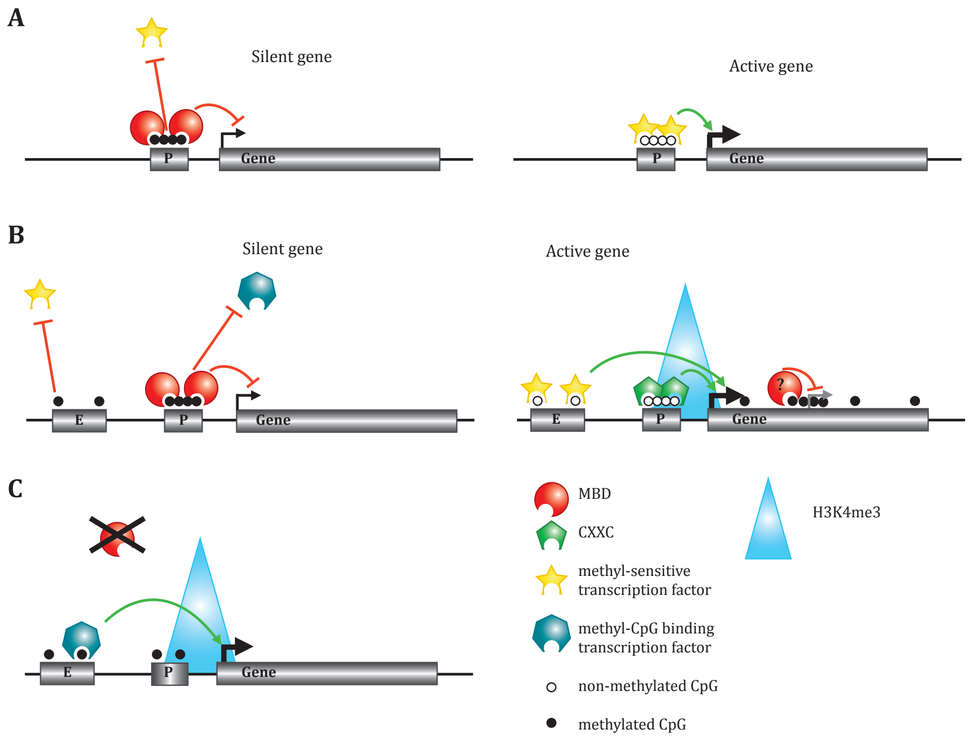


Figure 1. A. The "old" text-book model describing how DNA methylation regulates transcription. Methylated CpG island promoters recruit transcriptionally repressive MBD proteins and prevent transcription factor binding. Non-methylated CpG islands are bound by transcription factors. **B. New models describing regulation of transcription by DNA methylation**. Genes with methylated CpG island promoters are repressed by repressive MBD-containing complexes. In addition, methylation of an enhancer can block binding of a transcription factor. Most active genes with non-methylated CpG island promoters are bound by CXXC domain-containing activator complexes. In addition, transcription factors bind to non-methylated enhancers. Finally, gene bodies of active genes are highly methylated, which serves to repress cryptic transcription. **C. Uncoupling between DNA methylation and repression of transcription initiation**. In some cases, such as during early vertebrate development, some methylated low CpG dense promoters are actively transcribed. Transcriptionally repressive MBD proteins do not interact with these promoters for as yet unknown reasons. Furthermore, some low CpG dense DNA sequences (including enhancers and promoters) can be bound by activating transcription factors.

Three so-called methyl-CpG binding protein (MBP) families were identified in the 90s: the Methyl-CpG Binding Domain (MBD) containing family (MeCP2, MBD1-6), the Kaiso family that binds to mC via C2H2 zinc finger domains (Kaiso, ZBTB38 and ZBTB4), and the Set-and-Ring-Associated domain (SRA) family which only contains two members; UHRF1 and UHRF2 [20, 21]. Some of these proteins bind to mCpG with a high affinity *in vitro* in a DNA sequence-independent manner, such as MeCP2, MBD2 and UHRF1. The Kaiso family of proteins binds mCpG in a sequence-specific manner [22]. Several mammalian proteins (MBD3, MBD5 and MBD6) carry an MBD based on sequence homology but do not bind with a high affinity to methylated DNA *in vitro* [23].

Both MeCP2 and MBD2 associate with multi-subunit protein complexes containing histone deacetylases (HDACs). Since histone deacetylase activity is linked with gene silencing [24, 25], this implies a role for MBD2 and MeCP2 in repression of transcription through recruitment of HDACs. MeCP2 has been shown to associate with the Sin3/HDAC and N-CoR/SMRT co-repressor complexes [19, 26]. MBD2 and MBD3 interact, in a mutually exclusive manner, with the nucleosome remodelling and histone deacetylase (NuRD) complex [27]. In luciferase reporter assays, recruitment of these proteins leads to repression [10, 11, 28]. A vast number of recent studies have clearly linked genome-wide promoter CpG island hypermethylation to gene silencing in various cell types [29, 30]. Furthermore, reduction of DNA methylation levels in cells by 5-Aza-dC treatment or by depleting DNMTs reactivates transcription *in vivo* [31, 32]. These observations have led to a general textbook model in which promoter CpG island methylation serves as a recruitment signal for transcriptionally repressive methyl-CpG binding proteins, which ultimately results in gene repression in *cis*.

Proteins which interact with mC-containing DNA have diverse biological functions

During the last decade, quantitative mass spectrometry-based proteomics technology emerged as a powerful tool to study important biological questions. Within the research field of epigenetics, this technology has been applied to identify specific interactions between nuclear proteins and epigenetic histone and DNA modifications, including mCpG. These studies uncovered a large number of proteins in different cell types, which show specificity for different mCpG containing baits compared to their non-methylated counterpart [33-36]. Although such *in vitro* mass spectrometry-based approaches cannot discriminate direct from indirect interactors, a number of domains are consistently enriched in these experiments, suggesting that these domains directly read mCpG-containing DNA. In addition to the known mCpG-binding domains (MBD, zinc fingers and SRA), these include the homeobox and the winged-helix domains (including forkhead boxes). The interaction between the RFX5 winged-helix domain and mCpG-containing DNA was characterized in detail using Nuclear Magnetic Resonance (NMR). The K_d of this interaction was determined to be in the low micromolar range, which indicates that this interaction may be biologically relevant [36].

Although the above-mentioned mass spectrometry-based proteomics studies are unbiased and can be applied to a variety of different cell lines or even tissues, most of the published mCpG interaction screens were performed using a small number of mCpG containing baits. Recently, Hu and co-workers used a protein microarray consisting of 1321 transcription factors and 210 co-factor proteins to investigate direct interactions with 154 different human promoter sequences, each of which containing at least one mCpG [37]. The authors identified 47 proteins that bind to one or more of the methylated promoter sequences. One of the identified mCpG/CpG binding transcription factors is KLF4, which was also described as a novel mCpG binding protein in a previously published mass spectrometry-based study [36]. Hu *et al.* used conventional protein biochemistry, ChIP-bisulfite sequencing and luciferase reporter assays to further substantiate this finding [37]. Recently, the structure of the interaction between Klf4 and methylated DNA was solved [38]. KLF4 is one of the four Yamanaka reprogramming factors that can be used to generate induced pluripotent stem cells (iPSC) [39]. Since many genomic KLF4 binding sites are close to binding sites for the other reprogramming factors Oct4, Sox2, and Myc, KLF4 might play a pioneering role

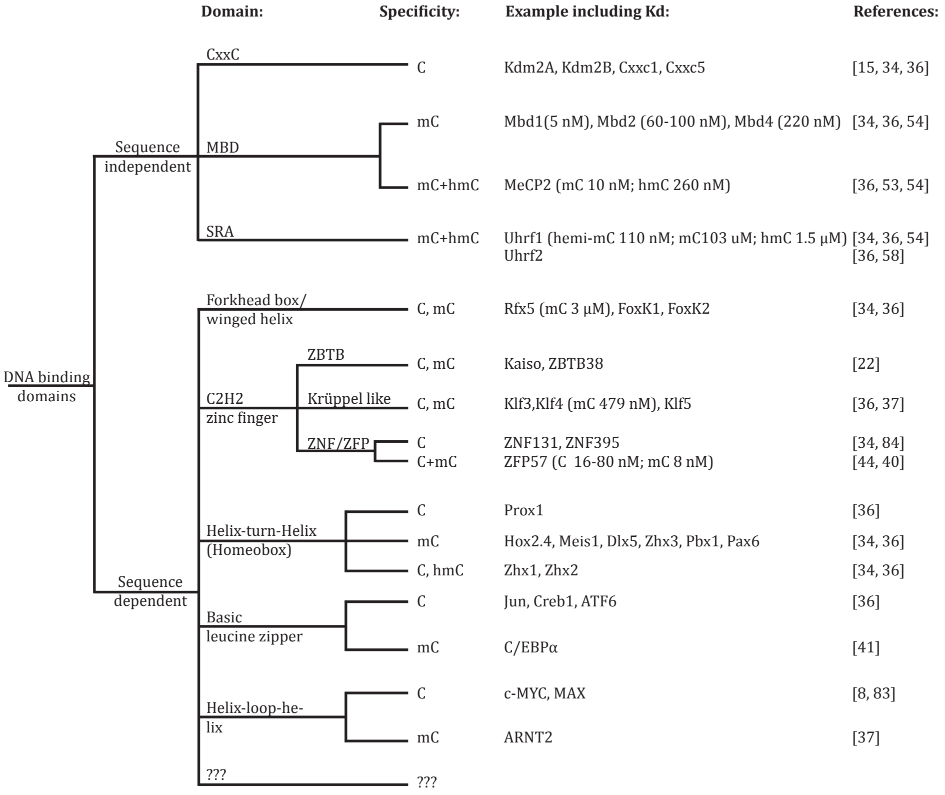


Figure 2: Schematic representation of the different classes of DNA binding domains that are implicated in C, mC and hmC binding binding: CXXC [15,34,36], MBD [34,36,53,54], SRA [34,36,54,58], forkhead box or winged helix [34,36], C2H2 zinc finger [22,34,36,37,40,44,83], helix-turn-helix (homeobox) [34,36], basic leucine zipper [36,41] and helix-loop-helix [8,37,84]. Both the sequence-dependent as well as the sequence independent branch show domains with binding specificities for multiple DNA modifications.

during the generation of iPSC. When KLF4 expression is induced in somatic cells, it may bind to its methylated target genes and recruit factors to decondense chromatin and demethylate DNA. This may finally enable the other pluripotency factors to bind to their target genes to induce stem cell-specific gene expression. However, further studies are required to investigate this hypothesis.

In addition to high-throughput screening methods, other, more targeted studies have identified additional proteins with affinity for mCpG-containing DNA. Examples include ZFP57, C/EBPα and ZBTB4 [40-42]. ZFP57 was originally discovered as a protein important for genomic imprinting [43]. In ZFP57 knock-out mice, imprinted, silent alleles lose DNA methylation and their expression is induced. These regions contain a consensus binding site for ZFP57 to which the protein can only bind when this binding site is methylated [44]. ZFP57 therefore seems to be essential for the transcriptional silencing of imprinted alleles via the recruitment of DNMT1 and Kap1/Setdb1 [44]. C/EBPα is a leucine-zipper containing transcription factor that is very important for lineage commitment of mammalian cells. Rishi and co-workers recently showed that C/EBPα interacts with the *cis*-regulatory element (CRE: TGACGTCA) when it is methylated. The authors further showed that *in vivo* binding of C/EBPα to such methylated elements is important for C/EBPα mediated activation of target gene transcription during differentiation of keratinocytes [41]. These examples illustrate that a DNA binding event driven by methylation of CpG dinucleotides can result in a different functional outcome depending on the biological function of the protein that binds this particular DNA sequence.

These findings prompt a re-evaluation of the original classification of MBPs. The MBD and SRA family of MBPs predominantly interact with mCpG in a DNA sequence-independent manner and their genome wide binding profiles correlate with mCpG density [12]. Many zinc finger-containing transcription factors may, in addition to binding their non-methylated consensus DNA binding site, interact with mCpGs in a sequence context that may differ from their known consensus DNA binding site. Three types of zinc finger proteins with putative mCpG binding ability have been identified: the ZBTB/POZ subfamily (including Kaiso), Krüppel-like zinc finger proteins (including KLF4) and the Znf/Zfp zinc finger protein subfamily (including ZFP57). Some of the mC-specific binding of these zinc finger-containing proteins may be explained by the fact that a methylcytosine structurally resembles a thymine when viewed from the major groove of the DNA helix. Methylation of a cytosine in a particular DNA sequence can thus reconstitute a TpG-containing consensus binding site for a zinc-finger containing transcription factor. This phenomenon has been shown, at least *in vitro*, for a number of proteins including ZBTB4 and RBP-J [33, 42]. Recent work has revealed additional domains that are capable of binding to mCpG in a sequence-dependent manner. These domains include the homeobox, forkhead box/winged helix and basic leucine zipper (Figure 2). Further studies are needed to characterize the K_d of these interactions and to determine their physiological relevance, for example by using ChIP-bisulfite sequencing [45]. Furthermore, ChIP-seq experiments in DNMT deficient cells or in cells in which DNA methylation is reduced using 5-Aza-dC can reveal which proportion of the binding sites for a particular transcription factor in the genome are DNA methylation-driven. In any case, the repertoire of proteins that specifically interact with methylated DNA sequences clearly extends beyond the three classes of proteins that were originally reported some twenty years ago. The known functions of these proteins are diverse, which implies that the biology of DNA methylation, particularly on DNA stretches with relatively low CpG density, encompasses more than just gene silencing. Further *in vivo* studies, however, are clearly needed to substantiate this hypothesis.

