# Understanding the role of CFP1 at CpG islands

David Alexander Brown

Department of Biochemistry

University of Oxford


**Doctor of Philosophy in**

**Chromosome and Developmental Biology**

# Declaration

I declare that I wrote and composed this thesis by myself and the research presented here is my own unless otherwise stated.

- David Brown 10th October 2014

# Acknowledgements

An enormous thank you to Rob for putting up with my shenanigans, you have challenged and inspired me. Although I am not always grateful, I frequently contemplate how lucky I am to be in this lab, in this department, and in this university.

Thank you to Lothar, who taught me my favourite techniques. I could not have done this work without your advice and feedback.

To Anca and Hannah, you have been incredible role models, and fantastic friends. Thank you and good luck in your new labs. I hope the difficult bits are short and worth it.

Hamish and Justin spent considerable time, helping me with ChIP and FRAP analysis for which I am very grateful.

Thank you to everyone in the Brockdorff and Klose labs. It's been a wonderful place to hangout.

To my parents who became the thesis help line, I cannot thank you enough.

I caught this bug from the two of you, and I'm happy to have it.

Emma, you mean the world to me.

I could not have done this without you.

Understanding the role of CFP1 at CpG islands David Brown
Doctor of Philosophy in Biochemistry Exeter College
Department of Biochemistry Trinity 2014

# Abstract

Vertebrate genomes are punctuated by CpG islands regions, which have an elevated frequency of CpG dinucleotides. CpG islands are associated with over 70% of mammalian promoters suggesting they may contribute to the regulation of transcription. However, despite being discovered over 30 years ago, the function of CpG islands is still not understood. Unlike the majority of the genome, CpG islands are resistant to DNA methylation. This provides a binding site for CFP1 which binds specifically to non-methylated DNA via its zinc-finger CXXC (zf-CXXC) domain. CFP1 is a subunit of the SET1 methyltransferase complex, and is thought to direct the activating histone modification H3K4me3 to CpG islands.

Interestingly, CFP1 also contains a PHD domain which is proposed to bind the H3K4me3 mark, potentially producing a feedback loop between H3K4me3 and the SET1 complex. Although the structural basis for discrimination of non-methylated CpGs is known, it is not clear how zf-CXXC proteins distinguish CpG islands amongst the irregular nucleosomal landscape which exists within the nucleus.

This thesis is focused on the role of CFP1 in the relationship between CpG islands, SET1 and H3K4me3. To address these questions, it was important to mechanistically dissect the contribution of the PHD and zf-CXXC domains. The proposal that the PHD domain of CFP1 binds selectively to H3K4me3 was confirmed by *in vitro* experiments, however this study demonstrates that the PHD domain is insufficient for stable interactions with chromatin. Using complementary genome-wide and live cell imaging approaches, the zf-CXXC domain shown to be required for PHD-dependent interactions. Genome-wide snapshots of binding interactions, together with spatial and temporal details, expose a surprising contribution of the SET1 complex to the nuclear mobility of CFP1, providing a new perspective on the role of CFP1 in H3K4 methylation.

# Table of Contents

# Table of Figures

# Foreword

## Genetic information

A genome is the sum total of an organism's genetic information. In our cells, DNA serves as a hard copy of this information, base pairing to form the famous double helix structure (Watson & Crick, 1953). Each strand of the double helix is a linear polymer of the nucleotides adenine (A), thymine (T), cytosine (C) and guanine (G). Inside the cell, sections of the DNA sequence called 'genes' are 'transcribed' by specific base pairing to produce an RNA copy. Many of these RNAs are then 'translated', according to an evolutionarily universal genetic code, dictating the sequence of amino acids required to synthesis a protein. In this way, the information required to synthesise any protein is encoded by the DNA sequence of its corresponding gene.

## The human genome project

In April 2003, sequencing of the human genome was practically complete providing a comprehensive map of our DNA. Annotation of this map with the positions and identities of genes involved massively parallel sequencing of RNA transcripts extracted from multiple human cell lines to reveal an approximate 'transcriptome', describing the complete set of human RNAs. For each RNA a corresponding gene was identified by mapping the RNA sequence back onto the genomic DNA sequence. Computation algorithms were also used to identify potential genes for which may have missed detection due to low levels of transcription or inactivity under the experimental conditions. Although refinement of the annotation is ongoing, the existing information is sufficient to catalogue essentially all human genes and their locations.

## Regulation of gene expression

Beyond serving as a template for RNA synthesis, genomic DNA also stores a vast amount of regulatory information including instructions such as how often a template ought to be copied or under what conditions. The distinction between coding templates or "structural genes" and regulatory DNA was recognised as early as 1959 by Pardee, Jacob and Monod who observed mutations in a regulatory sequence upstream of the E. coli β-galactosidase protein coding region that resulted in 'constitutive' or unregulated expression of β-galactosidase, which is normally 'induced' by the presence of a sugar called allolactose. This observation demonstrates the function of DNA elements in the control of gene expression.

Expression of our own genes is controlled by systems similar to those described in bacteria. Although there are exceptions, almost every cell in our bodies contains an identical copy of our genome, yet as multicellular organisms we are made up of organs, which are made up of tissues, which are made up of different types of cells. While the different cell types in our bodies have identical DNA with the same sequence and therefore the same genes, differences in which genes are expressed (active or on) and which are silent (off) determine the identity and fate of each cell. For this reason, understanding the regulation of gene expression is incredibly exciting because it may allow us to correct instances of pathological mis-expression including cancer and to direct expression allowing production of cell types, tissues or organs which are required for transplantation.

# Abbreviations

| | |
|---|---|
| °C | degree centigrade |
| A | ampere |
| aa | amino acids |
| ac | acetyl |
| APS | ammonium persulphate |
| ATP | adenosine triphosphate |
| bp | base pairs |
| BSA | bovine serum albumin |
| C- | carboxyl |
| CFP1 | CXXC-finger protein 1 |
| CGI | CpG island |
| Da | dalton |
| DAPI | 4',6-diamidino-2-phenylindole |
| DIC | differential interference contrast |
| DMSO | dimethyl sulfoxide |
| DNA | deoxyribonucleic acid |
| DNase | deoxyribonuclease |
| dNTP | deoxyribonucleoside triphosphate |
| DTT | dithiothreitol |
| EDTA | ethylenediamine tetraacetic acid |
| EGTA | ethyleneglycol tetraacetic acid |
| EMSA | electrophoretic mobility shift assay |
| ES | embryonic stem |
| fps | frames per second |
| FRAP | fluorescence recovery after photobleaching |
| h | hour |
| H3K4 | lysine 4 on Histone H3 |
| HAT | histone acetyltransferase |
| HDAC | histone deacetylase |
| HEPES | 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid |
| IPTG | isopropyl β-D-1-thiogalactopyranoside |
| kb | kilobase |
| kDa | kilodalton |
| L | litre |
| M | molar |
| MBD | methyl CpG binding domain |
| me | methyl |
| mg | milligram |
| min | minute |
| N- | amino |
| ng | nanogram |
| nm | nanometre |

| | |
|---|---|
| NMI | non-methylated island |
| OD | optical density |
| PAGE | polyacrylamide gel electrophoresis |
| PBS | phosphate buffered saline |
| PCR | polymerase chain reaction |
| pH | $-\log_{10}[H^+]$ |
| PHD | plant homeo domain |
| PMSF | phenylmethylsulfonyl fluoride |
| Pol II | RNA polymerase II |
| rcf | relative centrifugal force |
| RNA | ribonucleic acid |
| S | second |
| SAM | S-adenosyl-L-methionine |
| SDS | sodium dodecyl sulphate |
| SELEX | systematic evolution of ligands by exponential enrichment |
| SET | Su(var), E(z), Trx |
| TAE | Tris-acetate-EDTA |
| TEMED | tetramethylethylenediamine |
| Tris | tris(hydroxymethyl)aminomethane |
| UV | ultra violet |
| V | volt |
| zf-CXXC | zinc finger CXXC |
| µg | microgram |
| µl | microliter |
| µM | micromolar |
| µm | micrometer |

# Chapter 1 - Introduction

## 1.1 Transcription in eukaryotes

Gene expression depends on transcription of a DNA sequence to produce a complementary RNA. The template directed synthesis of protein-coding messenger RNAs is catalysed by RNA polymerase II (Pol II), which adds ribonucleotide bases processively in a 5' to 3' direction. Productive transcription initiates at a gene promoter, elongates though the gene body then releases the RNA transcript and finally pauses and disassembles at the gene terminator. *In vitro* reconstitution experiments have identified several core promoter sequences, which are able to support the initiation of transcription (Müller et al. 2007). Within these core promoters, mutagenesis studies have identified several conserved sequence elements which serve to position the Pol II enzyme on the promoter, dictating the transcription start site (TSS). In eukaryotes, the first core promoter element to be discovered was the TATAAA consensus sequence (TATA-box), which was found in core promoters from *Drosophila* (Goldberg 1979). Different sequence elements are found in different promoters (Butler & Kadonaga 2002). Other common examples include Initiator (Inr) which is located at the TSS (Kaufmann & Smale 1994), BRE (Lagrange et al. 1998), which is immediately upstream of TATA-boxes in some promoters and DPE (Burke & Kadonaga 1996), ~30 nt downstream of the TSS. Importantly, the Pol II enzyme makes no sequence specific contacts with the DNA template (Gnatt et al. 2001), and consistently purified Pol II is unable to initiate accurate transcription *in vitro* (Weil et al. 1979). Initiation of transcription is therefore dependent on additional proteins called general transcription factors (GTFs). Although no core promoter element is common to all core promoters (Kadonaga 2012), genome-wide studies suggest that most TSSs conform to a degenerate consensus sequence (Venters & Pugh 2013).

## 1.2 Pre-initiation complex assembly and the roles of the general transcription factors

Five different GTFs, TFIIB, TFIID, TFIIE, TFIIF and TFIIH are required to reconstitute initiation of transcription from a minimal core promoter based on a TATA-box element (Buratowski et al. 1989). The canonical model of transcription initiation is based on reductionist dissections of the mechanism *in vitro*, and on the structural models of the key steps which were elucidated by x-ray crystallography and electron microscopy (Kornberg 2007). In this model, recognition of the promoter is mediated by sequence-specific binding of TFIID subunit TBP to the TATA-box (Sawadogo & Roeder 1985). Inclusion of TFIIA further stabilises the interaction between TBP and the TATA-box (Buratowski et al. 1989). TFIIB mediates interactions between TFIID and the Pol II enzyme, constraining the position of the Pol II active site relative to the TATA-box (Bushnell et al. 2004). Further interactions between TFIIB and TFIIF stabilise binding of TFIIE and TFIIH to Pol II, forming the pre-initiation complex (PIC).

Assembly of the PIC stabilises local melting of the DNA duplex to provide a single strand template for RNA synthesis. This process is facilitated by the binding of TBP which distorts the B-form DNA (Nikolov et al. 1992). ATP-dependent separation of the duplex is catalysed by a helicase called XPB which is a subunit of TFIIH (Tirode et al. 1999). The helicase activity of XBP is stimulated by TFIIE. TFIIF stabilises the open complex by capturing the non-template strand (Yan et al. 1999). Transcription initiates with the formation of the first phosphodiester bond between ribonucleotide bases. Once the nascent RNA reaches 12-15 nucleotides, structural changes in the PIC result in a displacement and dissociation of TFIIB from the Pol II enzyme (Liu et al. 2010). Pol then dissociates from general initiation factors, as it transcribes away from the promoter, transitioning from initiation to early elongation. This canonical model highlights the distinct roles of each GTF in the initiation of basal transcription.

### 1.3 Regulation of PIC assembly by sequence specific transcription factors

The canonical model described above demonstrates that a minimal set of GTFs are competent to drive basal transcription *in vitro*. Initiation of transcription *in vivo* is additionally subject to extensive regulation by sequence-specific transcription factors (TFs). Eukaryotic cells contain thousands of different TFs including several large families which are structurally related and may share identical or closely related binding sites (Kadonaga 2004). In addition to a DNA binding domain, TFs are highly modular, and may contain activator or repressor domains which directly interact with the PIC (Kadonaga 2004). Transcriptional activators, which appear to be more common than transcriptional repressors, can increase the rate of transcription at specific genes by stabilising and in some cases driving PIC formation. For example, the related transcription factors c-Fos and c-Jun can activate transcription at promoters containing an AP-1 site (ATGAGTCAT), by directly stabilising the binding of TBP to the promoter (Metz et al. 1994; Franklin et al. 1995). Similarly p65, of the NFκB p50/p65 heterodimer that binds to a GGGRNYYYCC sequence (Chen et al. 1998), contains a C-terminal transactivation domain which binds to TBP (M. Lienhard Schmitz et al. 1995).

At the majority of eukaryotic genes, multiple transcription factors act in concert to regulate the level of transcription (Spitz & Furlong 2012). In higher eukaryotes this combinatorial regulation is particularly evident during development. In the fruit fly, *Drosophila* melanogaster, formation of anterior-posterior segmentation depends on the cooperative binding of bicoid and hunchback to the eve stripe 2 promoter (Ip et al. 1992). Similarly, during mammalian embryogenesis expression of the pluripotency factor Nanog depends on the cooperative binding of Sox2 and Oct4 to a composite binding site in the Nanog promoter (Rodda et al. 2005). The fundamental importance of transcription factors to the regulation of gene expression is highlighted by the ability of four artificially expressed TFs, Oct4, Sox2, c-Myc and Klf4, to induce pluripotency in otherwise differentiated adult cells (Takahashi & Yamanaka 2006).

## 1.4 Histones, nucleosomes and chromatin in the eukaryotic nucleus

Inside eukaryotic nuclei, DNA exists in the context of chromatin. This term was originally used by Walther Flemming to refer to nuclear material which could be stained dark using an aniline dye (Flemming 1882). The term is now generally used to refer to a composite material formed by nucleic acids and associated nuclear proteins. Among the proteins associated with DNA, histones proteins, contribute almost half the mass of chromatin and partially balance net negative charge of DNA (Kossel 1928). Spectroscopic differences reflecting amino acid composition correctly predicted the existence of five distinct histone proteins; H1, H2A, H2B, H3 and H4 (DeLange & Smith 1971), although variants of each histone have since been identified (Sarma & Reinberg 2005). Together the histones proteins serve to organise DNA by forming a regular, structure called a nucleosome.

The atomic structure of the nucleosome was solved by X-ray crystallography revealing that the DNA duplex (146-147 bp) is wound around a globular core of histone proteins (Luger et al. 1997). This histone core is an octamer formed from two copies each of H2A, H2B, H3 and H4 (Thomas & Kornberg 1975), which compact DNA into a linear array of nucleosomes (Kornberg & Thomas 1974). H1 is not part of the core nucleosome, but can interact with the connecting 'linker' DNA, causing further compaction (Sato et al. 1999). The linear repetition of a regular structure is supported by endonuclease digestion of chromatin with DNase I which produces discretely sized DNA fragments rather than a continuous range (Hewish & Burgoyne 1973). This was later confirmed directly by imaging of chromatin fibres using electron microscopy (Olins & Olins 1974). Recent studies suggest that despite regular spacing of the nucleosomes, chromatin fibre exists in a locally disordered and highly interdigitated state inside the interphase nucleus (Eltsov et al. 2008; Nishino et al. 2012).

As the cell enters mitosis, condensation of the chromatin resolves into individual chromosomes to enable accurate segregation of the genetic material (Weaver et al. 1985; Sundin & Varshavsky 1981). Unlike euchromatin, which decondenses shortly following mitosis, heterochromatin remains densely

packed, forming distinct domains within the nucleus (Heitz 1928). Since this initial observation, several compositional and functional differences between heterochromatin and euchromatin have been recognised. The heterochromatin typically maps to large contiguous blocks of DNA (Weiler & Wakimoto 1995). Although some genes are located within heterochromatic DNA, the average gene density is typically lower than in euchromatin regions (Hilliker et al. 1980; Brown 1966). Heterochromatic DNA is often highly repetitive (Lohe et al. 1993) and includes the pericentric and centromeric regions. In the mouse genome, these regions are characterised by tandem repetition of major and minor satellites, respectively (Wong & Rattner 1988; Joseph et al. 1989). Similarly, heterochromatin is also found at chromosome telomeres, which are comprised of the TTAGGG telomeric repeat. In line with the strong correlation between heterochromatic regions and repetitive DNA elements, most heterochromatin domains are a "constitutive" feature of the chromosome. A striking exception to this trend is seen in nuclei from mammalian females where heterochromatin is present on one copy of the X-chromosome but not the other (Barr & Bertram 1949). Examples of this "facultative" chromatin first suggested that the formation of heterochromatin is a regulated process and does not depend solely on the underlying sequence (Brown 1966).

### 1.5 Euchromatin, and heterochromatin

The retention of mitotic characteristics in heterochromatin was initially presumed to have a repressive effect on gene expression, in line with the global arrest of transcription during mitosis (Brown 1966; Heitz 1928). Formation of facultative heterochromatin on a single X-chromosome is associated with the inactivation of transcription as a mechanism of dosage compensation (Ohno et al. 1959), strengthening the correlation between heterochromatin and the repression of transcription. Inactivation of the maternal or paternal X-chromosome in somatic cells is a stochastic event, however once one copy of the X-chromosome has been inactivated, this choice is clonally inherited (Lyon 1961). Although X-inactivation is reversed in the germ line and not transmitted to the next generation, clonal inheritance of maternal or paternal X-inactivation can produce distinct

heritable phenotypes among genetically identical somatic cells. This has become a classic example of an epigenetic effect.

A second example of epigenetic differences leading to distinct heritable phenotypes was discovered based on specific mutations in *Drosophila* which could not be explained by classical genetics (Muller 1930). Spontaneous or artificial translocations which place a gene from a euchromatic region at the edge of a heterochromatin block, often result in a variegated phenotype showing expression of the gene in some cells but not others (Wallrath & Elgin 1995). Position effect variegation (PEV) occurs when a nearby block of heterochromatin encroaches across a gene in a subset of cells, causing stochastic gene silencing. Once stochastic silencing has been established in somatic cells, it is clonally inherited. Like X-inactivation, PEV produces distinct heritable phenotypes among genetically identical cells, providing another classic example of an epigenetic effect. Using a heat shock inducible hsp26 reporter system, variegated expression was shown to correspond to reduced accessibility to regulatory elements, while limited micrococcal nuclease digestion revealed that variegated hsp26 transgene exhibited regularly spaced nucleosomes (Wallrath & Elgin 1995). This early work highlighted the important role of chromatin structure in the regulation of gene expression.

### 1.6 Chromatin Remodelling

In addition to X-inactivation and PEV, a third non-mutually exclusive example of the role of chromatin in the regulation of gene expression was discovered by genetic experiments in *Saccharomyces cerevisiae*. Three independent groups studying the regulation of the HO gene in mating type switching, glucose repression of the SUC2-invertase gene, and the suppression of Ty transposable elements, discovered a common group of genes which were important for transcription (SWI/SNF) (Winston & Carlson 1992). The SWI/SNF genes encode subunits of a large (~2 MDa) complex which is conserved from yeast to human. *In vitro* experiments then demonstrated

that while TBP, a key GTF in the canonical model of transcription initiation, is unable to bind to a TATA-box sequence when it is incorporated in to nucleosomes (Klein & Struhl 1994), binding can be facilitated by the addition of the human SWI/SNF complex together with ATP (Imbalzano et al. 1994). A similar experiment using the Gal4 transcription factor demonstrated that addition of purified SWI/SNF, together with ATP, alleviated inhibition of sequence-specific binding by the formation of nucleosomes (Côté et al. 1994). In this context, SWI/SNF functions as a co-activator by actively repositioning nucleosomes which otherwise inhibit initiation of transcription (Knezetic & Luse 1986; Lorch et al. 1987; Carlson & Laurent 1994; Pazin & Kadonaga 1997).

### 1.7 Post-translational histone modifications

Beyond serving to compact DNA by organising it into nucleosomes, the core histones play a fundamental role in the regulation of transcription and other nuclear processes (Berger 2007). While interactions between H2A, H2B, H3 and H4 histone fold domains form the centre of the nucleosome, H3 and H4 each possess a relatively unstructured amino-terminal tail which extends between the gyres of the nucleosomal DNA (Luger et al., 1997). Similarly, the globular domains of H2A and H2B end in relatively unstructured C-terminal tails. Histone tails are subject to hundreds of different post-translational modifications. Many modifications seem to function by providing a new binding surface for protein-protein interactions (Jenuwein & Allis 2001; Yun et al. 2011). This system of histone protein modification followed by effector protein recruitment is a common feature of the fundamental processes that affect eukaryotic DNA, including transcription, replication, and repair (Wu & Grunstein 2000). Recent studies have demonstrated that different nucleosome modification states correlate with the level of transcriptional activity (Bernstein et al. 2012). The first modifications to be discovered were lysine acetylation and lysine methylation (Allfrey et al. 1964; Murray 1964), both of which are particularly relevant to this study. Histone lysine ubiquitylation was not discovered until 15 years later (Goldknopf & Busch 1977) but has since been recognised as an important factor in the regulation of transcription (Dover et al. 2002; Sun & Allis 2002).

## 1.8 Histone lysine acetylation

The majority of lysine residues in the core histone tails are subject to acetylation, including K4, K9 and K14 in histone H3, and K16 in histone H4 (Karmodiya et al. 2012). Acetylation neutralises the positive charge of lysine ε-amino group, disrupting electrostatic interactions of the histone tails. Acetylation of H4K16 has been shown to weaken the interactions between neighbouring nucleosomes, potentially reducing chromatin compaction (Shogren-Knaak et al. 2006). Acetylation of H3K9 and H3K14 is generally associated with active gene promoters (Pokholok et al. 2005) and their levels positively correlate with the level of transcription (Karmodiya et al. 2012). Histone acetylation corresponds to regions of increased DNase I sensitivity, suggesting that acetylation may play a role in transcriptional activation by creating a more accessible chromatin environment (Hebbes et al. 1994). This has been challenged by more recent experiments using DNA adenine methyl-transferase (dam) to probe accessibility *in vivo* (Wines et al. 1996). The Dam enzyme marks accessible GATC motifs by adding a small adenine-methylation tag that does not occur endogenously in most eukaryotes. Genome-wide profiles of dam accessibility suggest that chromatin compaction has a marginal effect on accessibility however the local positioning of nucleosomes may prevent access to specific sequences preventing TF or GTF binding (Morse 2007).

The first histone lysine acetyltransferase (HAT) to be purified (Brownell & Allis 1995) was found to be homologous to a known transcriptional co-activator GCN5 (Brownell et al. 1996), suggesting a role for histone acetylation in transcription. HAT activity of GCN5 was then shown to be essential for transcriptional activation providing the first evidence that chromatin modifications have a regulatory function (Wang et al. 1997; Kuo et al. 1998). Following the discovery of GCN5 acetyltransferase activity, HAT activity was detected in several other co-activator proteins including CBP/p300, TAFII250, and SRC-1 (Mizzen et al. 1996; Spencer et al. 1997; Bannister & Kouzarides 1996). In addition to direct biophysical effects, acetylated lysine residues provide a distinctive feature which can be recognised by bromodomain containing proteins (Zeng & Zhou 2002; Dhalluin et al. 1999).

Bromodomains, which bind specifically to acetylated lysine, are found in a large family of proteins including the majority of histone acetyltransferase enzymes and chromatin remodellers like SWI/SNF proteins associated with chromatin remodelling (de la Cruz et al. 2005; Hassan et al. 2001). The ability to place and bind to the same histone modification, is an interesting feature found in several chromatin modifying enzymes, suggesting the capacity to propagate or spread the modification to the equivalent histone tail in the same nucleosome, to neighbouring nucleosomes, or perhaps to nearby nucleosomes in trans. The Gcn5 bromodomain is thought to facilitate cross-tail acetylation, and acetylation of neighbouring nucleosomes (Li & Shogren-Knaak 2009).

## 1.9 Histone lysine methylation

Histone lysine methylation *in vivo* has been characterised on the K4, K9, K27, K36, and K79 residues of H3, on K20 of H4 and K26 on H1. Lysine residues can become mono-, di- or trimethylated (Holde 1988). Similar to lysine acetylation, the alteration of lysine residues by the addition of one or more methyl groups provides a distinctive feature for the recruitment of specific binding proteins. Unlike lysine acetylation, which is generally found on histones located at active promoters, lysine methylation has been associated with a range of different activities depending on the lysine modified, and on the degree of methylation (mono, di or tri).

### 1.9.1 Methylation of H3K9

In metazoan species trimethylation of H3K9 is generally considered a hallmark of pericentromeric heterochromatin. H3K9 methylation is catalysed by the SET domain of Suv39 enzymes (Tschiersch et al. 1994; Rea et al. 2000; Bernard & Allshire 2002). Combined deletion of Suv39h1 and Suv39h2 in mice results in a striking loss of H3K9me3 from pericentromeric heterochromatin although the formation of dense heterochromatic domains is unaffected. The double null mice have a drastically reduced viability and frequently develop tumors suggesting that pericentromeric H3K9me3 is important for genome stability (Peters et al. 2001). Suv39h was the first histone lysine

methyltransferase (HKMT) to be discovered, although its importance to heterochromatin formation was foreshadowed by its prior identification as a suppressor of position effect variegation (Cléard et al. 1997).

### 1.9.2 Histone lysine methyltransferases

The C-terminal region of Suv39 is closely related to two other *Drosophila* proteins, E(z), and Trx (Tschiersch et al. 1994; Carrington & Jones 1996). These proteins share the ability to methylate histone lysine residues (Czermin et al. 2002; Smith et al. 2004). The catalytic SET- (Suv39, E(z), Trx) domain is found in all HKMT enzymes, with the exception of Dot1, which methylates H3K79 within the globular histone fold, rather than a histone tail (Feng et al. 2002). E(z) and Trx enzymes are both critical for normal development. Trx (Trithorax) was first identified genetically as a spontaneous mutation affecting *Drosophila* segmentation (Ingham 1981). This trxZ11 allele contains a glycine-to-serine substitution in the SET domain that interferes with histone binding (Eissenberg & Shilatifard 2010). Inactivation of Trx results in the silencing of homeotic (hox) genes required for the differentiation of posterior segments. This leads to anatomical transformations (anteriorisation). Mutations in another gene called polycomb result in adventitious expression of hox genes, associated with posterior development, in anterior segments where they are normally repressed (posteriorisation) (Lewis 1949). Genes related to these homeotic transformation phenotypes can be divided between the Trx group (TrxG), which is generally associated with gene activation, and the Pc group (PcG), associated with gene repression (Lewis 1978). E(z) (enhancer of zeste) is a PcG protein, indicating that despite sharing the same catalytic mechanism (Marmorstein 2003; Trievel et al. 2002; Jacobs et al. 2002), E(z) and Trx perform mutually antagonistic functions during development (Schuettengruber et al. 2007; Pirrotta 1998; Pelegri & Lehmann 1994).

### 1.9.3 Methylation of H3K27

Lysine methylation by E(z) and its human homologue EZH2, is specific for H3K27 (Müller et al. 2002; Kuzmichev et al. 2002; Czermin et al. 2002; Cao et al. 2002). H3K27me3 co-localises with E(z) at broad (~10 kb) regions of the *Drosophila* genome including the hox gene clusters associated with segmentation (ANT-C, BX-C) (Schwartz et al. 2006). The methyltransferase activity of E(z) is required for silencing of the Ubx hox gene in *Drosophila* demonstrating that post-translational modification of histone tails plays an important role in the regulation of gene expression (Müller et al. 2002). Recapitulation of the E(z) mutant phenotype using a H3K27R point mutation highlights the importance of homeotic transformations as a fourth non-mutually exclusive example of the role of chromatin in gene expression (Pengelly et al. 2013).

### 1.9.4 Methylation of H3K4

Trx catalyses the methylation of H3K4 (Smith et al. 2004). H3K4 methylation was first recognised in 1999 when it was observed in *Tetrahymena*, yeast, and HeLa cells, a human cancer cell line (Strahl et al. 1999). In *Tetrahymena*, where DNA is divided between somatic macronuclei and germ line micronuclei, H3K4 methylation is exclusive to the macronucleus which is transcriptionally active. Similarly, immunofluorescent staining of H3K4me3 in mammalian cells reveals enrichment on metaphase chromosomes with the exception of the inactive X-chromosome in females (Boggs et al. 2002). Genome scale mapping in mouse and humans characterised H3K4me3 as a 'punctate' signal with sharp peaks of enrichment (~2 kb long, ~1% of the genome in total) against a uniform baseline (Bernstein et al. 2005). Despite comprising a small fraction of the genome, H3K4me3 peaks coincide with over 75% of protein coding gene promoters (Guenther et al. 2007), and the level of H3K4me3 is one of the best predictors of transcriptional state (active or inactive) and transcriptional output (Santos-Rosa et al. 2002; Bernstein et al. 2012). Although H3K4me3 levels correlate positively with transcriptional activity it is also observed at silent genes (Barski et al. 2007). In *Drosophila*, Trx is required to maintain hox gene expression (Ingham 1983) by antagonising PcG-dependent gene

repression. In mammalian cells H3K4me3 and H3K27me3 are not mutually exclusive and often overlap producing 'bivalent' regions of the genome (Bernstein et al. 2006; Barski et al. 2007). Bivalent promoters are particularly common during the early stages of embryogenesis (Vastenhouw & Schier 2012).

### 1.10 Methyl-lysine binding domains

Unlike acetylation, lysine methylation preserves the positive charge of the ε-amino nitrogen and therefore has a smaller effect on electrostatic interactions. However, a methylated lysine has distinct biophysical properties which may be recognised by specific protein domains. By controlling the binding of distinct effector proteins, methylation of distinct lysine residues, for example H3K4 or H3K27, can direct diverse and even antagonistic functions. H3K9 methylation provides a binding site for the chromodomain of heterochromatin protein 1 (HP1) (Bannister et al. 2001; Nielsen et al. 2002; Jacobs & Khorasanizadeh 2002). Analogous to a bromodomain, which binds selectively to acetylated lysines, the HP1 chromodomain binds selectively to di-, or trimethylated H3K9. Like Suv39h, HP1 was also identified as a suppressor of variegation (Suv205) reflecting its important role in the regulation of gene expression (Eissenberg et al. 1990). Importantly binding to H3K9me3 is not sufficient for recruitment of HP1 which must be stabilised by direct interactions with Suvar39, to bind to chromatin effectively (Steward et al. 2006). Chromodomains are found in several other proteins including polycomb in *Drosophila* and its mammalian homologues (CBX2, CBX4, CBX7 and CBX8) (Eissenberg 2012). When bound by a chromodomain, the methylated lysine sits within a hydrophobic pocket or 'aromatic cage' (Jacobs & Khorasanizadeh 2002). Selective binding of specific methylation states is determined by cation-π interactions between the lysine residues and the aromatic cage (Hughes et al. 2007).

In addition to the chromodomain, methyl-lysine residues can also recognised by several other structurally related domains, including Tudor-, MBT-, PWWP- and Agenet- domains (Sanchez & Zhou

2011). Importantly, some methyl-lysine residues are also recognised by plant homeotic domains (PHD) originally identified in *Arabidopsis* (Schindler et al. 1993). PHD domains are found across eukaryotic species and have been identified in over 200 human proteins (Letunic et al. 2012; Schultz et al. 1998). Unlike chromodomains and their relatives, the structure of the PHD domain is formed by a characteristic spacing of cysteine and histidine residues which coordinate two zinc ions (Schindler et al. 1993; Mellor 2006). The PHD domain of BPTF (bromodomain and PHD domain containing transcription factor) binds specifically to trimethylated H3K4 (H3K4me3) (Li et al. 2006; Wysocka et al. 2006). H3K4me3 is generally associated with transcription start sites, with higher enrichment at highly transcribed genes (Mikkelsen et al. 2007; Pokholok et al. 2005; Barski et al. 2007). The BPTF protein is a subunit of the nucleosome remodelling factor (NURF). Similar to the SWI/SNF complex, NURF is an ATP-dependent chromosome remodelling complex which is required for the induction hormone responsive genes in *Drosophila* (Badenhorst et al. 2005). Morpholino knockdown of BPTF in Xenopus also results in developmental defects, suggesting that H3K4me3-dependent recruitment of NURF by BPTF via its PHD domain is required for normal regulation of developmental genes (Wysocka et al. 2006). The TFIID subunit TAF3, also contains a PHD domain specific for H3K4me3 (Vermeulen et al. 2007), directly influencing transcription, by stabilising the pre-initiation complex (Lauberth et al. 2013).

PHD domains specific for di-, and trimethylated H3K4 have also been discovered in several other proteins associated with the regulation of transcription, including ING2 (Sanchez & Zhou 2011; Shi et al. 2006; Peña et al. 2006). ING2 is associated with transcriptional repression through indirect interactions with the histone lysine deacetylase enzymes (HDACs) 1 and 2 (Shi et al. 2006). PHD domains are also found in proteins which are not thought to be associated with the regulation of transcription. For example, the PHD domain of RAG2 is required for site directed V(D)J recombination (Elkin et al. 2005; Matthews et al. 2007), which occurs in the immune systems of jawed vertebrates. RAG2 binds specifically to H3 tails which are symmetrically dimethylated on

arginine 2 and trimethylated on lysine 4 (H3R2me2sK4me3) (Ramón-Maiques et al. 2007). Interestingly, these two modifications frequently occur in combination genome-wide (Yuan et al. 2012), while H3K4me3 and asymmetrically dimethylated H3R2 (H3R2me2a) are mutually exclusive (Kirmizis et al. 2007). Unlike chromodomains, PHD domains form two binding pockets separated by a conserved tryptophan (Sanchez & Zhou 2011). The ability to interrogate the modification states of two residues in combination may serve to increase the binding specificity of PHD domain proteins (Musselman & Kutateladze 2011).

### 1.11 DNA methylation

Like the histones which surround it, DNA itself is also subject to methylation. In vertebrate species, this occurs almost exclusively at the C5 carbon of cytosine bases which are followed by a guanine base (CpG) (Lister et al. 2009). In embryonic stem (ES) cells, cytosine methylation is also observed less frequently in the context of CpA, and still less CpT, however these modifications are rapidly lost during early differentiation (Ramsahoye et al. 2000; Lister et al. 2009). Interestingly, CpG dinucleotides are significantly depleted in vertebrate genomes (Gardiner-Garden & Frommer 1987). Spontaneous deamination of methyl-cytosine (meC) produces thymine, which is inefficiently repaired by glycosylase enzymes (Duncan & Miller 1980; Coulondre et al. 1978). Failure to repair all of the resulting T:G mismatches in the germ line may result in a gradual depletion of CpG dinucleotides through evolution (Antequera 2003).

Despite the overall depletion of CpGs, vertebrate genomes are punctuated by short regions where the CpG density and overall GC-content are higher than average (Bird 1986; Illingworth & Bird 2009). Crucially, these CpG islands (CGIs) are resistant to cytosine methylation (Bird et al. 1985; Bird 1986), suggesting that CpGs within CGIs may have escaped evolutionary depletion. Although CGIs are not found at all promoters, they encompass the majority of transcription start sites (TSSs), suggesting they may be involved in the regulation of transcription initiation. Genome-wide profiling of DNA

methylation revealed that methylation of CGI promoters is associated with tissue-specific gene silencing (Illingworth et al. 2008). Methylation of promoter DNA is known to be associated with transcriptional repression during differentiation (Gruenbaum et al. 1982; Keshet et al. 1985; Yisraeli et al. 1988). However, repression can still occur in the absence of DNA methylation, and is known to precede DNA methylation during X-inactivation and the silencing of transgenes (Mutskov & Felsenfeld 2004; Strunnikova et al. 2005).

CpGs in double stranded DNA can be methylated, hemi-methylated or non-methylated. Methylation is catalysed by two distinct classes of DNA methyltransferase (DNMT) enzymes. DNMT1 is a "maintenance" enzyme, specific for hemi-methylated CpGs (Bestor et al. 1988), and results in symmetrical methylation of base paired CpGs which are on opposite stands of the duplex (Holliday & Pugh 1975; Riggs 1975). Hemi-methylated CpGs are produced, when methylated CpGs are replicated, or a non-methylated CpG is methylated by a "*de novo*" DNMT enzyme (DNMT3a and DNMT3b in mammals) (Okano et al. 1999). The restoration of symmetrical methylation at hemi-methylated CpGs resulting from semi-conservative replication means that DNA methylation can act as a heritable epigenetic modification.

Due to the projection of the C5 methyl group into the major groove of the DNA duplex, DNA methylation is known to interfere with the binding of certain sequence-specific transcription factors (Iguchi-Ariga & Schaffner 1989). Conversely, a family of methyl-DNA binding proteins (MDBs) which bind specifically to DNA containing methylated CpGs (Boyes & Bird 1991; Ng et al. 2000; Lewis et al. 1992; Ohki et al. 2001), play a more general role in the repression of transcription by recruiting histone deacetylases (Fujita et al. 2000). By antagonising histone acetyltransferases associated with the initiation of transcription, deacetylases contribute to gene silencing. The molecular mechanisms connecting DNA methylation and post-translational histone modifications are highly complex (Rose & Klose 2014), however it is clear that there is considerable mutual reinforcement between DNA

methylation and histone modifications associated with heterochromatin and transcriptional repression. For example, MBD1 can interact directly with the Suv39h1-HP1 and SETDB1 methyltransferase complexes in addition to its specificity for methylated CpGs (Fujita et al. 2003; Sarraf & Stancheva 2004), potentially directing H3K9 methylation to methylated DNA. Conversely, genome-wide profiles suggest enrichment of DNA methylation and H3K4me3 are mutually antagonistic (Brykczynska et al. 2010; Illingworth et al. 2008), consistent with *in vitro* experiments suggesting that H3K4me3 blocks DNA methylation by preventing binding of DNMT3a to the H3 tail (Zhang et al. 2010).

### 1.12 Zinc finger-CXXC proteins

Vertebrate genomes are coarsely divided into large blocks of methylated DNA punctuated by non-methylated CGIs. It is not clear how CGIs remain free from methylation despite elevated levels of CpG dinucleotides. The discovery of MDB proteins, which bind specifically to DNA containing methylated CpGs, therefore prompted a search for proteins which could bind specifically to non-methylated DNA, potentially antagonising DNA methylation. The first of these proteins was discovered using phage-display to screen a library of human cDNAs (Voo et al. 2000). Phage were screened for proteins which could bind to DNA only when a CpG was present. This approach identified the first CpG binding protein (CGBP). Further characterisation revealed that CGBP binds specifically to non-methylated CpGs, via a conserved zinc finger domain previously identified in DNMT1, MLL1 and MBD1 (Cross et al. 1997; Ma et al. 1993; Bestor & Verdine 1994; Lee et al. 2001). The zf-CXXC domain, named for its characteristically spaced cysteine residues, is found in a small family of proteins, of which there are 12 in humans (Figure 1). To reflect its role in the characterisation of the zf-CxxC domain containing proteins, CGBP was later renamed CFP1 (CXXC finger protein 1).

Since the discovery of CFP1, selective binding to DNA containing non-methylated CpG dinucleotides has also been demonstrated for the zf-CXXC domains of MLL (Birke et al. 2002), MBD1 (Jørgensen et al. 2004), and DNMT1 (Pradhan et al. 2008), as well as KDM2A (Blackledge et al. 2010; Zhou et al. 2012), KDM2B (Koyama-Nasu et al. 2007; Farcas et al. 2012; Yamagishi et al. 2008) and a related protein FBXL19 by electrophoretic mobility shift assay (unpublished data, Klose Lab). Interestingly, the TET3 zf-CXXC is able to bind to non-CpG cytosine bases, provided they are non-methylated (Xu et al. 2012).

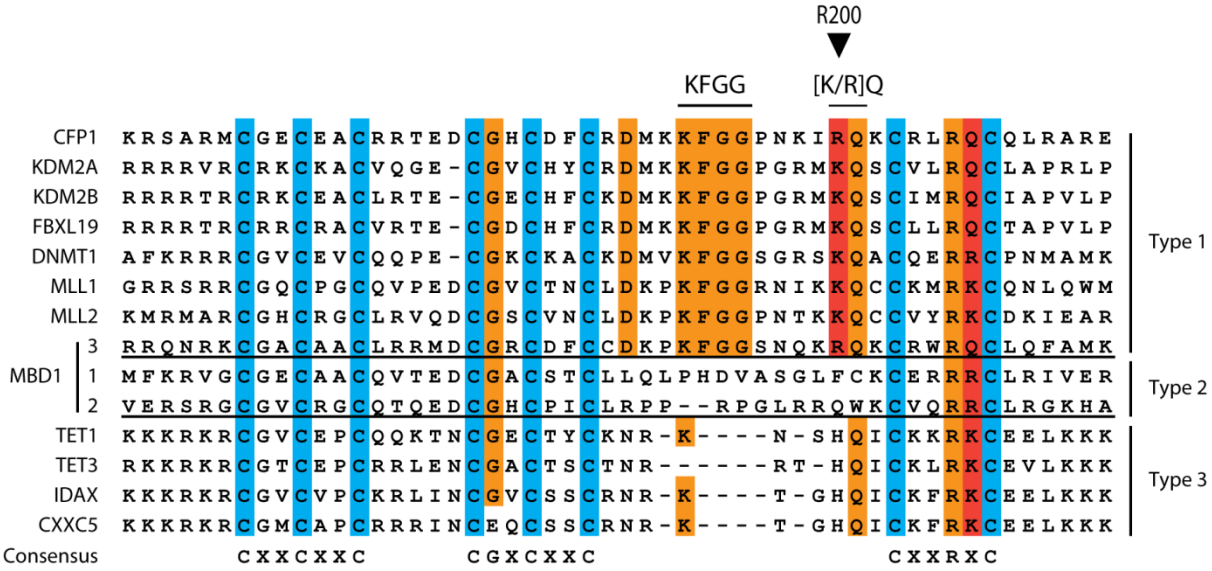



**Figure 1 – The family of zf-CXXC proteins human can be divided into three types**

Multiple sequence alignment of zf-CXXC domains found in human proteins reveals conservation, of sequence identity (highlighted orange), and similarity (red), across Type I zf-CXXC domains which bind to non-methylated DNA. Zinc-coordinating cysteines are highlighted cyan. In type II zf-CXXC domains, the KFGG and [K/R]Q motifs are not conserved, while in type III zf-CXXC domains, the KFGG loop is absent. This figure was adapted from Long et al. 2013 which shows mouse zf-CXXC proteins.

Crystallographic and NMR structures of the CFP1, MLL and DNMT1 zf-CXXC domains revealed the molecular basis for their selective binding to non-methylated CpGs (Xu et al. 2011; Cierpicki et al. 2010; Song et al. 2011). The eight cysteine residues which define the zf-CXXC domain coordinate two zinc ions generating an extended crescent-shaped structure. Duplex DNA interacts with the concave surface of the zf-CXXC crescent, which interacts with both major and minor grooves (Figure 2). CpG dinucleotides are recognised by a [K/R]Q motif at the tip of the zinc finger, which interrogates functional groups facing the major groove. Point mutations in the [K/R]Q ablate interactions between the zf-CXXC domain and DNA (Blackledge et al. 2010; Farcas et al. 2012; Cierpicki et al. 2010). CpGs which are methylated on the C5 cytosine carbon cause a steric clash with the constrained polypeptide backbone, preventing the zf-CXXC domain from binding to methylated CpGs (Xu et al. 2011; Cierpicki et al. 2010).

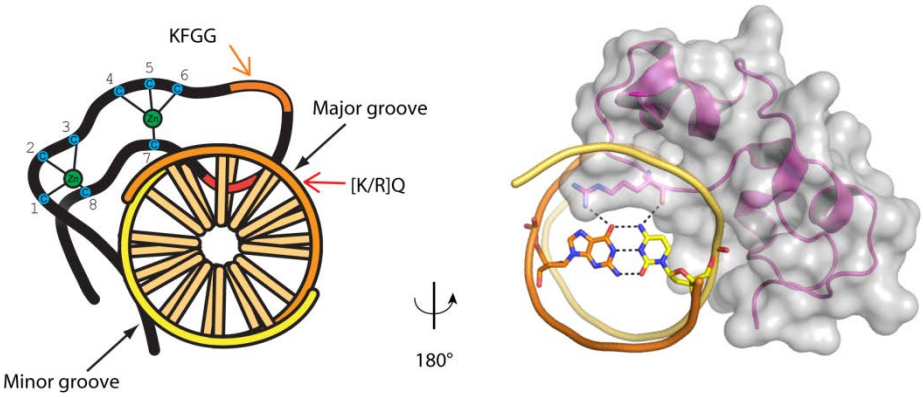



**Figure 2 – The structural basis for recognition of non-methylated CpGs by the zf-CXXC domain**
Schematic and structural models depicted the interaction of the zf-CXXC domain with DNA containing a non-methylated CpG dinucleotide. The locations of the KFGG and [K/R]Q motifs are indicated on the schematic, while the specific recognition of the C-G base pair, by the [K/R]Q arginine is highlighted in the structure by a stick representation.

The zf-CXXC domain is found in several enzymes which modify chromatin, including MLL1 and -2, the human homologues of Trx (Huntsman et al. 1999; FitzGerald & Diaz 1999; Voo et al. 2000). MLL1 and

-2 are both histone lysine methyltransferase enzymes specific for H3K4 (Milne et al. 2002; Jiang et al. 2013). Like Trx mutant *Drosophila*, heterozygous mice (+/-) MLL1 display homeotic transformations, while the homozygous mutation is lethal (Yu et al. 1995). The zf-CXXC family also contains two related histone demethylase enzymes KDM2A, and KDM2B, which remove methyl groups from H3K36 via a conserved Jumonji-C domain (Tsukada et al. 2006). Trimethylation of H3K36 is a co-transcriptional event catalysed by SET2 which interacts with Pol II during elongation (Strahl et al. 2002; Edmunds et al. 2008). Recognition of H3K36me3 by the chromodomain of Eaf3 directs repression of transcription due to deacetylation of H4K16 by the Rpd3s HDAC complex (Carrozza et al. 2005; Bell et al. 2007). At non-methylated CGIs, KDM2A maintains low levels of H3K36me1/2 demonstrating that CGIs possess a unique chromatin architecture which is continually maintained by the recruitment of zf-CXXC proteins (Blackledge et al. 2010).

### 1.13 CXXC-finger protein 1

As the first protein recognised to bind specifically to non-methylated CpGs (Voo et al. 2000), CFP1 was initially proposed to protect CGI regions from methylation by *de novo* DNMTs. In mice, CFP1 null embryos die before implantation (Carlone & Skalnik 2001). However, CFP1 mRNA is present in various embryonic and adult tissues, suggesting its essential function may not be limited to this early stage of development (Voo et al. 2000). As global levels of DNA methylation are reduced in CFP1 null ES cells (Carlone et al. 2005), CFP1 does not appear to protect CpGs from DNMTs.

Unlike the majority of Type 1 zf-CXXC proteins, CFP1 has no known catalytic activity. Nevertheless, transient overexpression of CFP1 was found to increase expression of reporter genes from non-methylated promoters (Voo et al. 2000). CFP1 null ES cells are viable but unable to differentiate, and exhibit a 3-fold increase in apoptosis (Carlone et al. 2005). Overall, these results suggest that CFP1 plays a vital role in development and an important but non-essential role in gene expression.

Immunoprecipitation of flag tagged CFP1, revealed its presence in a stable, multi-subunit complex of ~450 kDa (Lee & Skalnik 2005). Mass spectrometry identified all seven components of this complex as homologues of the yeast SET1 methyltransferase complex, which methylates H3K4 (Roguev et al. 2001; Miller et al. 2001; Lee & Skalnik 2005). Importantly, CFP1 is not essential for assembly of the SET1 complex, as a complex of reduced mass persists in constitutive CFP1-/- ES cells (Lee & Skalnik 2005). Analysis of specific histone modifications in the CFP1 null ES cell line revealed a global increase in H3K4me2, as well as a global decrease in H3K9me2 (Lee & Skalnik 2005). This global shift towards euchromatin was interpreted as a loss of SET1 inhibition in the absence of CFP1. Based on its ability to bind to non-methylated DNA, CFP1 was proposed to restrict the mammalian SET1 complex to euchromatin (Tate et al. 2010), which is consistent with enrichment of CFP1 at 81% of CGIs, and a strong correlation between the levels of CFP1 and H3K4me3 (Thomson et al. 2010). Despite the overlap between regions of non-methylated DNA, sites enriched for CFP1 and sites enriched for H3K4me3, the role of CFP1 in H3K4 methylation, and its mechanistic function at CpG islands remain poorly understood.

### 1.14 SET1 and the MLL methyltransferase enzymes

In yeast, the SET1 methyltransferase is singularly responsible for H3K4 methylation (Briggs et al. 2001). Importantly, the SET1 gene has duplicated and diversified in higher eukaryotes, producing a family of related enzymes (Aravind et al. 2011). In *Drosophila*, there are three H3K4 methyltransferases, Trithorax (Trx), Trithorax-related (Trr) and dSET1. Mammals have six related H3K4 methyltransferases; SET1A, SET1B, and MLL proteins 1-4 (Liu et al. 2009; Xiao et al. 2011; Aravind et al. 2011). SET1A and SET1B are most closely related to dSET1, while MLL1/2 are closest to Trx, and MLL3/4 to Trr (Figure 3). Interestingly, the MLL1 and MLL2 enzymes contain a Type 1 zf-CXXC domain, while SET1A/B and MLL3/4 do not, suggesting a functional distinction between these groups. The importance of this distinction to H3K4 methylation is not yet clear.

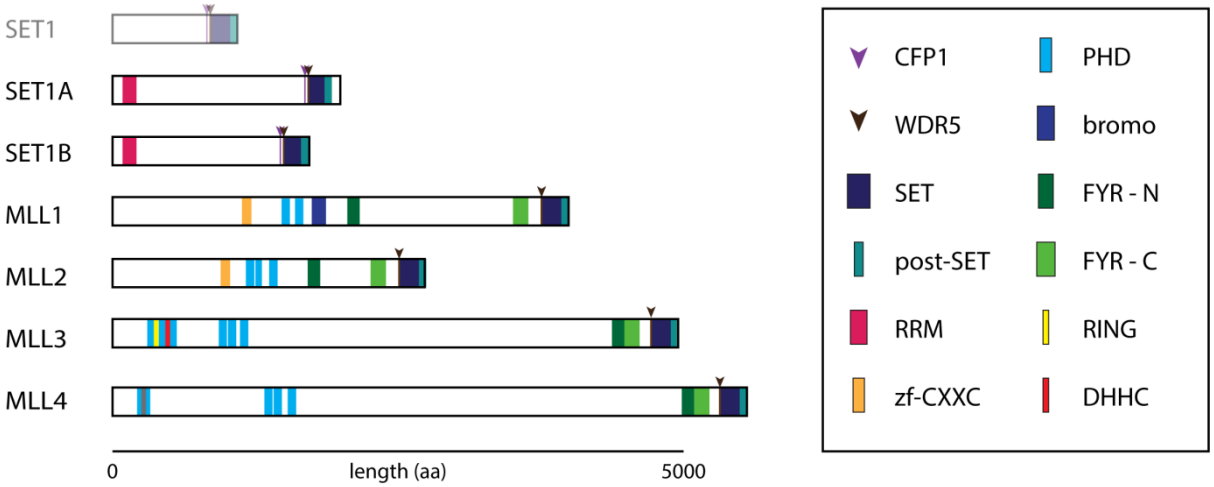



**Figure 3 – Domain architecture of H3K4 methyltransferases**

Schematic representations illustrate the domain architecture of the mammalian H3K4me3 methyltransferases. The yeast SET1 enzyme is shown at the top for comparison. CFP1 and WDR5 interaction motifs are indicated by arrowheads.

Recombinant SET1 is catalytically inactive (Roguev et al. 2001), unlike ASH1L, G9a, Suv39h1, and SET2 which also belong to the SET-family of histone methyltransferases (An et al. 2011; Patnaik et al. 2004; Tachibana et al. 2001; Strahl et al. 2002). The methyltransferase activity of SET1A/B and MLL1-4 depends on conserved protein-protein interactions. The mammalian SET1 complex consists of seven subunits, CFP1, Wdr82, SET1A/B, WDR5, RbBP5, ASH2L and Dpy-30 (Lee & Skalnik 2005; J.-H. Lee et al. 2007; van Nuland, Smits, et al. 2013). Importantly, CFP1 and Wdr82 interact with SET1A/B, but not with the MLL methyltransferases potentially reflecting divergent functions of the different H3K4 methyltransferases (Lee & Skalnik 2005; J.-H. Lee et al. 2007; Lee & Skalnik 2008). In contrast, WDR5, RbBP5, ASH2L and Dpy-30, are present in all SET1 and MLL complexes (Figure 4) (Ernst & Vakoc 2012; Dou et al. 2006).

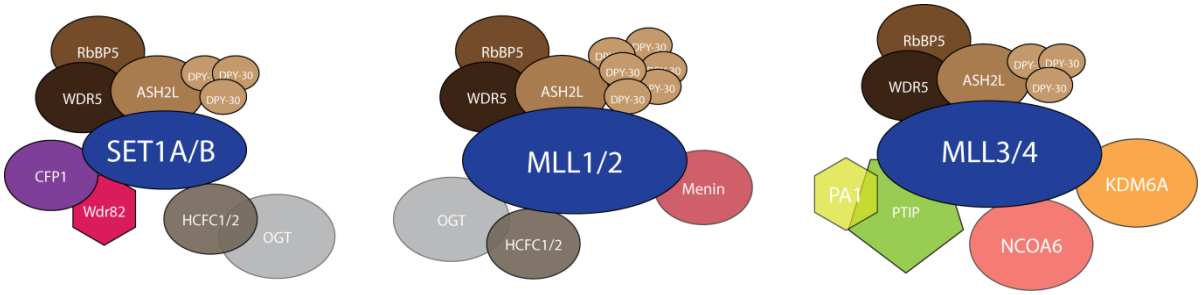



**Figure 4 –Composition of the mammalian H3K4 methyltransferase complexes**

The subunit compositions of the SET1A/B, MLL1/2 and MLL3/4 complexes are depicted by schematic representations.

## 1.15 WRAD and other components of the SET1 complex

Subunits of the SET1 complex have been studied by several groups, to try to understand how methylation of H3K4 is controlled. Among the non-catalytic subunits, WDR5, RbBP5, ASH2L and Dpy-30 (WRAD) form a stable sub-complex independently of the methyltransferase subunit (Lee & Skalnik 2005; Steward et al. 2006; Dou et al. 2006). The WRAD proteins interact with SET1A/B or MLL1-4 independently of CFP1 or Wdr82, forming a 'core' methyltransferase complex. Purifications from yeast strains with deleted subunits (Steward et al. 2006), and from insect cells expressing transgenic subunits in various combinations (Dou et al. 2006), indicated that RbBP5 and WDR5 are essential for assembly of the core complex. WDR5 and RbBP5 form a stable heterodimer, which binds cooperatively to the SET1 or MLL enzymes (Dehé et al. 2006). This interaction, as well as that between ASH2L and Dpy-30 had been observed prior to the discovery of the SET1 complex in yeast two-hybrid experiments (Miller et al. 2001). Stabilised by WDR5, RbBP5 supports assembly of the core complex by binding to a SPRY domain in ASH2L, which in turn interacts with Dpy-30 (Dou et al. 2006).

*In vitro*, an MLL1 complex lacking either RbBP5 or ASH2L is unable to catalyse H3K4 trimethylation, and produces H3K4me2 less efficiently (Dou et al. 2006). In yeast, this effect is recapitulated despite

ASH2L-independent association of the SET1 complex with chromatin (Schneider et al. 2005). This suggests that ASH2L may regulate H3K4 trimethylation by altering the product specificity of the catalytic subunit. By limiting the volume of the active site, a conserved tyrosine residue in the SET1 and MLL SET domains is thought to prevent further methylation of monomethyl-lysine (Collins et al. 2005; Patel et al. 2009). WRAD subunits may facilitate di- and trimethylation by overcoming the limitation imposed by this 'tyrosine switch'. Comparison of SET domains structures suggests that the substrate lysine and the methyl donor S-adenosylmethionine bind enter from opposite sides of the catalytic centre (Marmorstein 2003). This may allow some SET methyltransferases to add methyl groups progressively without releasing the substrate lysine. Interestingly, ASH2L interacts directly with the s-adenosylmethionine methyl-donor, and has been suggested to stimulate H3K4 methylation by forming a joint catalytic centre with the SET domain (Cao et al. 2010). Although the mechanistic details of H3K4 methylation are not certain, several studies suggest that the integrity of the core complex is critical to produce di-, and trimethylation of H3K4.

WDR5 is a WD40 repeat protein (Gori et al. 2001), and forms a characteristic 7-bladed β-propeller (Patel, Dharmarajan, et al. 2008; Song & Kingston 2008). The interaction between MLL1 and WDR5 maps to an 11 amino acid motif immediately upstream of the SET domain. This conserved WIN (WDR5 interaction) motif is present at the same location (n-SET) in the SET1A/B and MLL2-4 proteins. Interaction with WDR5 depends upon a specific arginine residue in the WIN motif, which projects into a central cleft in the propeller structure (Patel, Vought, et al. 2008; Dharmarajan et al. 2012). Interestingly, The WDR5 protein was previously reported to bind directly to the di-, and trimethylated K4 on the H3 histone tail (Wysocka et al. 2005). X-ray crystallography revealed that interaction with the H3 tail depends on H3R2, and is independent of H3K4me2/3 (Ruthenburg et al. 2006). Based on these interactions WDR5 is thought to play an important role in the interaction between H3K4 methyltransferases and chromatin. Interestingly, the WDR5-RbBP5 heterodimer dissociates upon binding to the H3 tail (Murton et al. 2010), suggesting the WDR5 subunit is

dispensable once the enzyme is bound to its substrate. Competition for the WDR5 cleft has recently been demonstrated *in vivo* using WIN and H3 peptides (Dharmarajan et al. 2012).An analogous function has also been proposed for the PRC2 component EED (Margueron et al. 2009). Like WDR5, EED is a WD40 repeat protein, which binds to both the H3 histone tail and to the methyltransferase EZH2 (Han et al. 2007).

Like CFP1, the WD40 repeat protein Wdr82 interacts with SET1A/B, but not with the MLL enzymes (Lee & Skalnik 2008). Wdr82 has been proposed to bind to chromatin containing ubiquityl-H2B, and to the initiating form of RNA polymerase (J.-S. Lee et al. 2007; Lee & Skalnik 2008), potentially linking SET1A/B to the early stages of transcription. Knock down of Wdr82 in several human cell lines results in reduced levels of H3K4me3, despite the presence of MLL1-4 (Wu et al. 2008). This was thought to show that SET1A/B is responsible for majority of H3K4 trimethylation. This appears to be the case in *Drosophila*, where dSET1 is the major H3K4 methyltransferases (Ardehali et al. 2011; Mohan et al. 2011). Although these studies reveal an important relationship between histone lysine ubiquitylation and histone lysine methylation, their findings are complicated by the additional involvement of Wdr82 in the PP1 phosphatase complex (Lee et al. 2010).

In *Saccharomyces cerevisiae*, deletion of the SET1 gene results in the complete absence of mono-, di-, and trimethylated H3K4, while these effects are reversed by expression of exogenous SET1 at wild type levels (Briggs et al. 2001). Deletion of Spp1, the yeast homologue of CFP1, results in a 2-fold reduction in the level of SET1 protein, and a 3-fold reduction in the level of H3K4me2 and H3K4me3 (Dehé et al. 2006). This suggests that CFP1 is not essential for SET1 methyltransferase activity, but plays a conserved role in the regulation of H3K4 methylation. In yeast, mutation of the SET1 tyrosine switch rescues deletion of the CFP1 homologue Spp1 (Takahashi et al. 2009), suggesting that CFP1 may regulate the product specificity of the SET1 enzymes. CFP1 contains several conserved protein domains, including a PHD domain at its N-terminal. Interestingly, this

domain is highly conserved in the yeast homologue Spp1, while the zf-CXXC domain is absent. Similarly, in *Drosophila*, the PHD domain is conserved while the zf-CXXC domain lacks the FGKK and IRQ motifs required to recognise non-methyl CpGs. The absence or divergence of the zf-CXXC domain in species which lack pervasive DNA methylation may indicate functions of CFP1 which are independent of DNA methylation. Despite clear evidence that CFP1 can bind specifically to non-methylated CpGs, function of CFP1 at CpG islands is unclear. Similarly, while CFP1 is known to interact with the SET1 complex *in vivo* and *in vitro*, its role in this complex is not understood.

### 1.16 Aims of this thesis

Transcription depends on the assembly of a pre-initiation complex (PIC) at the transcription start site. At a given promoter, this process may be regulated by a range of sequence specific transcription factors and by the surrounding chromatin environment which may facilitate or repress transcription. In vertebrate species, the majority of gene promoters overlap with CpG island regions, which are typically devoid of DNA methylation and rich in H3K4me3. Despite the apparent relationship between CpG islands and transcription start sites, it is not clear how this unique chromatin architecture is generated, or how it contributes to gene expression.

CpG islands are specifically recognised by a small family of proteins that contain a zf-CXXC domain. This family includes three proteins involved in the methylation of H3K4; CFP1, MLL1 and MLL2. While the MLL1/2 methyltransferases can modify H3K4 directly, CFP1 acts through the SET1A/B methyltransferases which lack a zf-CXXC domain. Despite an apparent redundancy between H3K4 methyltransferases which bind to CpG islands, CFP1, MLL1 and MLL2 are each essential for mammalian development, suggesting they must also perform unique roles. Although the biochemistry of these methyltransferase complexes is fairly well characterised *in vitro*, far less is known about their behaviour *in vivo*. To understand how CFP1 contributes to the function of CpG

islands, SET1A and H3K4me3, the central aims of this thesis are to determine how CFP1 interacts with the genome, and to mechanistically dissect the contribution of the zf-CXXC and PHD domains.

In addressing these questions, Chapters 3 and 4 investigate interactions between CFP1 and chromatin throughout the genome, identifying specific regions which are preferentially bound. Building on this analysis, Chapters 5 and 6 aim to determine the contribution of the CFP1 PHD domain to its interaction with chromatin, analysing its role in genome-wide localisation, and measuring the impact of its specific interactions on CFP1 motility in the nucleus. Chapter 7 examines the interaction between CFP1 and SET1 *in vivo*, to determine whether direct interaction with the SET1 complex is required for CFP1 to bind to chromatin. Together this work reveals novel distinctions between zf-CXXC-, and PHD-mediated interactions and connects their spatial and kinetic effects to the genome-wide localisation of CFP1, and the SET1 complex.

# Chapter 2 - Materials and Methods

## 2.1 DNA methods

### 2.1.1 Original cDNAs and plasmid expression vectors

Fragments and full length inserts used in this study were generated by PCR-amplification from cDNA templates and inserted into the required vectors by via ligation-independent cloning (LIC) (Aslanidis & de Jong 1990). The original cDNAs and plasmid vectors used in this study are shown below (Table 1 and Table 2 respectively). Constructs generated by PCR were verified by sequencing prior to use.

**Table 1- cDNAs**

| Origin | Protein | Length (bp) | NCBI Reference Sequence | Source |
|---|---|---|---|---|
| Homo sapiens | CFP1 | 1971 | NM_014593 | Full length IMAGE cDNA clone BC014940 |
| Homo sapiens | SET1A | 5124 | NM_014712 | Gene Art (Invitrogen) |
| Bioengineered | eGFP | 720 | U55763 | Gene Art (Invitrogen) |

**Table 2 - Plasmid vectors**

| Plasmid | Name | Function | Source |
|---|---|---|---|
| 177 | pNIC28 | Bacterial Expression<br>N-terminal 6His tag | Jin Zhou |
| 273 | pGTVL2 | Bacterial Expression<br>N-terminal GST fusion | SGC |
| 441 | pCAG FS2 LIC puro | Eukaryotic Expression<br>N-terminal FS2 fusion | Anca Farcas |
| 453 | pCAG eGFP LIC puro | Eukaryotic Expression<br>N-terminal GFP fusion | This project |
| 454 | pCAG LIC eGFP puro | Eukaryotic Expression<br>C-terminal GFP fusion | This project |

**Table 3 – PCR Primers**

| Primer | Name | Sequence | Function |
|---|---|---|---|
| 1720 | CFP1f FWD | TACTTCCAATCCATGcggcagaagtgccggctgcg | LIC cloning |
| 1721 | CFP1f REV | TATCCACCTTTACTGctaccgatgccgcttgtatcgct | LIC cloning |
| 2114 | CFP1 FWD | TACTTCCAATCCATGgagggagatggttcagac | LIC cloning |
| 2115 | CFP1 PHDf REV | TATCCACCTTTACTGctaccgctcccgtgacttcttgtg | LIC cloning |
| 871 | CFP1 REV | TATCCACCTTTACTGtcagcggtcggcactgga | LIC cloning |
| 2956 | CFP1 REV -stop | TATCCACCTTTACTCCCgcggtcggcactggagcgc | LIC cloning |
| 1809 | PHD* Y28A FWD | gagaatgcgcccatcGCctgcatctgccgcaaac | Mutagenesis |
| 1810 | PHD* Y28A REV | gtttgcggcagatgcagGCgatgggcgcattctc | Mutagenesis |
| 1811 | PHD* W49A FWD | gacaactgcaatgagGCgttccatggggactgc | Mutagenesis |
| 1812 | PHD* W49A REV | gcagtccccatggaacGCctcattgcagttgtc | Mutagenesis |
| 1990 | CXXC* R200A FWD | ggccccaacaagatcGCgcagaagtgccggctg | Mutagenesis |
| 1991 | CXXC* R200A REV | cagccggcacttctgcGCgatcttgttggggcc | Mutagenesis |
| 2162 | SID* K393A FWD | gcccagcccagctccGCGtattgctcagatgac | Mutagenesis |
| 2163 | SID* K393A REV | gtcatctgagcaataCGCggagctgggctgggc | Mutagenesis |
| 2164 | SID* Y394A FWD | cagcccagctccgcgGCTtgctcagatgactgtggc | Mutagenesis |
| 2165 | SID* Y394A REV | gccacagtcatctgagcaAGCcgcggagctgggctg | Mutagenesis |
| 2062 | SET1A FWD | tacttccaatccATGgatcaggaaggtggggg | LIC cloning |
| 2063 | SET1A REV | tatccacctttactTCAgtttagggagccccggcagc | LIC cloning |
| 1682 | XhoI EGFP FWD | tatttcCTCGAGgccgccaccATGgtgagcaagggcgagga | Cloning |
| 1683 | XmaI EGFP REV - stop | tatcgaCCCGGGcttgtacagctcgtccatgc | Cloning |

To amplify DNA inserts by PCR, 20-mer oligonucleotide primers were chosen, inclusively bracketing the desired sequence (Table 3). Where restriction sites were introduced, 6 bp were added to the 5' end of the primer to ensure efficient digestion of the PCR product by the restriction enzyme. For site directed mutagenesis, primers were designed in accordance with the Quikchange protocol

(Quikchange Agilent Technologies). Amplification of insert sequences for cloning was performed using Phusion High fidelity polymerase (#M0530 NEB). The human SET1A cDNA has a GC content of ~70% and did not amplify using the standard reaction conditions therefore amplification of SET1A was performed using the GC-rich HF reaction buffer. PCR reaction conditions and thermal cycling parameters are summarized below (Table 4 and Table 5 respectively).

**Table 4 - PCR reaction conditions**

| Master Mix | Volume (µl) / 50 µl Reaction | Final Concentration |
|---|---|---|
| $ddH_20$ | 32.5 | - |
| 5 x Phusion HF or GC-rich buffer | 10.0 | 1x |
| dNTP | 2.5 | 200 µM |
| DMSO | 1.5 | 3% |
| Phusion high fidelity polymerase | 0.5 | 1 U |
| 10 µM Primer 1 | 1.0 | 0.5 µM |
| 10 µM Primer 2 | 1.0 | 0.5 µM |
| 1 ng/µl plasmid DNA | 1.0 | 0.02 ng/µl |

**Table 5 - PCR Thermal cycling parameters**

| Step | Temperature (˚C) | Time (s) |
|---|---|---|
| Initial Melt | 98 | 30 |
| Melt | 98 | 10 |
| Anneal | 65 | 10 |
| Elongation | 72 | 30/ kb |
| Final Elongation | 72 | 420-600 |
| Hold | 10 | ∞ |

### 2.1.2 Gel electrophoresis

Size separation of DNA was achieved by electrophoresis at 100 V through a 0.5-2% agarose gel in TAE buffer (40 mM Tris-Acetate, 1m M EDTA, pH 8.0). DNA was visualised by UV illumination with 0.5 µg/ ml Ethidium Bromide, and imaged using a Molecular Imager GelDoc XR System (Bio-Rad).

### 2.1.3 Plasmids

CFP1f was expressed from a bacterial vector (pNIC-28), which introduced the N-terminal His-tag. For the generation of GFP-tagged stable cell lines, cDNAs were LIC cloned into p453 or p454 for N- or C-tagging respectively. These vectors were adapted from pCAG FS2 LIC puro (generated by Anca Farcas) by replacing the 479 bp sequence between XhoI and XmaI sites with an in frame enhanced GFP (subsequently referred to as GFP) (Zhang et al. 1996). Expression of the GFP-fusion proteins is driven by the robust CMV/IEE promoter (pCAG) which was empirically found to be effective in embryonic stem cells, remaining free of repressive methylation.

### 2.1.4 LIC cloning

The majority of plasmids generated in this study were constructed using ligation independent cloning (LIC) (Aslanidis & de Jong 1990). DNA inserts were amplified by PCR using specifically designed LIC primers (Table 3). These primers introduce specific sequences at the 5' and 3' ends of the insert, which are complementary to the sequences in the recipient vector. To prepare the recipient vector, 5 µg of vector DNA was linearised a total volume of 50 µl (Table 6). The vectors used in this study were designed for BaeI digestion. This restriction enzyme requires 20 µM S-adenosylmethionine (SAM). Digestion was performed overnight at 25°C. The amplified insert and the linearised vector DNA were both purified using the QIAquick kit according to the instructions from Qiagen. Purified DNA was eluted in 30 µl $ddH_20$. The insert and the vector DNA were then treated with T4 DNA polymerase (Fermentas 00072854). In the absence of dNTPs, the 3'->5' exonuclease activity of the polymerase dominates. By including a single dNTP (dCTP for inserts, dGTP for vectors) resection proceeds until the enzyme reaches a complementary base. The polymerase activity then adds the relevant nucleotide, stabilising the cohesive overhang. This treatment resects the 3' ends to expose the cohesive LIC sequences. After 30 minutes at 22°C, the polymerase is inactivated by heating the reaction to 75°C for 20 minutes. Following resection of the 3' ends, DNA was size selected using the Qiagen Gel Extraction Kit according to the manufacturer's instructions.

The recipient vector DNA was incubated for 20 minutes with a molar excess of insert DNA in a 20 µl volume. Due to the 12 bp cohesive ends DNA ligase is not required. Finally the hybridised DNA was transformed into chemically competent XL-10 Gold *E. coli*.

**Table 6 - BaeI digestion conditions**

| Reagent | Volume (µl)/ 50 µl Reaction | Final Concentration |
|---|---|---|
| $ddH_2O$ | 34 | - |
| 10x Cutsmart buffer (NEB) | 5 | X1 |
| 200 µM SAM | 5 | 20 µM |
| 5U/ µl BaeI (NEB) | 1 | 0.1 U |
| 1 µg/µl DNA | 5 | 100 ng/µl |

**Table 7 – T4 DNA polymerase treatment conditions**

| Reagent | Volume (µl)/ 50 µl Reaction | Final Concentration |
|---|---|---|
| PCR product/ linearised vector | 30 ul | - |
| 5x T4 Polymerase buffer (Fermentas) | 10 ul | 1x |
| 2 mM CTP or GTP | 2 µl | 4 nmol |
| 5 U/µl T4 Polymerase (Fermentas) | 2 µl | 10 U |
| $ddH_20$ | 1 µl | - |

## 2.2 Cloning and protein expression in bacteria

### 2.2.1 Heat-shock transformation of bacterial cells

Chemically competent XL-10 Gold *E. coli* were transformed using the heat-shock method (Sambrook & Russell 2001). Frozen aliquots of the competent *E. coli* were gently thawed on ice. For LIC cloning 1-7 µl of hybridised DNA cells was added to 100 µl *E. coli* at a competency of 106 colonies/ µg of DNA. For plasmid DNA, 1 ng was sufficient for efficient transformation. Heat-shock was performed at 42°C for 45 seconds, followed by 2 minutes on ice. The cells were then recovered in 1 mL of Luria-

Bertani (LB) media (10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl, dissolved in distilled H2O and pH adjusted to 7.0 prior to autoclaving) for 1 h at 37°C. Cultures were then plated on to LB agar containing 50 µg/ml kanamycin, or 100 µg/ml ampicillin, as appropriate for selection of transformed bacteria. For transformations using hybridised DNA, the entire 1 ml culture was plated, compared to 10% for transformations using intact plasmid DNA. After ~24h 10-100 colonies were typically observed compared to 0-5 colonies using untransformed *E. coli*, or cells transformed with linearised vector in the absence of insert DNA.

### 2.2.2 Bacterial expression of recombinant proteins

Proteins of interest were cloned into the pNIC28 or pGTVL2 vectors for bacterial expression. After verifying the inserted sequence by DNA sequencing, the resulting plasmids were transformed into SCG (BL-21 based) *E. coli*. LB media was inoculated with a single isolated SCG colony, and cultured at 37°C. Once the 2-3 L culture reached an optical density of 0.5 absorbance units (wavelength = 600 nm), the temperature was shifted to 30°C and expression was induced by adding 1 mM IPTG. For proteins which coordinate zinc, cultures were supplemented with $ZnCl_2$ to a final concentration of 250 µM to ensure proper folding. After 3 h the cells were harvested by centrifugation. To break open the cells, pellets were resuspended in lysis buffer and sonicated on ice using a microtip sonicator at 60% amplitude for six cycles (30 s on: 30 s off). After pelleting the insoluble fraction, soluble lysate was loaded onto a 500 µl column of Ni-NTA resin. The column was then washed with 20-40 column volumes of wash buffer. Bound protein was eluted in 500 µl fractions. Selected elutions were pooled and dialysed in 2 L of BC150 for >18 h at 4 ˚C with constant stirring.

**Lysis buffer**
20 mM Tris.HCl pH 8.0
500 mM NaCl
0.1% NP 40
20 mM imidazole

**Wash buffer**
50 mM $NaH_2PO_4$
300 mM NaCl
20 mM imidazole
Adjust pH to 8.0 using NaOH

**Elution buffer**
50 mM $NaH_2PO_4$
300 mM NaCl
250 mM imidazole
Adjust pH to 8.0 using NaOH

**Dialysis buffer (BC150)**
50 mM Hepes, pH 7.9
150 mM KCl
10% Glycerol
0.5 mM DTT (add fresh)

### 2.2.3 GST-pull down

The recombinant H3K4C and H3K4Cme3 histones used in this study were prepared by Jin Zhou. GST pull-downs were performed as described in Shi et al. 2006. Glutathione agarose beads (GE Healthcare) were washed with GST binding buffer prior to use. To immobilise each GST-fusion protein, 250 µl of a 100 µM stock was added to 50 µl of packed GSH beads, and placed at 4˚C for 1 h on a rotating wheel. To collect the beads, the mixture was centrifuged at 400 rcf and 4˚C for 5 minutes. To block non-specific binding, the beads were resuspended in GST binding buffer supplemented with 5% BSA, and rotated for 30 minutes at 4˚C. In each 200 µl reaction, 10 µg of recombinant histone was mixed with 50 µl of the blocked GST-fusion beads. After 2 h rotating at 4˚C, beads were collected by centrifugation at 400 rcf for 5 minutes at 4˚C. Following three washes with 1 ml of GST wash buffer, the beads were washed once in GST binding buffer, to reduce the salt concentration. The beads were then resuspended in 35 µl SDS loading buffer and boiled for 10 minutes at 95˚C to release any bound histone. After centrifugation at 400 rcf for 5 minutes, a 30 µl aliquot was taken as the GST pull-down.

**GST binding buffer**
50 mM Tris, pH 7.5
150 mM NaCl
10% glycerol
0.1% NP40

**GST wash buffer**
50 mM Tris, pH 7.5
150 mM NaCl
10% glycerol
0.5% NP40

### 2.2.4 Peptide-pull down

Biotinylated H3, H3K4me1, H3K4me2 and H3K4me3 peptides synthesised by GL Biochem (Shanghai) Ltd. were diluted to a working concentration of 100 µg/ml in PBS. Peptide pull-downs were performed following a protocol devised by the Gozani lab. In each 250 µl reaction, 10 µg of peptide was mixed with GST-fusion protein at a concentration of 100 µM. The mixtures were incubated overnight at 4˚C. Biotinylated peptides were captured by adding 20 µl of Sera-Mag Neutravidin Speed beads (Thermo Scientific). After several washes, the beads were resuspended in 40 µl of SDS loading buffer and boiled for 10 minutes at 95˚C to release bound fusion proteins. After magnetic capture of the beads, a 30 µl aliquot was taken as the peptide pull down.

**Peptide wash buffer**
50 mM Tris, pH 7.5
300 mM NaCl
0.05% NP-40

## 2.3 Generation of rabbit polyclonal antibodies

Two polyclonal α-CFP1 antibodies were generated and affinity purified from rabbit serum. Rabbit R3077 was immunised with a peptide antigen (CFP1p) while rabbit R3078 was immunised with a recombinant protein fragment (CFP1f).