The plot thickens: hmC, fC and caC enter the stage

Recently, the family of TET proteins was shown to convert mC to 5-hydroxymethylcytosine (hmC) [46]. Hydroxymethylcytosine is particularly abundant in embryonic stem cells and in brain cells [47, 48]. Further iterative TET-meditated oxidation results in the formation of 5-formylcytosine (fC) and 5-carboxycytosine (caC) [49]. Both fC and caC serve as substrates for Thymine-DNA glycosylase (TDG) [50], which, in combination with the base excision repair machinery, forms an active DNA demethylation pathway (reviewed in [51]).

The function of the oxidized versions of mC is currently unclear. Several groups have therefore pursued the identification of readers for hmC, fC and caC [36, 52], and the number of DNA repair-associated proteins found to interact with hmC, fC and caC reinforces the proposed link between oxidized mC derivatives and active DNA demethylation. In addition to DNA repair associated proteins (including helicases and glycosylases), transcription factors and chromatin modifying enzymes were found to interact specifically with mC derivatives. The number of identified readers for fC and caC greatly exceeds the number of readers for hmC. Furthermore, only limited overlap is observed with regard to proteins (in)directly interacting with each of the modified cytosine bases [36]. Most of the 'classic' MBD proteins have a lower affinity for hmC compared to mC. The exception to this is MeCP2, which was reported to bind to hmC, albeit with a slightly lower affinity compared to mC [53, 54]. Other groups, however, have not observed binding of MeCP2 to hmC [55, 56]. Mammalian MBD3, which does not bind mC with a high affinity, was reported to interact with 5hmC, [57] a finding that other studies have failed to reproduce [36, 54]. The SRA family of proteins, which consists of UHRF1 and UHRF2, binds to both mC and hmC. Whereas UHRF1 binds with a similar affinity to mC and hmC [36, 58], UHRF2 shows a clear preference for hmC [36, 59].

Chemically, mC and hmC differ both in their polarity as well as in their size. Whereas mC is hydrophobic, hmC can form hydrogen bonds. Furthermore, the hydroxymethyl group is more bulky than the methyl group and needs to be accommodated in a larger binding pocket. Apparently, the chemical differences between different oxidized mC derivatives result in quite distinct binding patterns. This is clear from examples such as RFX5, which potently binds mC and also interacts with C and fC but is strongly repelled by hmC and caC [36]. Further biochemical and structural studies are required to decipher the molecular mechanisms underlying these observations. Based on the distinct biochemical properties of the oxidized mC derivatives and their different, mostly non-overlapping readers, each of these modifications may have their own specific function(s). Given the seemingly very potent DNA damage response that is triggered by fC and caC, we expect that these modifications mainly function as DNA demethylation intermediates, whereas hmC may also play a role in transcription regulation.

Genome wide methylation profiling: lots of data, lots of surprises!

Numerous technologies can be used to profile DNA methylation patterns in cells. Of these methods, whole genome bisulfite sequencing provides single basepair resolution and is therefore considered the gold standard for genome-wide DNA methylation profiling. The disadvantage of most bisulfite-based methods is that they cannot discriminate between mC and its oxidized derivatives. Additional methods, such as glucosyltransferase treatment of hmC to protect and enrich it, as well as oxBS-seq and fCAB-seq, have been developed to facilitate hmC and fC profiling [60-62]. A comprehensive review of genome-wide (hydroxy)methylation profiling results is beyond the scope of this perspective but a number of major, surprising observations stand out. First of all, genome wide profiling has revealed that gene bodies are methylated [63-65]. A correlation can be found between the density of gene body methylation and gene expression, but the best correlation is between the density of gene body methylation and replication timing [66]. These results imply a possible link between DNA methylation in gene bodies and the regulation of gene expression or DNA replication. Evidence shows that gene body methylation inhibits transcription initiation from cryptic promoters [67]. Further studies are required to decipher the molecular mechanisms underlying these observations. Second, methylome profiling during early development has revealed that most of the differentially methylated regions (DMRs) in the genome map to enhancer sequences and not to promoters [68-72]. Furthermore, methylated (low CpG dense) promoters are not always silenced in germ cells and pluripotent cells but are sometimes actively being transcribed [73, 74]. These methylated genes are apparently not silenced by transcriptionally repressive MBD proteins but are instead bound by proteins that activate transcription or at least do not interfere with transcription [74-76]. Third, non-CpG methylation is prominent in embryonic stem cells and also in aging brain cells [1, 2]. The fact that non-CpG methylation appears not to be a random but rather a regulated process that is abundant in certain cell types or tissues indicates that this type of cytosine methylation may have unique function(s) that are yet to be discovered. The overall conclusion of these studies is that the function of DNA methylation extends beyond promoter CpG methylation, includes regulation of enhancer activity and may include regulation of replication, amongst other things. Furthermore, even well described methylation events such as promoter CpG methylation do not always result in gene silencing, especially when these promoters are characterized by a low CpG density.

Outlook

In this perspective we have reviewed the increasingly complex biology of DNA methylation in light of results from recent high-throughput proteomic and genomic approaches. The observations made in these studies indicate that the previously assumed strict correlation between DNA methylation and repression of transcription is in fact context dependent [77, 78]. This is obvious from proteomic screens which revealed that many transcription factors interact with mCpG-containing DNA sequences. This implies that in certain cases methylation of a CpG dinucleotide can induce activation of transcription initiation rather than repression, depending on the nature of the reader that interacts with this particular DNA sequence [79, 80]. We hypothesize that transcription factor binding to methylated DNA may be particularly relevant on enhancers or other DNA elements with a low CpG density. On DNA stretches with a high CpG density we expect that repressive MBDs will usually be the dominant mCpG readers (Figure 1C).

Interestingly, interactions with methylated DNA are highly dynamic during cellular differentiation. For example, the transcriptionally repressive MBD2/NuRD complex does not interact with methylated DNA in nuclear extracts derived from mouse embryonic stem cells grown in 2i medium, but does interact with methylated DNA in neuronal precursor cells or adult mouse brain nuclear extracts [36]. The molecular mechanism underlying this observation is not clear yet, but may involve an isoform switch for MBD2, which was recently shown to be important for lineage commitment and differentiation [81]. Other explanations could be lower expression levels of MBD2 or post-translational modifications that inhibit mCpG binding, something which has previously been shown for MeCP2 [82]. Also relevant in this context are the previously mentioned observations from a recent methylome profiling study in adult male germ cells, which revealed that many methylated promoters are in fact highly transcribed [74]. It should be noted, however, that these highly transcribed methylated promoters are characterized by a relatively low CpG content. In any case, a (partial) uncoupling between DNA methylation and transcriptional repression apparently exists in male germ cells, which may be explained by a low abundance of transcriptionally repressive readers or their inability to interact with methylated DNA in those cells. Interestingly, this (partial) uncoupling between DNA methylation and repression of transcription is also evident during early vertebrate development ([75] and unpublished observations).

As more putative mCpG interacting proteins are identified, the question is raised as to how binding specificity amongst different mCpG binders is achieved. Eventually, this may come down to affinity and protein abundance. Thus, in order to understand the biology of mCpG dynamics and its interactors, global quantitative methods are required to determine direct interactions and their K_ds for genomic mCpG-containing sequences. An important method to also use in this context is ChIP-bisulfite sequencing to proof that a protein of interest binds to methylated DNA sequences *in vivo*. These approaches have to be complemented with global profiling of absolute protein abundance. Eventually, such approaches will allow a quantitative modelling of mCpG and its functions in different sequence contexts during cellular differentiation and transformation.

ACKNOWLEDGEMENTS

We apologize to all people whose work could not be cited due to space limitations. We would like to thank Ozren Bogdanovic, Arie B. Brinkman, Susan L. Kloet and Arne H. Smits for critical reading of the manuscript. The Vermeulen lab is supported by grants from the Netherlands Organisation for Scientific Research (NWO-VIDI and Cancer Genomics Netherlands).

REFERENCES

1. Ramsahoye, B.H., et al., *Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a.* Proc Natl Acad Sci U S A, 2000. **97**(10): p. 5237-42.
2. Lister, R., et al., *Global epigenomic reconfiguration during mammalian brain development.* Science, 2013. **341**(6146): p. 1237905.
3. Jurkowska, R.Z., T.P. Jurkowski, and A. Jeltsch, *Structure and function of mammalian DNA methyltransferases.* Chembiochem, 2011. **12**(2): p. 206-22.
4. Arand, J., et al., *In Vivo Control of CpG and Non-CpG DNA Methylation by DNA Methyltransferases.* Plos Genetics, 2012. **8**(6).
5. Liang, G.G., et al., *Cooperativity between DNA methyltransferases in the maintenance methylation of repetitive elements.* Molecular and Cellular Biology, 2002. **22**(2): p. 480-491.
6. Iguchi-Ariga, S.M. and W. Schaffner, *CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation.* Genes Dev, 1989. **3**(5): p. 612-9.
7. Campanero, M.R., M.I. Armstrong, and E.K. Flemington, *CpG methylation as a mechanism for the regulation of E2F activity.* Proc Natl Acad Sci U S A, 2000. **97**(12): p. 6481-6.
8. Prendergast, G.C. and E.B. Ziff, *Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region.* Science, 1991. **251**(4990): p. 186-9.
9. Blattler, A. and P.J. Farnham, *Cross-talk between site-specific transcription factors and DNA methylation states.* J Biol Chem, 2013. **288**(48): p. 34287-94.
10. Bakker, J., X. Lin, and W.G. Nelson, *Methyl-CpG binding domain protein 2 represses transcription from hypermethylated pi-class glutathione S-transferase gene*

promoters in hepatocellular carcinoma cells. J Biol Chem, 2002. **277**(25): p. 22573-80.

11. Jiang, C.L., et al., *MBD3L1 and MBD3L2, two new proteins homologous to the methyl-CpG-binding proteins MBD2 and MBD3: characterization of MBD3L1 as a testis-specific transcriptional repressor.* Genomics, 2002. **80**(6): p. 621-9.
12. Baubec, T., et al., *Methylation-dependent and -independent genomic targeting principles of the MBD protein family.* Cell, 2013. **153**(2): p. 480-92.
13. Curradi, M., et al., *Molecular mechanisms of gene silencing mediated by DNA methylation.* Mol Cell Biol, 2002. **22**(9): p. 3157-73.
14. Huff, J.T. and D. Zilberman, *Dnmt1-independent CG methylation contributes to nucleosome positioning in diverse eukaryotes.* Cell, 2014. **156**(6): p. 1286-97.
15. Thomson, J.P., et al., *CpG islands influence chromatin structure via the CpG-binding protein Cfp1.* Nature, 2010. **464**(7291): p. 1082-6.
16. Farcas, A.M., et al., *KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands.* Elife, 2012. **1**: p. e00205.
17. Blackledge, N.P., et al., *Variant PRC1 Complex-Dependent H2A Ubiquitylation Drives PRC2 Recruitment and Polycomb Domain Formation.* Cell, 2014. **157**(6): p. 1445-59.
18. Kass, S.U., N. Landsberger, and A.P. Wolffe, *DNA methylation directs a time-dependent repression of transcription initiation.* Curr Biol, 1997. **7**(3): p. 157-65.
19. Jones, P.L., et al., *Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription.* Nat Genet, 1998. **19**(2): p. 187-91.
20. Defossez, P.A. and I. Stancheva, *Biological functions of methyl-CpG-binding proteins.* Prog Mol Biol Transl Sci, 2011. **101**: p. 377-98.
21. Bogdanovic, O. and G.J. Veenstra, *DNA methylation and methyl-CpG binding proteins: developmental requirements and function.* Chromosoma, 2009. **118**(5): p. 549-65.
22. Filion, G.J., et al., *A family of human zinc finger proteins that bind methylated DNA and repress transcription.* Mol Cell Biol, 2006. **26**(1): p. 169-81.
23. Laget, S., et al., *The Human Proteins MBD5 and MBD6 Associate with Heterochromatin but They Do Not Bind Methylated DNA.* Plos One, 2010. **5**(8).
24. Laherty, C.D., et al., *Histone deacetylases associated with the mSin3 corepressor mediate mad transcriptional repression.* Cell, 1997. **89**(3): p. 349-56.
25. Nagy, L., et al., *Nuclear receptor repression mediated by a complex containing SMRT, mSin3A, and histone deacetylase.* Cell, 1997. **89**(3): p. 373-80.
26. Lyst, M.J., et al., *Rett syndrome mutations abolish the interaction of MeCP2 with the NCoR/SMRT co-repressor.* Nat Neurosci, 2013. **16**(7): p. 898-902.
27. Le Guezennec, X., et al., *MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties.* Mol Cell Biol, 2006. **26**(3): p. 843-51.
28. Ng, H.H., et al., *MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex.* Nat Genet, 1999. **23**(1): p. 58-61.
29. Dong, S.M., et al., *Promoter hypermethylation of multiple genes in carcinoma of the uterine cervix.* Clin Cancer Res, 2001. **7**(7): p. 1982-6.
30. Kang, S.H., et al., *Transcriptional repression of the transforming growth factor-beta type I receptor gene by DNA methylation results in the development of TGF-beta resistance in human gastric cancer.* Oncogene, 1999. **18**(51): p. 7280-6.