### 2.3.1 Immunisation using a peptide antigen

The CFP1p peptide (CQLRARESYKYFPSSL) was conjugated to malemide-activated keyhole limpet hemocyanin (KLH) (Sigma) according to the supplier's instructions in a final concentration of 10% DMSO. KLH-CFP1p was then purified for immunisation using a HiTrap desalting column (GE life sciences). For affinity purification CFP1p was conjugated to agarose beads (Pierce Sulfolink) in a final concentration of 10% DMSO. The second α-CFP1p bleed from the R3077 rabbit was then affinity purified using these beads.

### 2.3.2 Immunisation with a protein fragment antigen

A 25 kDa His-tagged fragment of CFP1 corresponding to amino acids 206 to 360 (CFP1f) was expressed in SGC *E. coli* and purified on a Ni-NTA column. Peak elutions (4-7) were pooled and dialysed into BC150 yielding 2 ml of 0.5 mg/ml for immunisation of rabbit R3078. Following the manufacturer's guidelines 5 mg CFP1f was covalently immobilised on 500 µl Affi-Gel 10 resin (Bio-Rad) which binds to primary amines. This resin was used to affinity purify α-CFP1 polyclonal antibodies from the second bleed of R3078.

### 2.3.3 Affinity purification of polyclonal antibodies

Bleeds collected from immunised rabbits were incubated with the antigen resin for > 3 h at room temperature, or overnight at 4°C. A volume of 1 ml packed resin was used per 10 ml of serum. The resin was then washed with 25 ml 0.5 M NaCl followed by 25 ml PBS to remove non-specific interactions. Elutions were performed by adding 500 µl of 0.1 M Glycine pH 3 to the resin column to denature the antibody variable region. The flow through was immediately collected into Eppendorf tubes containing 50 µl of 1 M Tris pH 8 to neutralise the eluted protein. Eluted fractions were checked by SDS-PAGE analysis. Columns were neutralised with 1 M Tris pH 8, washed with PBS and stored at 4°C for future use.

## 2.4 Mammalian Tissue Culture

### 2.4.1 Growth and maintenance of mammalian cell lines

C127, HeLa and HEK293T cells were grown at 37°C and 5% $CO_2$ in DMEM (Gibco) supplemented with 10% foetal calf serum (FCS, Sera lab), 1000 U/ml penicillin and 1000 µg/ml streptomycin (Gibco, 15140-122). For mouse ES cells DMEM was supplemented with 15% FCS, non-essential amino acids (Gibco, 11140-035), 2 mM L-glutamine, 50 µM 2-mercaptoethanol, 0.01 ug/ml human LIF, 1000 U/ml penicillin and 1000 µg/ml streptomycin (Gibco, 15140-122). ES cells seeded onto plates gelatinised with 0.1% gelatine PBS, coated with a layer of mitomycin C inactivated SNLP feeder cells.

### 2.4.2 Generation of Stable Cell Lines

6-well plate wells were seeded with 0.24 x$10^6$ C127 cells, and grown for >18 h. The media was then changed to remove antibiotics which might interfere with the transfection. 3 µg of plasmid DNA was incubated for 15 minutes with 9 µl of Fugene HD (Promega) in 150 µl DMEM (Lonza). The Fugene-DNA mixture was then added to the C127 in 1 ml of media. After 24 h the transfected cells were washed twice in PBS, dissociated using trypsin (Invitrogen) and diluted 1/25 before plating. After a further 24 h the media was changed to include 2.5 µg/ml puromycin. Transfected cells were always grown in this concentration from this point on. Cell lines were first passaged (p1) after 6 days of selection; colonies were trypsinised and resuspended in individual wells of 48-well plates. Lines were screened by western blotting p4 nuclear extracts staining for endogenous protein and for GFP. Lines exhibiting appropriately sized bands that stained for both GFP and the endogenous protein were screened by live fluorescence microscopy and genotyped.

### 2.4.3 Isolation of genomic DNA

Genomic DNA was prepared from 1x$10^6$ cells. Cell pellets were washed PBS and resuspended in lysis buffer (50 mM Tris pH 8.0, 10 mM NaCl and 10 mM EDTA). 50 µl 10% SDS was added to lyse the cells

giving a final concentration of 0.5 %. Proteinase K (Sigma P4850-1ML) was added to a final concentration of 20 µg/ml, and DNAse-free RNAse (Sigma R6148-25ML) to 200 µg/ml. Phenol chloroform extraction of the genomic DNA was performed after 3 h at 65˚C. Diagnostic PCRs were performed with Phusion polymerase (NEB #M0530) using 50 ng of genomic DNA in a total volume of 50 µl. Control reactions used 0.5 ng plasmid DNA with or without polymerase. Bands of the correct size were gel extracted and verified by DNA sequencing.

### 2.4.4 Preparation of whole cell extract

Cell pellets were washed twice in 1xPBS then resuspended in 5x the pellet volume of Mammalian cell lysis buffer. After 30' rotating at 4˚C the insoluble fraction was pelleted by centrifugation at 16,100 g 4˚C for 10'. The supernatant was taken as whole cell extract.

**Mammalian cell lysis buffer**
1% NP-40
150 mM NaCl
10% Glycerol
50 mM Tris, pH 8.0
0.5 mM DTT (added fresh)
1 x EDTA free cOmplete Protease Inhibitor (Roche 11873580001) (added fresh)

### 2.4.5 Preparation of nuclear extract

Cells were harvested using trypsin and pelleted at 215 g for 3 minutes. The cells were washed once in PBS before transfer to a microcentrifuge tube. The cells were then spun at 1500 rcf, 4˚C for 5 minutes. The supernatant was then aspirated and the dry pellets frozen at -80˚C. When required the pellets were thawed and resuspended in 10 times the pellet volume of nuclear extraction buffer A, then incubated on ice for 10 minutes to swell. Swollen cells were spun at 1500 rcf, 4˚C and resuspended in 3 times the pellet volume buffer A with 0.1% NP-40. Nuclei were collected by centrifugation at 1500 rcf for 5 minutes at 4˚C and immediately resuspended in 1 pellet volume of nuclear extraction buffer B. After 1 h on ice the suspension was pelleted at 16000 rcf for 20 minutes

at 4˚C and the supernatant saved as nuclear extract. To ensure equivalent loading, extracts were quantified by Bradford assay and adjusted using buffer B to maintain salt concentration and buffer conditions.

**Nuclear extraction buffer A**
10 mM HEPES, pH 7.9
1.5 mM MgCl2
10 mM KCl
0.5 mM DTT (added fresh)
1 x EDTA free cOmplete Protease Inhibitor (added fresh)

**Nuclear extraction buffer B**
5 mM HEPES, pH 7.9
26% Glycerol
1.5 mM MgCl2
0.2 mM EDTA
0.5 mM DTT (added fresh)
1 x EDTA free cOmplete Protease Inhibitor (added fresh)

### 2.4.6 Chromatin preparation

C127 cell lines were expanded from $1x10^7$ to $5x10^7$ cells in two passages over 6 days. Cells were dissociated with trypsin and resuspended in 10 ml PBS, and fixed at 25˚C by adding 1 ml 37% formaldehyde, to give a final concentration of 1%. The fixation reaction was quenched after 15 minutes by adding glycine to a final concentration of 150 mM. These double cross-linked cells were then lysed by incubation in 10 ml LB1 for 10 minutes at 4˚C. The release nuclei were then washed for 10 minutes at 4˚C by resuspension in 10 ml LB2, before 50’ sonication in 1 ml LB3 to shear the DNA. After sonication each ml of chromatin was diluted by adding 150 µl LB3 containing 10% TritonX-100. The insoluble fraction was then pelleted by centrifugation at 16,100 g for 10 minutes at 4˚C and the supernatant containing soluble chromatin was transferred to a screw cap tube for storage at -80˚C.

To ChIP chromatin modifications such as H3K4me3, cells were fixed with 1% formaldehyde as above but without adding EGS. The additional cross linking step used to ChIP CFP1 was previously found to be required to capture binding of the zf-CXXC proteins KDM2A and KDM2B proteins to chromatin.

**LB1**
50 mM HEPES KOH, pH 7.9
140 mM NaCl
1 mM EDTA
10% Glycerol
0.5% NP40
0.25% Triton x 100

**LB2**
10 mM Tris HCl, pH 8.0
200 mM NaCl
1 mM EDTA
0.5 mM EGTA

**LB3**
10 mM Tris HCl, pH 8.0
100 mM NaCl
1 mM EDTA
0.5 mM EGTA
0.1% Na Deoxycholate
0.5% N-lauroylsarcosine

## 2.5 Analysis of cellular extracts and protein purifications

### 2.5.1 α-GFP immunoprecipitation

Nuclear extracts were diluted to 550 µl of 1 µg/µl with BC150 containing freshly added protease inhibitors and DTT. 50 µl was saved as input while 500 µl was added to 500 µl 2 µg/µl sheep α-GFP (provided by the Barr lab) in BC150 and incubated over night at 4˚C. The antibody was then precipitated by incubating with 15 µl (packed) of 2 mg/ml protein-G sepharose beads (GE Healthcare 17-0618-01) for 2 h at 4˚C. The beads were then pelleted at 400 rcf for 5 minutes and the supernatant taken as flow through. After washing three times with BC150, the immunoprecipitated fraction was collected by pelleting the beads at 16.2 rcf for 1 minute before boiling with SDS loading buffer. The beads were then pelleted again and the supernatant analysed by western blot.

### 2.5.2 SDS-PAGE

Proteins were separated based on their molecular weights using gel electrophoresis (Laemmli 1970). The stacking gel was composed of 5% Acrylamide/Bis solution (BioRad), 0.1% SDS, 0.1% ammonium persulphate (APS, Sigma), 0.1% N,N,N',N'-Tetramethylethylenediamine (TEMED, Sigma) and 125 mM Tris-HCl pH 6.8. The separating gel was prepared to different percentages of polyacrylamide (6-15%), depending on the size of the protein of interest. A typical casting scheme is shown in Table 8. All gels were cast and run using the Mini-Protean Tetra Cell system (BioRad). Protein samples were mixed with 1x SDS loading buffer (2% SDS, 100 mM Tris pH 6.8, 100 mM DTT, 10% glycerol and 0.1% bromophenol blue), and boiled at 95°C for 5 minutes. Gels were typically run at 200 V for 50 minutes in 1x SDS-PAGE running buffer (25mM Tris, 192mM glycine, 0.1% SDS). After Coomassie staining, protein gels were scanned (EPSON Perfection V750 PRO).

**Table 8 – 8% Polyacrylamide gel casting**

| Reagent | 5 ml Stacking Gel (ml) | 10 ml Separating Gel (ml) |
|---|---|---|
| ddH2O | 3.843 | 5.190 |
| 40% acrylamide | 0.623 | 2.000 |
| 1.5 M Tris HCl | 0.429 (pH 6.8) | 2.600 (pH 8.8) |
| 10% SDS | 0.050 | 0.100 |
| 10% APS | 0.050 | 0.100 |
| TEMED | 0.005 | 0.006 |

### 2.5.3 Western Blot

Following electrophoresis at 200 V for ≥1 h, semi-dry transfer was conducted at 200 mA and ≤23 V for 1.5 h onto a nitrocellulose membrane (0.45 µm). Membranes were blocked in 5% milk powder, 0.1% Tween20 PBS (MPBST) at 4˚C with gentle rocking for ≤18 h. Membranes were incubated with primary antibodies diluted in MPBST for 2 h at room temperature. After three washes in 0.1% Tween20 PBS (PBST) membranes were incubated with horse radish peroxidase conjugated secondary antibodies diluted in 1% BSA PBST for 1 h. After three washes in PBST membranes were

soaked in enhanced chemiluminescence (ECL) solution. X-ray films (Fuji) were exposed and developed to detect the antibody signals.

**Semi-Dry transfer buffer**
39 mM Glycine
48 mM Tris
0.037 % SDS
20% Methanol

## 2.6 Genome-wide profiling of binding interactions

### 2.6.1 Chromatin Immunoprecipitation

Chromatin immunoprecipitation were performed on freshly prepared chromatin. To examine range of fragment sizes produced by the sonication step a 50 µl aliquot from each chromatin preparation was de-crosslinked as described below and analysed by TAE electrophoresis on a 1.5% agarose gel. For genome-wide sequencing many IPs were performed in parallel and pooled. For each experiment the antibody coated Novex protein-A Dynabeads were prepared fresh. After making sure the stock was well suspended, 25 µl of 30 mg/ml Dynabeads/ IP was transferred to a 2 ml LoBind Eppendorf tube (#022431081). Beads were washed with 1ml of cold PBS supplemented with 0.5% BSA (filtered). Washed beads were divided into 1 ml LoBind Eppendorf tubes for conjugation to different antibodies in parallel. The appropriate volume of antibody (Table 9) was added to 25 µl of beads/ IP in 1ml 0.5% BSA, and rotated for 4˚C for 4 h. After attaching the antibody, wash the beads three times in 0.5% BSA as above, to remove unbound antibody.

**Table 9 – Antibodies used for ChIP sequencing**

| Antibody | Species | Source | Volume/ IP (µl) |
|---|---|---|---|
| α-CFP1f | Rabbit polyclonal | This Study | 5 |
| α-KDM2B | Rabbit polyclonal | Anca Farcas | 5 |
| α-H3K4me3 | Rabbit polyclonal | Klose Lab | 5 |
| α-Flag | Mouse monoclonal | Invitrogen (M2) | 3 |
| α-GFP | Mouse monoclonal | Invitrogen (3E6, A11120) | 3 |
| α-Pol II CTD | Mouse monoclonal | Covance (8WG16) | 3 |

Prior to immunoprecipitation, chromatin was diluted in ChIP dilution buffer. Chromatin was diluted 1/10 (or 1/50 for histone proteins or their post translational modifications) in sufficient volume to provide 1 ml of diluted chromatin/ IP, with a further 0.5 ml to serve as an input control. Immunoprecipitations were performed using 1 ml of diluted chromatin /25 µl of antibody coated beads, and rotated overnight at 4˚C. This is equivalent to $5x10^6$ cells (or $1x10^6$ for histone proteins).

**ChIP dilution buffer**
20 mM Tris-HCl, pH 8.0
150 mM NaCl
1 mM EDTA
1% Triton X 100

**Low salt buffer**
20 mM Tris-HCl, pH 8.0
150 mM NaCl
2 mM EDTA
1% Triton X 100
0.1% SDS

**High salt buffer**
20 mM Tris-HCl, pH 8.0
500 mM NaCl
1 mM EDTA
1% Triton X 100
0.1% SDS

**LiCl buffer**
10 mM Tris-HCl, pH 8.0
250 mM LiCl
1 mM EDTA
1% NP-40
1% Deoxycholate

To remove unbound chromatin, the beads were then washed according to the Upstate protocol (low salt buffer, high salt buffer, LiCl buffer, TE and TE once more). Each wash was performed in a volume of 1 ml, by inverting the tubes to mix and then rotating at 4°C for 4 minutes. Samples were then spun briefly to clear beads from the lid and placed on the magnetic rack to pellet the beads and remove supernatant before the next wash.

### 2.6.2 Elution of ChIP DNA

Washed beads were resuspended in 100 µl of freshly prepared elution buffer and shaken for 30 minutes. The tubes were placed at 65°C for 2 minutes then spun at 16,100 rcf to settle foam. To reverse the chemical cross-linking, 4 µl of 5M NaCl was added to 100 µl of the Input chromatin, and to each immunoprecipitated sample. To eliminate RNA, samples were incubated with 2 µl of 500 µg/ml RNaseH (Roche 11 119 915 001,) at 65°C overnight. To digest the protein, samples were incubated with 1 µl of proteinase K 20 mg/ml at 42°C for 1.5 hours. ChIP DNA was purified using Zymo Research ChIP DNA concentrator columns, following manufacturer's instructions, and eluted in 10 µl of ddH2O. For ChIP sequencing, replicate IPs were pooled at this stage. The concentration of double stranded DNA was quantified using PicoGreen.

**ChIP elution buffer**
100 mM $NaCO_3$
0.1% SDS

### 2.6.3 Sonication of ChIP DNA

To reduce bias due to preferential PCR amplification of short fragments, an additional sonication step was carried out on the immunoprecipitated DNA prior to library preparation. Samples were gently thawed on ice and spin briefly to collect liquid. Sonication of ChIP DNA (>10 ng) was performed in a total volume of 40 µl in 0.5 ml Eppendorf tubes. After sonicating on high power, 30s

on - 30s off, for 1 h, pausing at 30' to collect liquid to the bottom of the tubes, samples were collected for submission to WTCHG or library preparation.

### 2.6.4 Quantitative PCR

Quantitative PCR (qPCR) was then performed using Sybr Green (Quantace) on a Rotor-Gene 6000 (Corbett). Primers were designed with the Primer3 software using a product size range of 80-120 bp, minimum, optimal and maximal primer melting temperature (Tm) of 59, 60 and 61°C, and a maximal Tm difference of 1°C. Primer locations were checked using the UCSC genome browser. The qPCR primers used in this study were designed by Neil Blackledge.

**Table 10 – Primers for ChIP-qPCR**

| Primer | Name | Sequence |
|---|---|---|
| 376 | Suv420h1 CGI FWD | GACGTGGTTTCTTTGGTGGT |
| 377 | Suv420h1 CGI REV | CGGGAAGAGGCTGAGAGAC |
| 378 | Suv420h1 Body FWD | TGAGGCTCTCAGCAAGACTG |
| 379 | Suv420h1 Body Rev | ATCTTCCAGGAGAACGAGCA |
| 97 | NcoA2 CGI FWD | AGATGGGTGGCTGTGGTCTA |
| 98 | NcoA2 CGI REV | TCCCTTCACTTCCAGTCACC |
| 198 | NcoA2 Body FWD | GAGCATGGCAGGAGAGAAAG |
| 199 | NcoA2 Body REV | AAGCTGTCACCACCCTTGAC |

### 2.6.5 Ion Proton Sequencing Library Preparation

For Ion Proton Sequencing, DNA libraries were prepared using the NEBNext Fast DNA Library Prep Set (#E6270S/L) according to the supplier's instructions.

### 2.6.6 Genome-wide analysis of ChIP-Sequencing data

Single-end fastq files were concatenated for samples split over multiple lanes. Sam files were generated for each replicate using Bowtie to map fastq reads to the mm9 reference sequence. Sam files were then converted to sorted bam files. Bam files were filtered to remove unmapped reads

and read duplications which are most likely due to PCR amplification during the library preparation stage (Table 11 and Table 12).

| **Sample**<br><br>**Total reads** | **CFP1f rep.1**<br>**32,309,519** | **CFP1f rep.2**<br>**25,385,036** | **KDM2B rep. 1**<br>**35,355,934** | **KDM2B rep. 2**<br>**23,553,408** | **Pol II rep 1.**<br>**28,578,936** | **Pol II rep 1.**<br>**28,313,105** |
|---|---|---|---|---|---|---|
| Unmapped | 2,773,206 | 2,987,163 | 4,217,605 | 1,863,816 | 1,329,739 | 2,788,441 |
| Multiply mapped | 9,462,235 | 7,923,827 | 8,181,981 | 5,305,089 | 3,801,646 | 6,159,432 |
| Total mapped | 20,074,078 | 14,474,046 | 22,956,348 | 16,384,503 | 23,447,551 | 19,365,232 |
| PCR duplicates | 2,168,484 | 1,851,047 | 1,875,847 | 1,331,075 | 1,964,777 | 1,791,538 |
| Uniquely aligned reads | 17,905,594 | 12,622,999 | 21,080,501 | 15,053,428 | 21,482,774 | 17,573,694 |

Read numbers (x$10^6$ reads)

| **Sample**<br><br>**Replicate** | **GFP**<br>**1** | **GFP**<br>**2** | **WT**<br>**1** | **WT**<br>**2** | **CXXC***<br>**1** | **CXXC***<br>**2** | **PHD***<br>**1** | **PHD***<br>**2** | **DB***<br>**1** | **DB***<br>**2** | **SET1A**<br>**1** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Total | 31.0 | 40.5 | 26.9 | 31.5 | 37.1 | 35.4 | 38.3 | 34.9 | 34.8 | 33.8 | 30.3 |
| Unmapped | 2.1 | 2.3 | 1.9 | 2.2 | 2.4 | 3.2 | 3.8 | 2.4 | 2.3 | 3.1 | 2.9 |
| Multiply mapped | 7.7 | 9.4 | 6.4 | 8.0 | 9.2 | 8.9 | 10.8 | 8.7 | 8.6 | 9.0 | 6.9 |
| Total mapped | 21.2 | 28.7 | 18.6 | 21.3 | 25.5 | 23.3 | 23.7 | 23.8 | 23.9 | 21.8 | 20.5 |
| PCR duplicates | 0.7 | 1.0 | 0.7 | 1.0 | 0.9 | 0.7 | 1.0 | 0.7 | 1.0 | 0.7 | 12.2 |
| Uniquely aligned | 20.4 | 27.8 | 17.9 | 20.3 | 24.6 | 22.7 | 22.7 | 23.1 | 22.9 | 21.1 | 8.3 |

Read numbers (x$10^6$ reads)

### 2.6.7 Peak calling

Wild type CFP1 peaks were called using the MACS algorithm (Zhang et al. 2008). The effective genome size was set to 1.87x$10^9$, masking the highly repetitive regions in the mouse genome. Wild type and mutant GFP-CFP1 profiles, as well as GFP-SETA were called against the GFP profile as a negative control, while an input sample was used for CFP1, KDM2B, Pol II and H3K4me3.

## 2.7 Fluorescence microscopy

### 2.7.1 Formaldehyde fixation

Cells for microscopy were seeded at 0.06 x$10^6$/ml on No. 1.5H (170 µm ± 5 µm) coverslips (Marienfeld) and grown for 24h. Before fixation, cells were washed in PBS to rinse away growth serum proteins and reduce autofluorescence. All washes were performed by submerging the coverslip in two successive 500 ml volumes removing excess liquid in between. Immediately after washing, cells were transferred to a new dish containing 2% paraformaldehyde (pH 7) prepared fresh. This fixation was allowed to proceed for precisely 10 minutes. Cells were then washed in 0.05% Tween20 PBS (PBST.5) and permeabilised in 0.2% Triton X-100 PBS for 10 minutes. After washing in PBST.5 coverslips were blocked in 3% BSA + 5% normal goat serum (Sigma) PBS. Coverslips were then incubated with primary antibody diluted in block buffer for 1 h in a humidified chamber. After 3 washes in PBST.5, coverslips were incubated with the fluorophore conjugated secondary antibody diluted in block buffer for 1h. After 4 washes in PBST.5 the cells were fixed again as before. Following another wash in in PBST.5, cells were incubated with 2 µg/ml DAPI PBS for 10 minutes. Coverslips were then washed briefly in dd$H_20$ to remove salt, and equilibrated with mounting medium (Vectorshield). The coverslips were then mounted on the unfrosted side of SuperFrost microscopy slides and dried. After ~18 h the slides were sealed using clear nail polish, and cleaned for imaging.

### 2.7.2 Immunofluorescence microscopy

Fixed slides were imaged on an inverted widefield fluorescence microscope (Observer 1.z Carl Ziess) using a 63x/1.4 N.A. plan-apochromat oil-immersion objective. Cells were illuminated using a 120 W metal halide light source at maximum intensity and images detected on an AxioCamMR3. DIC images were taken using a proper Kohler illumination set up. While DAPI and GFP images were taken

through Zeiss filter sets 49 and 38 respectively. Exposure times were 10 ms for DIC and DAPI and 200 ms for GFP.

### 2.7.3 Live cell imaging

Cells were seeded onto microplates (Ibidi) at a density of 0.06 x$10^6$ cells/ml, and grown for 24h. Before imaging the the media was replaced with phenol red free media (Lonza) containing 10% FCS, 12.5 mM HEPES (Gibco) and 40 µM sodium pyruvate (Gibco) after one wash with PBS. Cells were imaged on an inverted Olympus microscope connected to a spinning disk confocal system (UltraView Perkin Elmer). The microscope was fitted with a 60x/1.4 N.A. plan apo oil objective which was used in conjunction with N=1.514 oil (Applied Precision). During imaging the temperature was maintained at 37 ˚C and cells were supplied with 1.5 ml $CO_2$/min in a humidified Tokai chamber. Images were captured on a HAMAMATSU EM-CCD camera. GFP images were taken with a 100 ms exposure time at 14% laser power and a detector sensitivity of 144, using a (W50) 525 nm emission filter.

## 2.8 FRAP

FRAP was conducted using the live imaging setup described above. Initial experiments were performed bleaching a 2 µm radius circle by raster scanning the 488 nm laser at 100% intensity. In these experiments, 50 pre-bleach and 1000 post-bleach images were collected at a rate of 8 fps. Each image was exposed for 100 ms. In subsequent FRAP experiments, pixels were binned 2x2, allowing clear images to be collected in a shorter exposure time (40 ms). To reduce the contribution of diffusion to the FRAP recovery, a circular spot (radius = 1 µm) was bleached using the 488 nm laser at 100% intensity for 40 ms. In fixed cells, these setting were found to produce an 80% reduction in fluorescence intensity within the bleached region. To monitor fluorescence recovery after photobleaching in live cells, 50 pre-bleach and 1000 post-bleach images were captured at a rate of 12 fps.

### 2.8.1 FRAP normalisation

Timelapse videos of FRAP experiments were saved as .tif files for processing. FRAP measurements were performed in MATLAB using the FRAP_analysis.m function developed for this project. Using a script called BatchFRAPanalysis.m, which calls the FRAP_analysis.m function, multiple files were processed in one go using the same parameters. The code for both of these is given in Appendix x. The FRAP_analysis.m function automates normalisation of FRAP measurements to account for the initial conditions (brightness of the cell and brightness of the spot) and for unintentional photobleaching over time (Mueller et al. 2012).

### 2.8.2 Biexponential Fitting

The following equation (Eq. 1) was used to determine the half recovery time $t_{1/2}$ under the chosen FRAP conditions without making assumptions about the behaviour of the fluorescent protein.

$$FRAP(t) = v - Ae^{-t/c} - Be^{-t/d} \qquad (1)$$

Where t is time, v is the plateau of the FRAP curve and A, B, c and d are fitted parameters. Fitting was performed in MATLAB using the BiExpCurveFit.m function developed for this project which uses the built in functions feval, to fit the data, and fzero to estimate $t_{1/2}$.

### 2.8.3 Kinetic modelling

Predicted diffusion coefficents (Df) were calculated using the mass diffusion equation (Eq. 2) (Sprague et al. 2004) based on mass relative to GFP.

$$\mathrm{Df} \, \alpha \, \mathrm{M}^{-1/3} \qquad (2)$$

GFP-CFP1 is 103.7 kDa and has a predicted Df of 9.6 $\mu m^2/s$. The COMPASS complex has a predicted mass of ~450 kDa, ~480 with a GFP tag. Therefore predicted Df of the GFP-tagged complex is 5.8 $\mu m^2/s$.

### 2.8.4 Pure Diffusion Model

The pure diffusion model is given by equation 3 (Soumpasis 1983).

$$FRAP(t) = e^{-\frac{A}{2t}}\left(I0\left(\frac{A}{2t}\right) + I1\left(\frac{A}{2t}\right)\right) \qquad (3)$$

I0, and I1, are zero, and first order, modified Bessel functions of the first kind, respectively.

$A = w^2/Df$, where w is the bleach radius and *Df* is the diffusion coefficient.

## 2.9 Software

Live cell imaging and FRAP was performed using Volocity. Image analysis was performed in FIJI and MATLAB. Structural models were produced in Pymol, using files from the protein databank. Figures were prepared in Adobe Illustrator CS6, saved as .tiff files in the RGB colour space, and inserted into Microsoft Word 2010. References were organised using the Mendeley citation manager.

# Chapter 3 - Production and validation of an antibody against CFP1

## 3.1 Introduction

Vertebrate genomes are punctuated by short regions of DNA called CpG islands (CGIs) (Cooper et al. 1983). These 1-2 kb regions are characterised by a high density of CpG dinucleotides and an elevated frequency of GC base pairs compared to the rest of the genome (Gardiner-Garden & Frommer 1987). As they are associated with over 70% of mammalian promoters and are often conserved between species, CGIs are proposed to play a role in gene expression (Long, Sims, et al. 2013; Deaton & Bird 2011; Saxonov et al. 2006). However, despite being discovered over 30 years ago, CGI function remains poorly understood.

Importantly, CGIs are resistant to DNA methylation (Bird et al. 1985; Bird 1986). While the majority of CpGs in the genome are modified by methylation of the cytosine base, CpGs within CGI regions tend to remain non-methylated, further distinguishing CGI DNA from the surrounding sequence (Bird 1986; Illingworth & Bird 2009). By resisting methylation, CGIs may represent a unique platform for interactions with DNA binding proteins (Blackledge & Klose 2011).

CXXC finger protein 1 (CFP1) is essential for mammalian embryogenesis and was the first protein to be discovered that binds specifically to non-methylated DNA (Carlone & Skalnik 2001; Voo et al. 2000). The CFP1 protein contains a zinc finger CXXC (zf-CXXC) domain which binds specifically to DNA containing non-methylated CpGs (Lee et al. 2001). Interaction with methylated CpGs is prevented by a steric clash with the additional methyl group, preventing CFP1 from binding (Xu et al. 2011). Specific binding to non-methylated DNA has since been demonstrated for several other zf-CXXC proteins including the histone lysine demethylases; KDM2A,  and KDM2B (Blackledge et al. 2010; Farcas et al. 2012).

Unlike the KDM2 proteins, CFP1 does not possess an intrinsic enzymatic activity, however it is known to form a complex with the SET1A/B histone lysine methyltransferases (Lee & Skalnik 2005). The SET1 enzymes catalyse mono, di, and trimethylation of lysine 4 on the histone H3 tail (H3K4) (Briggs et al. 2001; Lee & Skalnik 2005). Previous studies have shown that H3K4 trimethylation (H3K4me3) is often found at CGI regions suggesting that CFP1 may play a role in H3K4 methylation by recruiting SET1 (Clouaire et al. 2012; Wang et al. 2008; Birney et al. 2007; Thomson et al. 2010).

H3K4me3 is generally associated with active gene promoters however its contribution to transcription is not yet clear (Strahl et al. 1999; Santos-Rosa et al. 2002; Bernstein et al. 2012). The H3K4me3 mark is recognised by a family of proteins containing a plant homeodomain (PHD) (Sanchez & Zhou 2011). This family includes general transcription factors like the TFIID component TAF3 (Vermeulen et al. 2007; Lauberth et al. 2013), and nucleosome remodellers such as BPTF and ING4 (Wysocka et al. 2006; Li et al. 2006; Palacios et al. 2008). By directing placement of H3K4me3 to CGI regions, CFP1 may play a role in the activation of transcription through the recruitment of H3K4me3 binding proteins (Sims et al. 2007; Tate et al. 2010). This suggestion is supported by the observation that H3K4me3 levels are reduced at CGI regions in CFP1 null ES cells (Thomson et al. 2010).

Recent studies have shown that zf-CXXC proteins KDM2A and KDM2B bind to CGIs genome-wide, suggesting that their activities contribute to the unique chromatin architecture of CGI regions (Blackledge et al. 2010; Farcas et al. 2012). Similarly, CFP1 has recently been profiled genome-wide using chromatin prepared from mouse brain tissue (Thomson et al. 2010). Although peaks of CFP1 enrichment were observed at CGI regions, supporting the interaction of CFP1 with these elements *in vivo*, it is difficult to interpret the relevance of these profiles, as the brain is made up of many different cell types.

To appreciate its interaction with CGIs *in vivo*, CFP1 localisation must be characterised in a homogenous and experimentally tractable system. Attempts to study CFP1 binding genome-wide using the original published antibody and other commercially available alternatives were unsuccessful. Therefore an antibody was required which could be used to further study CFP1 function.

## 3.2 Generation of a CFP1 specific antibody

CFP1 was recently demonstrated to bind CGIs *in vivo* however its contribution to CGI function is not yet clear. Attempts to study CFP1 binding with existing antibodies were unsuccessful, necessitating a new antibody reagent capable of binding specifically to CFP1. To investigate the function of CFP1 at CGIs, a rabbit polyclonal antibody against CFP1 was generated and characterised.

Human CFP1 is 660 amino acids long and contains a plant homeodomain (PHD) and a zinc finger CXXC (zf-CXXC) domain (Figure 5). Both of these structurally conserved domains are also found in other chromatin binding proteins. To minimise cross reactivity with other PHD- or zinc finger containing proteins, these domains were avoided when choosing a CFP1 antigen. A short unique sequence within CFP1 sequence was chosen and synthesised as a peptide antigen for immunisation and antibody production.

Peptide antigens are poorly immunogenic because of their low molecular weight means they are rapidly filtered out of the host's blood stream after injection (Butler & Beiser 1973). Covalent attachment of peptide antigens to carrier proteins such as keyhole limpet hemocyanin (KLH) that have a high molecular weight (~400 kDa per monomer of a 20-meric complex) dramatically improves their immunogenicity. Typically, cysteine thiol groups in the peptide are conjugated to lysine amines in the KLH protein by a reactive linker molecule. Selection of a peptide with a single endogenous

cysteine at the N-terminal facilitated attachment to KLH while maintaining the identity of the CFP1 sequence.

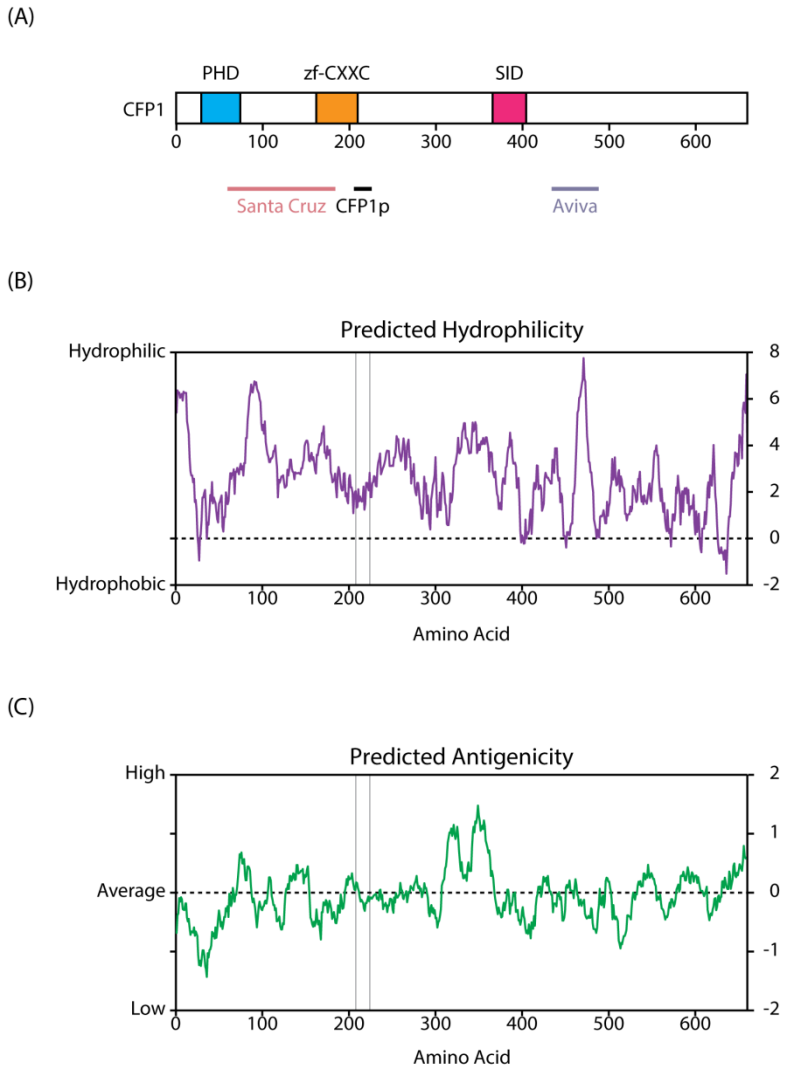



**Figure 5 - CFP1 domain architecture and antigen position**

(A) Schematic of human CFP1 depicting known structural and functional domains. The CFP1 peptide (CFP1p) sequence is indicated below the schematic along with the regions recognised by the commercial Santa Cruz (H-120) and Aviva (P050) antibodies. (B) The mean score for the predicted hydrophilicity of each residue is shown for a 15 amino acid sliding window across CFP1 (Parker et al. 1986). (C) Mean antigenicity was calculated for the same window (Welling et al. 1985). Note that a predicted antigenicity of 0 represents average antigenicity.

A 16 amino acid, peptide antigen (CFP1p) was selected based on its unique occurrence in CFP1 (using a protein BLAST search), and on predicted scores for antigenicity and solubility (Figure 5). Cysteines within the PHD and CXXC domains were avoided, as these residues are known to coordinate zinc and to be buried within the folded protein. Furthermore, sequences that were used to generate existing commercial CFP1 antibodies were also avoided as they fail to effectively recognise native CFP1 (unpublished observation). Sequence alignment of the nearly identical human and mouse CFP1 proteins revealed that the CFP1p sequence is perfectly conserved, suggesting an antibody raised against this antigen may be used in both species.

### 3.3 Characterisation of the α-CFP1p antibody

To determine whether immunisation with CFP1p resulted in the production of α-CFP1 antibodies, whole cell extracts from HeLa cells were probed with serum from the immunised rabbit (Figure 6). As a control, the same extracts were also probed with pre-immune serum. Whole cell extracts were prepared from untransfected HeLa cells, cells transfected with an empty expression vector or from cells expressing flag tagged CFP1. A western blot of these extracts with the CFP1p-immunised serum detected a specific band corresponding to Flag-CFP1. This band was not detected in extracts from cells transfected with the empty vector, or by probing with pre-immune serum suggesting that immunisation with CFP1p successfully stimulated the generation of CFP1 specific antibodies.

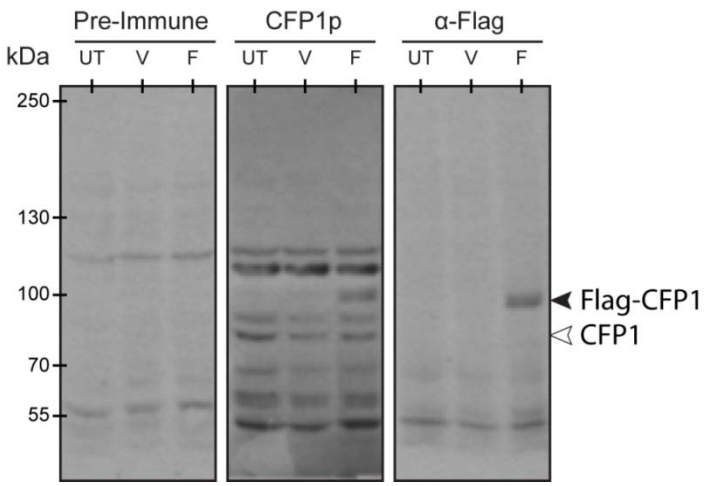



**Figure 6 – Rabbit serum collected after immunisation contains antibodies against the CFP1 protein**
Western blot analysis comparing rabbit serum collected before (Pre-Immune), or after immunisation with KLH-CFP1p (CFP1p). Whole cell extracts were prepared from untransfected cells HeLa cells (UT), cells transfected with an empty expression vector (V), or cells overexpressing Flag tagged CFP1 (F). Staining with α-Flag in parallel confirmed the presence of full length Flag-CFP1 in transiently transfected HeLa cells (F). Flag-CFP1 was also detected by the CFP1p, indicating that the serum contains antibodies capable of binding to CFP1. The expected position of endogenous CFP1 (88 kDa) is indicated by the unfilled arrowhead.