31. Chiurazzi, P., et al., *In vitro reactivation of the FMR1 gene involved in fragile X syndrome.* Hum Mol Genet, 1998. **7**(1): p. 109-13.
32. Robert, M.F., et al., *DNMT1 is required to maintain CpG methylation and aberrant gene silencing in human cancer cells.* Nat Genet, 2003. **33**(1): p. 61-5.
33. Bartels, S.J., et al., *A SILAC-based screen for Methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein.* PLoS One, 2011. **6**(10): p. e25884.
34. Bartke, T., et al., *Nucleosome-interacting proteins regulated by DNA and histone methylation.* Cell, 2010. **143**(3): p. 470-84.
35. Mittler, G., F. Butter, and M. Mann, *A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements.* Genome Res, 2009. **19**(2): p. 284-93.
36. Spruijt, C.G., et al., *Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives.* Cell, 2013. **152**(5): p. 1146-59.
37. Hu, S., et al., *DNA methylation presents distinct binding sites for human transcription factors.* Elife, 2013. **2**: p. e00726.
38. Liu, Y., et al., *Structural basis for Klf4 recognition of methylated DNA.* Nucleic Acids Res, 2014. **42**(8): p. 4859-67.
39. Lewitzky, M. and S. Yamanaka, *Reprogramming somatic cells towards pluripotency by defined factors.* Curr Opin Biotechnol, 2007. **18**(5): p. 467-73.
40. Liu, Y., et al., *An atomic model of Zfp57 recognition of CpG methylation within a specific DNA sequence.* Genes Dev, 2012. **26**(21): p. 2374-9.
41. Rishi, V., et al., *CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes.* Proc Natl Acad Sci U S A, 2010. **107**(47): p. 20311-6.
42. Sasai, N., M. Nakao, and P.A. Defossez, *Sequence-specific recognition of methylated DNA by human zinc-finger proteins.* Nucleic Acids Res, 2010. **38**(15): p. 5015-22.
43. Mackay, D.J., et al., *Hypomethylation of multiple imprinted loci in individuals with transient neonatal diabetes is associated with mutations in ZFP57.* Nat Genet, 2008. **40**(8): p. 949-51.
44. Quenneville, S., et al., *In embryonic stem cells, ZFP57/KAP1 recognize a methylated hexanucleotide to affect chromatin and DNA methylation of imprinting control regions.* Mol Cell, 2011. **44**(3): p. 361-72.
45. Brinkman, A.B., et al., *Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk.* Genome Research, 2012. **22**(6): p. 1128-1138.
46. Tahiliani, M., et al., *Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1.* Science, 2009. **324**(5929): p. 930-5.
47. Kriaucionis, S. and N. Heintz, *The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain.* Science, 2009. **324**(5929): p. 929-30.
48. Munzel, M., et al., *Quantification of the sixth DNA base hydroxymethylcytosine in the brain.* Angew Chem Int Ed Engl, 2010. **49**(31): p. 5375-7.
49. Ito, S., et al., *Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine.* Science, 2011. **333**(6047): p. 1300-3.
50. Maiti, A., et al., *TDG excision of fC may be a predominant element of pathways for active DNA demethylation.* Faseb Journal, 2013. **27**.
51. Wu, H. and Y. Zhang, *Reversing DNA methylation: mechanisms, genomics, and*

biological functions. Cell, 2014. **156**(1-2): p. 45-68.

52. Iurlaro, M., et al., *A screen for hydroxymethylcytosine and formylcytosine binding proteins suggests functions in transcription and chromatin regulation.* Genome Biol, 2013. **14**(10): p. R119.
53. Mellen, M., et al., *MeCP2 binds to 5hmC enriched within active genes and accessible chromatin in the nervous system.* Cell, 2012. **151**(7): p. 1417-30.
54. Hashimoto, H., et al., *Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation.* Nucleic Acids Res, 2012. **40**(11): p. 4841-9.
55. Valinluck, V., et al., *Oxidative damage to methyl-CpG sequences inhibits the binding of the methyl-CpG binding domain (MBD) of methyl-CpG binding protein 2 (MeCP2).* Nucleic Acids Res, 2004. **32**(14): p. 4100-8.
56. Khrapunov, S., et al., *Unusual characteristics of the DNA binding domain of epigenetic regulatory protein MeCP2 determine its binding specificity.* Biochemistry, 2014. **53**(21): p. 3379-91.
57. Yildirim, O., et al., *Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells.* Cell, 2011. **147**(7): p. 1498-510.
58. Frauer, C., et al., *Recognition of 5-hydroxymethylcytosine by the Uhrf1 SRA domain.* PLoS One, 2011. **6**(6): p. e21306.
59. Zhou, T., et al., *Structural Basis for Hydroxymethylcytosine Recognition by the SRA Domain of UHRF2.* Mol Cell, 2014.
60. Booth, M.J., et al., *Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution.* Science, 2012. **336**(6083): p. 934-7.
61. Li, Y., et al., *Selective capture of 5-hydroxymethylcytosine from genomic DNA.* J Vis Exp, 2012(68).
62. Song, C.X., et al., *Genome-wide profiling of 5-formylcytosine reveals its roles in epigenetic priming.* Cell, 2013. **153**(3): p. 678-91.
63. Ogoshi, K., et al., *Genome-wide profiling of DNA methylation in human cancer cells.* Genomics, 2011. **98**(4): p. 280-7.
64. Aran, D., et al., *Replication timing-related and gene body-specific methylation of active human genes.* Hum Mol Genet, 2011. **20**(4): p. 670-80.
65. Hellman, A. and A. Chess, *Gene body-specific methylation on the active X chromosome.* Science, 2007. **315**(5815): p. 1141-3.
66. Jjingo, D., et al., *On the presence and role of human gene-body DNA methylation.* Oncotarget, 2012. **3**(4): p. 462-74.
67. Maunakea, A.K., et al., *Conserved role of intragenic DNA methylation in regulating alternative promoters.* Nature, 2010. **466**(7303): p. 253-7.
68. Hogart, A., et al., *Genome-wide DNA methylation profiles in hematopoietic stem and progenitor cells reveal overrepresentation of ETS transcription factor binding sites.* Genome Res, 2012. **22**(8): p. 1407-18.
69. Hon, G.C., et al., *Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues.* Nat Genet, 2013. **45**(10): p. 1198-206.
70. Ziller, M.J., et al., *Charting a dynamic DNA methylation landscape of the human genome.* Nature, 2013. **500**(7463): p. 477-81.

71. Xie, W., et al., *Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells.* Cell, 2013. **153**(5): p. 1134-1148.
72. Kulis, M., et al., *Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia.* Nat Genet, 2012. **44**(11): p. 1236-42.
73. Seisenberger, S., et al., *The dynamics of genome-wide DNA methylation reprogramming in mouse primordial germ cells.* Mol Cell, 2012. **48**(6): p. 849-62.
74. Hammoud, S.S., et al., *Chromatin and Transcription Transitions of Mammalian Adult Germline Stem Cells and Spermatogenesis.* Cell Stem Cell, 2014.
75. Bogdanovic, O., et al., *Temporal uncoupling of the DNA methylome and transcriptional repression during embryogenesis.* Genome Res, 2011. **21**(8): p. 1313-27.
76. Fouse, S.D., et al., *Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation.* Cell Stem Cell, 2008. **2**(2): p. 160-9.
77. Baubec, T. and D. Schubeler, *Genomic patterns and context specific interpretation of DNA methylation.* Curr Opin Genet Dev, 2014. **25**: p. 85-92.
78. Jones, P.A., *Functions of DNA methylation: islands, start sites, gene bodies and beyond.* Nat Rev Genet, 2012. **13**(7): p. 484-92.
79. Boyes, J. and A. Bird, *Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein.* EMBO J, 1992. **11**(1): p. 327-33.
80. Hsieh, C.L., *Dependence of transcriptional repression on CpG methylation density.* Mol Cell Biol, 1994. **14**(8): p. 5487-94.
81. Lu, Y., et al., *Alternative Splicing of MBD2 Supports Self-Renewal in Human Pluripotent Stem Cells.* Cell Stem Cell, 2014.
82. Tao, J., et al., *Phosphorylation of MeCP2 at Serine 80 regulates its chromatin association and neurological function.* Proc Natl Acad Sci U S A, 2009. **106**(12): p. 4882-7.
83. Solomon, D.L., B. Amati, and H. Land, *Distinct DNA binding preferences for the c-Myc/Max and Max/Max dimers.* Nucleic Acids Res, 1993. **21**(23): p. 5372-6.
84. Donaldson, N.S., et al., *Kaiso regulates Znf131-mediated transcriptional activation.* Exp Cell Res, 2010. **316**(10): p. 1692-705.

Chapter 8

Discussion

DNA modifications

Biological processes are complex. Regulation of transcription is complex. The research described in this thesis has aimed to uncover, at least part of, the function of epigenetic modifications and one of the complexes binding to them. In the previous chapter, we have summarized what is known about DNA methylation and described its well-known link with repression of transcription initiation. In addition, based on the results obtained in **chapter 3** in which many transcription factors were identified as putative novel mC-binding proteins, **chapter 7** speculates that DNA methylation may not exclusively be associated with repression of transcription, but that the biological outcome of DNA methylation is also strongly dependent on sequence and CpG density.

Chapter 3 also describes readers for the oxidized derivatives of methylcytosine: hydroxymethyl-, formyl- and carboxylcytosine. Hydroxymethylcytosine interactors were identified in mouse embryonic stem cells (mESCs), neuronal precursor cells (NPC) and adult mouse brains. A number of these interactors, such as Wdr76 and Thy28, were common in all developmental stages and in different cell lines (unpublished data). These are most probably sequence-independent DNA binding proteins, just like the ubiquitous mC readers MBD1, 4 and MeCP2, as described in **chapter 7**. We also performed DNA affinity purifications in mESC nuclear extract with methylated and hydroxymethylated oligonucleotides containing the HoxC4 CpG island sequence, which in mESC is hydroxymethylated (unpublished data) [1]. Although proteins binding to the unmodified HoxC4 sequence showed a large overlap with the C-binders using the artificial sequence, a number of novel mC and hmC readers were identified. For mC we identified, in addition to Mbd1, 4 and MeCP2, Hic1, Hic2 and Rex1 as specific binders, which all have C2H2-type zinc fingers and probably bind to the methylated HoxC4 CpG island in a sequence-specific manner. In addition to the previously identified hmC-binding proteins Carf, Dhm1 and Thy28, the hmC-containing HoxC4 DNA was bound by Recq4, Nufip2 and Zfp206. Since these proteins are known to be sequence-specific DNA binders, our method thus reveals both sequence-dependent and -independent mC and hmC binding proteins.

The functions of the proteins binding to the different DNA modifications are still unclear. The method we described in **chapter 2** and **3** does not distinguish direct from indirect binders, as evidenced by the specific enrichment of the entire MBD2/NuRD complex binding to methylated DNA in NPC. However, this approach can be adjusted by incorporation of cross-linkable nucleotides in the DNA sequence to identify direct binders [2]. To study whether recruitment of proteins to target sites in the genome is driven by the modifications, chromatin immunoprecipitation (ChIP) experiments combined with enrichment strategies for the different modifications need to be performed.

More *in vivo* experiments will be required to unveil functions of fC and caC readers in the DNA damage response or in regulating genome stability. Examining the levels and the genomic patterns of fC and caC and their readers in cells that are depleted for the readers of these modifications may indicate whether the readers play a role in removal of the oxidized bases from specific functional DNA elements such promoters and enhancers. Other assays may include examining the levels of the different modifications in the same knock-out cell lines using mass spectrometry as described in **chapter 3**. Further experiments that may be pursued are colony formation assays after TET1 overexpression in the knock-out lines. If the hmC/fC/caC binding factors play a role in active DNA demethylation pathways, then the sustained high levels of hmC, fC and caC in the knock-out cells may induce more DNA damage compared to wild-type cells, leading to lesser (growth of) colonies. Comet-assays can be applied to test whether indeed more DNA damage is present in these cells.

Recruitment by histone modifications

In **chapter 4** we have identified putative readers for H3, H3K4me3 and H3K9me3 in liver, brain and kidney tissue. In contrast to readers for DNA modifications that vary extensively between cell types, readers for histone marks seem to be much more ubiquitously expressed. In some cases a tissue-specific reader was observed, e.g. CHD5 as a reader for H3K4me0 in brain cells. These tissue-specific readers will probably have yet to be discovered additional functions, since their ubiquitously expressed paralogues can be incorporated in the same complexes, making tissue-specific readers seem redundant.

Also within a tissue multipe readers for the same modification are expressed. Perhaps the apparent redundancy in function between different readers for the same histone modification is related to the fact that histone modifications themselves are much more diverse and dynamic than DNA methylation seems to be. With a few exceptions [3], DNA methylation patterns are quite stable and mainly change during cellular differentiation, while histone modification patterns change quickly after cellular signalling or during the cell cycle [4, 5]. Histone modifications therefore seem to act on a short time-scale to attract or stabilize binding of RNA polymerase and general transcription factors when a gene needs to be switched on. Different readers for the same modification may thus act on the target genes of different cellular signals.

In recent years, genome wide binding profiles of complexes that were believed to be associated with repression of transcription have been shown to overlap with H3K4me3 [6-9]. Similarly, we describe in **chapter 6** that ZMYND8 and MBD3 co-localize with H3K4me3 *in vivo*, while binding of ZMYND8 to the histone H3 N-terminus is inhibited by H3K4me3 *in vitro*. Even though co-enrichment of a protein with certain histone marks by ChIP-seq only shows a correlation and not a causative relationship between these two, these complexes may not be as repressive as we tend to think and may fine-tune rather than stably repress transcription. Depletion of these factors often leads to both up- and down-regulation of sets of genes [9], suggesting that these complexes are not clearly repressive or activating towards transcription. Explanations may be the repression of noisy, stochastic transcription, regulation of transcription cycles, or, when present in gene bodies, repression of transcription initiation from cryptic promoters [10-12] thereby ensuring higher levels of the full-length transcript. We thus need to change or view of these complexes and start addressing them as transcription regulation complexes rather than transcriptional repressive complexes.

Another issue complicating the biology of histone marks is that readers for histone modifications do not occupy all loci that are enriched for this modification. This indicates that target gene specificity is not exclusively determined by the histone modification a chromatin reader binds to. A histone mark may stabilize chromatin binding of a protein complex that obtains its specific binding pattern from a sequence specific transcription factor it transiently interacts with. The presence of attractive histone marks at these loci could lock the complex in position, especially with the presence of multiple histone binding domains in a single complex. The target-gene specificity of histone reading protein complexes thus is determined by DNA binding of the sequence-specific transcription factors it interacts with. A schematic representation of this model is shown in Figure 1. An example of such a combination of DNA sequence-specific recruitment and histone binding modules was described in **chapter 6**: the ZMYND8 protein, which mediates the interaction between a number of putative DNA-binding proteins (ZNF687, ZNF592 and ZNF532) and the histone binding modules-containing NuRD complex.

In summary, recruitment of transcription regulation complexes is a joint effort of specificity and stability. Specificity is obtained through interactions with DNA sequence-specific transcription factor and stability through binding of histone modifications by multiple reader domains. The specific targeting of histone reader complexes and different combinations of reader domains will reduce the redundancy between different readers of the same histone modification.

A combination of DNA-sequence and DNA modification dependent binding

As described above, we hypothesize that ZMYND8 mediates the interaction the putative DNA sequence specific Z3 module and the NuRD complex for specific recruitment of the complex to hypomethylated DNA. However, ZMYND8 may not only be recruited to chromatin by histone marks or DNA sequence specific transcription factors. Curiously, ZMYND8 was also identified as a binder of formylcytosine in **chapter 3.** In these affinity-enrichments ZMYND8 was accompanied by ZNF687, but not by any NuRD subunits. Since we know that these two proteins interact, we do not know which of them directly binds fC, though we would assume this to be ZNF687. The function of ZMYND8 in relation to fC is not known yet. As described in the introduction of this thesis, multiple NuRD subunits seem to have a role in the DNA damage response. Interestingly, so do many fC and caC readers. Could fC binding by ZMYND8/ZNF687 thus be the first step in NuRD recruitment to sites of DNA damage? But then, why was NuRD not enriched in these purifications? In any case, ZMYND8 binding to fC should be confirmed *in vivo* before a role can be assigned to it. To do so, one would need to do fC-specific sequencing on the DNA enriched by ZMYND8-ChIPs. Since fC is not very abundant, an fC enrichment step would need to be performed, which may be challenging.

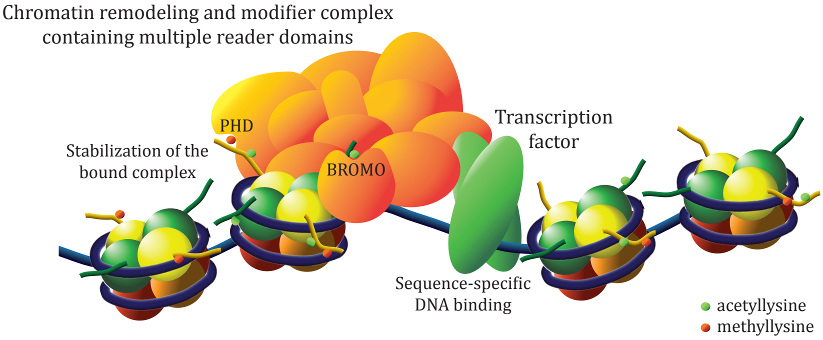

Chromatin remodeling and modifier complex containing multiple reader domains
PHD
BROMO
Stabilization of the bound complex
Transcription factor
Sequence-specific DNA binding
● acetyllysine
● methyllysine

Figure 1. Schematic model of cooperation between sequence-specific binding of transcription factors and stabilization of the interaction by multivalent interactions with histone modifications.

A role for ZMYND8 in leukemia?

Recently, ZMYND8 has been identified as a fusion-gene with RelA in a case of infant acute erythroid leukemia (AEL) [13]. RelA is also known as the NFκB transcription factor p65 [14]. The fusion consists of an almost complete ZMYND8 protein followed by the entire RelA protein. However, at this point it remains unclear whether the ZMYND8-RelA translocation induces oncogenic transformation. Assuming that this translocation product can act as an oncogene, multiple molecular mechanisms are possible. First of all, the RelA protein may hijack the strong nuclear localisation signal of ZMYND8, thereby circumventing inhibition by cytosolic IκB which may result in hyperactive NFκB signalling. This is a hallmark of many leukemias, for example in translocation events involving MLL [15]. Another mechanism to stimulate NFκB signalling would be if the fusion-protein would recruit the repressive NuRD and BHC complexes to the IκB gene, thereby repressing it. Normally, this gene is switched on shortly after activation of the NFκB signalling pathway as a negative feedback loop [14]. Repression of IκB expression, mediated by the NuRD and BHC complexes, may disable this negative feedback loop.

Of course, NFκB signalling has many more target genes than its own inhibitor. Given the fact that NFκB signalling plays an important role in cell survival and apoptosis, these target genes may also play a role in the development of leukemia. To investigate whether the fusion protein mainly localizes to ZMYND8/NuRD target genes or to NFκB target genes, ChIP-seq of both single proteins and the fusion protein should be performed, preferentially in haematopoietic cells. In addition, RNA-seq may reveal which genes are dysregulated in cells expressing the fusion gene.

Notably, the ZMYND2 protein, which also contains a MYND domain, is better known as the ETO part of the acute myeloid leukemia (AML)-causing AML-ETO gene fusion. Interestingly, this fusion protein requires the MYND-domain of ZMYND2/ETO for interaction with the N-CoR repressive complex [16]. In addition, the ZMYND8 interactor ZNF687 has also been identified as a RUNX1 fusion in AML [17]. RUNX1 is another name for AML. Thus, the RUNX1-ZNF687 fusion protein interacts via ZMYND8 with NuRD and the BHC complex, which is a similar scenario compared to AML-ETO interacting with N-CoR. These facts all may suggest that the ZMYND8-RelA fusion might indeed function as a cancer driving gene and cause acute erythroid leukemia.

THE NuRD complex does not exist

As evidenced by the many commonly occuring mutations of epigenetic factors in cancer [18], regulation of transcription regulation complexes is very important. The enzymatic activities and interactions of the NuRD complex are probably regulated by multiple mechanisms, just like its genomic recruitment by transcription factors. One of the factors that could have an activity-regulating role might be CDK2AP1 (also called DOC-1). In **chapter 5** we have shown that this protein is a *bona fide* subunit of the complex. CDK2AP1 is very small and doesn't have any known functional domains. The stoichiometry of this subunit in the NuRD complex varies between 1 and 2 copies per complex, depending on the cell type (unpublished data). However, since CDK2AP1 was shown to form homodimers [19], measured stoichiometries lower than two probably represent an average of two populations of the NuRD complex: the complexes with a CDK2AP1 dimer and the ones without. Since the stoichiometries of 1 and 2 mentioned above were obtained in different cell types in which also the transient interactors of the NuRD complex differ, this could suggest that CDK2AP1 regulates these interactions.

Interestingly, CDK2AP1 is, just like GATAD2A, one of the proteins showing the highest ratios in SILAC-based affinity purifications of ZMYND8, indicating that it may be close to the interaction surface between NuRD and ZMYND8. However, knock-down of CDK2AP1 did not significantly alter the association of ZMYND8 with MBD3-GFP (**chapter 6**). Another piece of evidence indicating that CDK2AP1 has an important role within the NuRD complex, is that an orthologue of CDK2AP1 exists in Drosophila melanogaster, an invertebrate organism that has a NuRD complex with much lower complexity than mammals. Although chapter 5 clearly shows that CDK2AP1 is a *bona fide* NuRD subunit, it does not elaborate on possible functions for this small protein. More functional studies including RNA-seq, MBD3/NuRD or ZMYND8 ChIPs and post-translational modification analyses of the NuRD complex in CDK2AP1 knock-out cells may shed light on the role of this small protein for the NuRD complex.

As mentioned in the previous paragraph, NuRD complexes with and without CDK2AP1 may be present in the cell. This is also true for other subunits. The NuRD complex contains 7 different subunits, and in humans two to three paralogues for almost every subunit are known. When we assume that all subunits are present in the stoichiometry described by Smits *et al.* [20] (Table 1), we can calculate the number of possible NuRD complex compositions.

Table 1

Stoichiometry	Paralogues	Possibilities
1	CHD3, CHD4 or CHD5	3
2	GATAD2A or GATAD2B (mutually exclusive)	2
3	MTA1, MTA2 and MTA3 (111, 222, 333, 122, 112, 223, 233, 331, 113, 123)	10
6	RBBP4 and RBBP7 (444444, 444447, 444477, 444777, 447777, 477777, 777777)	7
1	MBD2 or MBD3	2
2	HDAC1 and HDAC2 (11, 12, 22)	3
2	CDK2AP1	1

Multiplying all different possibilities would result in 2100 theoretical compositions of the NuRD complex. This calculation does not take into account that NuRD sub-complexes may exist that lack one or two subunits, such as CDK2AP1-less complexes. So how can we assign functions to a protein complex when its composition is so heterogeneous? When studying the NuRD complex, we should study each of the subunits in every experiment, to see whether the studied function requires all subunits of the complex or only a subset. Furthermore, testing each paralogue of the involved subunits will be useful to assign paralogue-specific functions. This approach will extend every cell biological assay to a laborious and time-consuming experiment, but it will result in clarification of the roles of different NuRD subunit paralogues. Therefore, development of high-throughput methods for studying the complex in various cellular functions, such as DNA damage response, would be desirable.