Detection of Flag-CFP1 suggests that the rabbit serum collected after immunisation with CFP1p contains specific antibodies against CFP1. Critically, several additional bands were detected by the immunised serum, which were not observed using the pre-immune. Detection of multiple bands may indicate the presence of non-CFP1 binding antibodies in the polyclonal serum leading to the detection of unrelated proteins. Alternatively the additional bands may result from cross reactivity of the α-CFP1p antibodies with other proteins. To enrich for CFP1 specific antibodies and deplete non-specific or low affinity antibody species, the α-CFP1p serum was affinity purified using the CFP1p antigen.

To purify CFP1 specific antibodies, the CFP1p antigen was immobilised by covalent attachment to agarose beads. The α-CFP1p serum was incubated with the CFP1p-beads overnight to allow binding. Unbound components of the serum were then washed away. Antibodies that remained bound to the CFP1p-beads were then sequentially eluted and analysed by SDS-PAGE (Figure 7A). Coomassie staining revealed the characteristic IgG heavy and light chains. The peak fractions were then pooled and quantified (Figure 7B).

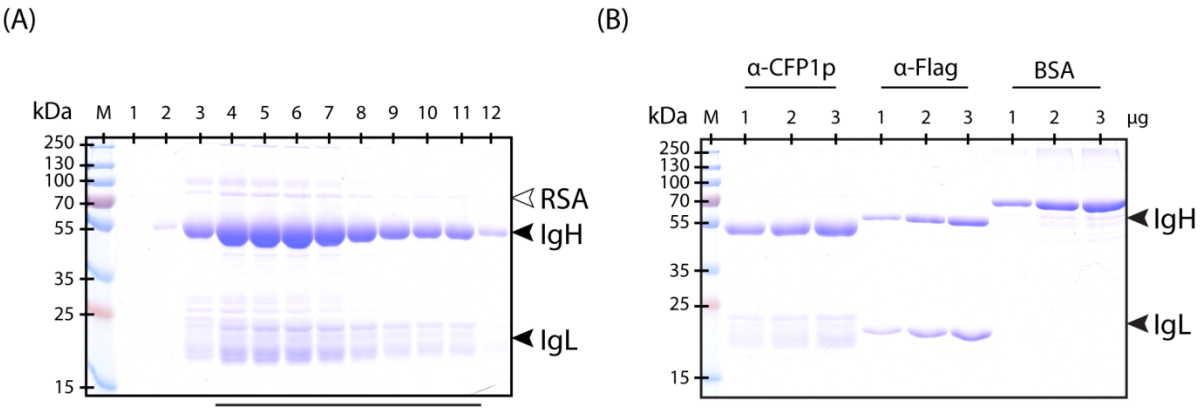



**Figure 7 – Affinity purification of α-CFP1p**

(A) Sequential elutions from CFP1p-beads analysed on a Coomassie stained 12% acrylamide SDS-PAGE gel. M: indicates protein markers and consecutive elutions are numbered 1-12. Fractions 4-11 (underlined) were pooled for quantitation. Filled arrows indicate the heavy (IgH) and light (IgL) chains respectively. Rabbit serum albumin (RSA) is indicated by the unfilled arrow. (B) Quantitation of purified α-CFP1p antibodies via Coomassie stained 12% acrylamide SDS-PAGE gels with standards of known concentration. Purified α-CFP1p was run alongside the α-Flag monoclonal antibody and BSA for quantitative comparison.

The reactivity of α-CFP1p for endogenous CFP1 before and after affinity purification was compared by western blot (Figure 8). Extracts were prepared from human and mouse cell lines to confirm that α-CFP1p recognises its epitope in the highly conserved mouse protein. Importantly, the purified α-CFP1p antibody still detected CFP1, but the reactivity to non-specific bands was noticeably reduced, suggesting that affinity purification resulted in effective depletion of cross reactive antibodies. The flow through, collected after incubation of the serum with the CFP1p-beads, was also analysed in parallel by western blot (Figure 8). Strongly immunoreactive bands are still present in this fraction including two bands observed in the pre-immune serum. Therefore affinity purification using the CFP1p-beads completely depleted this CFP1-independent immunoreactivity yielding a highly specific α-CFP1p antibody reagent.

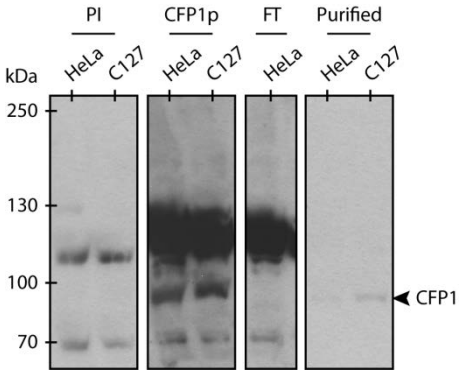



**Figure 8 – Purification of α-CFP1p improves specificity by removing cross reactive antibodies**
Western blot analysis using purified α-CFP1p. Nuclear extracts prepared from human (HeLa) and mouse (C127) cells were probed with rabbit serum before (Pre-immune) or after (Immunised) injection with CFP1p. Purification of the immunised serum was assessed by probing the same extracts with pooled elutions from the CFP1p-beads (Purified) or the unbound flow through (FT). The expected position of endogenous CFP1 (~88 kDa) is indicated by the black arrowhead.

### 3.4 Chromatin immunoprecipitation using the α-CFP1p antibody

ChIP-sequencing of CFP1 in mouse brain tissue suggests that CFP1 binds exclusively to CGI regions of the genome (Thomson et al. 2010), consistent with several other zf-CXXC proteins (Farcas et al. 2012; Blackledge et al. 2010). Attempts by our lab to analyse CFP1 localisation with the commercial antibody used in this original study have failed. Initial characterisation of the α-CFP1p antibody suggests it is able to recognise denatured CFP1 after separation by SDS-PAGE. However, for α-CFP1p to function well in ChIP experiments, it must also bind specifically to native CFP1.

To determine its ability to recognise native CFP1, α-CFP1p was used to immunoprecipitate sonicated chromatin prepared from v6.5 mouse ES cells (Figure 9). Immunoprecipitated DNA was then isolated for analysis by quantitative PCR (q-PCR). As CFP1 binds to CGIs, amplicons were chosen to quantify immunoprecipitated DNA at two CGI promoters, Suv420h1 and NcoA2. Control amplicons were designed at a non-CGI promoter, Fabp7, and within the Suv420h1 gene body. CFP1 enrichment at the CGI promoters was then estimated relative to the Su420h1 gene body.

Chromatin immunoprecipitation (ChIP) using the α-CFP1p antibody revealed a ~5-fold enrichment of CFP1 at the CGI promoters, however no enrichment was observed at the non-CGI Fabp7 promoter (Figure 9). These results are consistent with the proposal that CFP1 binds specifically to CGI chromatin (Lee et al. 2001; Thomson et al. 2010; Xu et al. 2011), and suggest that the α-CFP1p antibody is able to recognise native CFP1.

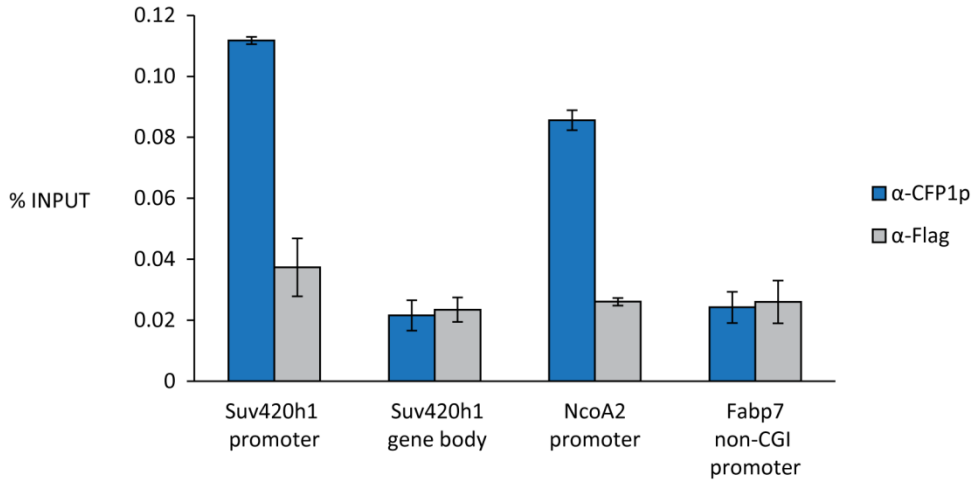



**Figure 9 – CFP1 is enriched at CpG island promoters**

ChIP-qPCR analysis of CFP1 using α-CFP1p. Chromatin prepared from v6.5 mouse ES cells was immunoprecipitated with affinity purified α-CFP1p or with α-Flag as a negative control. The percentage of input DNA recovered was quantified at four loci by qPCR. The means of three biological replicates are shown with standard error bars. qPCR amplicons were designed by Neil Blackledge and ChIP-qPCR was performed by Anca Farcas.

Although α-CFP1p appears to recognise CFP1 under ChIP conditions, and indicates that CFP1 is bound to CGIs, empirical evidence from previous work in the Klose lab has shown that ChIP signal enrichments lower than ~10-fold are generally uninterpretable by ChIP- sequencing. Therefore, the ~5-fold enrichment observed with the α-CFP1p antibody may not be suitable for investigating the binding of CFP1 genome-wide. The weak enrichment of CFP1 could reflect a low abundance of the

protein at these regions, or stem from inefficient or low affinity binding by the antibody. To try to produce an antibody more suitable for ChIP-seq a second CFP1 antigen was designed.

### 3.5 Generation of a second CFP1 specific antibody using a fragment antigen

Although α-CFP1p can detect denatured CFP1 by western blot, immunoprecipitation requires recognition of the native protein. In the native state certain epitopes may be altered, buried within the folded protein, or occluded by protein-protein interactions. If the CFP1p epitope is altered, buried, or occluded this may explain the weak enrichment seen by ChIP-qPCR.

Immunisation with a longer fragment of CFP1, could present the immune system with a greater variety of epitopes, increasing the chances of generating specific antibodies which recognise the native protein. Therefore a second antigen was designed corresponding to a 155 amino acid fragment of the CFP1 sequence (Figure 10). As for CFP1p, highly conserved regions were avoided when selecting the CFP1f sequence. A BLAST search of mouse proteins using the CFP1f sequence returned no closely related sequences, and the predicted solubility and antigenicity scores were above average (Figure 10B-C). Due to its length, the fragment antigen (CFP1f) was not synthesised, but was expressed in *E. coli* and then purified for immunisation.

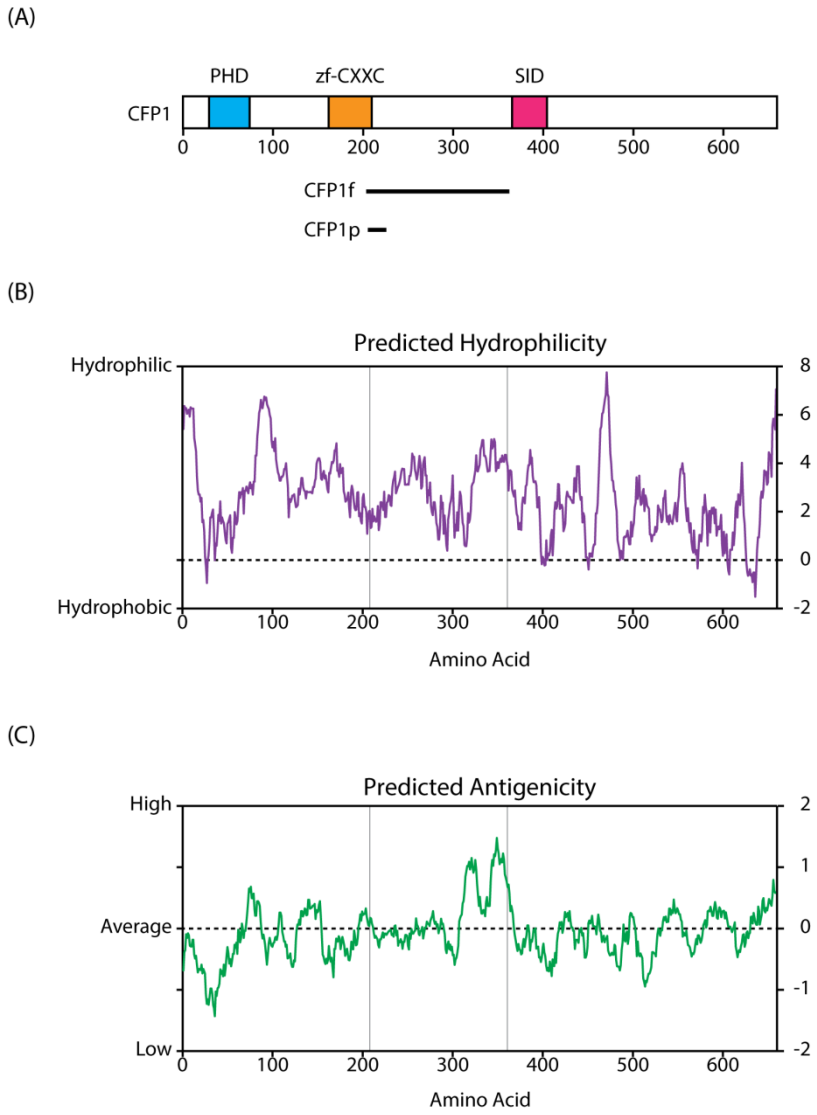



**Figure 10 –The CFP1 fragment antigen**

(A) Schematic of the human CFP1 protein. The CFP1 sequence chosen for expression as a fragment (CFP1f) is underlined, with the sequence of the CFP1 peptide antigen (CFP1p) for comparison. (B) The mean score for the predicted hydrophilicity of each residue is shown for a 15 amino acid sliding window (Parker et al. 1986). (C) Mean antigenicity is shown for the same window (Welling et al. 1985). A predicted antigenicity of 0 represents average antigenicity.

The CFP1f sequence was cloned into an IPTG-inducible, bacterial, expression vector introducing a 6xHis tag for nickel-NTA purification. Expression of a glutathione S-transferase (GST) tagged version was also tested in parallel (Supplementary Figure 1). Both constructs expressed well, implying that

the CFP1 fragment is soluble. For immunisation the His tagged fragment was expressed without GST so that the antigen was almost entirely comprised of CFP1 (Figure 11A).

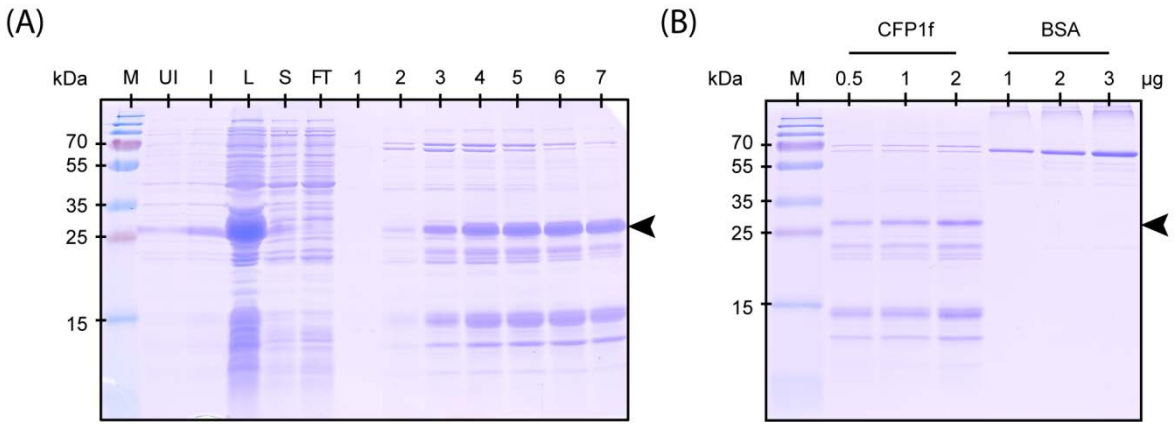



**Figure 11 – Expression and purification of a CFP1 fragment**

Expression and purification of the CFP1f antigen was analysed by SDS-PAGE. (A) 6His-CFP1f was expressed in *E. coli* and purified using a Ni-NTA column. Samples were resolved on an 8% polyacrylamide gel and visualised by staining with Coomassie Brilliant Blue. M: Protein Markers, UI: Uninduced sample before addition of IPTG, I: IPTG induced sample, L: Lysate after sonication, S: Soluble supernatant after centrifugation of lysate, FT: Flow through from the Ni-NTA resin column, 1-7: consecutive elutions from the Ni-NTA column. The arrowhead indicates an IPTG induced protein band of the expected size (~25 kDa). (B) Quantitation of purified 6His-CFP1f for immunisation.

### 3.6 Characterisation of the α-CFP1f antibody

To determine whether CFP1f immunisation resulted in the production of α-CFP1 antibodies, nuclear extracts were probed with serum collected before and after immunisation (Figure 12). A major band corresponding to the expected position of endogenous CFP1 is visible on the membrane probed with immunised serum but absent using the pre-immune serum. This suggests that immunisation with CFP1f successfully stimulated the production of α-CFP1 antibodies.

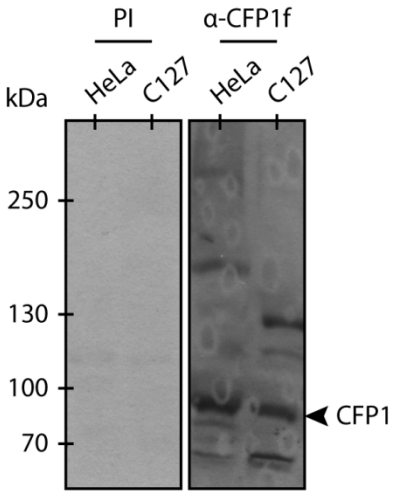



**Figure 12 – Serum collected after immunisation with CFP1f contains antibodies against CFP1**
Western blot analysis of nuclear extract comparing pre-immune (PI) or CFP1f immunised (α-CFP1f) rabbit serum. Nuclear extracts prepared from human (HeLa) and mouse (C127) cells. The expected position of endogenous CFP1 (~88 kDa) is indicated by the black arrowhead. The immunised serum detected a band of the expected size in both nuclear extracts suggesting it contains antibodies which recognise endogenous CFP1.

In addition to detecting CFP1, the immunised serum detected several additional bands not seen in the pre-immune blot (Figure 12). To enrich for CFP1 specific antibodies, the α-CFP1f serum was affinity purified using the CFP1f antigen. To affinity purify CFP1 specific antibodies, the CFP1f antigen was covalently attached to a resin bead matrix. The α-CFP1f serum was then incubated with the CFP1f –beads. After several wash steps, to remove unbound components of the serum, antibodies which remained bound were sequentially eluted and analysed by SDS-PAGE (Figure 13A).

The elution profile of α-CFP1f was broader than for α-CFP1p. The concentration of protein in the first sixteen fractions was quantified using Bradford reagent. As the concentration remained high even in the 16th elution, a further 10 fractions were collected and analysed (the first seven of these 17-23 are shown in Figure 13A). Reasoning that the population of antibodies which eluted early may differ

from the population which eluted late, early and late eluting fractions were pooled separately for analysis (Figure 13B).

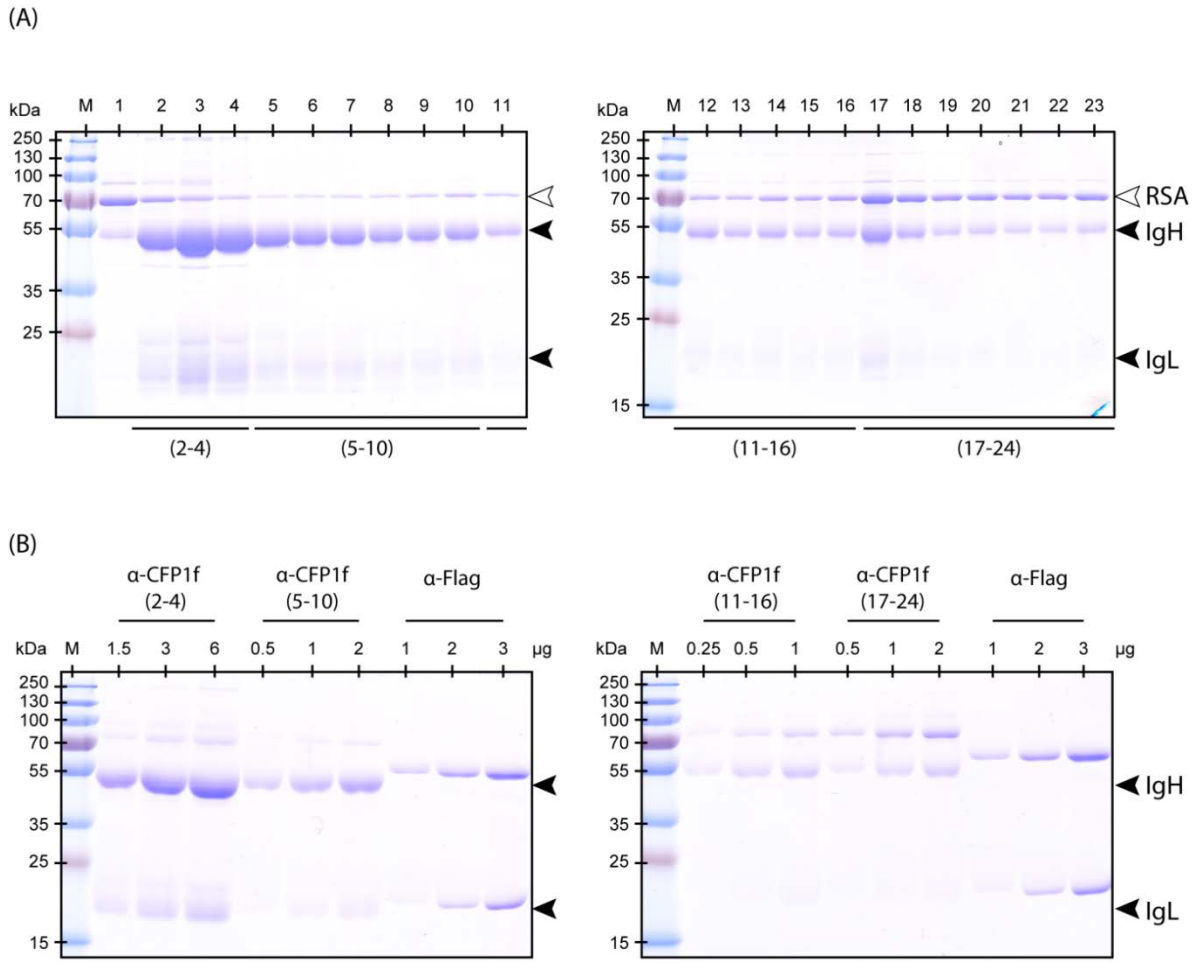



**Figure 13 – Affinity purification of α-CFP1f**

(A) Sequential elutions from CFP1f-beads were analysed by SDS-PAGE. Samples were resolved on a 12% polyacrylamide gel and visualised by staining with Coomassie Brilliant Blue. M: Protein Markers, 1-24: consecutive elutions from the CFP1f-bead matrix. Filled arrows indicate the heavy (IgH) and light (IgL) chains respectively. Rabbit serum albumin (RSA) is indicated by the unfilled arrow. (B) Quantification of pooled elutions.

The affinity purified α-CFP1f antibody was first assessed by western blot. To compare the sensitivity and specificity of pooled α-CFP1f elutions, nuclear extracts were probed with equal concentrations

of the early (2-4) or late (17-24) elutions (Figure 14). In parallel, the same extracts were probed with an equal concentration of affinity purified α-CFP1p, allowing a direct comparison of the antibodies produced by each immunisation.

An ~88 kDa band corresponding to endogenous CFP1 was detected by α-CFP1p, and α-CFP1f, but was more intense in α-CFP1f blots, possibly reflecting a higher affinity, or a greater polyclonality generated by immunisation with the fragment antigen. Compared to the early α-CFP1f elutions, late elutions detected fewer bands and only a slight reduction in intensity suggesting a higher degree of specificity in the later elutions (Figure 14B). For this reason the late α-CFP1f elutions were chosen for subsequent experiments and are from here on referred to simply as α-CFP1f.

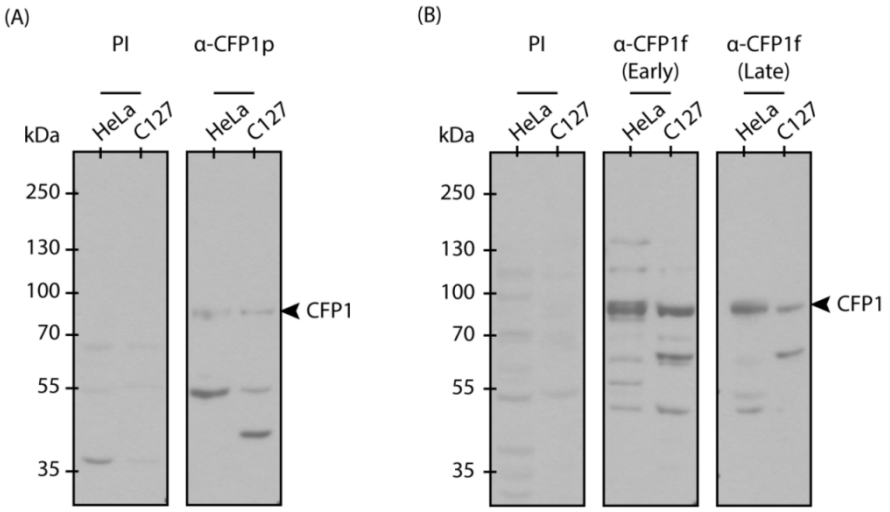



**Figure 14 – Comparison of affinity purified CFP1 antibodies**

(A) α-CFP1p and (B) α-CFP1f antibodies were compared by western blot analysis. α-CFP1p was raised against a peptide antigen, while α-CFP1f was raised against a 25 kDa fragment. Nuclear extract (50 µg) from HeLa or C127 cells was resolved on an 8% polyacrylamide gel and probed using pre-immune (PI) serum, or affinity purified antibody (1 ng/ml final concentration). Early and late elutions from the α-CFP1f affinity purification were tested in parallel. The expected position of endogenous CFP1 (~88 kDa) is indicated by a black arrowhead.

### 3.7 Chromatin immunoprecipitation of endogenous CFP1 using α-CFP1f

Western blot analysis suggests that α-CFP1p and α-CFP1f are both capable of binding specifically to endogenous CFP1 in its denatured state. Despite its apparent specificity for CFP1, initial attempts to study CFP1 bound to chromatin suggest that the α-CFP1p antibody performs poorly in ChIP experiments. Therefore, to investigate chromatin bound CFP1, ChIP experiments were carried out using the α-CFP1f antibody (Figure 15). Chromatin was prepared from C127 cells and immunoprecipitated using α-CFP1f, or α-GFP as a negative control. Immunoprecipitated DNA was then quantified by qPCR.

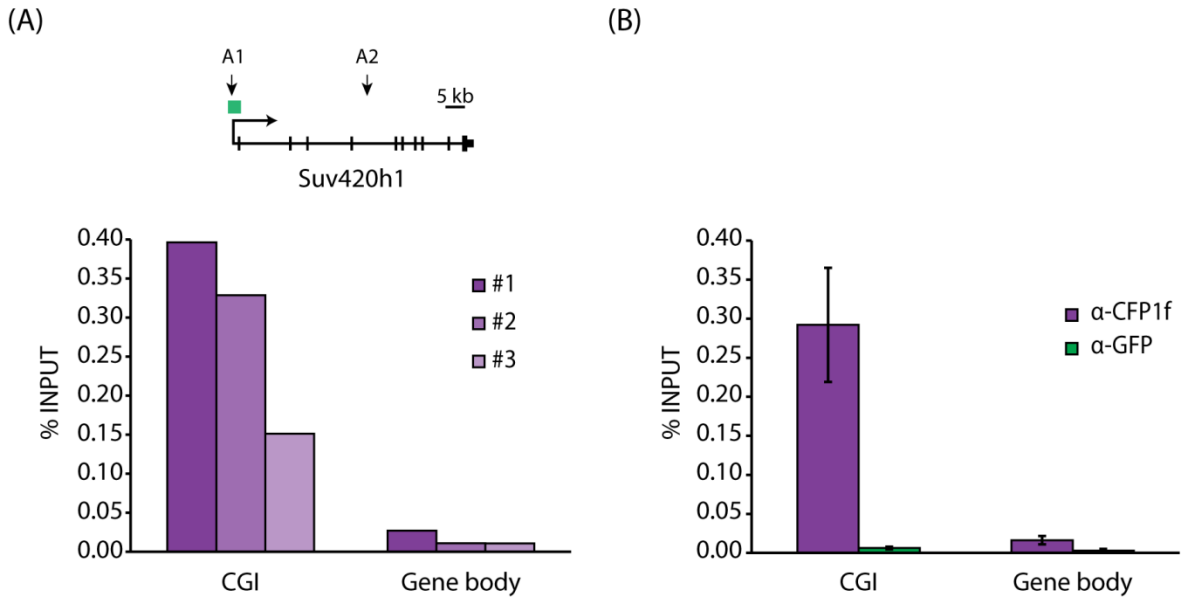



**Figure 15 – Chromatin immunoprecipitation using the α-CFP1f antibody**

ChIP-qPCR analysis of CFP1 using α-CFP1f. Chromatin prepared from C127 cells was immunoprecipitated using α-CFP1f. Immunoprecipitated DNA was then quantified at two amplicons within the Suv420h1 locus, (A1) within the CGI found at the promoter region, and (A2) within the body of the gene as a negative control. The position of each amplicon is shown on the schematic above the chart. The CpG island region is shown in green below the schematic. (A) ChIP signal from biological triplicates (#1-3). (B) Mean ChIP signal for α-CFP1f or α-GFP as a negative control. Bars

indicate the standard error of biological triplicate experiments. Theα-CFP1f ChIP signal is ~20-fold higher in the CGI region than within the gene body.

As CFP1 binds to CGIs, immunoprecipitated DNA was quantified at a CGI promoter, and at a non-CGI control region within the body of the gene. Based on biological triplicate experiments, CFP1 levels were ~20-fold higher in the CGI region than within the gene body (Figure 15A). CFP1 levels in the gene body were close to the levels of the α-GFP negative control (Figure 15B). These results are consistent with the specific binding of CFP1 to CGI regions and indicate that α-CFP1f is able to recognise native CFP1 under the conditions used for chromatin immunoprecipitation.

Although ChIP-signal at the CGI region was more strongly enriched using α-CFP1f than with α-CFP1p, the two antibodies were not directly compared, as the α-CFP1p ChIP experiments were performed in a different cell line. Importantly, based on previous work in the lab to characterise other zf-CXXC proteins, the (>10-fold) enrichment observed using α-CFP1f is sufficient to suggest this antibody is suitable for investigation of CFP1 binding genome-wide by ChIP-sequencing.

### 3.8 Immunofluorescence imaging of endogenous CFP1

Initial ChIP experiments suggest that α-CFP1p and α-CFP1f antibodies are both able to recognise native CFP1. CGI chromatin was more strongly enriched using the α-CFP1f antibody, perhaps indicating that this antibody bound to CFP1 more effectively than α-CFP1p. To compare their ability to bind native CFP1 inside the nucleus, α-CFP1f and α-CFP1p antibodies were used to stain CFP1 for imaging by immunofluorescence microscopy (Figure 16).

The α-CFP1p antibody preferentially stains nuclei, which were visualised by counterstaining with DAPI (Figure 16A). Staining with α-CFP1f was also predominantly nuclear, but lower concentrations were required to achieve a clear signal (0.5 µg/ml compared to 5 µg/ml for α-CFP1p) (Figure 16B). Raising the concentration of α-CFP1f to 1 µg/ml increased nuclear staining towards saturating levels,

with minimal staining of the cytoplasm. In contrast, an equal concentration of a commercial α-CFP1 antibody (Sigma SAB2103447) stained both nuclei and cytoplasm. These results suggest that the α-CFP1f antibody is sufficiently sensitive and specific to detect native CFP1 inside the nucleus.

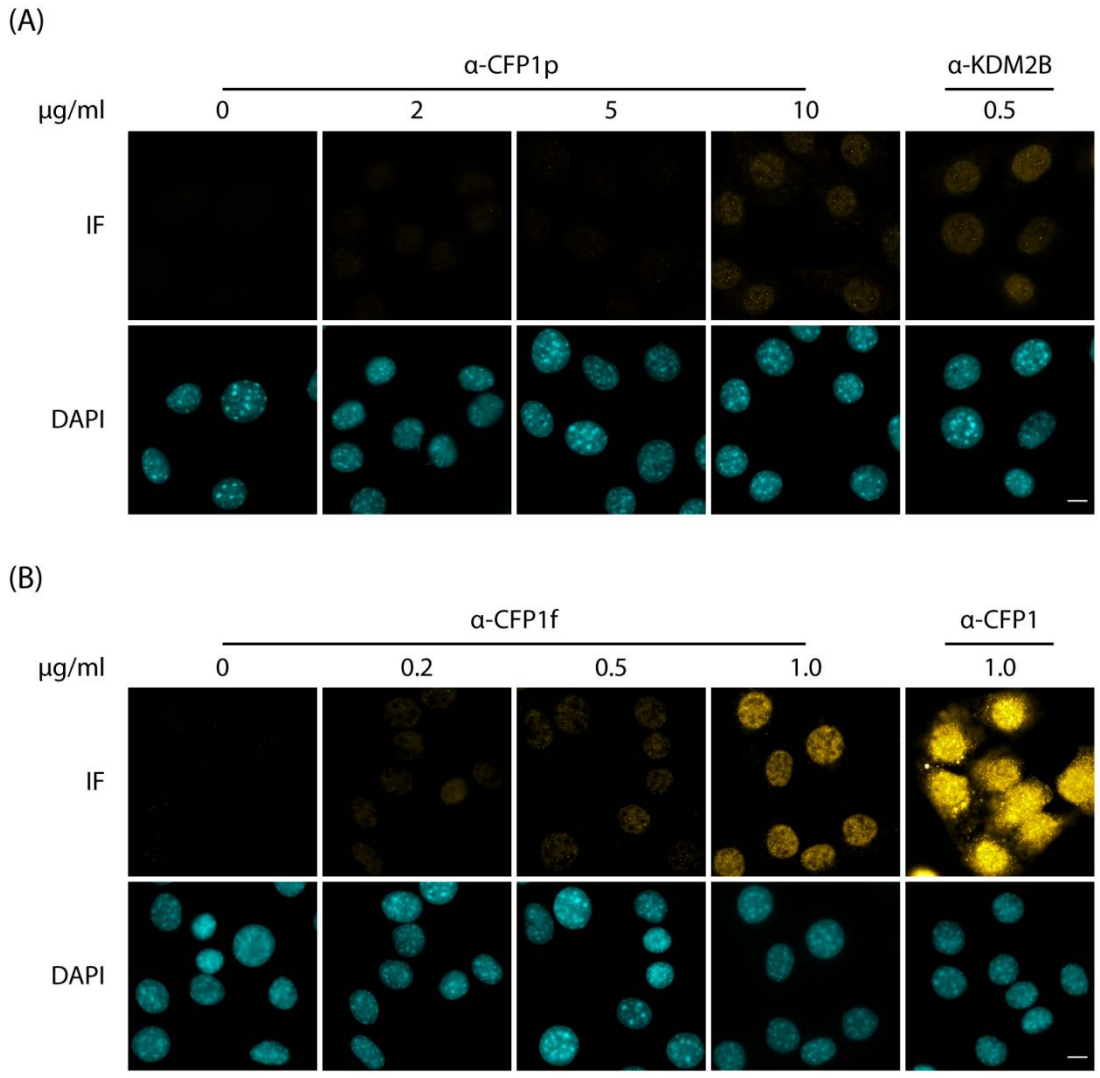



**Figure 16 – Immunofluorescence imaging of endogenous CFP1**

C127 cells were stained with two antibodies raised against CFP1. α-CFP1p was raised against a 16 aa peptide, while α-CFP1f was raised against a 25 kDa polypeptide fragment. Nuclei were counterstained with DAPI (below). (A) Titration of α-CFP1p (0, 2, 5, 10 µg/ml) with 0.5 µg/ml KDM2B as a positive control. (B) Titration of α-CFP1f (0, 0.2, 0.5, 1 µg/ml) with a commercial α-CFP1 antibody (1 µg/ml, Sigma SAB2103447) for comparison. Scale bars in the bottom right images represent 5 µm.

## 3.9 Discussion

CGIs are a conspicuous feature of vertebrate genomes. As well as being recognisable at the level of DNA sequence, CGIs are resistant to DNA methylation, and are often further distinguished from the surrounding chromatin by specific histone modifications. Recent work in the Klose lab has demonstrated that the zf-CXXC protein KDM2A directly contributes to the unique chromatin architecture of CGIs, by removing H3K36 methylation (Blackledge et al. 2010). Similarly KDM2B was shown to direct the mono-ubiquitylation of histone H2A at CGI regions through its interaction with a variant PRC1 complex (Farcas et al. 2012). These examples illustrate the ability of CGI sequences to orchestrate specific chromatin environments through the recruitment of zf-CXXC proteins.

CFP1 belongs to the zf-CXXC protein family, and is thought to direct the placement of H3K4me3 at CGIs through its interaction with the SET1 (Clouaire et al. 2012; Wang et al. 2008; Birney et al. 2007; Thomson et al. 2010). Although its potential role in H3K4me3 placement is particularly intriguing, systematic study of CFP1 function at CGIs has been limited by the lack of a robust antibody reagent. This chapter describes two attempts to overcome this limitation, raising and characterising rabbit polyclonal antibodies against two different CFP1 antigens. Immunisation with a peptide antigen yielded a specific antibody against CFP1 (α-CFP1p) which was suitable for western blot analysis and ChIP-qPCR, but not for IF or ChIP-sequencing. Using a larger fragment antigen, a second specific antibody was produced (α-CFP1f) which performed much better in IF experiments. ChIP-qPCR experiments suggest that the α-CFP1f may also be suitable for investigation of CFP1 localisation genome-wide by ChIP-sequencing which is the subject of the next Chapter.

# Chapter 4 - Exploring the genome-wide localisation of CFP1 by ChIP-Sequencing

## 4.1 Introduction

CFP1 is essential for mouse embryogenesis (Carlone & Skalnik 2001) and has been proposed to contribute to normal development by influencing chromatin modification and gene regulation. The CFP1 protein contains a zinc finger CXXC (zf-CXXC) domain which binds specifically to DNA containing non-methylated CpG dinucleotides (Lee et al. 2001). CFP1 also interacts with histone lysine methyltransferase enzymes SET1A and SET1B which specifically methylate lysine 4 on the histone H3 tail (Lee & Skalnik 2005; J.-H. Lee et al. 2007). Based on these findings CFP1 has been proposed to direct H3K4 methylation to CpG island regions (CGIs) which contain a high density of non-methylated CpGs (Tate et al. 2010; Thomson et al. 2010; Clouaire et al. 2012).

Previous attempts to study the role of CFP1 in H3K4 methylation have been limited by the failure of existing antibodies to effectively recognise CFP1 under ChIP conditions (unpublished observations, Klose Lab), preventing investigation of CFP1 localisation genome-wide. Chapter 3 describes the production and characterisation of an antibody that appears to work well in ChIP experiments (α-CFP1f). To understand how CFP1 interacts with chromatin throughout the genome, its localisation was investigated using this new reagent.

## 4.2 ChIP sequencing of CFP1 using the α-CFP1f antibody

Initial ChIP-qPCR experiments suggested that like other zf-CXXC proteins, CFP1 is enriched at CGIs however such analysis was limited to a few candidate loci. To provide a genome-wide profile of its interactions with chromatin and enable a comprehensive analysis of CFP1 localisation, ChIP-sequencing was performed using the α-CFP1f antibody. Chromatin was prepared from mouse C127 cells (a transformed epithelial cell line) and immunoprecipitated using the α-CFP1f antibody. ChIP

DNA from two biological replicates was submitted for library preparation and sequencing by the Oxford Genomics Centre (Wellcome Trust Centre for Human Genetics) using the Illumina HiSeq2000 platform. Over 25 million sequencing reads were obtained per replicate of which approximately half represented unique unambiguous alignments to the mouse genome (Table 13). To investigate CFP1 localisation, genome-wide profiles were generated for each replicate from the uniquely aligned reads.

**Table 11 – Alignment of CFP1 ChIP-sequencing reads**

| **α-CFP1f ChIP (Total reads)** | **Rep. 1 (32,309,519)** | **Rep. 2 (25,385,036)** | **Total (57,694,555)** |
|---|---|---|---|
| Unmapped | 9% | 12% | 10% |
| Multiply mapped | 29% | 31% | 30% |
| Total mapped | 62% | 57% | 60% |
| PCR duplicates | 7% | 7% | 7% |
| Uniquely aligned reads | 55% (17,905,594) | 50% (12,622,999) | 53% (30,528,593) |

## 4.3 The CFP1 ChIP-seq profile is punctate, consisting of short regions of enrichment against a low background

Visual inspection of the replicate profiles revealed punctate peaks of enrichment (~0.3-1 kb) against a low, uniform background, consistent with specific binding to discreet regions of the genome (Figure 17). Peaks were observed in the same locations and at the same relative intensities in both biological replicates supporting the reproducibility of the α-CFP1f ChIP-sequencing (Figure 17A). To understand how these sites of CFP1 enrichment are distributed across the genome, a peak set was generated for each biological replicate using the MACS algorithm (Zhang et al. 2008). The resulting peak sets contain coordinates for every CFP1 peak above a probability threshold ($p<10^{-5}$). Only peaks detected in both biological replicates were considered for further analysis. The majority of CFP1 peaks met this criterion with 67%, (5,654/8,431) of peaks detected in both replicates (Figure 17B).

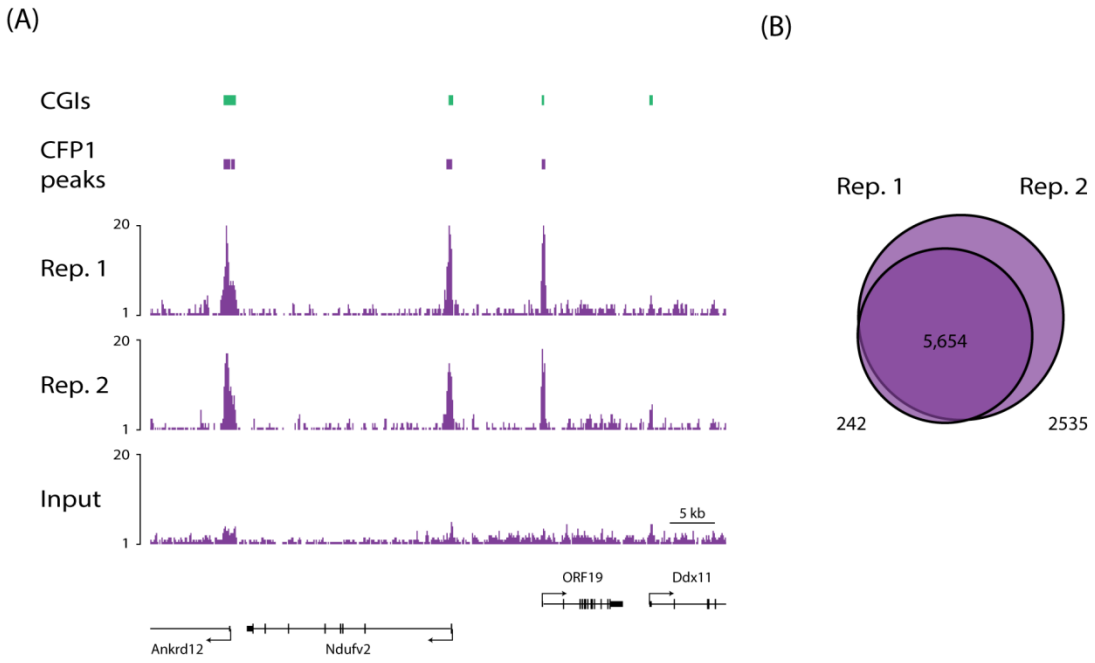



**Figure 17 - CFP1 interacts preferentially with short regions of the genome producing punctate peaks of enrichment**

(A) The CFP1 ChIP-sequencing profiles for biological replicates are shown with input DNA (Input) as a negative control. Genes are shown below the profiles with arrows indicating the direction of transcription. Algorithm predicted CpG islands (green) and CFP1 peaks identified by MACS (purple) are indicated above the profiles. (B) The genome-wide overlap between peaks detected in each replicate is depicted as a proportional Venn diagram.

## 4.4 Peaks of CFP1 enrichment mark a minority of CpG island regions

CFP1 has previously been proposed to bind virtually all CGIs in mouse brain tissue (Thomson et al. 2010). To determine the relationship between CFP1 and CGIs in C127 cells, the CFP1 peak set was compared to publicly available coordinates for each CGI in the mouse genome (Figure 18A and B). As expected for a zf-CXXC protein, over 90% of CFP1 peaks correspond to a CGI (Figure 18C). However, CFP1 peaks were only detected at 33% of CGI (Figure 18C), suggesting a difference between CGIs with and without CFP1. A comparison of CpG density across CGIs with or without CFP1 peaks suggested that some CGIs are not recognised by CFP1 despite having a high density of CpGs (Figure 18D).

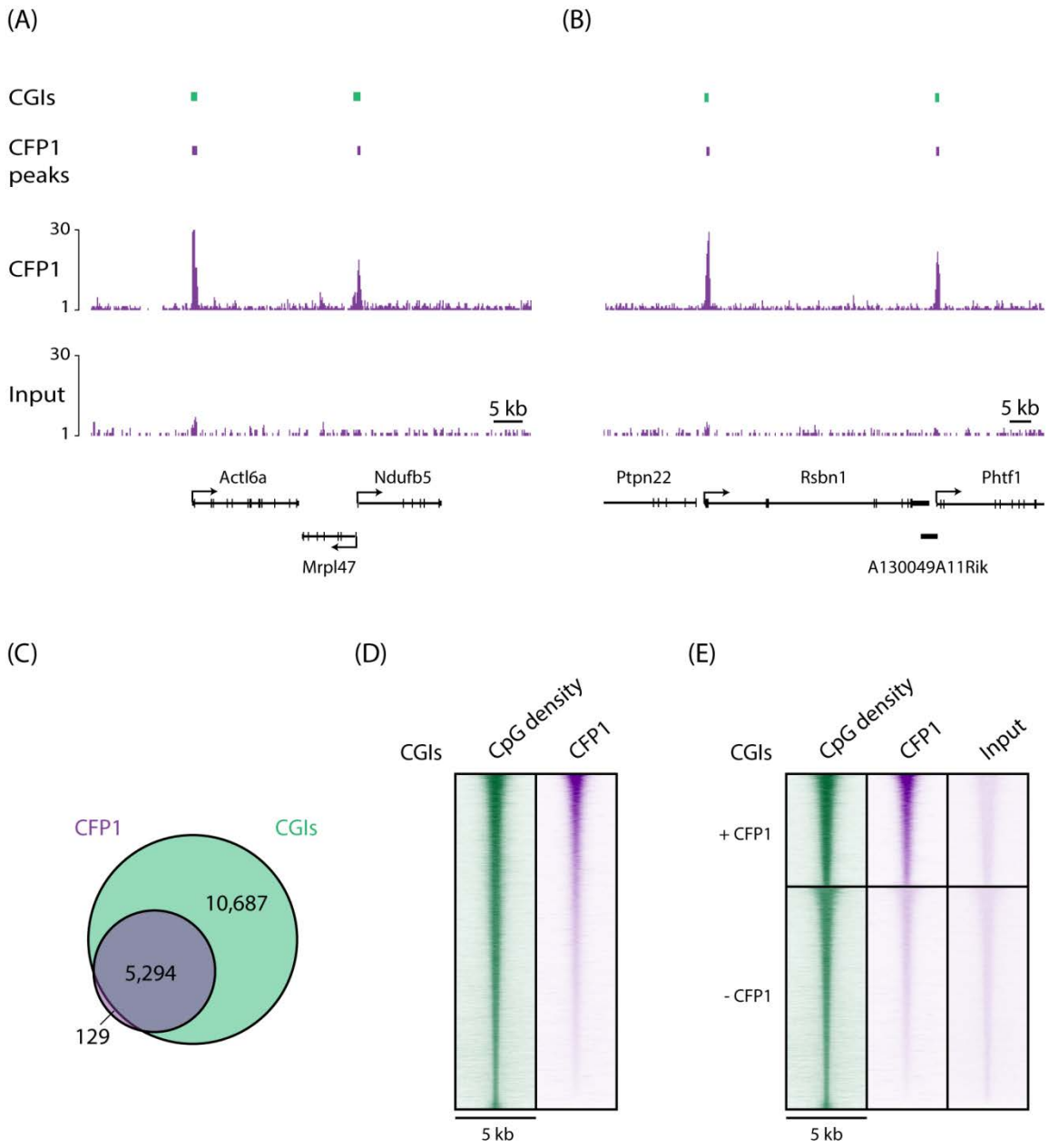



**Figure 18 – CFP1 interacts with minority of CpG islands**

(A and B) Snap shots from the UCSC genome browser showing the profile of CFP1 enrichment alongside input DNA (Input) as a negative control. Genes are shown below the profiles. Algorithm predicted CpG islands (CGIs) are indicated above the profiles (green). CFP1 peaks based on the profile are also shown (purple). (C) The genome-wide overlap of CFP1 peaks and predicted CGIs is depicted as a proportional Venn diagram. (D and E) Heat maps illustrate variation in CpG density or CFP1 enrichment across all CGIs ranked by CFP1 reads. In (E), CGIs have been divided into those which correspond to a detected peak of CFP1 (+ CFP1) and those that do not (CFP1 -). Enrichment of input chromatin is shown as a negative control.

## 4.5 Profiling of non-methylated DNA reveals that 40% of CGIs are methylated in C127 cells

CGIs were originally mapped using an algorithm to detect GC-rich regions with a high CpG density (Gardiner-Garden & Frommer 1987). Although CGIs are generally resistant to CpG methylation, the methylation status of individual CpG islands can vary between different cell types (Illingworth et al. 2008; Deaton et al. 2011; Long, Sims, et al. 2013). As the CFP1 zf-CXXC domain is unable to bind methylated DNA, CGI methylation may affect the genome-wide distribution of CFP1 by reducing the number of non-methylated CGIs available.

To understand why some CGIs were not recognised by CFP1 despite a high CpG density bioCAP-sequencing was performed to determine the profile of non-methylated DNA genome-wide. This technique was previously developed by the Bird and Klose labs (Illingworth et al. 2008; Blackledge et al. 2012) and utilises a biotinylated zf-CXXC domain to affinity purify DNA containing non-methylated CpGs. Genomic DNA was extracted from C127 cells, sheared by sonication, and then affinity purified by bioCAP. DNA fragments captured by the bioCAP protocol were submitted for library preparation followed by Illumina sequencing (performed by the WTCHG). The bioCAP purification from the sonicated DNA was performed by Hannah Long.

Alignment of bioCAP reads to the mouse genome produced a genome-wide profile of non-methylated DNA in C127 cells (Figure 19A). Using this profile, non-methylated islands (NMIs) were defined independently of algorithm predicted CGIs by detecting peaks in the bioCAP profile (Figure 19A). CGI and NMI peak sets were then compared revealing that only 60% of CGIs were recognised as NMIs in C127 cells (9,505/16,026) implying the remaining 40% are methylated in this cell line (Figure 19B). This implication was supported by a comparison of bioCAP signal across all predicted CGIs (Figure 19C).

Importantly, BioCAP-sequencing of the C127 cell line identified over 4,000 NMIs (70%) that did not correspond to a predicted CGI (Figure 19B). Previous work in the Klose lab determined that 50-60% of NMIs identified in mouse testes were not predicted by the existing CGI algorithm (Long, Sims, et al. 2013). To verify the presence and extent of non-methylated DNA away from predicted CGIs, bioCAP signal was compared across all detected NMIs (Figure 19D). Although enrichment of bio-CAP signal was lower at non-CGI NMIs than at predicted CGIs, the level of bioCAP signal across non-CGI NMIs was clearly higher than the background level in the input control. CGI prediction algorithms do not predict NMIs which fall below a threshold CpG density (Stadler et al. 2011). Nevertheless these regions represent a non-negligible fraction of the cell's non-methylated DNA, demonstrating the importance of profiling non-methylated DNA experimentally for investigation of zf-CXXC protein localisation.

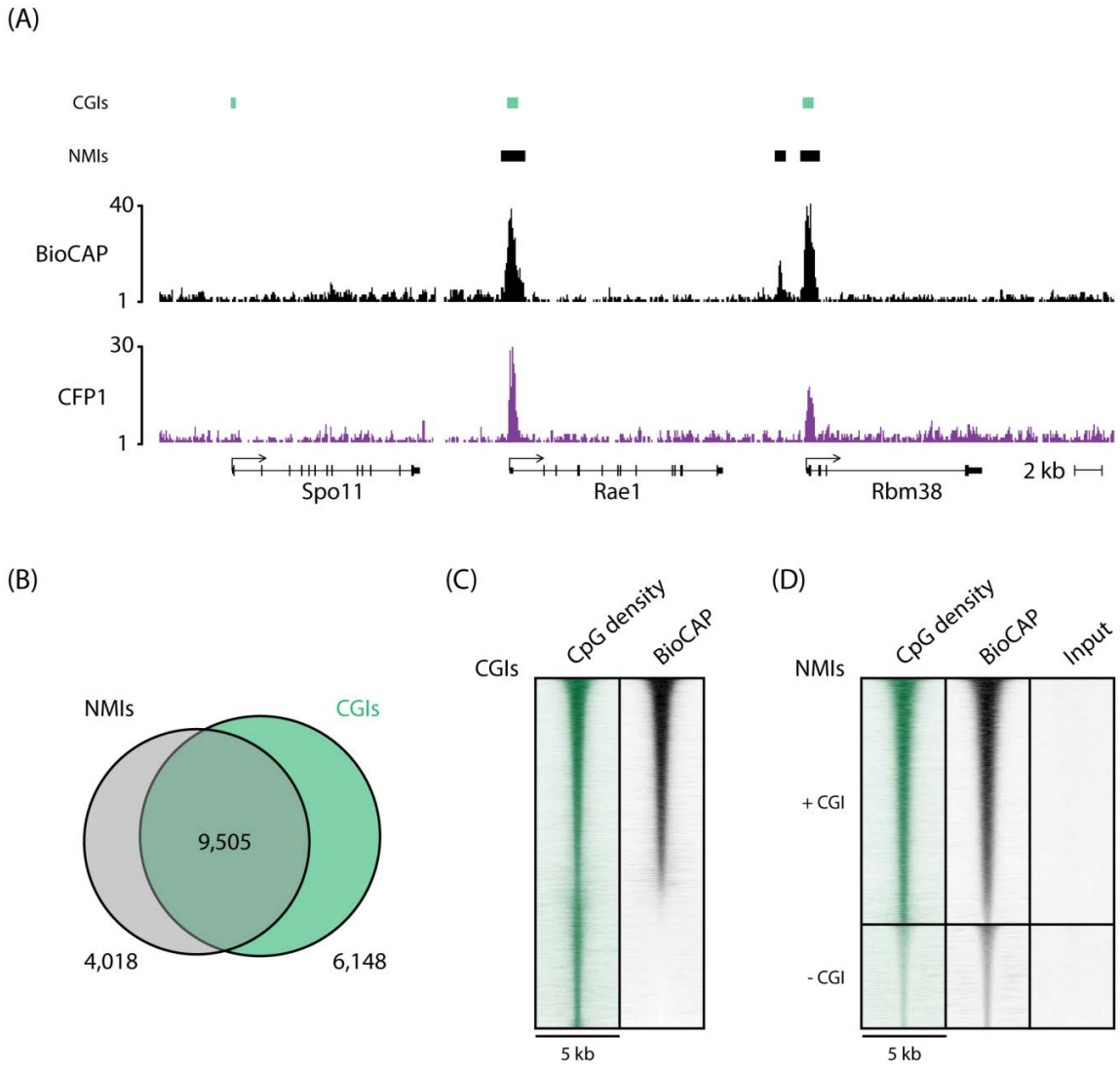



**Figure 19 – In C127s, non-methylated DNA is found at just 60% of computationally predicted CGIs**

(A) Profiles for non-methylated DNA (BioCAP) and CFP1 enrichment are shown across a region containing three protein coding genes (below the profiles, arrows indicate direction of transcription). Computationally predicted CpG islands (CGIs) are indicated above the profiles (green). Non-methylated islands (NMIs) based on the bioCAP profile are also shown (black). The scale is given in the bottom right corner. (B) The genome-wide overlap between predicted CGIs and experimentally determined NMIs is depicted as a proportional Venn diagram. (C) Heat maps illustrate variation in CpG density and bioCAP signal across all CGIs ranked by the bioCAP read count. (D) Heat maps illustrate variation in CpG density, bioCAP signal and bioCAP input DNA as a negative control, across all NMIs. NMIs which corresponded to a predicted CGIs (+ CGI) were separated from those that did not (- CGI). Both groups were then ranked by bioCAP signal.

### 4.6 Unlike other zf-CXXC proteins, CFP1 binds to a minority of non-methylated islands

Unexpectedly, inspection of the CFP1 profile suggests that it does not bind all NMIs despite the presence of non-methylated CpGs (Figure 20A). This observation is supported by the clear discrepancy between the fraction of CGIs that are non-methylated (60%), and the fraction containing a CFP1 peak (33%) (Figure 18C and Figure 19B). Consistent with this discrepancy, CFP1 peaks were only detected at 40% of NMIs genome-wide (5,358/13,572) (Figure 20B).

The suggestion that CFP1 interacts with a restricted set of NMIs contrasts with previous work profiling other zf-CXXC proteins such as KDM2A and KDM2B (Farcas et al. 2012; Zhou et al. 2012; Blackledge et al. 2010). Both KDM2A and KDM2B were shown to bind NMIs throughout the genome. Importantly, these experiments were performed in different cell lines. Therefore it was not clear whether binding to a limited subset of NMIs might extend to other zf-CXXC proteins in C127 cells, or may distinguish CFP1 from other proteins in this family. To allow a comparison of the two zf-CXXC proteins KDM2B was profiled in the same cell line (Figure 20A).

Visual inspection of the KDM2B profile in C127 cells revealed clear peaks of enrichment at of NMI regions (Figure 20A). Consistent with previous studies in other cell lines (Farcas et al. 2012), KDM2B peaks were detected at the majority 85% of NMIs genome-wide suggesting that the limited binding of CFP1 is not shared by all zf-CXXC proteins (Figure 20C). Comparison of the regions enriched with each zf-CXXC protein, revealed that KDM2B is present at nearly all CFP1 peaks while CFP1 is only found at a subset of KDM2B peaks.

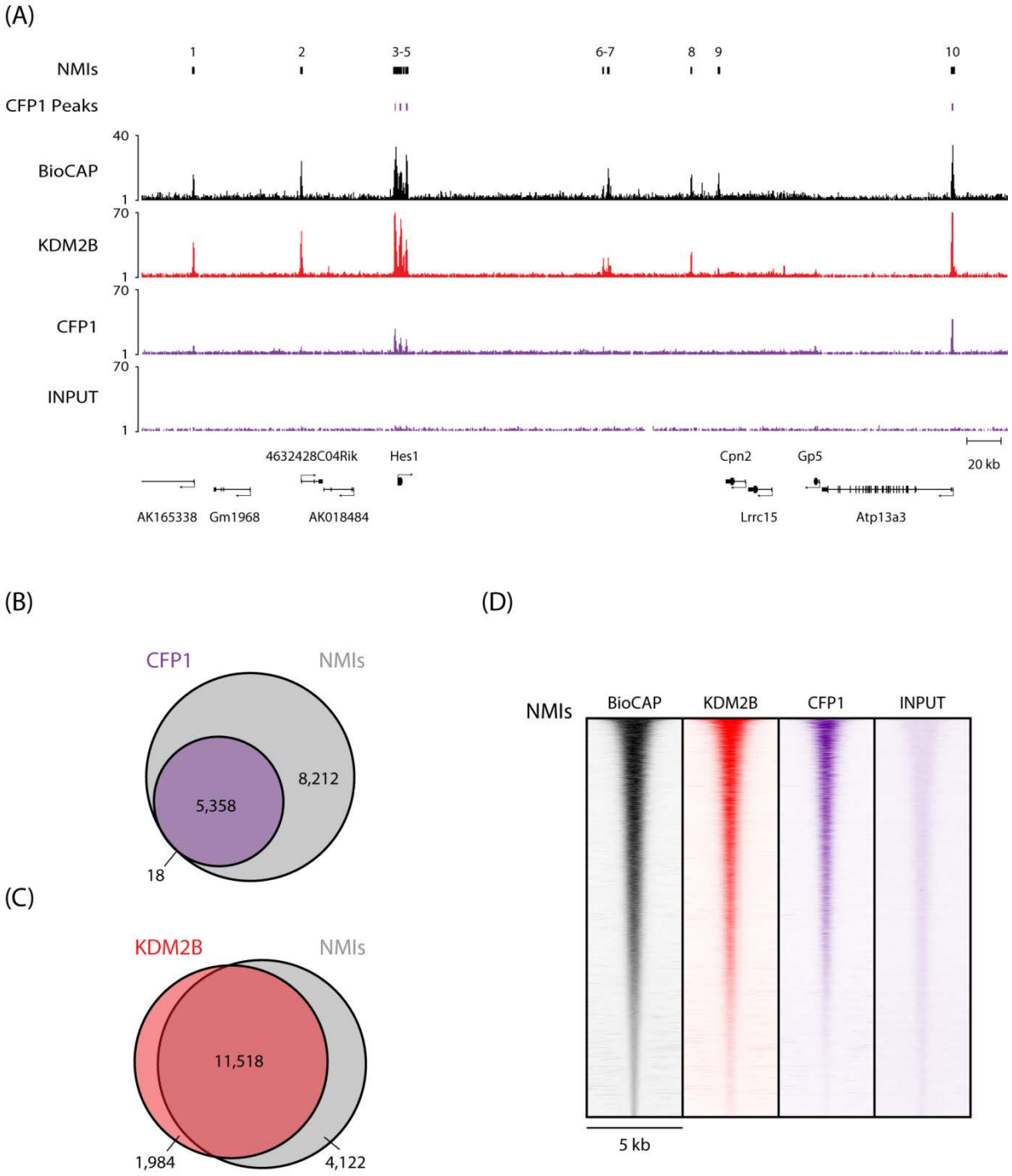



**Figure 20 – Unlike other zf-CXXC proteins CFP1 binds a subset of non-methylated islands**

(A) Profiles of non-methylated DNA (BioCAP), and enrichment of zf-CXXC proteins KDM2B and CFP1 are shown across a region of the mouse genome that contains ten non-methylated islands (NMIs). Each NMI is numbered above the profiles. Although peaks of KDM2B enrichment were detected at all ten NMIs, only four peaks of CFP1 enrichment (purple) were detected in this region. Importantly KDM2B, CFP1 and input control profiles were generated from equal numbers of sequencing reads

and are shown on the same scale. Genes are indicated below the profiles with arrows indicating the direction of transcription. The scale is given in the bottom right corner. (B and C) Genome-wide overlap between CFP1 or KDM2B peaks and NMIs are depicted as proportional Venn diagrams. (D) Heat maps illustrate variation in bioCAP signal, and in KDM2B, or CFP1 enrichment across C127 NMIs ranked by bioCAP read count. Enrichment of input chromatin is shown on the right as a negative control.

To understand why some NMIs are bound by KDM2B but not CFP1, their enrichments were compared across NMIs sorted by bioCAP reads (Figure 20D). As expected, enrichment of both zf-CXXC proteins diminished with bioCAP signal, however enrichment of KDM2B declined more slowly suggesting it might able to bind to NMIs with a lower density of non-methylated CpGs than CFP1.

### 4.7 CFP1 binds preferentially to non-methylated islands located at gene promoters

Examples of NMIs enriched for CFP1 and KDM2B, and NMIs with KDM2B but not CFP1, can easily be found by visually examining their genome-wide profiles (Figure 20A). CFP1 appears to favour NMIs located at gene promoters. To investigate this observation genome-wide, peaks containing a transcription start site (TSS) were counted for both CFP1 and KDM2B (Table 14). CFP1 peaks were overrepresented at TSSs, and were overrepresented to a greater extent than KDM2B. The skew of CFP1 towards NMIs containing a TSS (promoter NMIs) is easily visualised by plotting the distribution of peaks against distance from TSS (Figure 21A).

**Table 12 – Genome-wide distribution of CFP1 and KDM2B peaks**

| Peak Set (peaks) | TSS | Gene Body | Intergenic |
|---|---|---|---|
| BioCAP (13,572) | 70% (9,588) | 11% (1,576) | 18% (2,408) |
| CFP1 (5,653) | 90% (5,092) | 6% (311) | 4% (250) |
| KDM2B (11,653) | 77% (4,335) | 11% (608) | 12% (710) |

This was further examined by comparing KDM2B and CFP1 enrichments at C127 NMIs with or without a TSS (Figure 21B). KDM2B enrichment levels were higher than the input control at nearly all NMIs, and reflected the abundance of non-methylated DNA (BioCAP). Like KDM2B, CFP1 enrichment diminished with non-methylated DNA at promoter NMIs. Unlike KDM2B, CFP1 was not present above the input control at the majority of non-promoter NMIs.

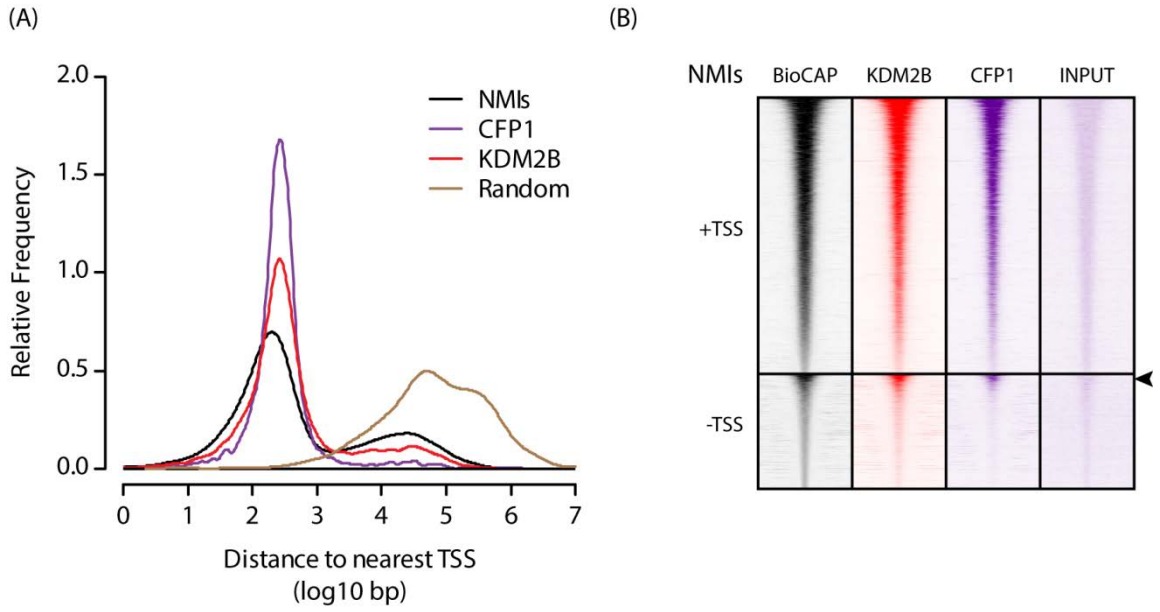



**Figure 21 - CFP1 binds preferentially to NMIs containing a transcriptional start site**

(A) The distribution of NMI, CFP1, KDM2B peak sets, and a control set of random genomic intervals is depicted relative to annotated transcriptional start sites (TSSs) by plotting relative peak frequency against distance to the nearest TSS. Distance was measured from the centre of each peak to the nearest annotated TSS (The refGene data set used for this analysis contained 23,203 TSSs). (B) Heat maps illustrate the variation in bioCAP signal and in enrichment of KDM2B, and CFP1 across all C127 NMIs. Input DNA is also shown as a negative control. NMIs were divided into promoter NMIs, which contain a TSS (+TSS) and NMIs without a TSS (-TSS). Both groups were ranked by BioCAP signal. Although CFP1 enrichment is almost exclusive to promoter NMIs (+TSS), it also occurs at a minority of –TSS sites (black arrowhead). This small fraction may represent NMIs containing an unannotated TSS, addressed in Figure 23C.

## 4.8 ChIP-sequencing reveals that 89% of promoter NMIs are occupied by RNA polymerase II in C127s

As CFP1 binds preferentially to promoter NMIs, it was important to decipher whether this preference was related to promoter activity, or whether it relies on underlying sequence characteristics associated with promoter regions. To investigate the effect of promoter activity on CFP1 binding, RNA polymerase II (Pol II) was profiled in C127 cells by ChIP-sequencing. A genome-wide profile was obtained by aligning Pol II reads to the mouse genome (Figure 22A and B). To enable a rough classification of active or inactive genes in C127 cells, the genome-wide profile was used to identify peaks of Pol II enrichment. 12,111 peaks were present in both biological replicates. The majority (72%) of Pol II peaks were associated with an annotated TSS (Figure 22C). In some cases these peaks tailed into the gene body presumably reflecting high levels of transcription (Figure 22B). Pol II enrichment was detected at 70% of NMIs and 89% of promoter NMIs (Figure 22D and E respectively).

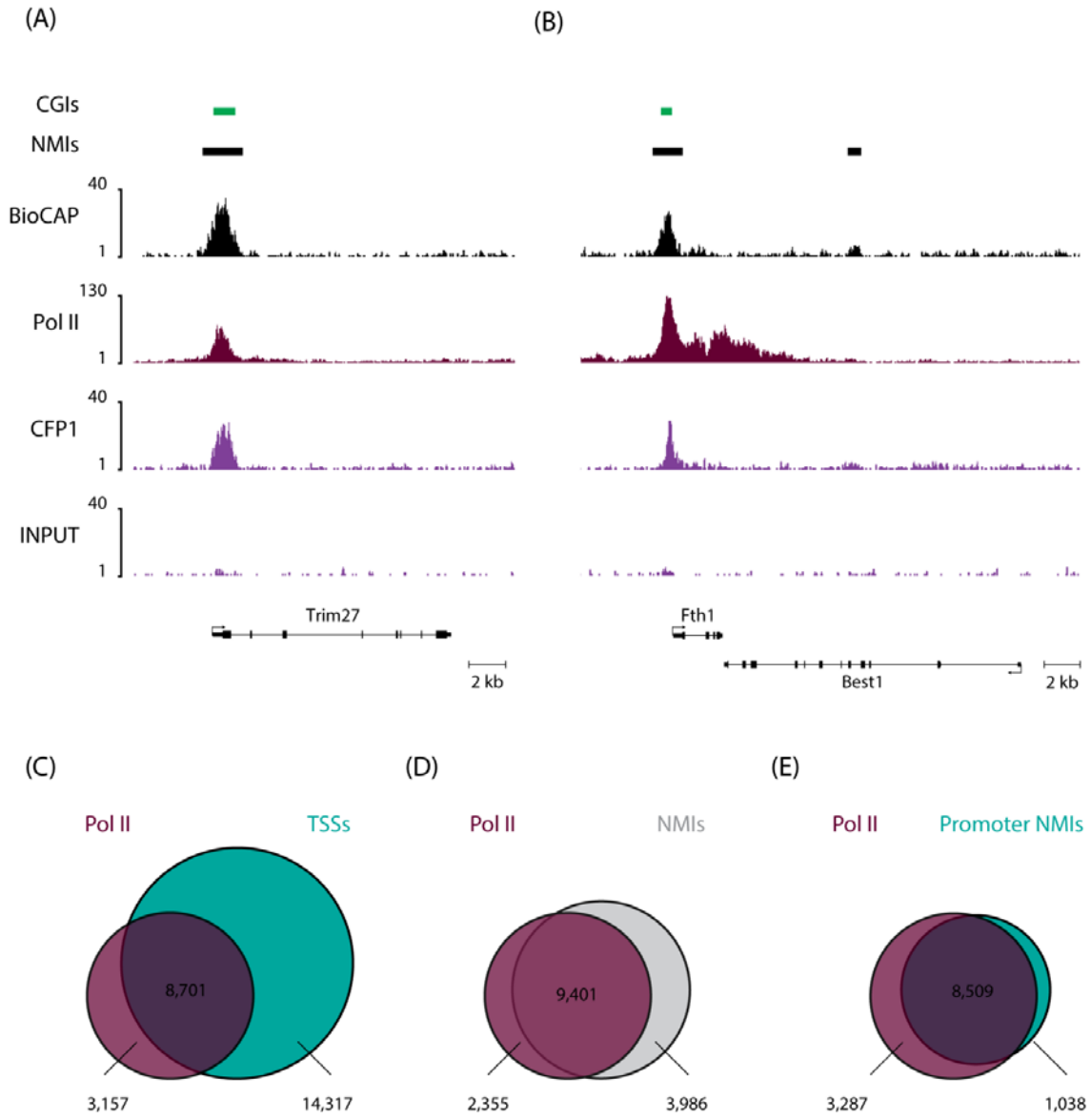



**Figure 22 – RNA polymerase II is found at the majority of non-methylated island promoters**
(A and B) Sequencing profiles for non-methylated DNA (BioCAP), Pol II and CFP1 are shown at two NMI promoters. Genes are shown below the profiles with arrows indicating the direction of transcription. Scales are given in the bottom right hand corner (C) (D) and (E) Proportional Venn diagrams depict the genome-wide overlap between Pol II and TSSs (C), Pol II and NMIs (D) and between Pol II and promoter NMIs (E).

### 4.9 CFP1 peaks are associated with high levels of RNA polymerase II

Non-methylated DNA appears to be essential for CFP1 enrichment, yet the preferential enrichment of CFP1 at gene promoters suggests additional factors, perhaps including polymerase occupancy

itself, may contribute to its genome-wide localisation. To investigate this possibility the relationship between Pol II and CFP1 was examined. CFP1 peaks are almost exclusively found at sites of Pol II enrichment (Figure 23A), but are only found at 53% of promoter NMIs (Figure 23B), despite the detection of Pol II at 89% of promoter NMIs (Figure 22E). This suggests that despite frequently overlapping on the genome, the presence of Pol II at a promoter NMI is not sufficient to drive enrichment of CFP1. To visualise the relationship between Pol II levels and CFP1, their enrichments were compared across all C127 NMIs (Figure 23C). Although CFP1 enrichment was generally higher at NMIs with higher levels of Pol II, the levels of non-methylated DNA (BioCAP) and KDM2B were also higher at these sites.

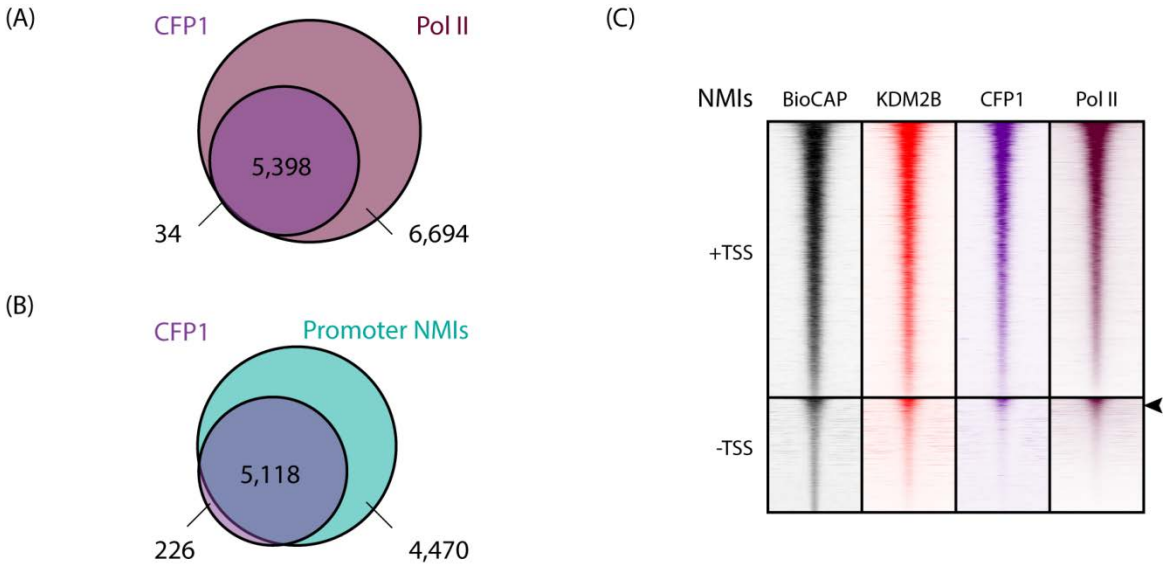



**Figure 23 - CFP1 binds preferentially to active non-methylated island promoters**

(A) The genome-wide overlap between peaks of CFP1 and RNA Polymerase II is depicted a proportional Venn diagram. (B) This Venn diagram shows the overlap between CFP1 peaks and non-methylated islands (NMIs) containing a transcription start site (Promoter NMIs). (C) Heat maps illustrate variation in BioCAP, KDM2B, CFP1 and Pol II across all C127 NMIs. NMIs were divided into Promoter NMIs, which contain a TSS (+TSS) and NMIs without a TSS (-TSS). Both groups were then ranked by Pol II read count. As in Figure 21B, CFP1 enrichment is observed at a minority of non-promoter NMIs (-TSS) (black arrowhead). The enrichment of Pol II at these sites supports the suggestion that these sites represent promoter NMIs containing an unannotated TSS.