For any biochemical assay, purification of the complex would be required. Since it is not yet possible to reconstitute the complex from single subunits *in vitro*, this is almost impossible. Purification of a single composition of the NuRD complex would require sequential purifications for a single paralogue of each subunit. When applying this 7-step purification approach (one step per subunit to select a single paralogue), still one could not distinguish complexes containing one RBBP4 and five RBBP7 proteins from complexes containing one RBBP7 and five RBBP4 proteins, for example. Furthermore, purification of a complex using so many steps would require an enormous amount of protein material to start with. An alternative solution may be to express the paralogues in insect-cells using the baculovirus-based MultiBac system [21]. This would enable purification of the complexes that have a single paralogue of each subunit incorporated using a single affinity purification step, for example by expressing CHD4, GATAD2A, MTA2, RBBP4, MBD2, HDAC1, and CDK2AP1. These complexes can then be used to test their effect on *in vitro* transcription or in HDAC activity assays, for example, to study the effect of different histone modifications or the presence of sequence-specific DNA-binding proteins on the activity of the complex.

CONCLUSION

In summary, this thesis describes many novel DNA and histone readers. The modifications that these readers bind to regulate gene expression in ways that we still do not completely understand. The modifications and their readers play an important role during development, as evidenced by the existence of tissue-specific readers. The chromatin associated complexes they are part of contain multiple catalytic activities that need to be carefully balanced in order to regulate and fine-tune gene expression. Disturbance of the balance between transcription activating and repressing complexes will result in disease, such as cancer. Many of the genes that display high mutation rates in cancer are involved in transcription regulation in one way or the other [18]. These include transcription factors, histone modifying enzymes, proteins involved in DNA (hydroxy)methylation and readers for histone modifications. This indicates that dysregulation of transcription at any level can be disastrous for normal cell functioning.

Methylation-dependent or methylation-sensitive binding of transcription factors to their consensus sequence is most likely the first step in the recruitment of the multi-subunit complexes, such as the NuRD complex, that these factors interact with. This is followed by stabilization through multivalent interactions with histone tails carrying variable modifications, to attain the required biological outcome. Since the histone and DNA modifications act in conjunction with the DNA sequence underneath them, general mechanisms based on single model genes often turn out to be too simplistic. The interplay between DNA sequence and DNA and histone modifications for recruitment of chromatin-associated complexes requires much more attention, although the nucleosome pull-downs performed by Bartke *et al.* and Van Nuland *et al.* are a good start for deciphering the contributions of DNA sequence and histone modifications [22, 23]. Bioinformatics approaches will also be useful to decipher the different aspects that play a role in recruitment, since recent appearance of large datasets enables Meta-analysis of specific promoter features [24].

REFERENCES

1. Pastor, W.A., et al., *Genome-wide mapping of 5-hydroxymethylcytosine in embryonic stem cells.* Nature, 2011. **473**(7347): p. 394-7.
2. Pfaffeneder, T., et al., *Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA.* Nat Chem Biol, 2014. **10**(7): p. 574-81.
3. Kangaspeska, S., et al., *Transient cyclical methylation of promoter DNA.* Nature, 2008. **452**(7183): p. 112-5.
4. Schram, A.W., et al., *A dual role for SAGA-associated factor 29 (SGF29) in ER stress survival by coordination of both histone H3 acetylation and histone H3 lysine-4 trimethylation.* PLoS One, 2013. **8**(7): p. e70035.
5. Varier, R.A., et al., *A phospho/methyl switch at histone H3 regulates TFIID association with mitotic chromosomes.* EMBO J, 2010. **29**(23): p. 3967-78.
6. Shimbo, T., et al., *MBD3 localizes at promoters, gene bodies and enhancers of active genes.* PLoS Genet, 2013. **9**(12): p. e1004028.
7. Gunther, K., et al., *Differential roles for MBD2 and MBD3 at methylated CpG islands, active promoters and binding to exon sequences.* Nucleic Acids Res, 2013. **41**(5): p. 3010-21.
8. Whyte, W.A., et al., *Enhancer decommissioning by LSD1 during embryonic stem cell differentiation.* Nature, 2012. **482**(7384): p. 221-5.
9. Li, Y., et al., *Global transcriptional and translational repression in human-embryonic-stem-cell-derived Rett syndrome neurons.* Cell Stem Cell, 2013. **13**(4): p. 446-58.
10. Blake, W.J., et al., *Noise in eukaryotic gene expression.* Nature, 2003. **422**(6932): p. 633-7.
11. McAdams, H.H. and A. Arkin, *Stochastic mechanisms in gene expression.* Proc Natl Acad Sci U S A, 1997. **94**(3): p. 814-9.
12. Vollmers, C., et al., *Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome.* Cell Metab, 2012. **16**(6): p. 833-45.
13. Panagopoulos, I., et al., *Fusion of ZMYND8 and RELA genes in acute erythroid leukemia.* PLoS One, 2013. **8**(5): p. e63663.
14. Perkins, N.D., *Integrating cell-signalling pathways with NF-kappaB and IKK function.* Nat Rev Mol Cell Biol, 2007. **8**(1): p. 49-62.
15. Kuo, H.P., et al., *Epigenetic roles of MLL oncoproteins are dependent on NF-kappaB.* Cancer Cell, 2013. **24**(4): p. 423-37.
16. Liu, Y., et al., *Structural basis for recognition of SMRT/N-CoR by the MYND domain and its contribution to AML1/ETO's activity.* Cancer Cell, 2007. **11**(6): p. 483-97.
17. Nguyen, T.T., et al., *Identification of novel Runx1 (AML1) translocation partner genes SH3D19, YTHDf2, and ZNF687 in acute myeloid leukemia.* Genes Chromosomes Cancer, 2006. **45**(10): p. 918-32.
18. Kandoth, C., et al., *Mutational landscape and significance across 12 major cancer types.* Nature, 2013. **502**(7471): p. 333-9.
19. Ertekin, A., et al., *Human cyclin-dependent kinase 2-associated protein 1 (CDK2AP1) is dimeric in its disulfide-reduced state, with natively disordered N-terminal region.* J Biol Chem, 2012. **287**(20): p. 16541-9.
20. Smits, A.H., et al., *Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics.* Nucleic Acids Res, 2013. **41**(1): p. e28.

21. Bieniossek, C., T.J. Richmond, and I. Berger, *MultiBac: multigene baculovirus-based eukaryotic protein complex production.* Curr Protoc Protein Sci, 2008. **Chapter 5**: p. Unit 5 20.
22. Bartke, T., et al., *Nucleosome-interacting proteins regulated by DNA and histone methylation.* Cell, 2010. **143**(3): p. 470-84.
23. van Nuland, R., et al., *Multivalent engagement of TFIID to nucleosomes.* PLoS One, 2013. **8**(9): p. e73495.
24. Eijkelenboom, A., et al., *FOXO3 selectively amplifies enhancer activity to establish target gene regulation.* Cell Rep, 2013. **5**(6): p. 1664-78.

Appendix

Nederlandse samenvatting voor niet-ingewijden
List of abbreviations
Curriculum vitae
List of publications
Dankwoord

NEDERLANDSE SAMENVATTING VOOR NIET-INGEWIJDEN

De organisatie van een cel

Een menselijke **cel** kan omschreven worden als een grote fabriek: een gebouw vol machines en georganiseerd volgens een bepaald plan. In een cel zijn de **eiwitten** de machines, en vaak werken deze niet alleen, maar samen in groepen. Een groep van eiwitten die aan elkaar binden noemen we een **eiwitcomplex** (Figure 1A). De meeste eiwitten zijn pas actief als ze deel uitmaken van een eiwitcomplex. **Eiwit-eiwit interacties** zijn dus bepalend voor de functie van een eiwit. Een eiwit dat deel uitmaakt van zo'n eiwitcomplex, noemen we een **subunit** van dat complex. Omdat de eiwitten in een cel de machines zijn die bepaalde functies uitvoeren, bepaalt de combinatie van alle eiwitten in die cel wat voor soort cel het is (bijvoorbeeld een lever- of hersencel).

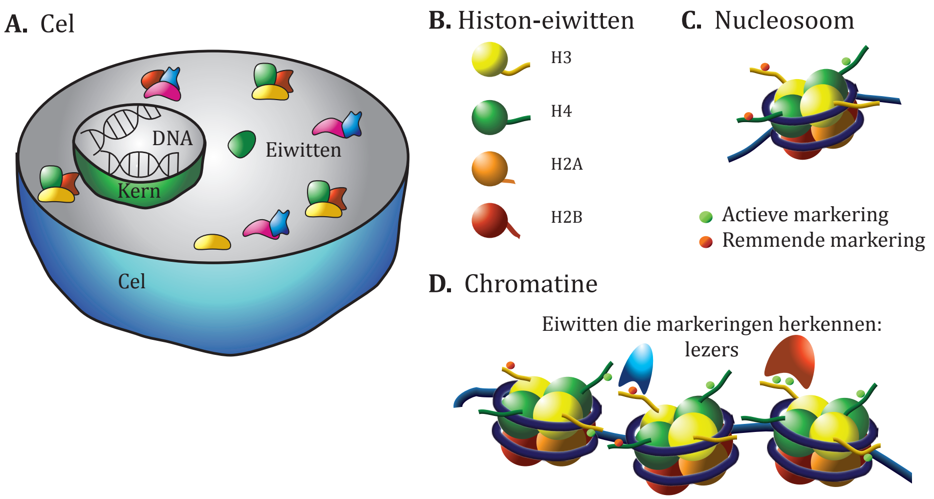
In de **kern van de cel** wordt de functie van de cel geregeld, door te bepalen welke eiwitten gemaakt moeten worden. Dit gebeurt aan de hand van een plan: het **DNA**. Het DNA bevat de informatie om alle eiwitten te maken. De informatie voor het maken van één eiwit zit in een **gen** (meervoud **genen**). Anders gezegd: een gen is een stukje van het DNA dat de bouwinstructies voor één eiwit bevat. Als er een nieuw eiwit gemaakt moet worden, wordt er een kopie van dat gen gemaakt. Het kopiëren noemen we **transcriptie**. De kopie noemen we **RNA**. Het RNA wordt buiten de celkern gebruikt om het eiwit te maken. Op deze manier blijft het DNA veilig in de kern van de cel, zodat het niet kwijt kan raken of kapot kan gaan.

De code van DNA bestaat uit vier verschillende structuren: A, G, C en T. De volgorde waarin deze letters staan, noemen we de **sequentie.** Alle cellen in ons lichaam bevatten dezelfde DNA sequentie, maar toch hebben cellen heel verschillende functies. Een rode bloedcel moet bijvoorbeeld zuurstof vervoeren en een levercel moet giftige stoffen afbreken. De cellen moeten dus de eiwitten maken die nodig zijn voor de functie van die cel. Om dat te doen worden genen gemarkeerd. Genen die niet nodig zijn worden uitgezet door het begin daarvan chemisch te veranderen (modificeren). Dit gebeurt bijvoorbeeld door de koppeling van een **methyl**groep aan de C's in het DNA. In 2009 en 2011 werden ook nog andere modificaties ontdekt, waarvan de functie nog niet bekend is: hydroxymethyl, formyl en carboxyl.

Niet alleen het DNA zelf wordt op die manier versierd. Het DNA ligt opgerold om eiwitten heen. Deze eiwitten heten **histonen** en kunnen op hun staart gemarkeerd worden. Die steekt uit en is dus makkelijk herkenbaar (Figure 1B). Op de histonen kunnen veel meer verschillende markeringen worden aangebracht dan op het DNA. Bijvoorbeeld methyl, acetyl, phospho en ubiquitine. Niet alleen de soort markering is belangrijk, maar ook of die aan het begin, eind of ergens midden op de staart zit. Het eiwitcomplex, bestaande uit acht histonen plus het DNA dat daaromheen gedraaid zit, heet een **nucleosoom** (Figure 1C).

De markeringen op het DNA en de histonen bepalen samen hoe een cel functioneert zonder het DNA te veranderen. Deze markeringen worden daarom **epigenetische modificaties** genoemd. Modificaties zijn epigenetisch als ze bij een celdeling worden gekopieerd, en transcriptie beïnvloeden zonder de DNA sequentie aan te passen.

Als die markeringen geen effect hebben op de DNA sequentie, hoe zorgen ze er dan voor dat genen aan of uit staan? Er zijn eiwitten die een bepaalde markering herkennen en daaraan binden. We noemen deze eiwitten de 'lezers'. De eiwitten



Figuur 1: DNA is opgeslagen in de celkern als chromatine. A. Een cel met daarin een kern. Het DNA zit in de kern en de meeste eiwitten (en eiwitcomplexen) daarbuiten. **B.** Er zijn vier verschillende histon-eiwitten. **C.** Acht histon-eiwitten met daaromheen een stukje DNA noemen we een nucleosoom. **D.** DNA dat in nucleosomen zit plus alle daaraan bindende eiwitten noemen we chromatine.

die gemethyleerd DNA herkennen zorgen ervoor dat dit DNA heel compact wordt opgeborgen, zodat het niet gekopieerd kan worden. En aan de histon-markeringen kunnen zowel eiwitten binden die de transcriptie stimuleren, als eiwitten die dit remmen, afhankelijk van de positie en soort van de markering. In hoofdstuk 3 en 4 hebben we veel nieuwe lezers geïdentificeerd.

Een van de eiwitcomplexen die zo'n methyl-DNA bindend eiwit bevatten, is het **NuRD** complex. Dit staat voor Nucleosoom Remodelerend en histon Deacetylerend complex. Oftewel, een complex dat nucleosomen over het DNA heen-en-weer kan schuiven en acetylgroepen van histonen kan verwijderen. Elk van deze activiteiten wordt uitgevoerd door een specifiek eiwit in het complex. Het **MBD2** eiwit in dit complex kan methyl-DNA herkennen. Zoals eerder beschreven is gemethyleerd DNA vaak inactief. Dit wordt bijvoorbeeld veroorzaakt door dit complex, omdat het verwijderen van de acetylgroepen van histonen, het aantal actieve (acetyl-)markeringen vermindert. Er bestaat echter ook een versie van het NuRD complex dat geen MBD2, maar MBD3 bevat. Dit eiwit kan niet aan gemethyleerd DNA binden. De vraag is daarom hoe het MBD3/NuRD complex naar genen wordt gerekruteerd. Met behulp van massa spectrometrie (hieronder in detail beschreven) hebben wij nieuwe interactie-partners van MBD3/NuRD gevonden. Dat werk is beschreven in hoofdstuk 6.

Er zijn ook eiwitten die aan actieve histon modificaties binden en de gemarkeerde genen kopieren. Dat zijn bijvoorbeeld algemene transcriptie factoren, zoals TFIID. Vaak werken deze eiwitten samen met **DNA sequentie-specifieke transcriptie factoren.** Deze sequentie-specifieke transcriptie factoren herkennen een bepaald woord in het DNA en ze regelen vaak één bepaalde functie in de cel. Alle genen die nodig zijn voor die functie bevatten het herkenningswoord. Op die manier heb je enkel een hoop kopieën van 1 sequentie-specifieke transcriptie factor nodig om al die genen aan te zetten. Eén van die functies is de stam-cel identiteit. Een stam cel kan heel vaak delen en heeft het potentieel om in allerlei andere celtypen te veranderen. Wanneer een cel het signaal krijgt om in een neuron (hersencel) te veranderen, wordt de stam-cel sequence-specifieke transcriptie factor niet meer gemaakt en alle genen waar hij aan bind gaan uit. Daarentegen wordt er nu een sequentie-specifieke transcriptie factor gemaakt die bijvoorbeeld alle neuron-specifieke genen aan zet. Op die manier kunnen heel snel veel verschillende genen uit- of aangezet worden.

Het geheel van DNA, histonen, modificaties en alle eiwitten die daar aan binden (zoals het NuRD complex en transcriptie factoren) wordt **chromatine** genoemd (Figure 1D). Vaak wordt dit ingedeeld in twee soorten: compact als de genen in-actief zijn (**heterochromatine**) of open, omdat de genen vaak worden gekopieerd (**euchromatine**).

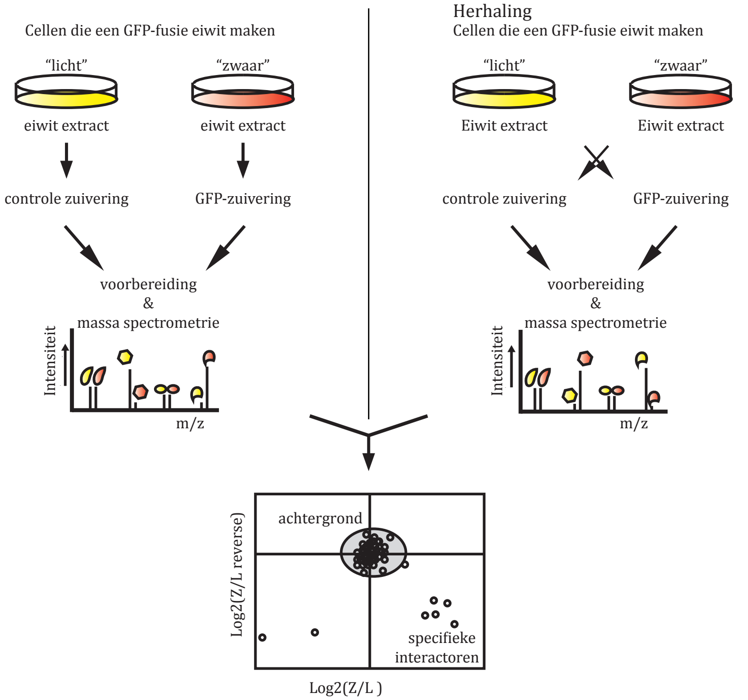
Om te bestuderen in welke complexen onze eiwitten functioneren, zuiveren we ze. Hiervoor maak ik een Groen Fluorescerend Eiwit (**GFP**) aan mijn eiwit vast. We noemen het geheel van een eiwit (waarin we geïnteresseerd zijn) met daaraan GFP een **fusie-eiwit**. Door in een reageerbuis de DNA sequentie te maken die de code voor het fusie-eiwit bevat en dit DNA in cellen te stoppen, kunnen de cellen het GFP-fusie eiwit maken. Met behulp van een GFP-specifiek **antilichaam** kan ik het eiwit waarin ik geïnteresseerd ben zuiveren. Alle eiwitten die aan mijn eiwit binden, omdat ze in hetzelfde eiwitcomplex zitten, zullen meegezuiverd worden. Om te testen welke eiwitten er aan mijn eiwit binden, kunnen we western blot gebruiken. Bij deze techniek gebruik je een antilichaam dat heel specifiek één eiwit herkent. We binden dan de eiwitten uit het sample op een membraan en testen of het antilichaam aan die eiwitten bindt. Een voordeel van western blot is dat het een snelle techniek is: na een dag weet je het resultaat. Maar het heeft ook een aantal nadelen. Ten eerste moet je al een idee (**voorkennis**) hebben over welke eiwitten er in het sample zouden zitten. Ten tweede moet er dan ook nog een **specifiek** antilichaam tegen dat eiwit beschikbaar zijn. Omdat mensen zo'n 20 000 verschillende eiwitten hebben waarvan sommige ook nog heel erg op elkaar lijken, is dat niet altijd het geval. En ten derde is het veel werk om op die manier **veel verschillende** eiwitten te bekijken. Een techniek zonder deze nadelen is massa spectrometrie.

Massa spectrometrie

Voor het onderzoek in deze thesis is veel gebruik gemaakt van massa spectometrie. Dit is een techniek die heel nauwkeurig kan meten hoeveel een molecuul weegt. **Eiwitten** zijn opgebouwd uit 20 verschillende bouwblokken, de **aminozuren**, die in een lange keten achter elkaar worden gezet. Bijna al deze aminozuren kunnen in massa onderscheiden worden. Om massa spectrometrie te doen, knippen we eerst het eiwit in een reageerbuis in stukjes van zo'n 10 tot 30 aminozuren (een **peptide**). We stoppen deze peptiden in de massa spectrometer. Die meet de precieze massa van elk peptide, en dus de zeer waarschijnlijke aminozuur-samenstelling. Verder bepaalt de massa spectrometer een deel van de aminozuur-volgorde. Omdat we de DNA sequentie van mensen precies weten, kennen we ook (bijna) alle eiwitten die door een cel gemaakt kunnen worden. Al die informatie bevindt zich in de database. Nu zoeken we in deze database, met behulp van de combinatie van de samenstelling van de peptiden en de gedeeltelijke volgorde die we weten, naar het hele eiwit. Op deze manier kunnen we bijna zonder aannames bepalen welke eiwitten er in ons sample zitten. Daarbij komt, dat in een vrij korte meting van twee-en-een-half uur ongeveer 3000 eiwitten in een sample geïdentificeerd kunnen worden. En omdat zoveel eiwitten kunnen worden geïdentificeerd, weten we vaak niet meer naar welke eiwitten we op zoek waren. We moeten dus meerdere samples met elkaar vergelijken.

Stel: je wilt zien welke eiwitten er in een complex zitten. Daarvoor zuiveren we één eiwit van dat complex uit een cel-extract. Om te controleren welke eiwitten gewoon plakkerig zijn en daardoor mee-gezuiverd worden, doen we ook een controle-zuivering (alle componenten bij elkaar maar zonder antilichaam tegen GFP). Als je nu allebei die samples in de massa spectrometer stopt, dan identificeer je heel veel eiwitten, waarvan een groot deel hetzelfde is in de twee samples. Maar als je wilt weten of de hoeveelheid van één bepaald eiwit hoger is in de specifieke zuivering van jouw eiwit, dan is dat wat lastiger. Niet elke massa spectrometrie-meting is hetzelfde en daardoor kan de **intensiteit** van hetzelfde eiwit in twee verschillende metingen niet direct met elkaar vergeleken worden. Hiervoor moet eerst een normalisatie gedaan worden. Daarbij komt, dat je beide samples drie keer moet meten, om te bepalen welke eiwitten statistisch gezien meer voorkomen in de specifieke zuivering. In een aantal gevallen gebruik ik deze methode, die we **LFQ** noemen, omdat hiermee meer dan twee situaties met elkaar vergeleken kunnen worden.

Er zijn ook andere manieren om de hoeveelheden van elk eiwit in verschillende situaties met elkaar te vergelijken. Hierbij wordt gebruik gemaakt van het inbouwen van stabiele (niet-radioactieve) **isotopen**. Deze isotopen zijn net iets zwaarder dan de van nature meest voorkomende variant. Als je dus de natuurlijke (**lichte**) en de isotoop-ingebouwde (**zware**) peptiden samen in de massa spectrometer doet, kun je die van elkaar onderscheiden. Je ziet dan in het spectrum twee



Figuur 2: Schematische weergave van een eiwitzuivering met gebruik van SILAC.

pieken; die van het lichte en die van het zware peptide. Verder veranderen de stabiele isotopen niets aan de biologische of chemische eigenschappen van het eiwit. Deze isotopen kunnen op verschillende punten tijdens de proefopzet ingebouwd worden. Wij hebben veel gebruikt gemaakt van **SILAC** (Stable Isotope Labeling by Amino acids in Cell culture, oftewel het labelen met aminozuren die stabiele isotopen bevatten tijdens het kweken van de cellen).

Terwijl de cellen groeien gebruiken ze de aangeleverde aminozuren om nieuwe eiwitten te maken. De isotopen worden in die cellen dus in alle eiwitten ingebouwd. Voor de zuivering van het eiwitcomplex wordt de proefopzet als volgt: De lichte cellen worden gebruikt voor de controle-zuivering en de zware cellen worden gebruikt voor de specifieke zuivering (see Figure 2). Na de zuivering kunnen de samples gecombineerd worden. Hierna wordt het sample (door het combineren blijft er maar één sample over) voorbereid voor massa spectrometrie. Omdat alle eiwitten van de specifieke zuivering zwaar zijn, kunnen in het massa spectrum per peptide twee signalen worden gezien; de lichte (negatieve controle) en de zware (specifieke zuivering). De intensiteit van deze twee signalen kan wel direct met elkaar worden vergeleken, want ze zijn in dezelfde massa spectrometrie meting gevonden. Eiwitten die specifiek binden aan jouw gezuiverde eiwit hebben een ratio van de zware intensiteit/lichte intensiteit van meer dan twee. Eiwitten die een zwaar/licht ratio van één hebben, zijn niet interessant en worden ´achtergrond'-eiwitten genoemd. Zij zijn namelijk met dezelfde concentratie aanwezig in de controle als in de specifieke zuivering. Om met meer zekerheid te weten of dit de eiwitten zijn die we willen identificeren, herhalen we het experiment. In de herhaling wisselen we echter lichte en zware cellen. Dus nu doen we de controle-zuivering met zwaar extract en de specifieke zuivering met licht extract. De ratio's van beide experimenten kunnen in een puntenwolk-grafiek met elkaar vergeleken worden. En alleen de eiwitten die in de hoek rechtsonder terecht komen zijn specifieke interactoren van ons GFP-fusie eiwit. Ze zijn duidelijk te onderscheiden van de achtergrondeiwitten die een dichte groep in het midden van de grafiek vormen.

Deze methode heeft ook zo zijn voor- en nadelen. Doordat de samples vroeg in het stappenplan worden gecombineerd, is de meting van de concentratieverschillen uiteindelijk nauwkeuriger. Daarnaast zijn er maar twee massa spectrometrie-meting nodig om de verschillen tussen twee situaties te meten. Het nadeel is dat er maar twee, of in speciale gevallen drie, verschillende situaties met elkaar vergeleken kunnen worden.

Hierboven heb ik een begrijpelijke uitleg gegeven over mijn onderzoeksvragen en de technieken die ik gebruikt heb om ze te beantwoorden. Hierna geef ik in het kort per hoofdstuk aan wat de uitkomsten van mijn onderzoek zijn.