### 4.10 The relationship between CFP1 and Pol II is obscured by their underlying correlations with non-methylated DNA

CFP1 levels were higher at NMIs with high levels of Pol II suggesting a positive correlation (Figure 23C). To examine this relationship in more detail, read counts were compared across all NMIs (Figure 24A). These comparisons were based on read densities to account for the variation in NMI length. Read density was calculated as the number of reads per million divided by the length of the NMI in base pairs (reads/million/bp). A moderate positive correlation was observed between the levels of CFP1 and Pol II. This was confirmed by calculating the Pearson's product-moment correlation coefficient ($r= 0.63$, $n=13,570$, $p < 2.2x10^{-16}$). However the broad scatter of these data suggest the levels of CFP1 and Pol II are only weakly related ($R^2= 0.40$).

To aid interpretation of these values, the same analyses were performed examining the relationship between KDM2B and Pol II (Figure 24B). Once again a moderate positive correlation was observed ($r=0.58$, $n=13,570$, $p < 2.2x10^{-16}$) although marginally weaker than between CFP1 and Pol II. Similarly the scatter was slightly broader ($R^2=0.34$). The fact these two different zf-CXXC proteins share a similar relationship with Pol II is unsurprising given their strong correlation with one another ($r=0.70$, $n=13,570$, $p < 2.2x10^{-16}$) (Figure 24C).These analyses suggest that the relationship between CFP1 and Pol II is slightly stronger than between KDM2B and Pol II. Due to the large sample size ($n=13,570$) this small difference is highly significant ($p < 7.9x10^{-11}$) and unlikely to occur by chance. However, as the difference is small it is unlikely that a stronger relationship with Pol II is sufficient to explain the limited interaction of CFP1 with a subset of NMIs.

Although the effect of differently sized NMIs was normalised by comparing read densities rather than read counts, variation in CpG density across NMIs was not accounted for. As there is a strong correlation between CpG density and transcriptional activity (Deaton & Bird 2011), Pol II is likely to correlate with zf-CXXC proteins due to a shared correlation with CpG density. To determine the

underlying relationship between Pol II and non-methylated DNA, read densities for Pol II and bioCAP were compared across all NMIs (Figure 24D). This relationship was significantly stronger (r=0.70, n=13,570, $p < 2.2x10^{-16}$), and more linear ($R^2$=0.49) than between Pol II and either of the zf-CXXC proteins. As expected, both zf-CXXC proteins were strongly related to bioCAP read density (Figure 24E and F).

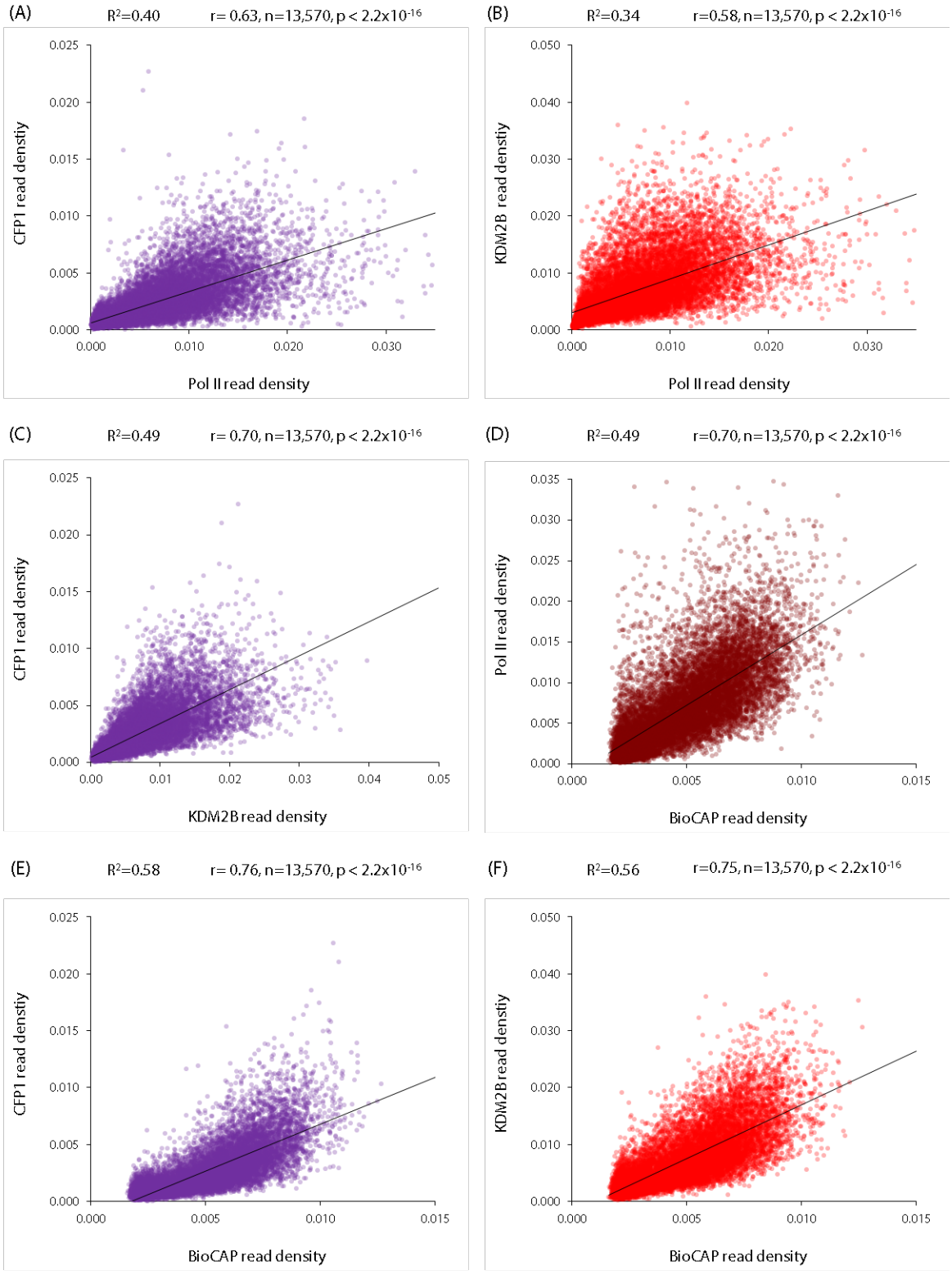



**Figure 24 – CFP1 enrichment correlates more strongly with non-methylated DNA than with Pol II**
Scatter plots with linear regression lines. The $R^2$ value of the linear regression and the Pearson's correlation coefficient (r) is given above each plot.

## 4.11 Discussion

This chapter describes the genome-wide localisation of endogenous CFP1. Previous attempts in the Klose lab to study CFP1 by ChIP-seq were limited by low affinity or low specificity antibodies. Using the α-CFP1f reagent developed in the previous chapter, these issues were overcome. Although genome-wide data is available for CFP1 from whole mouse brain (Thomson et al. 2010), the present study provides the first profile of CFP1 in an established cell line providing a suitable resource for further investigations.

In this study, chromatin was prepared from a different tissue type and immunoprecipitated using a different antibody, however the punctate character of the CFP1 ChIP-seq profile is consistent with the published profile from whole mouse brain (Thomson et al. 2010). Together with structural and biochemical information about the strict requirement of zf-CXXC domains for non-methylated CpGs (Xu et al. 2011; Lee et al. 2001; Blackledge et al. 2010; Cierpicki et al. 2010), these profiles support the division of the genome into a peaks of CFP1 enrichment against intervening regions with low CFP1 levels.

Importantly, while CFP1 peaks were observed at virtually all NMIs (~81%) in brain tissue, they were only observed at 40% of NMIs in C127 cells despite a comparable number of NMIs in total (13,572 in C127s compared to 13,348 in mouse brain). A simple explanation for this difference is that while CFP1 may bind a subset of NMIs in a given cell type, this selectivity is masked in the profile from a heterogeneous tissue like the brain. Further work is currently underway to address the localisation of CFP1 in different cell types. Although the profile from whole brain tissue demonstrates that CFP1 is capable of binding to the majority of NMIs genome-wide, its profile in C127 cells suggests that in a uniform cell type, enrichment is limited to a specific subset. This has several important implications for CFP1 function.

Preferential binding to a subset of NMIs may suggest an additional level of specificity beyond the recognition of non-methylated CpGs (this possibility is examined in detail in Chapters 5 and 6). Based on its interaction with SET1A and SET1B (Lee & Skalnik 2005; J.-H. Lee et al. 2007), which methylate H3K4 (Briggs et al. 2001), CFP1 was previously thought to direct H3K4 methylation to non-methylated CGIs throughout the genome, contributing to the unique chromatin architecture observed at CGI regions (Thomson et al. 2010; Zhou et al. 2012; Blackledge & Klose 2011; Tate et al. 2010). In contrast, selective enrichment of CFP1 at a specific subset of NMIs may suggest it preferentially directs H3K4 methylation to these regions. Interestingly, the preferential binding observed for CFP1 was not shared by KDM2B, which bound to NMIs genome-wide consistent with previous work in the Klose lab (Farcas et al. 2012). This suggests that while KDM2B exerts a global function affecting all NMIs, CFP1 may function at specific sites depending on the identity of the cell. Despite an apparent preference for binding to promoter NMIs, without further investigation it is not clear how CFP1 is able to bind selectively to a subset of NMIs. The following Chapter explores the interaction between CFP1 and chromatin in more detail to try to address this question.

# Chapter 5 - Dissecting the contribution of the PHD domain to CFP1 localisation

## 5.1 Introduction

CFP1 binds to non-methylated DNA via its zf-CXXC domain. However, CFP1 also contains a plant homeo domain (PHD) at its N-terminus. PHD domains are found in a large family of proteins and often mediate interactions with specific histone modifications on nucleosomes (Zhang 2006; Mellor 2006). It is not known whether the PHD domain of CFP1 contributes to its interaction with chromatin *in vivo*, however a multiple sequence alignment of CFP1 homologues from diverse eukaryotic species revealed conservation of the PHD domain from yeast to humans (Figure 25). The PHD domain of Spp1, a yeast homologue of CFP1, has been shown to bind specifically to H3K4me3 (Shi et al. 2007). Based on the sequence conservation this interaction was proposed to occur in higher eukaryotes (Tate et al. 2010). This chapter investigates the contribution of the PHD domain to the genome-wide localisation of CFP1.


**Figure 25 – The CFP1 PHD domain is conserved from yeast to humans**

Multiple sequence alignment. CFP1 PHD domains from; *Homo sapiens*, *Mus musculus*, *Capra aegagrus hircus* (Goat), *Xenopus laevis*, *Danio rerio*, *Crassostrea gigas* (Pacific Oyster), *Drosophila melanogaster*, *Anopheles gambiae* (Mosquito) and the PHD domain from *Saccharomyces cerevisiae* Spp1, were aligned using ClustalW. Identities are highlighted yellow, and orange, similarities in red and zinc coordinating residues in cyan.

## 5.2 Expression and purification of the CFP1 PHD domain

To determine whether CFP1 is able to bind H3K4me3 via its PHD domain, an N-terminal fragment containing the PHD domain was expressed in bacteria and purified for *in vitro* experiments (Figure 26A and B). The first 92 amino acids of CFP1 were chosen for expression based on conservation with the *S. cerevisiae* homologue Spp1, and on the PHD fragments used in previous studies examining the histone binding properties of the ING2 and MLL PHD domains (Chang et al. 2010; Gozani et al. 2003). A fragment of the CFP1 cDNA encoding the selected 92 amino acid fragment (PHDf) was cloned into an inducible prokaryotic expression vector, then expressed in bacteria and purified. This vector also introduced an N-terminal 6xHis tag, enabling nickel NTA purification, and a glutathione-S-transferase (GST) tag, to facilitate histone pull down experiments. As a negative control, the GST tag was independently expressed and purified (Figure 26C and D).

## 5.3 The CFP1 PHD domain binds to H3K4me3 *in vitro*

Using the purified PHD fragment (GST-PHDf), GST pull-downs (Shi et al. 2006) were conducted to determine whether the CFP1 PHD domain recognises H3K4me3 *in vitro*. In these experiments, GST-PHDf was immobilised on agarose beads coated with glutathione. As positive controls, the ING2 PHD domain and the tandem chromodomain from CHD1 which bind specifically to H3K4me3, were also expressed, purified, and immobilised on glutathione coated beads (Sims et al. 2005; Gozani et al. 2003) (Figure 26E-H). The ING2 PHD domain expression construct was a kind gift from Or Gozani and the CHD1 tandem chromodomains (TCD) from Cheryl Arrowsmith (Nady et al. 2008).

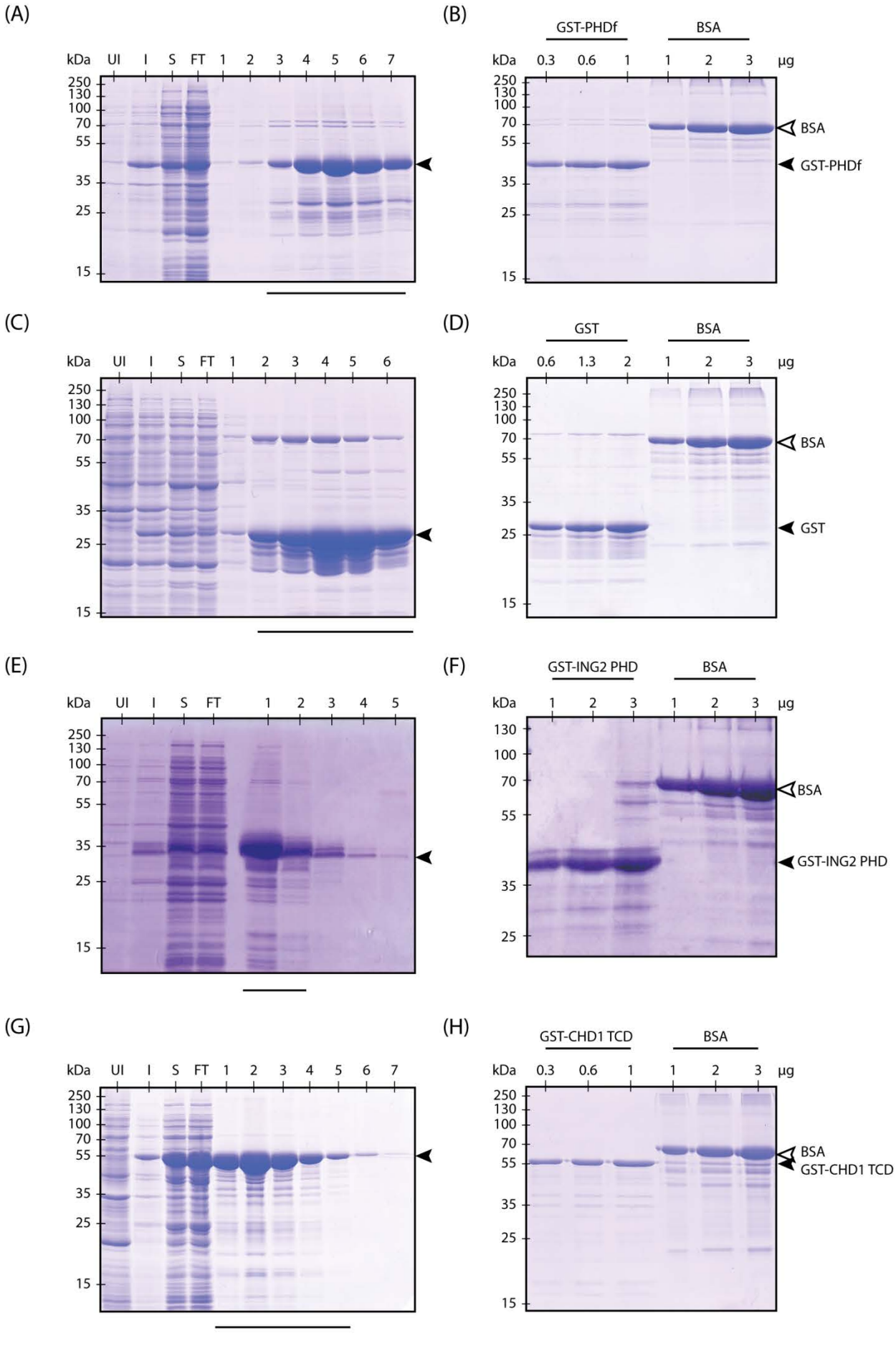

(A)
kDa UI I S FT 1 2 3 4 5 6 7
250
130
100
70
55
35
25
15
(B)
GST-PHDf
BSA
kDa 0.3 0.6 1 1 2 3 µg
BSA
GST-PHDf
(C)
kDa UI I S FT 1 2 3 4 5 6
(D)
GST
BSA
kDa 0.6 1.3 2 1 2 3 µg
BSA
GST
(E)
kDa UI I S FT 1 2 3 4 5
(F)
GST-ING2 PHD
BSA
kDa 1 2 3 1 2 3 µg
BSA
GST-ING2 PHD
(G)
kDa UI I S FT 1 2 3 4 5 6 7
(H)
GST-CHD1 TCD
BSA
kDa 0.3 0.6 1 1 2 3 µg
BSA
GST-CHD1 TCD

**Figure 26 – Expression and purification of the CFP1 PHD domain**

(A) Expression and Ni-NTA purification of GST- PHDf (39 kDa). UI: Uninduced, I: IPTG induced, S: soluble, FT: Flow through from Ni-NTA purification. 1-7: consecutive elutions. The eluted fractions (3-7) pooled for dialysis are indicated by the horizontal, black bar below the gel. (B) Quantitation of GST-PHDf after dialysis. Purified GST-PHDf was run with bovine serum albumin (BSA) as a quantitative comparison. (C) Expression and purification of GST (30 kDa). Fractions 2-6 were pooled for dialysis (black horizontal bar). (D) Quantitation of GST as in (B). (E) Expression and purification of GST-ING2 PHD (37 kDa). (F) Quantitation of GST-ING2 PHD after dialysis. (G) Expression and purification of GST-CHD1 TCD (52 kDa). (H) Quantitation of GST-CHD1 TCD after dialysis.

To determine whether the CFP1 PHD domain recognises H3K4me3, a recombinant methyl lysine analogue (MLA) was generated by specific alkylation of recombinant H3 in which the fourth lysine was mutated to cysteine (H3K4C), producing H3K4Cme3. MLA offers precise control over the methylation state of a specific lysine residue (Simon et al. 2007), ensuring trimethylation of the K4 position. The H3K4C control was also alkylated to produce aminoethyl-cysteine at the K4 position more closely resembling unmodified lysine. H3K4Cme3 nucleosomes have been used for previous *in vitro* studies (van Nuland, Schram, et al. 2013), and are recognised by the TAF3 PHD which also binds H3K4me3 (Lauberth et al. 2013). The H3K4C and H3K4Cme3 proteins used in this study were produced by Jin C. Zhou.

To determine whether the CFP1 PHD domain recognises H3K4me3, GST-pull downs using H3K4C or H3K4Cme3 were performed in parallel (Figure 27) and bound histone was analysed by western blot (Figure 27D). This demonstrated that the CFP1 PHD domain binds to H3K4me3 more robustly than the unmodified H3 tail, suggesting that CFP1 is able to recognise the H3K4me3 mark (Figure 27F). Under these conditions the ING2 PHD domain bound robustly to both H3K4me3 and unmodified H3, nevertheless a preference was observed for the trimethylated histone. Similarly, a low level of binding was observed between the CHD1 tandem chromodomains and the unmodified H3. When the salt concentration was raised to 300 mM NaCl, the ING2 PHD domain bound specifically to

H3K4me3, however the CFP1 PHD domain did not bind under these conditions (not-shown). Non-specific binding was not observed in the GST control. Together these results suggest that the CFP1 PHD domain binds specifically to H3K4me3 albeit with a lower affinity than the ING2 or CHD1 domains.

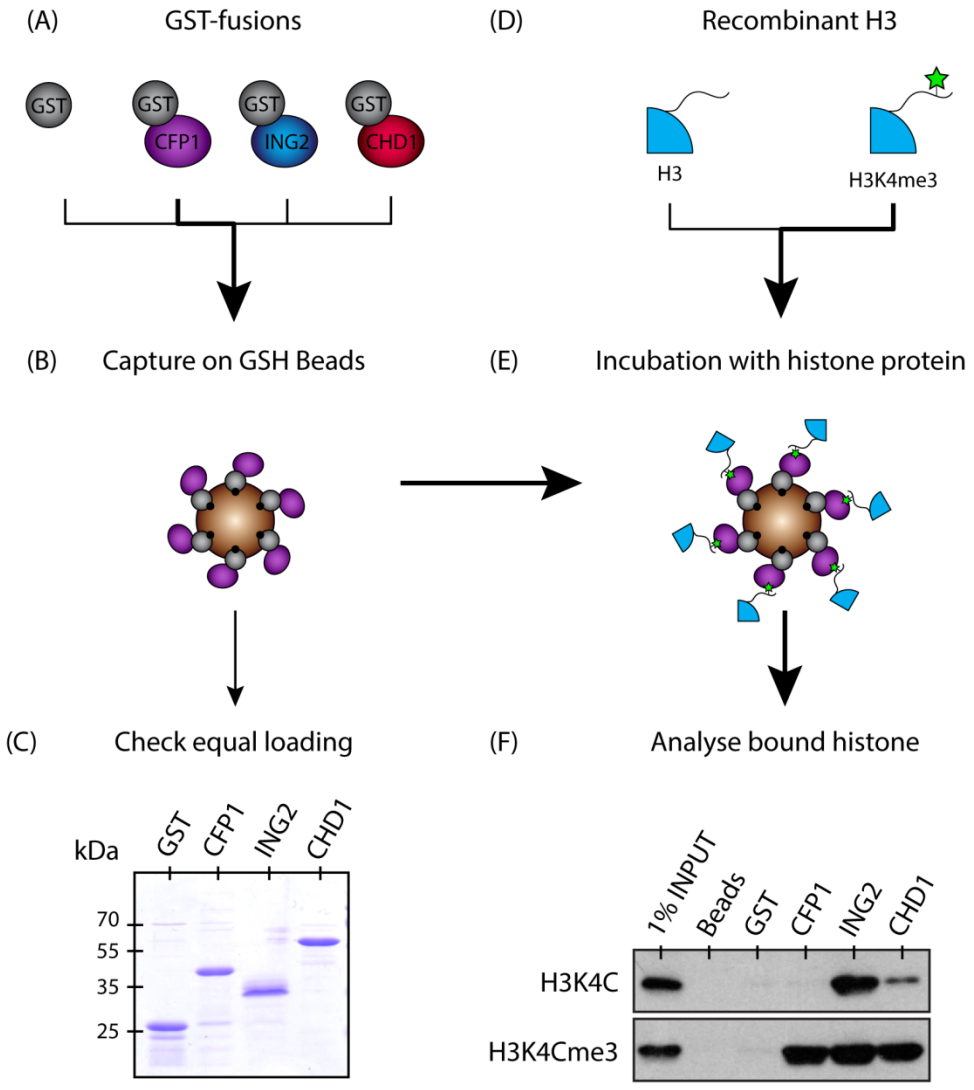



**Figure 27 – The CFP1 PHD domain binds selectively to H3K4me3**

Schematic of GST-pull down experiments (A) GST (grey), GST-PHDf (purple), GST-ING2 PHD (blue) and GST-CHD1 TCD (red) were expressed and purified. (B) Equimolar amounts of each fusion were incubated with GSH beads (brown) to capture the GST tagged proteins. (C) Equal loading of each GST-fusion is verified by SDS-PAGE. The equivalent of 6% INPUT was separated on an 8% polyacrylamide gel and stained with Coomassie. (D) The H3K4me3 analogue H3K4Cme3 was

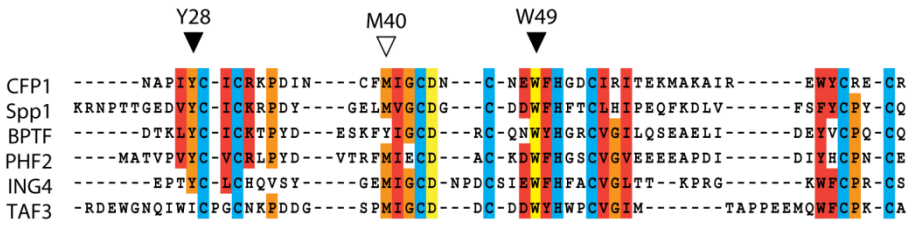
generated by specific alkylation of recombinant H3K4C, an alternative alkylation produced H3K4C as an unmodified H3 control (This work was performed by Jin Zhou). (E) Beads coated with each GST-fusion were divided between incubation with H3K4C or H3K4me3. (F) After wash steps to remove unbound histone, interaction with H3 was analysed by western blot staining with α-H3.

While this work was in progress, CFP1 was also identified as H3K4me3 binding protein by another lab using an unbiased proteomic screen based on mass spectrometry (Eberl et al. 2012). Using peptide pull down assays, this group found that like Spp1, CFP1 recognises H3K4me3, and that this interaction is abrogated by removal of the PHD domain. This independent evidence supports the specific binding of the CFP1 PHD domain to H3K4me3 as suggested by the GST-pull down experiments.

## 5.4 Designing independent point mutations to disrupt the CFP1 PHD domain

Although the PHD and zf-CXXC domains of CFP1 have been characterised *in vitro*, their relative contributions to CFP1 localisation, cellular distribution and chromatin binding kinetics remains unknown. Exclusive localisation of CFP1 to non-methylated islands might suggest that the zf-CXXC domain alone is sufficient for proper localisation of CFP1. However genome-wide profiling of CFP1 described in the previous chapters revealed preferential localisation to active gene promoters. As active promoters are typically marked by H3K4me3 (Heintzman et al. 2007), preferential enrichment at these sites may suggest that the CFP1 PHD domain contributes the observed localisation. To test this possibility, a mutant versions of CFP1 was required which would disrupt the activity of PHD in order to examine its relative contribution to CFP1 binding to chromatin *in vivo*.

(A)




(B)

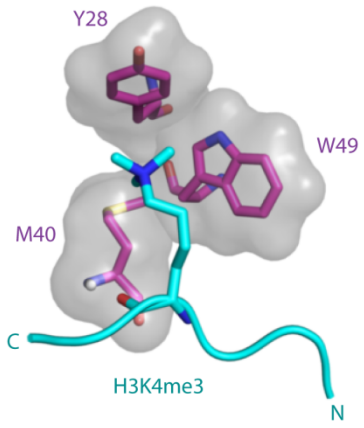



**Figure 28 – Multiple sequence alignments reveal key residues in the CFP1 zf-CXXC and PHD domains**

(A) Multiple sequence alignment of the PHD domains from various H3K4me3 binding proteins reveal several positions of conserved sequence identity (orange or yellow for 100%), and similarity (red). Zinc coordinating residues are highlighted in cyan. Amino acids Y28, M40 and W49 which form the K4me3 binding site are indicated above the alignment. (B) A structural model of the CFP1 aromatic cage is based on a crystal structure of the BPTF PHD domain in complex with an H3K4me3 peptide (PDB ID - 2FUU). The H3 peptide (cyan) K4me3 sits within an aromatic cage comprised of Y28, M40 and W49 (purple). N: blue, O: red, S: yellow.

To design a subtle point mutation that would prevent the CFP1 PHD domain from binding to H3K4me3, its sequence was compared to PHD domains from other proteins which bind specifically to H3K4me3. Crystal structures with a bound H3K4me3 peptide are available for PHD domains from several proteins exposing the residues which contact the K4me3 ligand (Li et al. 2006; Sanchez & Zhou 2011; Palacios et al. 2008). These structures reveal the position of K4me3 within a conserved aromatic cage of the PHD domain (Figure 28C). As PHD domains share a common structure, the

equivalent residues were identified in CFP1 by performing a multiple sequence alignment (Figure 28D). Based on the sequence alignment, and on mutations used in previous studies (Shi et al. 2006; Sanchez & Zhou 2011; Li et al. 2006; Peña et al. 2008; Peña et al. 2006) a point mutation was designed in the CFP1 PHD domain converting the central tryptophan (W49) of the aromatic cage to an alanine (W49A). A second point mutation was also designed converting the tyrosine 'lid' of the aromatic cage to alanine (Y28A).

### 5.5 Point mutations in the aromatic cage abrogate CFP1 PHD binding to H3K4me3

To verify that the newly designed CFP1 point mutations abrogated H3K4me3 binding they were analysed using a peptide pull down assay. In these experiments binding to mono- and dimethylated K4 was also tested to determine the impact of each mutation on binding to these methylation states. The aromatic cage mutants W49A and Y28A were each expressed and purified as GST tagged fragments and compared to the wild type PHDf (Figure 29A and B). Peptide pull-down experiments made use of four synthetic peptides corresponding to N-terminal H3 tail (amino acids 1-21) (Figure 29C). The four peptides differed only at the K4 position, which was either unmodified, mono-, di- or tri-methylated. Each peptide was biotinylated at its C-terminal. Equal concentrations of wild type or mutant GST-PHDf were incubated with each peptide (Figure 29D). Peptides were then captured by streptavidin coated beads, which bind tightly to biotin. After washing to remove unbound protein, GST-PHDf that remained bound was analysed by western blot using an antibody against the GST tag (Figure 29E). To control for nonspecific interactions with the streptavidin beads, a mock pull down was performed by omitting biotinylated peptides.

Peptide pull down experiments replicated the preferential binding of wild type CFP1 to H3K4me3 previously observed by GST-pull down, providing further evidence for the recognition of H3K4me3 by the CFP1 PHD domain. Crucially, both mutants exhibited equal binding to all four peptides, unlike the wild type PHDf which bound more strongly to H3K4me2 and H3K4me3 (Figure 29E). This

suggests that the specific interaction with H3K4me3 is abrogated by mutation of Y28 or W49. CHD1 is reported to bind di- and trimethylated H3K4 with a 3-fold higher affinity than to monomethylated H3K4 (Sims et al. 2005). Under these experimental conditions, the CHD1 TCD bound robustly to H3K4me1, H3K4me2 and H3K4me3 but not to the unmodified H3 peptide. Compared to CHD1 TCD, the CFP1 PHD domain bound with greater selectivity for di- and trimethyl H3K4. This selectivity no longer observed when the aromatic cage was mutated.

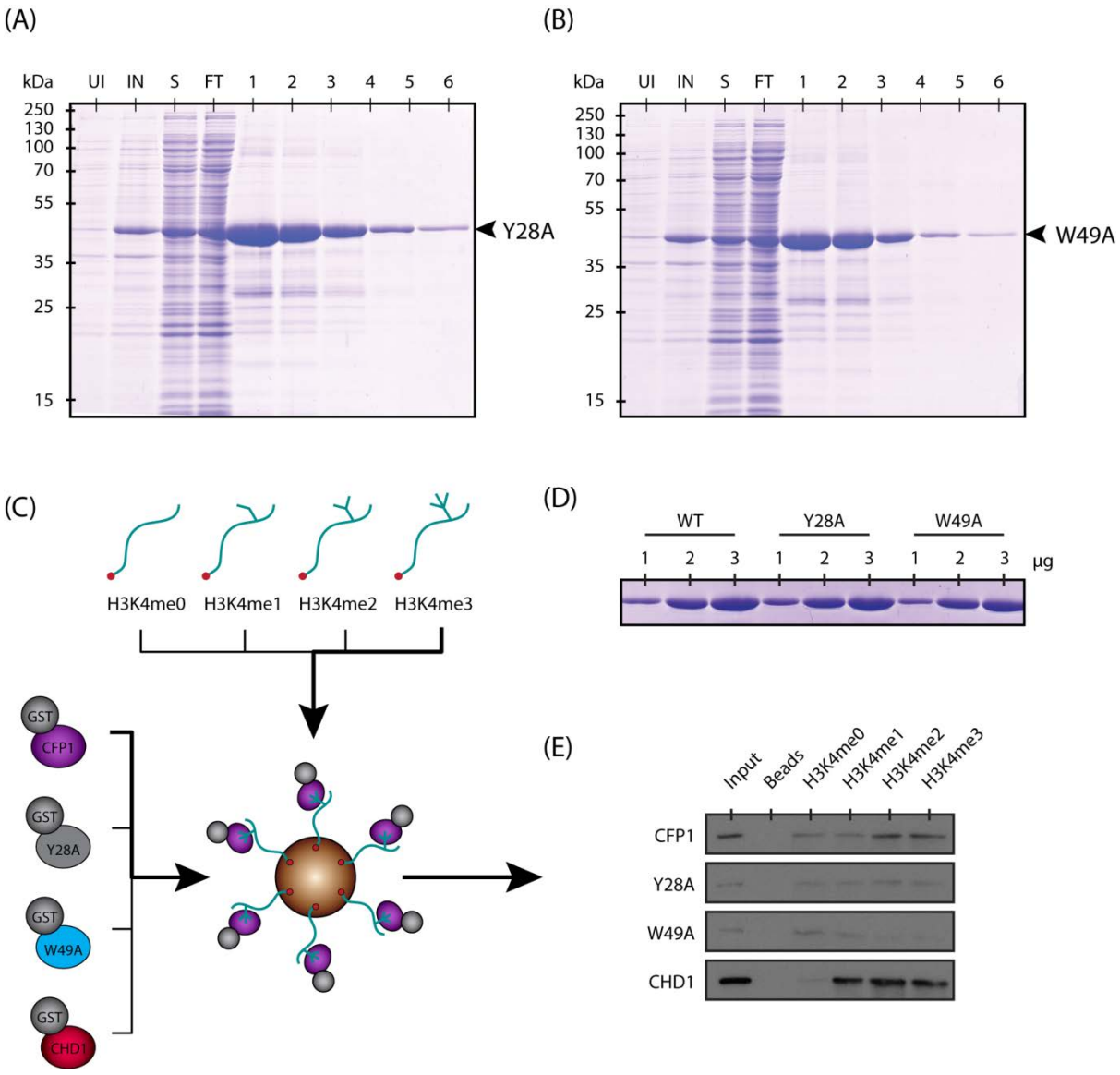


**Figure 29 – Aromatic cage mutations in the CFP1 PHD domain abrogate binding to H3K4me3**
(A) and (B) Expression and Ni-NTA purification of (A) Y28A, and (B) W49A, mutant GST-PHDf. UI: Uninduced, I: IPTG induced, S: soluble, FT: Flow through from Ni-NTA purification. 1-6: consecutive elutions. (C) Schematic diagram of the peptide pull down assay. Each peptide was incubated with each GST fusion and then captured on the surface of streptavidin coated beads (brown). After several wash steps to remove unbound molecules, captured fusion proteins were analysed by western blot.(D) The final concentrations of wild type and mutant GST-PHDf used in the peptide pull down were compared by SDS-PAGE. (E) Interactions between each peptide and GST-fusion were analysed by western blot, staining with α-GST. No interactions were detected between the GST-fusions and the streptavidin beads (Beads).

## 5.6 Developing a system to express CFP1 transgenes in mammalian cells lines

*In vitro* evidence suggests that the zf-CXXC and PHD domains of CFP1 offer two distinct modes of interaction with chromatin. The zf-CXXC domain binds to non-methylated CpG dinucleotides and the PHD domain binds to the H3K4me3 histone modification. As H3K4me3 and islands of non-methylated CpG are frequently found together on the genome it is important to understand how H3K4me3 and non-methylated DNA contribute to CFP1 chromatin binding activity *in vivo*. To examine the genome-wide localisation, cellular distribution, and binding dynamics of wild-type and mutant forms of CFP1, a system was developed to express CFP1 transgenes as GFP fusions in mammalian cells. For this purpose a GFP-tag was chosen as it enables live cell imaging experiments (discussed in Chapter 6) and serves as a high affinity epitope for chromatin immunoprecipitation (ChIP).

To generate stable cell lines expressing CFP1 and mutant derivatives, a new plasmid vector was designed including several features to streamline the process (Figure 30A). The pCAG vector which drives expression from the robust CMV/IEE promoter was adapted to include an N-terminal GFP tag followed by a ligation independent cloning (LIC) site. The inclusion of a LIC site enabled PCR based gene cloning so that multiple cDNAs could be simply and rapidly cloned into the same expression

vector. Downstream of the LIC site, the inclusion of an internal ribosome entry site followed by a puromycin resistance cassette results in polycistronic expression of the GFP fusion protein, and the puromycin n-acetyltransferase, when the vector is transfected into mammalian cells. Therefore, selection of puromycin resistant cells lines should also ensure stable expression of the upstream GFP fusion transgene.

## 5.7 Expression of GFP-tagged CFP1 in C127 cells

To begin mechanistically dissecting the chromatin binding activity of CFP1 *in vivo*, mouse C127 cells were chosen as they are easy to grow as monolayers allowing the production of large numbers of cells for chromatin immunoprecipitation experiments and highly amenable to light microscopy by virtue of their flat nuclei (Lowy et al. 1978; Markaki et al. 2012; Markaki et al. 2010). The C127 line was derived from a RIII mouse mammary tumour and is therefore aneuploid but genetically stable and retains an epithelial morphology. To test the capacity to generate stable C127 cell lines expressing GFP-CFP1 using the newly designed expression plasmid, GFP-CFP1 or GFP only vectors were transfected into C127 cells. Following puromycin selection each transfection yielded distinct, stably resistant colonies indicating random integration of the plasmid into the mouse genome. Clonal populations grown from resistant colonies proliferated normally suggesting that expression of a GFP or GFP-CFP1 transgene is non-toxic. Stable clones were then screened by western blotting to confirm the expression of full length fusion protein (Figure 30B). Initial screening of cells transfected with the GFP-CFP1 expression plasmid detected a specific band of the expected size (~115 kDa). Visual inspection by fluorescence microscopy verified that GFP-CFP1 was uniformly expressed in the NG1 clone. Therefore the newly designed stable expression vectors appear to function well and the NG1 GFP-CFP1 clone was chosen for further investigation using ChIP (fluorescence imaging of GFP-CFP1 is discussed in Chapter 6).