Samenvatting van mijn thesis

In **hoofdstuk 1** geef ik in meer detail een inleiding over de markering van DNA en histonen en over de eiwitten die hieraan binden. Ook leg ik de verschillende massa spectrometrie methoden uit die ik gebruik. In **hoofdstuk 2** wordt stap voor stap een methode uitgelegd, waarmee je eiwitten kunt identificeren die specifiek aan DNA met methylgroepen binden. Ik heb deze methode ook gebruikt, zoals beschreven is in **hoofdstuk 3**. In dit hoofdstuk beschrijven we hoe we eiwitten hebben geïdentificeerd die aan verschillende DNA markeringen kunnen binden. Sinds een paar jaar zijn er, naast methylering, nieuwe markeringen bekend: hydroxymethyl, formyl en carboxylgroepen.

Het is nog niet duidelijk wat hun functie is voor de cel. Omdat we denken dat de meeste functies in de cel worden uitgevoerd door eiwitten, wilden we weten wat voor soort eiwitten aan de nieuwe markeringen zouden binden. We hebben dit onderzocht voor embryonale stamcellen, neuronale voorganger cellen en muizenhersenen, omdat de markeringen en ook de eiwitten verschillende concentraties hebben in verschillende celtypen en vooral veel aanwezig zijn in de hersenen. We hebben in dit hoofdstuk veel nieuwe eiwitten geïdentificeerd en deze blijken ook voor een groot deel weefsel- (of celtype) specifiek te zijn. Hoogstwaarschijnlijk dienen formylcytosine en carboxylcytosine vooral voor het verwijderen van DNA methylatie, waar nog geen enzym voor bekend is. Methylcytosine en hydroxymethylcytosine hebben daarnaast ook een rol in het reguleren van transcriptie.

We hebben een zelfde soort proefopzet gebruikt om in muizenlever, -nieren en -hersenen te testen welke eiwitten aan welke histon-markering binden, zoals beschreven is in **hoofdstuk 4.** Opmerkelijk genoeg vonden we hier minder weefsel-specifieke eiwitten dan voor de DNA markeringen. Interessant was dat een aantal eiwitten hetzelfde bindingspatroon vertonen als de bekende NuRD-subunits. Na nadere inspectie bleken deze eiwitten ook interactie aan te gaan met het NuRD-complex. In **hoofdstukken 5 en 6** heb ik de interactie van een van deze eiwitten en het CDK2AP1 eiwit met het NuRD-complex in meer detail gekarakteriseerd. Hiervoor heb ik massa spectrometrie gebruikt, maar ook andere technieken. We weten nu dat het ZMYND8 eiwit (gevonden in hoofdstuk 4 en in detail bestudeerd in hoofdstuk 6) het NuRD complex verbindt met sequentie-specifieke transcriptie factoren. Ook zit ZMYND8 op dezelfde plekken van het DNA als NuRD. We denken dat het op die manier MBD3/NuRD naar bepaalde plekken op het DNA zou kunnen sturen.

In **hoofdstuk 7** beschrijf ik de nieuwe inzichten die we hebben gekregen omtrent DNA-bindende eiwitten. De basis hiervoor zijn zowel mijn studie van eiwitten die aan gemarkeerd DNA binden, alsmede studies van anderen waarin onderzocht wordt welke stukken DNA gemarkeerd zijn. DNA methylatie blijkt een minder eenduidig signaal te zijn dan we altijd dachten. Naast het remmen van transcriptie, blijken methyl-markeringen voor sommige activerende transcriptie factoren ook nodig om aan DNA te kunnen binden.

In **hoofdstuk 8** geef ik een meer algemene conclusie van mijn onderzoek en beschrijf ik op welke punten het onderzoek nog niet volledig is. Een van de belangrijkste conclusies is dat histon-markeringen op zichzelf waarschijnlijk niet genoeg zijn om eiwitten naar het DNA te trekken, maar dat voor de specifieke regulatie van genen ook DNA sequentie-specifieke transcriptie factoren nodig zijn. Ik geef in dit hoofdstuk ook aan welke vragen nog openstaan en op welke manieren deze wellicht beantwoordt zouden kunnen worden.

LIST OF ABBREVIATIONS

BER	Base Excision Repair
BHC	Braf-HDAC Complex
caC	carboxylcytosine
CDK2AP1	Cyclin-dependent kinase 2 Associated Protein 1 (also called DOC-1)
CGI	CpG island
CHD	Chromodomain-Helicase and DNA-binding protein
ChIP	Chromatin ImmunoPrecipitation
CHIP-seq	ChIP followed by whole genome sequencing
CID	Collision Induced Dissociation
CpG	CG-dinucleotide
CXXC	CxxC domain-containing protein, recognize non-methylated CGI
DBD	DNA binding domain
DNMT	DNA Methyl-Transferase
DOC-1	Deleted in Oral Cancer 1 (also called CDK2AP1)
EMSA	ElectoMobility Shift Assay
ESI	Electrospray Ionisation
fC	formylcytosine
FDR	False Discovery Rate
H3K4me3	Histone H3 trimethylated at lysine 4
H3K9,14ac	Histone H3 acetylated on lysine 9 and 14
HAT	Histone Acetyl-Transferase
HDAC	Histone De-Acetylase
hmC	hydroxymethylcytosine
HMT	Histone Methyl-Transferase
iBAQ	intensity-Based Absolute Quantification
KDM	Lysine (K) DeMethylase
LC-MS/MS	Liquid Chromatography-tandem Mass Spectrometry
LFQ	Label-Free Quantification (in this thesis MaxLFQ)
LSD	Lysine-Specific Demethylase
MBD	Methyl-CpG Binding Domain
MBD2	Methyl-CpG Binding Domain-containing protein 2
MBD3	Methyl-CpG Binding Domain-containing protein 3
MBP	Methyl-C Binding Protein
mC	methylcytosine
mCpG	methylcytosine in CpG context
MeCP2	Methyl-CpG binding protein 2
mESC	mouse Embryonic Stem Cell
MTA	Metastasis associated protein
MYND	Myeloid, Nervy, Deaf; protein-protein interaction domain
NFKB	Nuclear Factor KB
NPC	Neuronal Precursor Cells
NuRD	Nucleosome Remodeling and Deacetylase complex
PHD	Plant Homology Domain, domain commonly binding to the H3 tail
PPI	Protein-Protein Interaction
PRC	Polycomb Repressive Complex
PTM	Post-Translational Modification
PWWP	domain commonly binding H3K36me3
RbAp	Retinoblastoma Associated protein
RBBP	RetinoBlastoma Binding Protein
SILAC	Stable Isotope Labelling by Amino acids in Cell culture
STAGE-tips	Stop-And-Go-Extraction tips
TDG	Thymine-DNA Glycosylase
TET	Ten-Eleven Translocation
TNFα	Tumor Necrosis Factor α
UHRF	Ubiquitin-like, containing PHD and RING finger domains
Z3	a protein module consisting of ZNF687, ZNF592 and ZNF532
ZNF	Zinc finger containing protein

CURRICULUM VITAE

Cornelia Gijsbertha (Nelleke) Spruijt werd geboren op 8 maart 1987 in Amersfoort. Na haar basisschoolperiode op de Oranje Nassauschool in Nijkerk, begon zij in 1999 aan het VWO op het Christelijk College Groevenbeek. Hier deed ze de profielen Natuur&Techniek en Natuur&Gezondheid en behaalde ze in juni 2005 haar diploma. In september dat jaar begon zij aan haar Bachelor Scheikunde aan de Universiteit Utrecht. Ze rondde deze studie in 2008 af en startte in september, ook aan de UU, haar Master Biomolecular Sciences. Tijdens haar master deed ze twee stages; de eerste in de groep van Paul van Bergen en Henegouwen, waar ze werkte aan de herkenning van ubiquitine en ubiquitine-gelijkende domeinen door ubiquitine-interactie motieven. Haar tweede stage deed Nelleke in de groep van Michiel Vermeulen. Hier bestudeerde zij het nieuwe NuRD-subunit DOC-1, wat leidde tot haar eerste publicatie. Tijdens haar Master nam Nelleke ook deel aan het X-track honours programma, waarvoor zij een onderzoeksvoorstel over hydroxymethylcytosine schreef. Na het *cum laude* afronden van haar Master in augustus 2010 begon Nelleke als OIO in de groep van Michiel Vermeulen, waar ze het onderzoeksvoorstel omtrent hydroxymethylcytosine ten uitvoer bracht. Daarnaast bleef ze aan het NuRD-project werken. Tijdens haar promotie heeft Nelleke ook meerdere samenwerkingen tot een goed einde gebracht.

LIST OF PUBLICATIONS

C.G. Spruijt and M. Vermeulen (2014). "DNA methylation: old dog, new tricks?" Nature Structural and Molecular Biology **21**(11): 949-954

T. Pfaffeneder, F. Spada, M. Wagner, C. Brandmayr, S. K. Laube, D. Eisen, M. Truss, J. Steinbacher, B. Hackner, O. Kotljarova, D. Schuermann, S. Michalakis, O. Kosmatchev, S. Schiesser, B. Steigenberger, N. Raddaoui, G. Kashiwazaki, U. Muller, **C.G. Spruijt**, M. Vermeulen, H. Leonhardt, P. Schar, M. Muller and T. Carell (2014). "Tet oxidizes thymine to 5-hydroxymethyluracil in mouse embryonic stem cell DNA." Nature Chemical Biology **10**(7): 574-581.

H.I. Baymaz, **C.G. Spruijt** and M. Vermeulen (2014). "Identifying nuclear protein-protein interactions using GFP affinity purification and SILAC-based quantitative mass spectrometry." Methods in Molecular Biology **1188**: 207-226.

M. Escamilla-Del-Arenal, S.T. da Rocha, **C.G. Spruijt**, O. Masui, O. Renaud, A.H. Smits, R. Margueron, M. Vermeulen and E. Heard (2013). "Cdyl, a new partner of the inactive X chromosome and potential reader of H3K27me3 and H3K9me2." Molecular and Cellular Biology **33**(24): 5005-5020.

C.G. Spruijt, H.I. Baymaz and M. Vermeulen (2013). "Identifying specific protein-DNA interactions using SILAC-based quantitative proteomics." Methods in Molecular Biology **977**: 137-157.

C.G. Spruijt*, F. Gnerlich*, A.H. Smits, T. Pfaffeneder, P.W.T.C. Jansen, C. Bauer, M. Munzel, M. Wagner, M. Muller, F. Khan, H.C. Eberl, A. Mensinga, A.B. Brinkman, K. Lephikov, U. Muller, J. Walter, R. Boelens, H. van Ingen, H. Leonhardt, T. Carell and M. Vermeulen (2013). "Dynamic readers for 5-(hydroxy)methylcytosine and its oxidized derivatives." Cell **152**(5): 1146-1159.

H.C. Eberl, **C.G. Spruijt**, C.D. Kelstrup, M. Vermeulen and M. Mann (2013). "A map of general and specialized chromatin readers in mouse tissues generated by label-free interaction proteomics." Molecular Cell **49**(2): 368-378.

S.J. Bartels***, C.G. Spruijt***, A.B. Brinkman, P.W.T.C. Jansen, M. Vermeulen and H.G. Stunnenberg (2011). "A SILAC-based screen for Methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein." PLoS One **6**(10): e25884.

C.G. Spruijt*, S.J. Bartels*, A.B. Brinkman, J.V. Tjeertes, I. Poser, H.G. Stunnenberg and M. Vermeulen (2010). "CDK2AP1/DOC-1 is a *bona fide* subunit of the Mi-2/NuRD complex." Molecular Biosystems **6**(9): 1700-1706.

*These authors contributed equally

DANKWOORD

Opeens is het dan zover… je promotie. Als je net begint lijkt 4 jaar enorm lang. Maar als je bezig bent, druk bent met alle experimenten en het naar je zin hebt, dan is het zo voorbij. Ik heb het enorm naar mijn zin gehad in de groep en op de afdeling, en wil alvast iedereen daarvoor hartelijk bedanken. Maar natuurlijk wil ik ook een aantal mensen meer persoonlijk bedanken.

Allereerst wil ik natuurlijk Michiel bedanken, omdat hij me de mogelijkheid heeft geboden om in zijn lab stage te lopen en daarna als PHD te blijven. Ik weet nog goed hoe ik de eerste uitbreiding was van de 'Michiel-en-Pascal-groep'. We zaten met zijn drietjes in één kantoor. Michiel, hartelijk bedankt voor je begeleiding tijdens mijn stage en PHD traject. Je hebt me de vrijheid geboden alle technieken te leren die ik wilde en waar nodig heb je me wat bijgestuurd. Ik heb het altijd erg naar mijn zin gehad in je groep. Ik wens je heel veel succes toe in Nijmegen, de eerste resultaten zien er in elk geval veelbelovend uit!