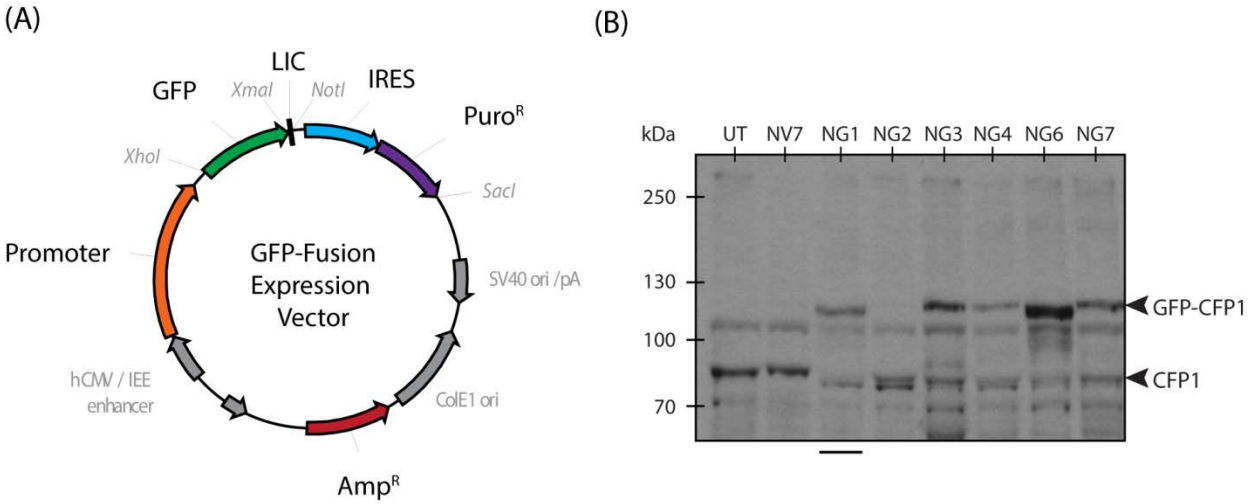



**Figure 30 – Generation of a plasmid vector for stable expression of GFP fusion proteins**
(A) Schematic diagram of the GFP fusion vector for mammalian expression. LIC: Ligation Independent Cloning site, IRES: Internal Ribosomal Entry Site. An N-terminal eGFP tag was introduced between XhoI and XmaI sites. The GFP-CFP1 fusion protein was generated by cloning human CFP1 cDNA into the LIC site. (B) Western blot screening of nuclear extract from stably resistant C127 clones after 10 days under puromycin selection. The membrane was probed with α-CFP1p which detects both endogenous CFP1 (76 kDa) and the GFP-CFP1 fusion (108 kDa). Note that both proteins run ~15 kDa higher than expected. UT: untransfected C127s, NV7: a stable C127 clone expressing free GFP from the empty vector, NG1-7: stable C127 clones transfected with the GFP-CFP1 expression vector. The NG1 clone (underlined) was selected for further characterisation.

## 5.8 GFP-CFP1 is enriched at promoter NMIs

To confirm that the normal binding activity of CFP1 is unaffected by addition of the GFP tag, chromatin extracted from the GFP-CFP1 cell line was immunoprecipitated using an antibody against GFP. As a control for non-specific interactions, α-GFP ChIP was also performed in the parental C127 line. GFP-CFP1 levels were measured at two promoter NMIs corresponding to peaks of endogenous CFP1 (as determined in Chapter 4) and at matched control regions in the body of each gene (Figure 31A and B). To serve as a positive control, α-KDM2B ChIP was performed and quantified in parallel

(Figure 31C and D). In both cases, GFP-CFP1 was substantially enriched at the NMI relative to the control region. This suggests that the GFP-CFP1 fusion localises correctly to NMI regions. KDM2B was also enriched at both promoters and similar enrichments were observed in C127 and NG1 lines. Despite enrichment of KDM2B, no enrichment was seen in the α-GFP ChIP in parental C127s, demonstrating the specificity of the GFP antibody and supporting its suitability for ChIP. In the NG1 line, GFP-CFP1 levels at the gene body regions are similar to the background levels in the negative control (C127). This suggests that like the endogenous protein, GFP-CFP1 is extremely selective for non-methylated DNA, and that the levels measured in the gene body regions likely reflect experimental noise rather than meaningful biological interactions.

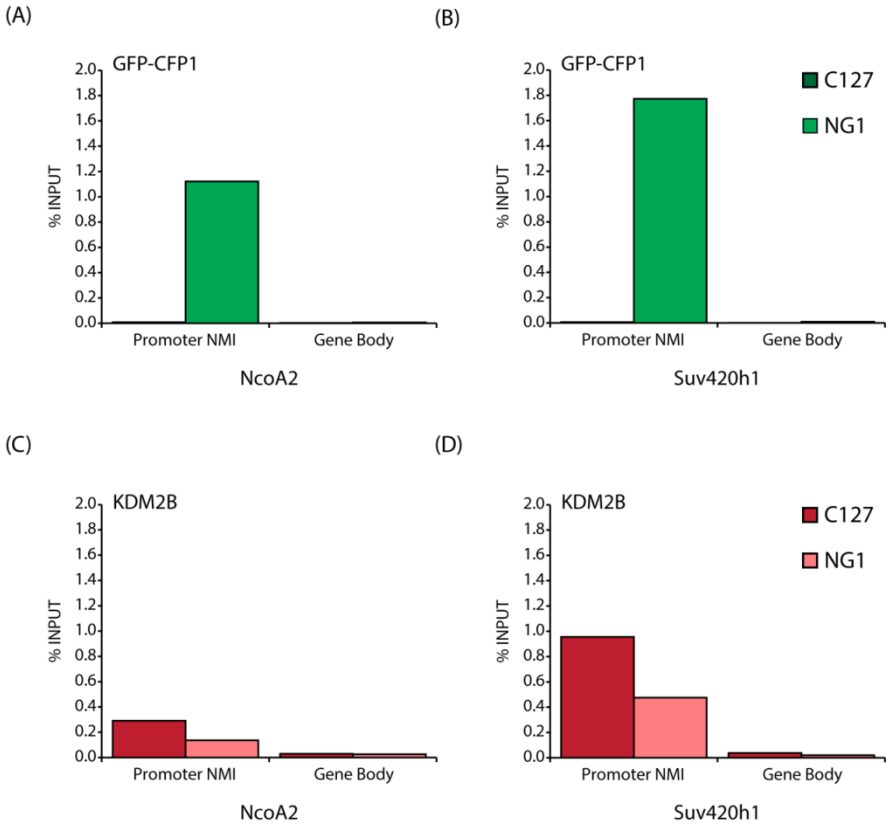


**Figure 31 – GFP-CFP1 is enriched at CpG islands**

Chromatin was prepared from NG1 cells which express GFP-CFP1 and immunoprecipitated using α-GFP or α-KDM2B as a positive control. Both IPs were also performed using chromatin from untransfected C127 cells in which GFP is absent. Precipitated DNA was quantified at two non-methylated islands (NMIs) found at gene promoters and at matched control regions within each gene body. (A) α-GFP ChIP-qPCR at the NcoA2 locus. (B) α-GFP ChIP-qPCR at the Suv420h1 locus. (C) α-KDM2B ChIP-qPCR at the NcoA2 locus. (D) α-KDM2B ChIP-qPCR at the Suv420h1 locus.

## 5.9 The genome-wide localisation of GFP-CFP1 closely resembles the endogenous protein

ChIP analysis at a selected number of target NMIs suggests that the GFP-CFP1 fusion protein binds to NMIs in a similar manner to endogenous CFP1. To enable dissection of the interaction between CFP1 and chromatin, it was important that the wild type GFP-CFP1 fusion could recapitulate the genome-wide localisation of endogenous CFP1. To validate GFP-CFP1 localisation at the genome scale α-GFP ChIP was performed on chromatin from C127 cells expressing GFP-CFP1 or GFP alone. ChIP DNA from two biological replicates was then submitted for library preparation and sequencing (WTCHG). The genome-wide localisation of GFP-CFP1 was then compared to profile of endogenous CFP1 described in Chapter 4 (Figure 32).

Importantly, the profile of wild type GFP-CFP1 closely resembled endogenous CFP1 (Figure 32A). Comparison of enrichment across all NMIs suggested a clear positive relationship between GFP-CFP1 and the endogenous protein (Figure 32B). To assess the strength of the relationship, the Pearson's correlation coefficient was calculated for read densities across all NMIs, revealing a strong positive correlation ($r=0.80 \pm 0.006$, $n=13,570$, $p<2.2x10^{-16}$). Appropriately, for independent measurements of the same protein, this correlation is significantly stronger than between two different zf-CXXC proteins across all NMIs ($r=0.70 \pm 0.009$, $n=13,570$, $p<2.2x10^{-16}$ for CFP1 against KDM2B). The close agreement between CFP1 and GFP-CFP1 profiles cross-validates the α-CFP1f antibody, generated in Chapter 3, and supports use of the GFP-fusion to investigate CFP1 localisation genome-wide.

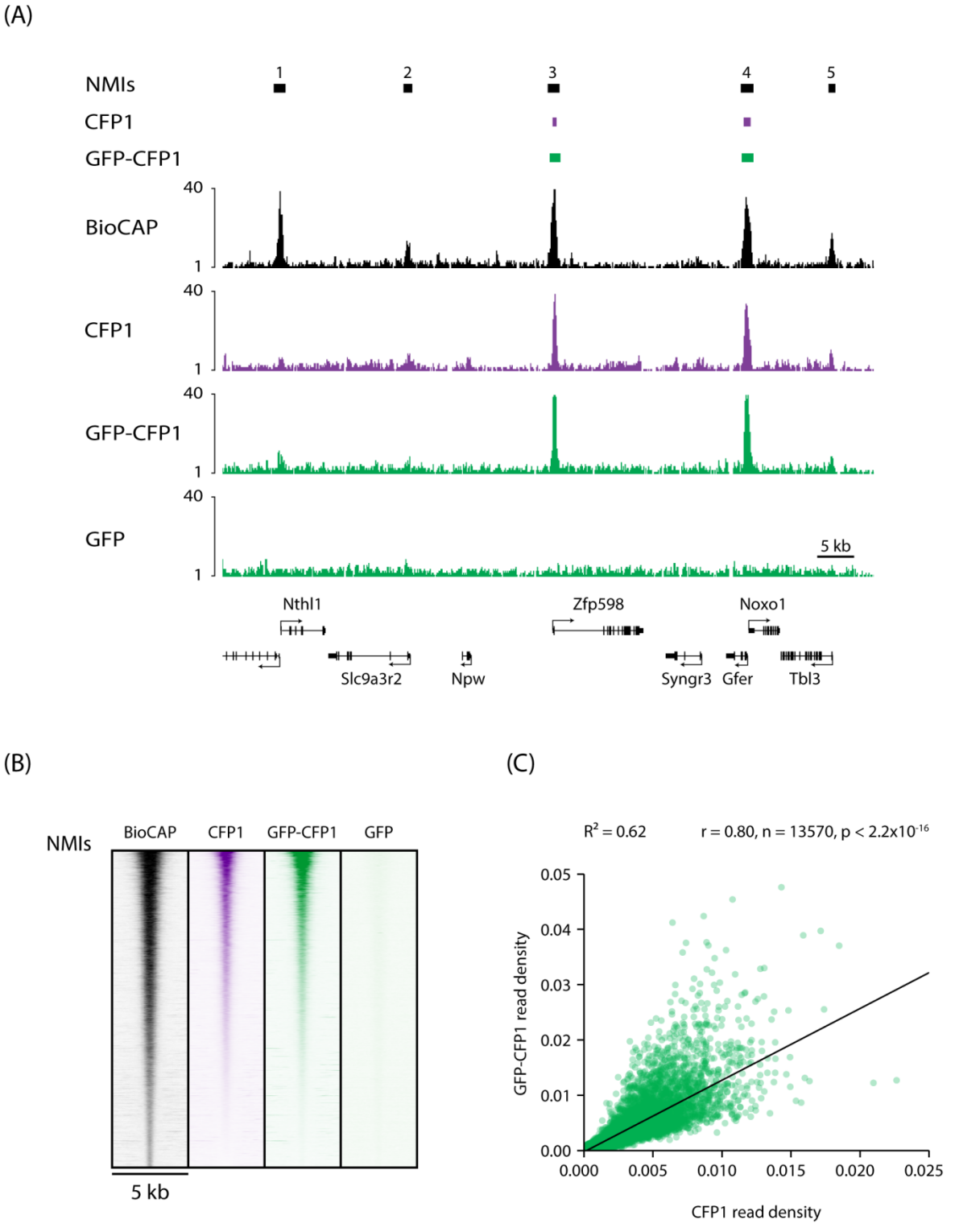



**Figure 32 – Like the endogenous protein, GFP-CFP1 binds to a minority of NMIs**

(A) Profiles of non-methylated DNA (BioCAP), endogenous CFP1 protein (CFP1f), GFP-CFP1 and free GFP are shown across a ~100 kb region of chromosome 17. This region contains five NMIs (numbered above). Peaks of endogenous CFP1 are observed NMIs 3 and 4. Like the endogenous

protein, GFP-CFP1 peaks are observed at NMIs 3 and 4. (B) Heat maps illustrate variation in bioCAP signal, CFP1f, GFP-CFP1 and free GFP as a negative control, across NMIs ranked by their bioCAP signal. (C) The relationship between GFP-CFP1 and endogenous CFP1 levels at C127 NMIs is depicted as a scatter plot. The Pearson's correlation coefficient (r = 0.8) suggests a strong positive correlation.

### 5.10 CFP1 binds selectively to a subset of non-methylated regions

The previous chapter found that unlike KDM2B, peaks of CFP1 enrichment were limited to a minority of NMIs. However it was not clear from this analysis whether peaks were only detected at a specific subset of NMIs due to selective binding of CFP1 or because of a technical limitation of the antibody used. ChIP-sequencing of the GFP-CFP1 fusion protein using α-GFP allowed an independent assessment of CFP1 localisation which did not rely on the α-CFP1f reagent.

Visual inspection of the GFP-CFP1 profile revealed that, like the endogenous protein, enrichment of GFP-CFP1 was not observed at all NMIs (Figure 32A). A comparison of endogenous CFP1 and GFP-CFP1 enrichment across all NMIs supported the initial observation that CFP1 enrichment is limited to a subset of NMIs (Figure 32B). To understand how sites of CFP1 enrichment are distributed across the genome, peaks of GFP-CFP1 enrichment were called using the MACS algorithm (Zhang et al. 2008), enabling the localisation of GFP-CFP1 to be compared to the endogenous protein.

In total, 8,444 GFP-CFP1 peaks were identified, of which 6,946 peaks (82%) were identified in both replicates (Figure 33A). A comparison of CFP1 and GFP-CFP1 peak sets genome-wide revealed 75% of GFP-CFP1 peaks were also detected in the profile of endogenous CFP1 (Figure 33B). Conversely, 92% of endogenous CFP1 peaks were confirmed by the fusion protein. Consistent with the endogenous peak set, GFP-CFP1 peaks were almost exclusively observed at detectable NMIs, supporting the proposal that the zf-CXXC domain is required for CFP1 enrichment on chromatin (Figure 33C).

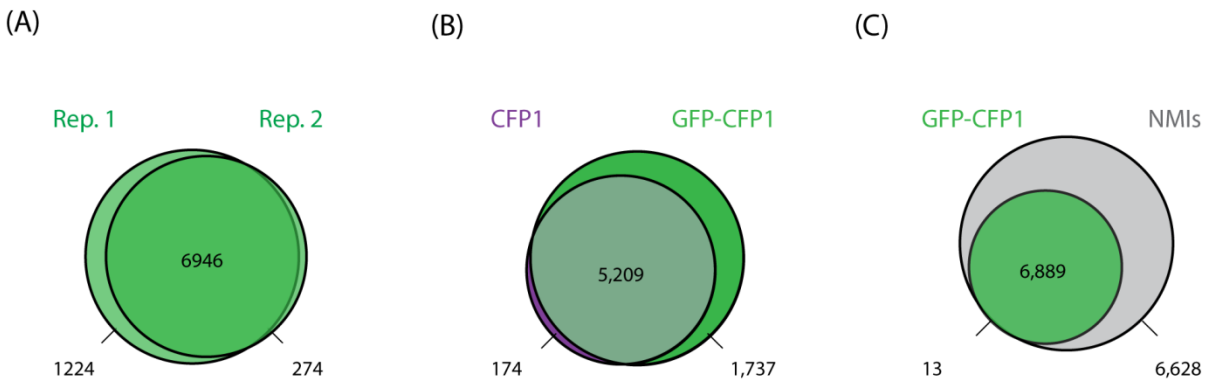



**Figure 33 - GFP-CFP1 binds to a subset of NMIs**

Proportional Venn diagrams depict the genome-wide overlap between various peak sets. (A) Intersection of GFP-CFP1 peaks detected in biological replicate ChIP experiments reveals an 82% overlap. Further analysis was performed on these replicated peaks. (B) 75% of replicated GFP-CFP1 peaks corresponded to peaks of endogenous CFP1. (C) GFP-CFP1 peaks were almost exclusive to NMI regions, but were only observed at 51%.

## 5.11 Mechanistic dissection of CFP1 binding to chromatin

CFP1 contains a zf-CXXC domain which binds specifically to non-methylated CpG dinucleotides (Xu et al. 2011). This domain appears to be essential for CFP1 binding based on its exclusive localisation to non-methylated regions (Figure 33C). Consistent with the ChIP-seq profile of endogenous CFP1, GFP-CFP1 enrichment is only observed at a subset of NMIs *in vivo*, suggesting that additional factors may be involved in its interaction with the genome. *In vitro* characterisation of the CFP1 PHD domain suggests it is able to recognise H3K4me3, potentially providing a second mode of interaction with chromatin.

To understand how CFP1 interacts with chromatin *in vivo*, stable C127 cell lines were generated which express mutant versions of GFP-CFP1 at comparable levels to the wild type GFP-CFP1 (Figure 34). One of the mutant transgenes expressed includes a single point mutation in the zf-CXXC domain which has been shown to disrupt interaction with non-methylated DNA (CXXC*). The W49A PHD mutant (PHD*) described above was expressed in another line, and a third cell line expressed double

mutant CFP1 combining both zf-CXXC and PHD mutations (DB*). The α-GFP antibody was used to profile chromatin binding of the CFP1 transgene in the wild type and mutant GFP-CFP1 lines by ChIP-sequencing of biological duplicates. Following alignment to the mouse genome, sequencing reads from each experiment were randomly downsampled resulting in an equal number of unambiguously aligned reads across the different cell lines. This allowed the profiles of wild type or mutant GFP-CFP1 to be compared.

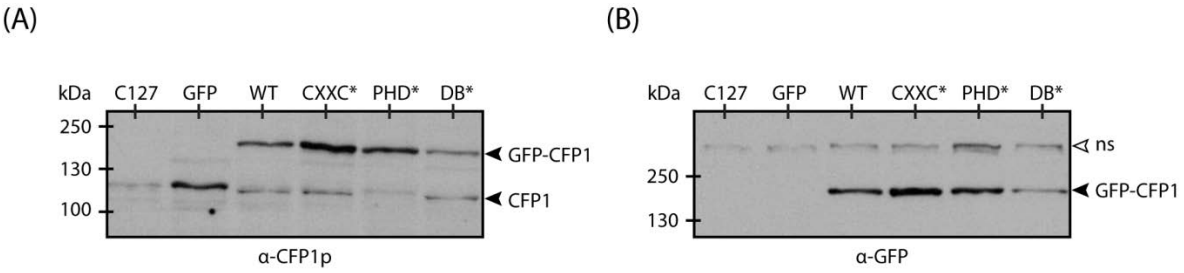



**Figure 34 – Expression of wild type and mutant GFP-CFP1**

Whole cell extract from parental C127, GFP only, WT, CXXC*, PHD* and DB* cell lines were analysed by western blot using (A) α-CFP1p and (B) α-GFP antibodies. Endogenous CFP1, GFP-CFP1 and a non-specific cross reactive band (ns) are indicated by arrowheads.

Loss of signal from CFP1 peaks was immediately obvious from a visual inspection of the mutant profiles (Figure 35A). Wild type CFP1 peaks were completely absent in the zf-CXXC mutant, suggesting that the zf-CXXC domain is absolutely required for enrichment of CFP1 at NMI chromatin. Unlike mutation of the zf-CXXC domain, mutation of the PHD domain did not prevent binding to non-methylated DNA as clear peaks of enrichment were still observed at NMI regions (Figure 35A). However, PHD* enrichment levels were substantially lower than wild type GFP-CFP1.

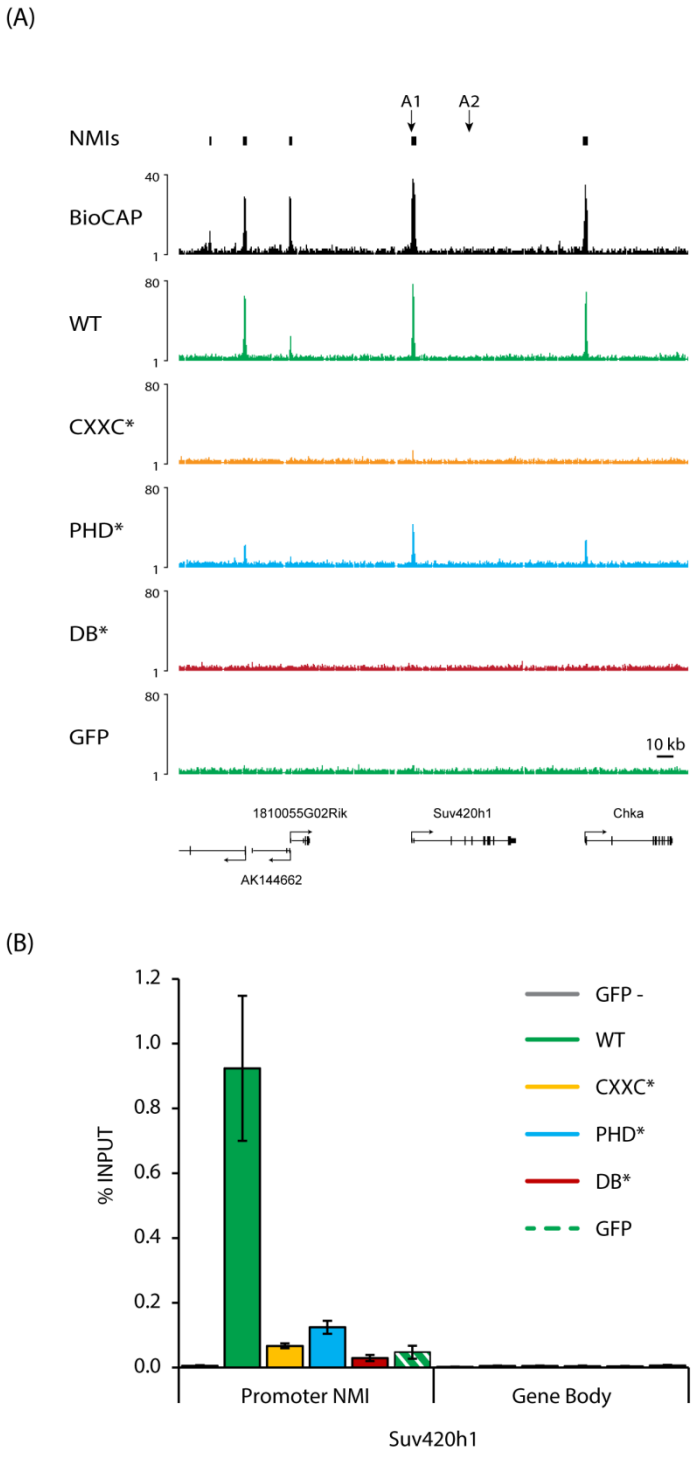



**Figure 35 – CFP1 binding is abolished by mutation of the zf-CXXC domain, and substantially reduced by mutation of the PHD domain**

(A) Profiles of non-methylated DNA (BioCAP), wild type or mutant GFP-CFP1 and free GFP are shown across a region of chromosome 19 containing the Suv420h1 locus. Chromatin was prepared in parallel from WT, CXXC*, PHD*, DB*, and GFP only cell lines and ChIP-sequenced using α-GFP. CXXC* contains the R200A zf-CXXC mutation preventing binding to non-methylated DNA. PHD* contains

the W49A PHD mutation which prevents binding to H3K4me3. DB* contains W49A & R200A mutations. Profiles were downsampled to 35.8 million reads. Promoter NMI (A1) and gene body (A2) amplicons used for ChIP-qPCR are indicated above the profiles. (B) Quantitation of GFP-CFP1 enrichment at the Suv420h1 locus by qPCR (n=5, standard error bars).

To validate and quantify the effect of CXXC* and PHD* mutations on CFP1 binding observed in the ChIP-sequencing experiments, enrichments of wild type and mutant GFP-CFP1 were measured at the Suv420h1 locus by q-PCR (Figure 36B). Based on five biological replicates, CXXC* enrichment at the promoter NMI (A1), was ~13-fold lower than wild type GFP-CFP1. A two–tailed Student's t-test indicated that this difference is highly unlikely to have occurred by chance ($p<0.01$) consistent with the critical role of the zf-CXXC domain in the interaction between CFP1 and chromatin. Background levels of wild type or mutant GFP-CFP1, measured at a control amplicon (A2) within the gene body, did not differ significantly from free GFP ($p>0.05$, using a two tailed Student's t-test on five matched biological replicates) suggesting that the differences observed at the promoter NMI reflect specific differences in chromatin binding . At the NMI promoter, no significant difference ($p>0.05$) was observed between CXXC* and the free GFP control, indicating that specific binding to NMI chromatin is completely abrogated by the zf-CXXC mutation (R200A). Complete loss of binding in the CXXC* mutant supports the proposal that the zf-CXXC domain is absolutely required for specific enrichment of CFP1. Although binding to the NMI region was not completely abrogated by mutation of the PHD* domain, a 6-fold reduction (85% loss) in binding was observed suggesting that the PHD domain also contributes significantly ($p<0.01$) to the interaction between CFP1 and chromatin.

Given that q-PCR based analysis quantitatively validates the loss of binding in the mutants observed in ChIP-seq at the Suv420h1 locus, the effects of CXXC* and PHD* mutations were examined in more detail at the genome scale by comparing average enrichments of wild type or mutant GFP-CFP1 across all NMIs (Figure 36C). The average enrichment in the CXXC* line was similar to the free GFP line suggesting that the zf-CXXC domain is required throughout the genome for CFP1 binding to NMI

chromatin. The average enrichment in the PHD* line was below that of wild type GFP-CFP1 but above free GFP suggesting that the PHD domain clearly contributes to CFP1 binding at NMIs genome-wide and suggests that the zf-CXXC domain is insufficient to dictate normal CFP1 occupancy on chromatin. Similarly, when binding of the CXXC* and PHD* mutants was compared across all wild type CFP1 peaks there was a clear loss of binding in the zf-CXXC mutant and a degree of retention in the PHD mutant (Figure 36D). This analysis suggests that while the CFP1 zf-CXXC domain is required for normal chromatin binding, it is insufficient for normal levels of CFP1 binding at target sites.

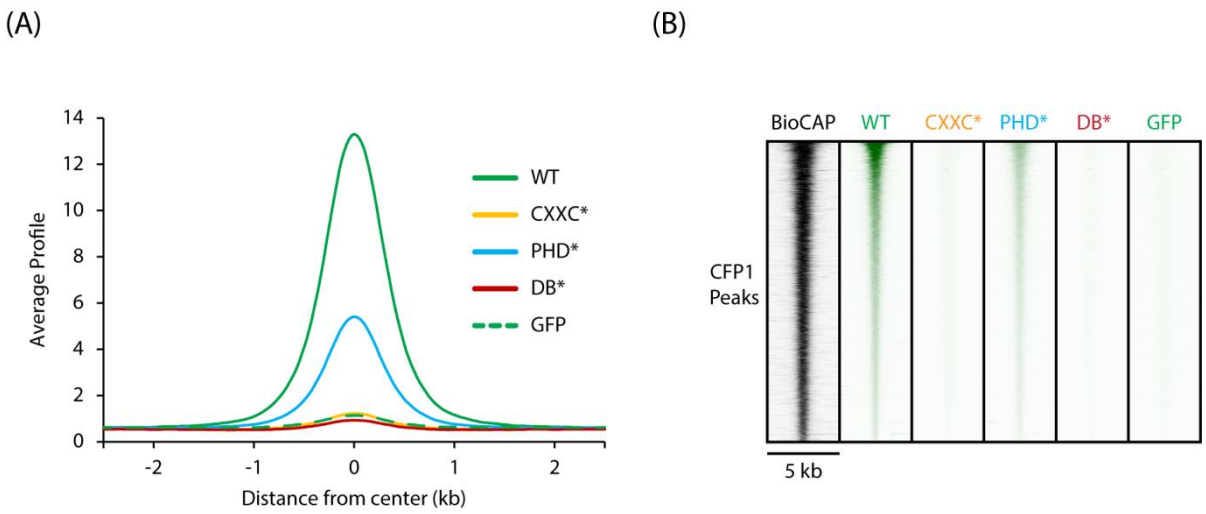



**Figure 36 – The CFP1 PHD domain contributes to binding genome-wide**

(A) The average profile of GFP-CFP1 is shown across all non-methylated islands for wild type or mutant GFP-CFP1. (B) Heat maps illustrate the variation in bioCAP signal, and enrichment of wild type or mutant GFP-CFP1 across all CFP1 peaks, sorted by bioCAP read count.

## 5.12 Genome-wide profiling of H3K4me3 reveals CFP1 at a subset of H3K4me3 regions

Comparison of wild type and CXXC* GFP-CFP1 suggests that CFP1 requires an intact zf-CXXC domain to bind to chromatin, however the PHD domain was also found to contribute significantly to the observed level of enrichment (Figure 36). This contribution is consistent with *in vitro* evidence that the CFP1 PHD domain binds specifically to H3K4me3 (Eberl et al. 2012), and with previous work *in vivo* suggesting that 93% of CFP1 peaks correspond to peaks of H3K4me3 enrichment (Thomson et

al. 2010). To understand how CFP1 localisation is influenced by this histone modification, H3K4me3 was profiled in C127 cells by ChIP-sequencing (Figure 36D). Visual inspection of the H3K4me3 profile revealed clear peaks of enrichment at promoter regions. The MACS algorithm detected 10,790 H3K4me3 peaks of which 76% were within 1 kb of an annotated TSS (Table 15). The observed association with TSSs is consistent with previous reports suggesting that H3K4me3 is a landmark of promoter regions (Bernstein et al. 2005; Guenther et al. 2007; Heintzman et al. 2007). H3K4me3 peaks were almost exclusively observed at NMI regions, consistent with a contribution of non-methylated DNA to H3K4me3 deposition (Table 15 and Figure 37B). Conversely however, H3K4me3 peaks were not detected at all NMIs indicating that non-methylated DNA is not sufficient for enrichment of H3K4me3.

**Table 13 – Summary of overlap between H3K4me3 peaks and other genomic features**

| | % of H3K4me3 Peaks | % with H3K4me3 Peaks |
|---|---|---|
| TSSs | 76 | 36 |
| NMIs | 82 | 65 |
| Promoter NMIs | 75 | 84 |
| GFP-CFP1 Peaks | 59 | 91 |
| Pol II Peaks | 78 | 69 |

To investigate the interaction between CFP1 and H3K4me3, the overlap between their peak sets was assessed genome-wide (Table 15). H3K4me3 peaks were observed at almost all CFP1 peaks (Figure 37C), consistent with the proposal that CFP1 directs H3K4me3 placement through its interaction with the SET1A/B methyltransferases (J.-H. Lee et al. 2007; Lee & Skalnik 2005). Conversely, CFP1 peaks were observed at a smaller fraction of H3K4me3 peaks indicating that enrichment of H3K4me3 is not sufficient for enrichment of CFP1. This is consistent with the total loss of binding in the zf-CXXC mutant GFP-CFP1 despite containing an intact PHD domain. Furthermore as H3K4me3 is observed in its absence, enrichment of CFP1 is not essential for H3K4me3.

To examine the relationship between CFP1 and H3K4me3 in more detail, CFP1 read density was compared to H3K4me3 density across all TSSs (Figure 37D). Surprisingly, although H3K4me3 enrichment was observed at the majority of CFP1 peaks, the correlation between CFP1 and H3K4me3 was very weak (r= 0.31 +/-0.01, n= 14,627, $p<2.2x10^{-16}$). This is a counterintuitive result considering *in vitro* and *in vivo* evidence which suggest that the CFP1 PHD domain binds specifically to H3K4me3. Similarly, the weak relationship between the magnitude of CFP1 and H3K4me3 seems inconsistent with the proposal that CFP1 directs H3K4me3 deposition.

Poor correlation between the magnitude of CFP1 and H3K4me3 enrichment may indicate an issue with the initial H3K4me3 ChIP. Although the punctate profile and specific enrichment at gene promoters is consistent with previous reports, the number of uniquely aligned H3K4me3 reads (~2,000,000) was very low. ChIP-sequencing of H3K4me3 C127 to a greater depth (~20,000,000 reads) is necessary to assess the correlation between CFP1 and H3K4me3.

Despite the low H3K4me3 read count in this initial comparison, its weak correlation with CFP1 may indicate that their relationship is more complicated than expected. To aid interpretation this result, H3K4me3 read density was also compared to Pol II density across all TSSs (Figure 37D). Consistent with previous studies (Heintzman et al. 2007), a strong positive correlation (r= 0.62 +/-0.01, n= 14,627, $p<2.2x10^{-16}$) was observed between enrichment of H3K4me3 and Pol II across all TSSs. While the correlation between CFP1 and H3K4me3 may be underestimated due to the low H3K4me3 read count, strong correlations, as between H3K4me3 and Pol II, are still evident. This suggests that despite specific binding of H3K4me3 by the CFP1 PHD domain *in vitro*, NMIs with a high level of H3K4me3 are not always highly enriched with CFP1.

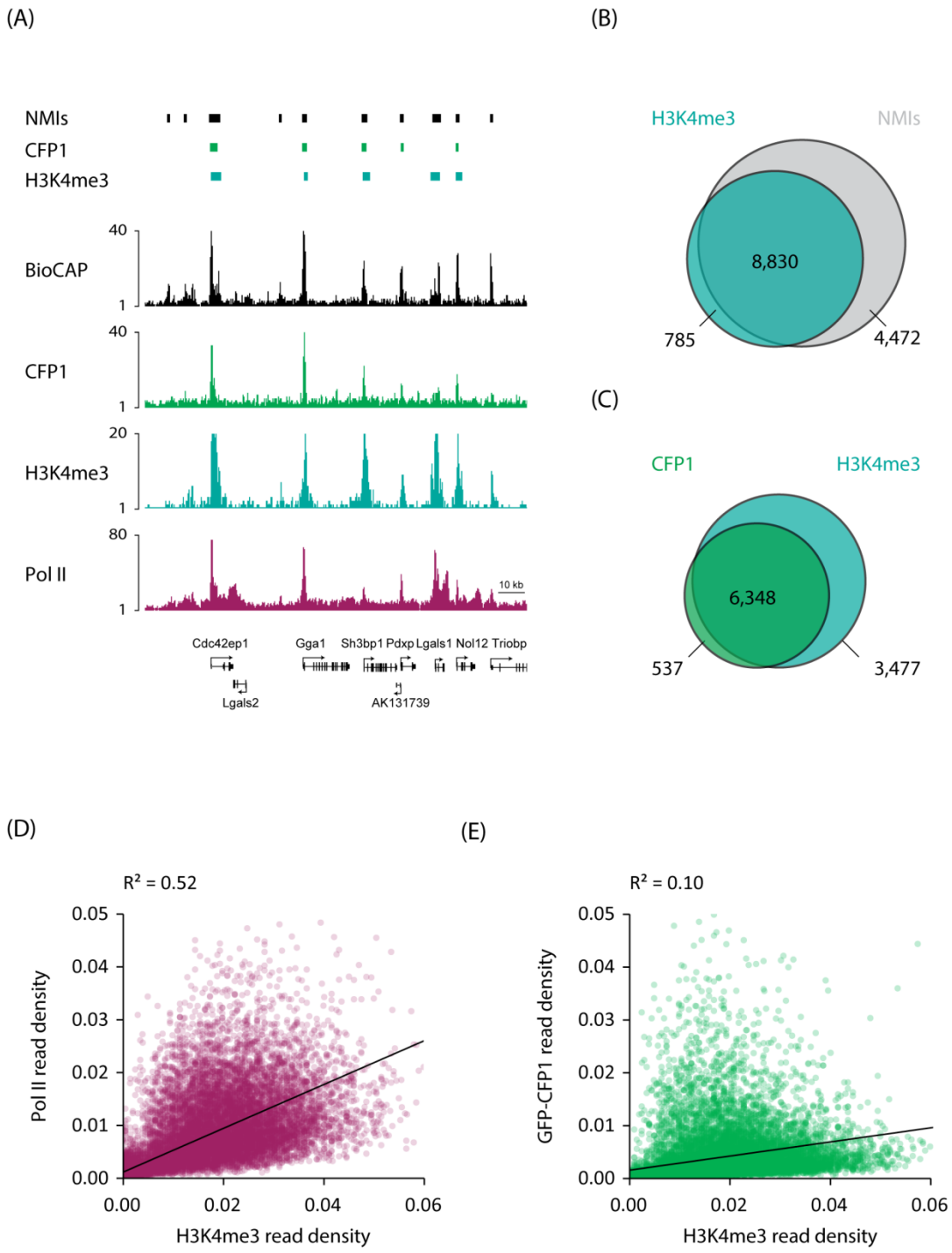



**Figure 37 – H3K4me3 is associated with active TSSs but correlates poorly with CFP1 levels**

(A) Profiles of non-methylated DNA (BioCAP), GFP-CFP1, H3K4me3 and Pol II are shown across a ~200 kb region. This region contains nine annotated TSSs and ten NMIs. Peaks of H3K4me3 generally correspond to promoter NMIs. (B) The genome-wide overlap between peaks of H3K4me3 enrichment and non-methylated islands (NMIs) is depicted as a proportional Venn diagram. (C) Overlap between CFP1 and H3K4me3. (D) The relationship between Pol II and H3K4me3 levels at

C127 NMIs is depicted as a scatter plot. The Pearson's correlation coefficient (r = 0.62) suggests a moderate positive correlation. (E) The relationship between CFP1 and H3K4me3 levels at C127 NMIs. The Pearson's correlation coefficient (r = 0.32) suggests a weak positive correlation.