Ten tweede wil ik Marc bedanken, omdat de Vermeulen groep (en daarmee ikzelf) toch ook gegroeid is dankzij de vele gezamenlijke werk- en literatuurbesprekingen. Op gebieden waar Michiel minder ervaring had, konden we altijd rekenen op jouw advies en hulp. Heel erg bedankt voor alle jaren advies, steun en goede ideeën.

Dan kom ik nu bij mijn paranimfen: Pascal en Arne. Pascal, door de jaren heen heb ik je vaak gevraagd om mass spec advies en om hulp bij het meten van heel wat samples. Daarnaast was het met jou altijd gezellig op 't lab en tijdens borrels. Ik hoop dat je naast het runnen van de Fusion ook nog tijd hebt om lekker wat experimenten in het lab te doen. Arne, mede-OIO. Ik heb altijd het gevoel gehad dat je pas een paar jaar na mij was begonnen, maar jij zit ook al in je 4de jaar! Ook met jou kan ik altijd goed opschieten, met je grappen en je serieuze kanten. Ik kan met jou ook vooral de theoretische en analyse-kant van experimenten goed bespreken. En dankzij jou kan de hele groep nu scatterplots maken met R. Tijdens mijn stage moest ik Michiel nog voor elke plot lastig vallen… Je gaat als een trein door je promotie en ik weet zeker dat je ook daarna goed terecht zult komen. Volgens mij kan je met iedereen goed overweg en netwerken is dan ook een ander sterk punt van je. Je hebt me tijdens meetings al aan heel wat mensen voorgesteld. Helaas verhuisde de groep naar Nijmegen en heb ik het laatste jaar veel minder contact gehad met jullie. Maar jullie weten dat ik mijn best heb gedaan zo vaak mogelijk langs te komen. En als er een borrel werd georganiseerd waren de leden van de Vermeulengroep vaak de laatsten die vertrokken. Al met al, jongens, is het met jullie altijd gezellig. Heel erg bedankt dat jullie mijn paranimfen willen zijn!

Ik wil ook graag alle leden van mijn beoordelings- en AIO-commissie bedanken. Boudewijn en Geert, tijdens de eerste AIO evaluatie hebben jullie me geleerd beslissingen te maken. Helaas is dat nog wel eens moeilijk, omdat alles zo interessant is. Albert, Frank, Gert-Jan, Joost en Harmjan, bedankt voor het lezen en beoordelen van mijn proefschrift.

Of course, I want to thank the rest of the Vermeulen group. Even after the move to Nijmegen I tried to have as much contact with all of you as possible. Danny, je was er maar een jaar, maar daarin hebben we veel contact gehad. Irem, what a coincidence that we have our birthday on the same day! I always liked it to be roomies on meetings or to have dinner together. Please hold on to your perseverance and finish your PHD well! I know you can do it. Radhika, I really wish your post-doc would have been easier. However, I think you really made the best of it together with Anneloes. I hope your project will be wrapped up into a nice story soon. Remember to step up for yourself whenever you need. Don't let others spoil your valuable time when you're busy! I wish you all the best for your future, where-ever you may go. Anneloes, ik bewonder je loyaalheid aan Radhika en het EMSY project. Ik hoop dat het mooi wordt afgesloten en dat je een leuke nieuwe baan vindt. Susan, my fellow NuRD-buddy! Thanks for the discussions of data on our shared complex and thanks for all your advice for my America trip. I hope someday you will appreciate all the possibilities that the move to Nijmegen brought for the group, for example when you publish a nice story with loads of ChIP-seq data you analyzed yourself! Matthew and Raghu, even though the two of you only worked in Nijmegen, for me, you are really my colleagues. Every time we meet feels like I know you already for a long time. I wish you all the best for your projects and you can ask me anything concerning DNA pull-downs or other experiments. Marijke, je was een collega in Utrecht, en nu weer in Nijmegen. Ik ben heel blij voor jou en voor de groep dat je nu voor Michiel werkt! Ik hoop dat je nog heel veel belangrijke bijdrages aan mooie papers zult leveren. Ino, wat leuk dat je terug gekomen bent om je PHD bij Michiel te doen! Heel veel succes met je MBD2 project. Xiaofei en Cristina, I didn't meet you that many times, but I am sure you will be valuable to the group. I wish you all the best with your projects.

In the Timmers group I would also like to thank many people. Sjoerd, Markus and Nikolai Mischerikow, thanks for being around and answering silly questions during my first year in the lab. Nikolay Outchkourov, I like your calm way of thinking and the way you saved me during the paintball game. Gianpiero, it was really nice to work with you. I hope you are having a good time in London. Andrée, het was erg fijn om je in de groep te hebben. Ik hoop dat je gelukkig bent met je keuze de wetenschap te verlaten en ik wens je alle goeds voor de toekomst. Petra, ik heb altijd bewondering gehad voor je efficiënte werkwijze en hoe je toch tijd had om anderen te helpen. Bedankt voor alle hulp en adviezen. Ik hoop dat alle jaren keihard werken snel zullen uitbetalen voor je. Rick, onze mini-werkbesprekinkjes waren altijd erg nuttig. Ik vond het fijn om bijna dagelijks met jou even te bespreken wat we aan het doen waren, zodat we elkaar onbewust ook tips en trucs konden geven of kritisch laten nadenken over controles. Daarnaast was het erg gezellig met jou en Pascal in één lab. En nu zit je in Amerika. Ik wens je een heel succesvolle post-doc periode en goede toekomst. Maria, je bent er bijna! En het gaat helemaal goedkomen! Ik vind het jammer dat ik niet vaker een praatje over experimenten kon maken met je, alleen maar omdat ik niet zoveel van gist weet. Ik hoop dat je een leuke baan vindt voor na je PHD en ik weet zeker dat je doorzettingsvermogen je daarbij zal helpen! Roy, altijd enthousiast! Dat is goed om je door een PHD heen te slaan en ik denk dat het je zal helpen om ver te komen in de wetenschap! Want enthousiasme en inzet zijn al de helft van het slagen van een project, en daar ontbreekt het bij jou niet aan. Elfi, wat leuk om ook een MD in de groep te hebben! Je hebt nieuwe methoden en denkwijzen aan ons laten zien en ik geloof dat dat voor ons allemaal verrijking is. Ik hoop dat je een mooie plek vindt om in de kliniek aan de slag te gaan en dat je je PHD naar tevredenheid af kunt sluiten. Simona, always cheerful! It was really nice to have such a fresh PHD student running around, and to discuss your and my experiments together. I think we learned a lot from each other. Keep up the good work and one day you'll find some very nice and world-changing X-links in your project! Hetty, zonder jou wordt het lab echt een zootje. Jij regelt zoveel, daar is iedereen je dankbaar voor. Richard, je weet echt ontzettend veel van biochemie en eiwitzuiveringen. Bedankt voor je hulp en het synthetiseren van peptiden.

Of course, I would also like to thank all bachelor and master students that have helped me on my project or kept up the good atmosphere in the group. Deepani, thanks for winning the bet for me! Your experiments showed that my theory was right. I admire the way you developed during your internship. From never holding a pipet, to designing and performing your own clonings and experiments. I think you will get very far in science! Moritz, I realize it must have been difficult to start up a project when your supervisor was so busy with another one. But you did well and I am still using the cell lines you made! You were eager to learn new techniques and I think this is a quality that will help you through your PHD as well. Good luck in Munich with all your experiments! Of course Caroline, Corina, Rik and Lisa, thanks for the nice atmosphere and very good borrels!

Ik wil natuurlijk ook de andere mensen bedanken die veel tijd bij de massa spectrometer doorbrengen. Allereerst, Harmjan. Bedankt voor je altijd rustige en opbouwende commentaar. Zelfs als ik iets fout deed bleef jij rustig en wist je een oplossing. Ik hoop dat het Proteins@Work een succes wordt en dat het je genoeg uitdaging blijft bieden. Vincent, je bent altijd enthousiast en ervoor in om nieuwe dingen te proberen. Dankzij jouw experimenten, proberen wij soms ook nieuwe technieken zoals dimethyl en cross-linking. Ga zo door en dan komen er vast meer mooie mass spec studies bij jullie uit de groep! And Sibel and Robert, we didn't have so much contact, but I hope you'll both have a nice time working at MCR.

Natuurlijk mag ik de mensen niet vergeten die niet direct met onderzoek te maken hebben, maar wel een belangrijke ondersteunende rol: de secretaresses, ICT, Marian Bömer en Marjoleine, Cheuk, Marcel en Huub. Andrea, Sandy, Betty, Cristina, Marianne en Mirjam: bedankt voor het beantwoorden van al mijn administratieve vragen, het versturen van tientallen pakketjes en het regelen van allerlei zaken waar wij lab-mensen geen weet van hebben! Wim, Marc, Eric, Dennis en de rest van het ICT team: bedankt voor het oplossen van alle computerproblemen, problemen met dataopslag of met nieuwe software. Wat moeten we zonder jullie! Marian Bömer, bedankt voor het beantwoorden van al mijn vragen over declaraties en besteedbaar budget. En Marjoleine en Cheuk, bedankt dat jullie alle bestellingen voor ons doen, en uitzoeken waar het fout is gegaan als de inkoop weer eens een bestelling heeft tegengehouden. Ook Marcel, bedankt voor al het schone glaswerk! Huub, bedankt dat jij de regels in de gaten houdt, zodat we allemaal veilig kunnen werken. En we zouden haar haast vergeten: Lenie! Bedankt dat je altijd de troep achter ons opruimt, zelfs als we een iets te wilde borrel hebben gehad!

In addition to all these people, I would like to thank everybody in the department for your 'gezelligheid' and your help. Being such a large department makes life easier, since there is always someone that can help you because they have experience with a technique that is new to you. During my PHD, multiple people have helped me one way or another. Thanks for sharing and spreading your expertise over the departent. I know I may forget some people, but I would like to specifically mention Tobias, Holger and Livio at this point. Furthermore I would like to thank everybody that came up with ideas or suggestions at the big work discussion, because even if I did not continue working on them, they helped me think in new directions.

Of course I should also acknowledge Shabaz! Without you I would not have started my internship or PHD with Michiel.

Furthermore, I would like to thank all the people whom I have collaborated with in the Netherlands and abroad. The work described in this thesis would not have been possible without collaborations and sharing of data. I have always liked the collaborations and discussions about possible follow-up experiments. Often it feels like I know you very well without actually meeting you, just because we had so much contact.

Natuurlijk ben ik tijdens mijn werk niet alleen gesteund door mijn collega's maar ook door familie. Het begon al vroeg met tante Ans en ome Kees, die mijn interesse voor de natuur gewekt hebben. Sommigen van mijn biologen-collega's vinden mij meer bioloog dan ze zelf zijn, en dat terwijl ik chemicus ben! Ik was altijd welkom bij jullie op de boerderij en vind het nog steeds heerlijk om een rondje door de Arkemheense polder te fietsen.

Ik wil ook graag mijn schoonouders bedanken voor hun interesse. Het is erg lastig om uit te leggen wat ik precies onderzoek, maar ik hoop dat de Nederlandse samenvatting meer duidelijk maakt.

Lobke en Tjeerd, ik ben blij dat wij het zo goed met elkaar kunnen vinden! Ik hoop dat jullie allebei snel een baan vinden waar jullie het net zo naar je zin hebben als ik het hier tijdens mijn PHD heb gehad. En Tjeerd, bedankt voor je hulp met het opstarten van InDesign. En Berit, het voelt bijna alsof je mijn zusje bent. Ik vond het altijd erg gezellig om na het werk bij jou op de Uithof te komen eten. Ik vind het jammer dat je nu zo ver weg woont, maar ik wens je een mooie toekomst en hopelijk snel meer werk als dierenarts.

Papa en mama, bedankt dat jullie me altijd gesteund hebben. Als de trein weer eens vertraging had, wachtten jullie met eten tot ik ook thuis was. En hoewel het wetenschappelijke deel jullie misschien boven de pet ging, waren jullie wel degelijk geïnteresseerd in hoe het op het werk ging. Ik hoop dat jullie door het lezen van de Nederlandse samenvatting toch een beetje begrijpen waaraan ik de afgelopen jaren heb gewerkt. Ik heb het thuis altijd erg fijn gehad en ik hoop dat jullie nu met zijn tweetjes ook nog een hele mooie toekomst tegemoet gaan.

En als laatste natuurlijk Maarten, mijn grote liefde. Tijdens onze hele studie en mijn PHD zijn wij al samen en heb je mij gesteund als ik laat thuis was vanwege een uitgelopen experiment of als ik op zaterdag naar het werk moest. Bedankt dat je er altijd voor me bent…

Nelleke