## 5.13 Discussion

To dissect the interaction between CFP1 and chromatin, the effects of independent point mutations in the zf-CXXC and PHD domains were investigated *in vivo* by ChIP-sequencing. The zf-CXXC domain appears to be a prerequisite for chromatin binding, as a point mutation preventing recognition of non-methylated CpGs by this domain abrogated binding genome-wide. Interestingly, a mutation in the CFP1 PHD domain drastically reduced its enrichment at NMIs throughout the genome, demonstrating a contribution of this domain to chromatin binding *in vivo*. Together these results suggest that the zf-CXXC domain restricts CFP1 to regions of the genome containing non-methylated DNA, while the PHD is required for normal levels of binding at these sites. The apparent role of the CFP1 PHD domain in stabilisation but not recruitment is consistent with previous work suggesting that BPTF which is recruited to chromatin by site specific transcription factors is stabilised on chromatin by recognition of the H3K4me3 via its PHD domain (Wysocka et al. 2006).

*In vitro* experiments described in this Chapter suggest that the CFP1 PHD domain binds specifically to di- and trimethylated H3K4. Specific recognition of H3K4me3 by the CFP1 PHD domain has since been reported by another lab (Eberl et al. 2012). As CFP1 is also thought to contribute to H3K4 methylation through its interaction with methyltransferase enzymes SET1A and B (Lee & Skalnik 2005; J.-H. Lee et al. 2007), recognition of H3K4me3 could potentially produce an amplification loop. A similar loop has previously been proposed for the mixed lineage leukemia (MLL1) proto-oncogene (Chang et al. 2010). MLL1 methyltransferase contains a zf-CXXC domain, and three PHD domains, one of which (PHD3) binds specifically to H3K4me3. Consistent with the effect of mutating the CFP1 PHD domain, mutations of MLL1 PHD3 which abrogated recognition of H3K4me3 *in vitro* were reported to reduce MLL1 occupancy at CGI promoters.

Interestingly, while PHD domains are also found in KDM2A and KDM2B, previous work in the Klose lab revealed that despite conservation of the zinc coordinating residues, the KDM2A PHD domain does not bind to histones (Zhou et al. 2012). Similarly, the PHD domain of KDM2B is unlikely to bind to chromatin as the hydrophobic residues which generally form the binding pocket are not conserved. Specific binding to H3K4me3 as well as non-methylated CpGs may help to localise CFP1 to a specific subset of NMI, potentially explaining why the genome-wide localisation of CFP1 differs from KDM2B. Despite several lines of evidence suggesting that the CFP1 PHD domain binds specifically to H3K4me3, the relationship between the level of CFP1 and the level H3K4me3 was very weak. Given that CFP1 enrichment was not observed at some NMIs despite enrichment of H3K4me3, the presence of non-methylated CpGs and H3K4me3 appears not to be sufficient for CFP1 binding.

CFP1 co-purifies with SET1A and SET1B enzymes suggesting a relatively stable interaction with the evolutionarily conserved SET1 complex (Lee & Skalnik 2005; J.-H. Lee et al. 2007; Miller et al. 2001; Ardehali et al. 2011). Like CFP1, most of the other subunits which make up the SET1 complex contain putative or bona fide binding domains for chromatin or the transcriptional machinery itself. These interactions may indirectly contribute to the localisation of CFP1 through its interactions with SET1A or SET1B. ASH2L which associates with both SET1A and B, binds weakly to DNA in a non-sequence specific fashion through a winged helix motif (Chen et al. 2011), while the Wdr82 subunit is thought to bind the RNA polymerase C-terminal domain (Lee & Skalnik 2008). Modulation of CFP1 localisation through its interactions with the SET1 complexes may explain its selective enrichment at a specific subset of NMIs and the weak relationship observed between enrichment of CFP1 and the H3K4me3 mark. The interaction between CFP1 and SET1A *in vivo* is explored in Chapter 7. Chapter 6 investigates the contribution of the zf-CXXC and PHD domains to other aspects of CFP1 function.

# Chapter 6 - Dissecting the cellular distribution and dynamics of CFP1 by fluorescence microscopy

## 6.1 Introduction

In the previous chapters the localisation of CFP1 has been considered in a single dimension relative to sequence of the genome. However inside the nucleus, genomic DNA is wound around histone octamers forming a flexible array of nucleosomes (Kornberg & Thomas 1974; Thomas & Kornberg 1975). Traditionally, the nucleosome array was proposed to fold into a regular higher order structure however recent studies have demonstrated that even when condensed during mitosis, chromatin fibres exist in a highly disordered and interdigitated state (Eltsov et al. 2008; Nishino et al. 2012), forming an entangled polymer of genetic information. Attempts to reconcile data from microscopy and chromosome conformation capture techniques have resulted in a shift towards a probabilistic model of interphase chromatin organisation (Fudenberg & Mirny 2012). Understanding how the transcriptional machinery is able to recognise gene promoters amidst this apparent disorder remains a significant challenge for chromatin biology.

CpG island regions (CGIs) are distinguished from the surrounding sequence by their elevated GC content and a high density of CpG dinucleotides (Bird 1987). However as discussed in previous chapters, CGIs which remain non-methylated (NMIs) are associated with a unique chromatin architecture (Blackledge et al. 2010; Thomson et al. 2010). This may distinguish NMIs from the surrounding chromatin allowing them to be recognised by transcriptional machinery or by other regulatory factors. To understand how CFP1 interacts with NMIs in the context of the surrounding nucleosomal landscape, genome-wide snap shots of binding interactions revealed by ChIP-seq must be complemented by spatial and temporal details. This chapter investigates the contribution of the zf-CXXC and PHD domains to the distribution and dynamics of CFP1 in single cells to understand how it interacts with chromatin *in vivo*.

## 6.2 GFP-CFP1 is nuclear but non-nucleolar, and forms distinct foci in the nucleus

To begin exploring the distribution of CFP1 inside the nucleus, C127 cells stably expressing GFP-CFP1 were examined by fluorescence microscopy (Figure 38). Fixed cells were counterstained with DAPI revealing stereotypical bright heterochromatic domains as well as weakly staining nucleoli. GFP-CFP1 was clearly localised to the nucleus, unlike free GFP which was distributed throughout the cell.

Interestingly, bright GFP-CFP1 foci were observed in the majority of nuclei (Figure 38B). No foci were observed in cells expressing free GFP. To ensure that the foci did not result from the introduction of an N-terminal GFP tag, a complementary cell line was generated which stably expressed CFP1 tagged at the C-terminal (CFP1-GFP). CFP1-GFP recapitulated the nuclear distribution of CFP1, and confirmed the presence of bright CFP1 foci (Figure 38B). To estimate the frequency of CFP1 foci within a population of cells, the number of foci in each cell was recorded across multiple fields of view. CFP1 foci were observed in ~60% of cells (total=183 cells) regardless of which end of CFP1 was tagged with GFP (Figure 38C). Approximately 30% of cells contained more than 1 focus, with some cells containing as many as 7 (Figure 38D). These results suggested that the bright foci observed are a general feature of CFP1 distribution in C127 cells, however further work was necessary to understand their relevance to CFP1 function.

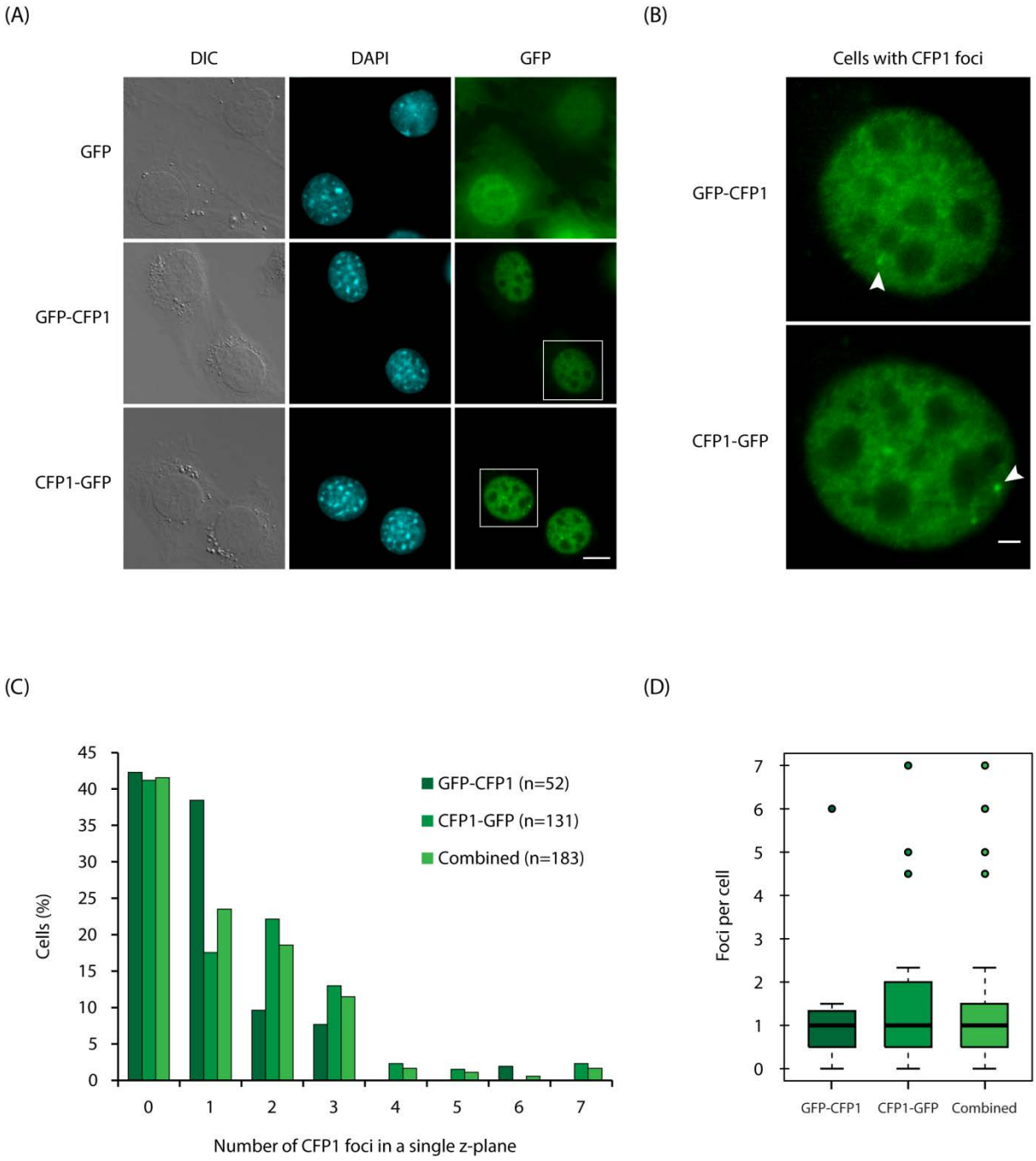



**Figure 38 – GFP-CFP1 is nuclear but non-nucleolar and forms bright, punctate foci**

(A) C127 cells stably expressing GFP-tagged CFP1 were fixed, counterstained with DAPI, and compared by fluorescence microscopy. Cells expressing free GFP are shown for comparison. A 10 µm scale bar is provided in the lower right hand corner. (B) Examples of bright foci are indicated by white arrowheads. The scale bar indicates 2 µm. (C) CFP1 foci were counted in images of 52 cells expressing GFP-CFP1, and 132 cells expressing CFP1-GFP. (D) An average of 1.2 CFP1 foci were observed per cell.

## 6.3 GFP-CFP1 recapitulates the spatial distribution of endogenous CFP1 inside the nucleus

To verify that GFP-CFP1 behaves like endogenous CFP1, C127 cells were immunostained with the α-CFP1f antibody generated and characterised in Chapter 3 (Figure 39). The α-CFP1f staining suggested that endogenous CFP1 was predominantly nuclear but non-nucleolar, in the parental C127 cell line, validating the distribution suggested by imaging GFP-CFP1. Similarly, endogenous CFP1 formed bright foci in parental C127s. To directly compare GFP-CFP1 distribution to endogenous CFP1, α-CFP1f immunostaining of the GFP-CFP1 cell line was performed in parallel. GFP-CFP1 foci in each cell (n=68) were counted and compared to the number of foci visible using the α-CFP1f stain (Figure 39C). Slightly higher numbers of foci were observed in the α-CFP1f staining, perhaps because recognition of CFP1 by primary and secondary antibodies amplifies the signal resulting in greater contrast than the GFP-CFP1 images. Nevertheless, a positive linear relationship was observed between the fluorescence intensities of GFP-CFP1 and α-CFP1f across the foci observed ($R^2$=0.90, n=259 foci). This relationship was not substantially affected by normalisation to the average fluorescence intensity of the whole nucleus (Figure 39B).

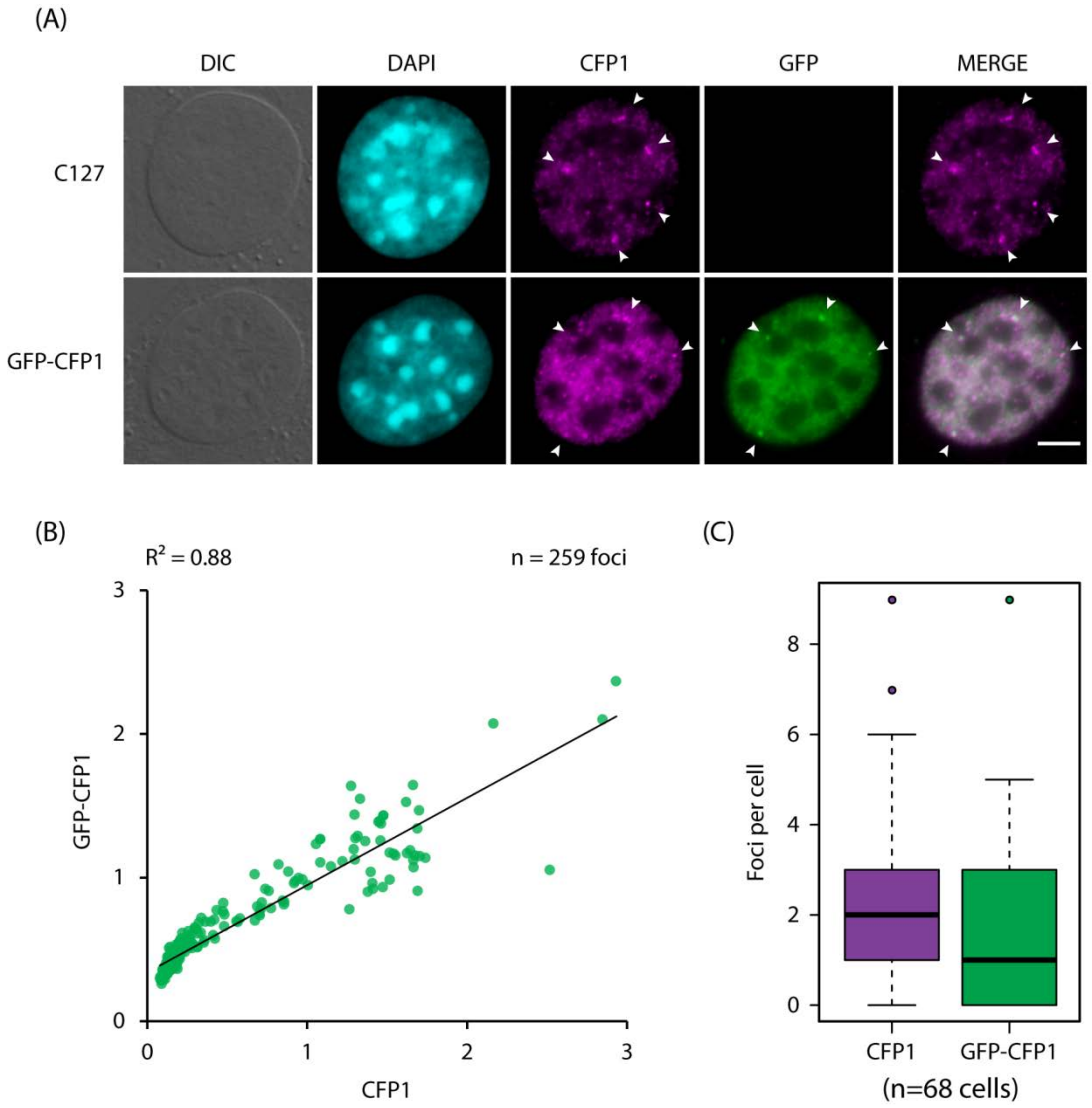



**Figure 39 – Bright CFP1 foci are formed by endogenous CFP1 as well as the GFP-fusion**

(A) C127 cells were fixed with 2% formaldehyde for α-CFP1f immunostaining. Staining was performed on parental C127 cells to examine the distribution of endogenous CFP1, and on the GFP-CFP1 cell line to compare α-CFP1f staining to the GFP signal. Bright foci are indicated by white arrowheads. The scale bar represents 5 µm. (B) The fluorescence intensities measured at each focus displayed a positive linear relationship, consistent with the visible similarity between α-CFP1f staining and the GFP-CFP1. (C) The number of foci visible in the α-CFP1f or GFP-CFP1 signal was counted per cell. The variation in numbers of foci is shown as a box and whisker plot.

### 6.4 CFP1 localises to euchromatic domains inside the nucleus

To investigate the spatial distribution of CFP1 at higher resolution, fixed C127s expressing GFP-CFP1 were imaged using three-dimensional structured illumination microscopy (3D-SIM). This technique improves the resolution to ~120 nm in the xy plane and ~300 nm in the z axis. Super-resolution 3D-SIM imaging highlighted the exclusion of CFP1 from peripheral and pericentric heterochromatin as well as nucleoli. Spatial exclusion of CFP1 from heterochromatin is consistent with wide field images from previous studies which suggest that CFP1 co-localises with acetylated H3 and H4 histones which are hallmarks of euchromatin (Lee & Skalnik 2002; Lee & Skalnik 2005). At super resolution CFP1 appeared non-randomly distributed with respect to the DAPI counterstain, localising to the interface between chromatin and the inter-chromatin space. Importantly, the observed distribution of CFP1 is consistent with the distribution of H3K4me3 reported by previous studies which imaged C127 cells at super-resolution (Markaki et al. 2010; Smeets et al. 2014).

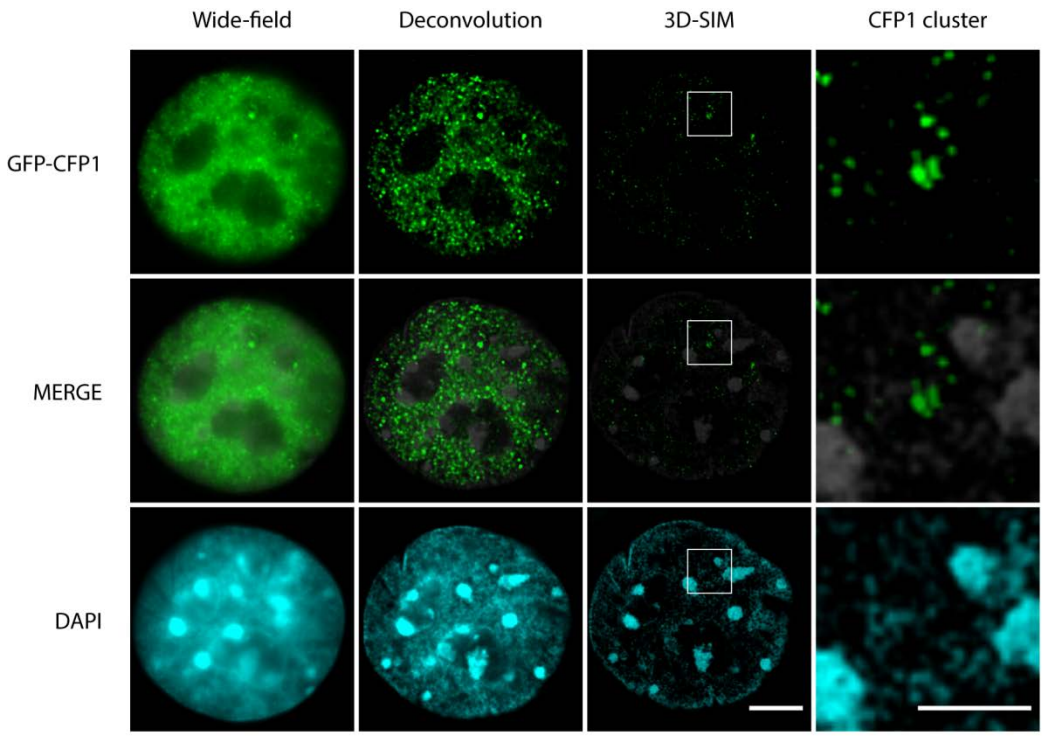


**Figure 40 – CFP1 forms bright foci where several smaller foci cluster together**

Three dimensional structured illumination microscopy (3D-SIM) imaging of formaldehyde fixed C127 cells expressing GFP-CFP1. Wide-field and deconvolved images of the same cell were generated from the raw 3D-SIM data to illustrate the improvement in resolution. A close up of a bright CFP1 focus is shown on the right, corresponding to the boxed region. Each image depicts a single z-plane. Scale bars indicate 5 µm, or 2 µm for the inserts.

While the diffuse background glow of nuclei imaged by wide field microscopy suggests that CFP1 is broadly distributed across chromatin, super resolution 3D-SIM imaging suggests that CFP1 only interacts with a small fraction of nuclear chromatin (Figure 40). The ability to discern discrete spots of CFP1 is consistent with CFP1 enrichment at ~5,000 sites, comprising ~0.1% of the genome. Improved resolution reveals the bright CFP1 foci are comprised of several smaller foci clustered together. This suggests that the bright foci visible in live and fixed cells do not represent huge enrichment of CFP1 at a single site, but indicate enrichment at several sites clustered in close proximity. Further experiments are now required to determine how the spatial organisation of CFP1 inside the nucleus relates to the location of non-methylated islands, and to understand why some CFP1 sites appear to cluster together.

## 6.5 Mutation of the PHD domain eliminates clustering of CFP1 sites

To determine whether the zf-CXXC and PHD domains influence the nuclear distribution of CFP1, C127 cell lines expressing wild type and mutant GFP-CFP1 were compared by wide field fluorescence microscopy (Figure 41). Wild type and mutant GFP-CFP1 were previously shown to be expressed at comparable levels in the cell lines used for this analysis. Compared to the wild type fusion, mutation of the zf-CXXC domain (CXXC*) had no discernible effect on CFP1 distribution. Surprisingly, CFP1 clusters were completely absent in cells expressing the PHD mutant (PHD*). Similarly, no clusters were seen in cells expressing the combined zf-CXXC, and PHD double mutant (DB*). To check that

this effect was not specific to the W49A mutation, a new stable cell line was generated expressing GFP-CFP1 with the alternative PHD mutation (Y28A) characterised in Chapter 5. Western blot analysis suggests that the level of Y28A GFP-CFP1 expression is comparable to the wild type fusion protein (Figure 41B and C). In agreement with the W49A mutation, no clusters were observed in the Y28A line suggesting that mutation of the PHD domain results in loss of CFP1 clusters (Figure 41).

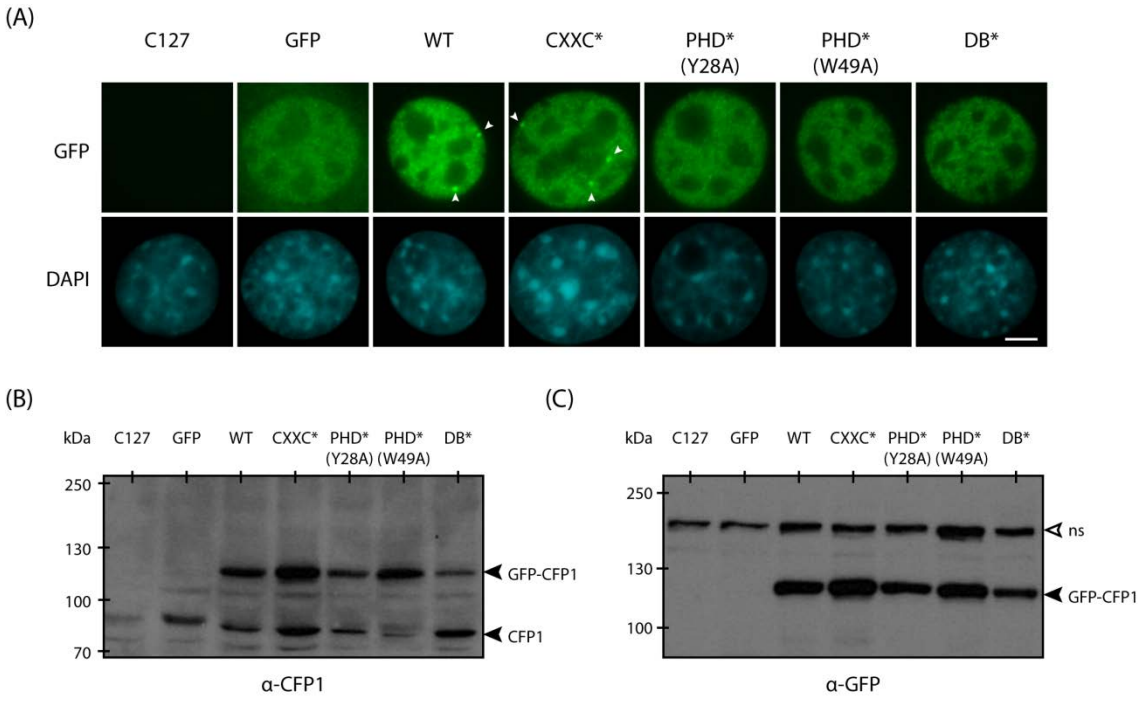



**Figure 41 - Mutation of the PHD domain eliminates nuclear CFP1 clusters**

(A) Stable cell lines expressing wild type or mutant GFP-CFP1 were fixed, counterstained with DAPI, and imaged by immunofluorescence microscopy. The GFP signal was amplified by immunostaining with a monoclonal antibody against GFP, and a secondary antibody conjugated to Alexa488. Parental C127 cells, and cells expressing free GFP, were imaged in parallel to control for non-specific immunostaining and non-specific localisation. Cells were counterstained with DAPI. Examples of bright CFP1 clusters are indicated by white arrowheads. Clusters were not observed in the PHD* or DB* cell lines. Scale bar indicates 5 µm. (B and C) Nuclear extracts from the cell lines imaged in (A) were analysed by western blot.

The severe effect of mutations in the PHD domain on CFP1 localisation was unexpected given that ChIP-seq analysis showed specific enrichment of PHD* CFP1 at non-methylated islands. To further investigate the impact of PHD* mutations on the nuclear distribution of CFP1, wild type and PHD* mutant GFP-CFP1 were compared by 3D-SIM (Figure 42). This analysis revealed that while mutation of the PHD domain resulted in the absence of CFP1 clusters, discreet spots of CFP1 were still observed, consistent with the residual ability of PHD* CFP1 to bind to chromatin containing non-methylated CpGs.

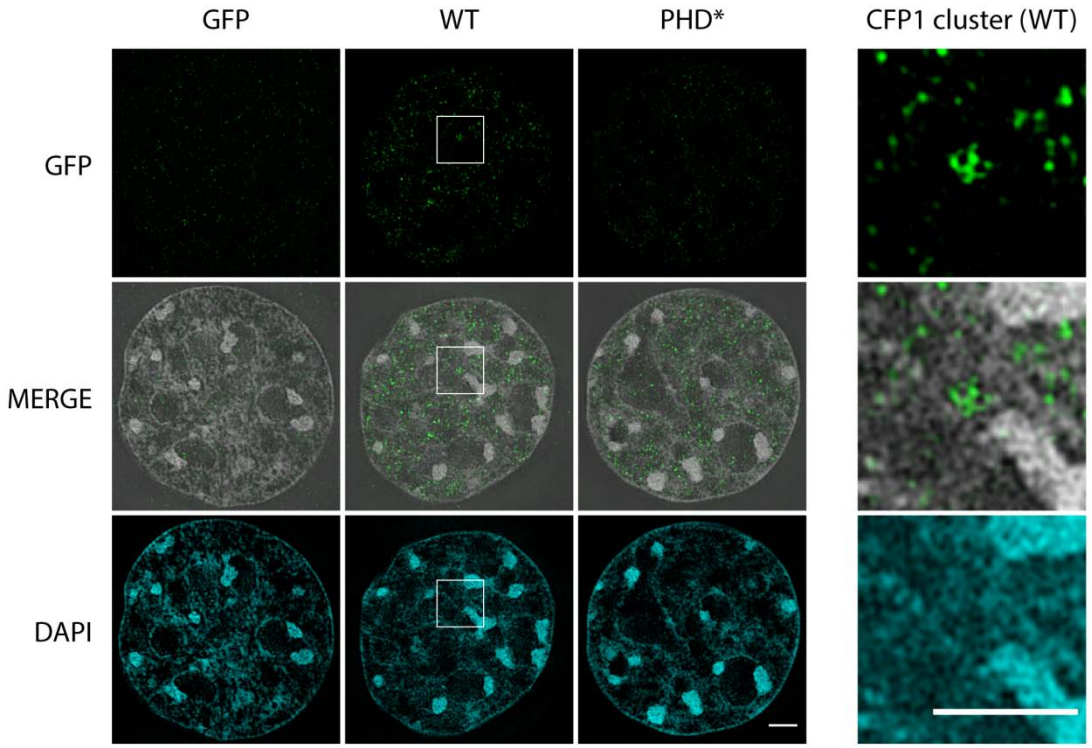



**Figure 42 – Mutation of the PHD domain abolishes CFP1 clusters but does not affect its global distribution in the nucleus.**

3D-SIM imaging of formaldehyde fixed C127 cells expressing free GFP, wild type GFP-CFP1 or GFP-CFP1 containing a mutation in the PHD domain (PHD*). A close up of a bright CFP1 cluster is shown on the right, corresponding to the boxed region. Each image depicts a single z-plane. Scale bars indicate 5 µm, or 2 µm for the inserts. Although clusters were abolished by mutation of the PHD domain, PHD* is still strictly nuclear forming discreet spots outside of heterochromatic regions.

## 6.6 Recognition of H3K4me3 by the CFP1 PHD domain may play a role in the clustering of CFP1 sites

Despite evidence from this study and from Eberl et al. that the CFP1 PHD domain is able to recognise H3K4me3 (Eberl et al. 2012), ChIP-sequencing experiments described in the previous chapter suggest that the PHD domain is dispensable for the specific recognition of non-methylated islands. Similarly, the correlation between CFP1 levels and the level of H3K4me3 was very weak, suggesting H3K4me3 may play a minor role in the genome-wide localisation of CFP1. Conversely, as the PHD domain has a recognisable effect on the distribution of CFP1, interaction with H3K4me3 may be important for the spatial organisation of CFP1 inside the nucleus.

To determine whether PHD-dependent effects on CFP1 localisation were dependent on H3K4me3, cells expressing GFP-CFP1 were transiently transfected with a plasmid vector driving overexpression of a H3K4 demethylase (RBP2). Previous work in the Klose lab has demonstrated that overexpression of RBP2 results in acute global depletion of the H3K4me3 mark. The distribution of GFP-CFP1 in H3K4me3 depleted cells, identified by staining RBP2, was then compared to the distribution in their untransfected neighbours by fluorescence microscopy. Consistent with the importance of the PHD domain, no foci were observed in cell expressing RBP2 (Figure 43). Unfortunately, due to the low transfection efficiency of C127 cells, very few H3K4me3 depleted cells were observed in this initial experiment. Future experiments will aim to validate these results with higher transfection efficiency.

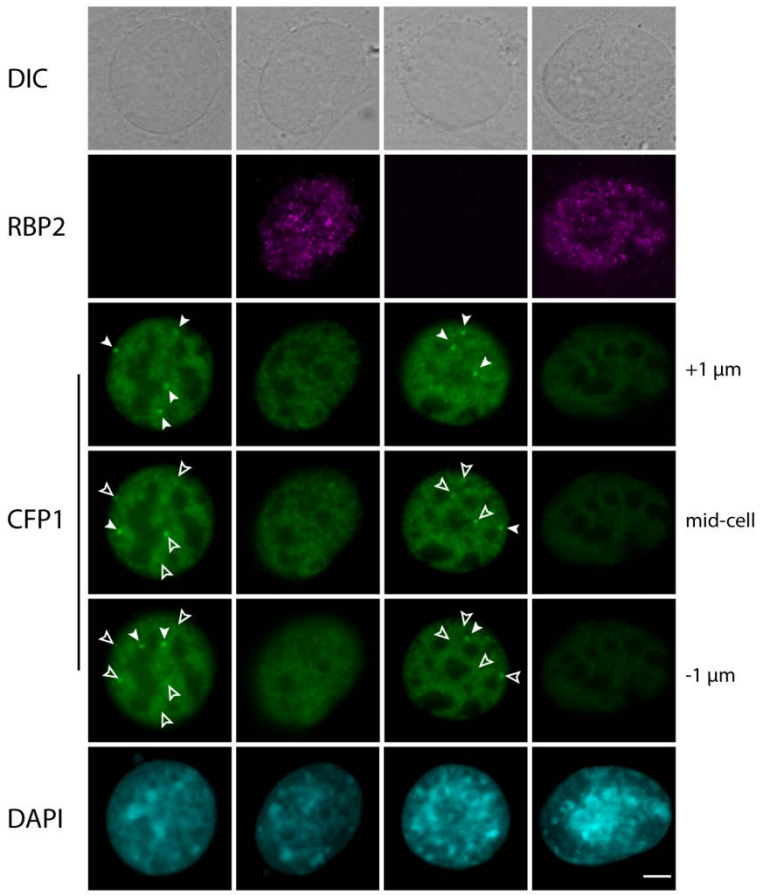



**Figure 43 – CFP1 clusters were absent in cells overexpressing the H3K4me3 demethylase RBP2**
C127 cells stably expressing GFP-CFP1 were transfected with Flag-RBP2. After 24 h cells were fixed and stained with α-Flag to identify cells expressing the demethylase. Optical sections through the mid-cell and 1 µm either side are shown from representative cells. Foci within the section are indicated by filled arrowheads, while the positions of foci in other sections are indicated by clear arrowheads.

## 6.7 CFP1 is highly dynamic inside the nucleus

CFP1 interacts with the genome at NMI regions (Thomson et al. 2010) and appears to localise to euchromatic domains within the nucleus (Lee & Skalnik 2002; Lee & Skalnik 2005). These investigations were carried out using formaldehyde fixed cells, offering a snap shot of CFP1 localisation. To understand how CFP1 interacts with chromatin *in vivo*, its nuclear dynamics were investigated in live cells. Having demonstrated that GFP-CFP1 effectively recapitulates the localisation of endogenous CFP1 genome-wide, and its spatial distribution throughout the cell, the

GFP-CFP1 line was examined in live cell experiments using a spinning disk fluorescence microscope (Figure 44). This instrument is well suited to rapid imaging with high sensitivity thus minimising photo-damage to the cells. Importantly, the distribution of CFP1 observed in live cells, closely resembled that of the fixed cell images. This included the presence of CFP1 clusters, which were stably positioned over many hours (data not shown).

The nuclear dynamics of GFP-CFP1 was investigated by measuring the fluorescence recovery after photobleaching (FRAP). For each individual FRAP measurement, GFP molecules within a small region of the nucleus were permanently bleached using a brief pulse of excitation at high laser intensity. The fluorescence signal in the bleached region relative to its pre-bleach intensity was monitored over time, describing the movement of unbleached, GFP-CFP1 into the region from elsewhere in the nucleus. Interactions with a relatively static structure such as chromatin delay fluorescence recovery as the protein must dissociate from its binding site, before it is able to diffuse around the nucleus equilibrating with an unbleached protein. Initial experiments bleaching a circular region within the nucleus, revealed that CFP1 is highly dynamic (Figure 44), with fluorescence recovering almost completely after 20 seconds. FRAP experiments were also performed on regions containing a bright CFP1 cluster (Figure 44B). Partial fluorescence recovery of the bright clusters is visible after 10 seconds revealing that CFP1 turnover also occurs at these sites despite their stable positioning in the nucleus.

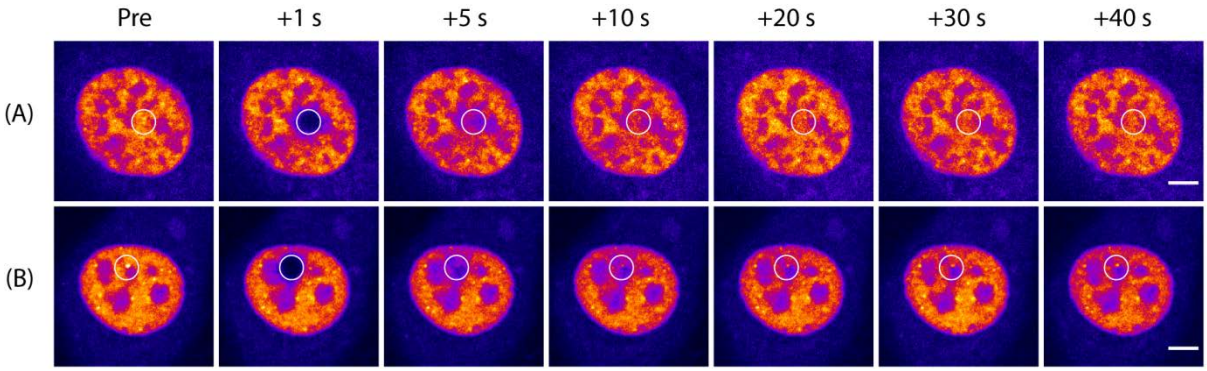


**Figure 44 – CFP1 is highly dynamic inside the nucleus**

The nuclear dynamics of CFP1 was investigated in live cells using fluorescence recovery after photobleaching (FRAP). Recovery of GFP-CFP1 fluorescence was monitored after photobleaching a circular region (radius =2 µm) of the nucleus (A) or a region containing a CFP1 cluster (B). Scale bars = 5 µm. Fluorescence in the bleached region approached its initial pre-bleach intensity within 1 minute indicating that CFP1 is highly dynamic within the nucleus.

To control for potential effects of attaching GFP to its N-terminal, GFP-CFP1 dynamics was initially compared to the C-terminal fusion (CFP1-GFP) (Figure 45A). Average fluorescence recovery curves, revealed a close agreement between the fusions, regardless of which terminus was linked to GFP (Figure 45). To aid interpretation of CFP1 dynamics, FRAP measurements of free GFP were performed in parallel. Importantly, GFP-CFP1 (Mw=104 kDa) recovered substantially slower than free GFP (Mw = 27 kDa). This difference is larger than expected based on the relative mass of the two proteins, implying that the recovery of GFP-CFP1 is slower than predicted by free diffusion. On average, the recovery of bright CFP1 clusters was slower than the surrounding regions (Figure 45B), supporting the suggestion that clusters represent an aggregation of interaction sites similar to those elsewhere in the nucleus.

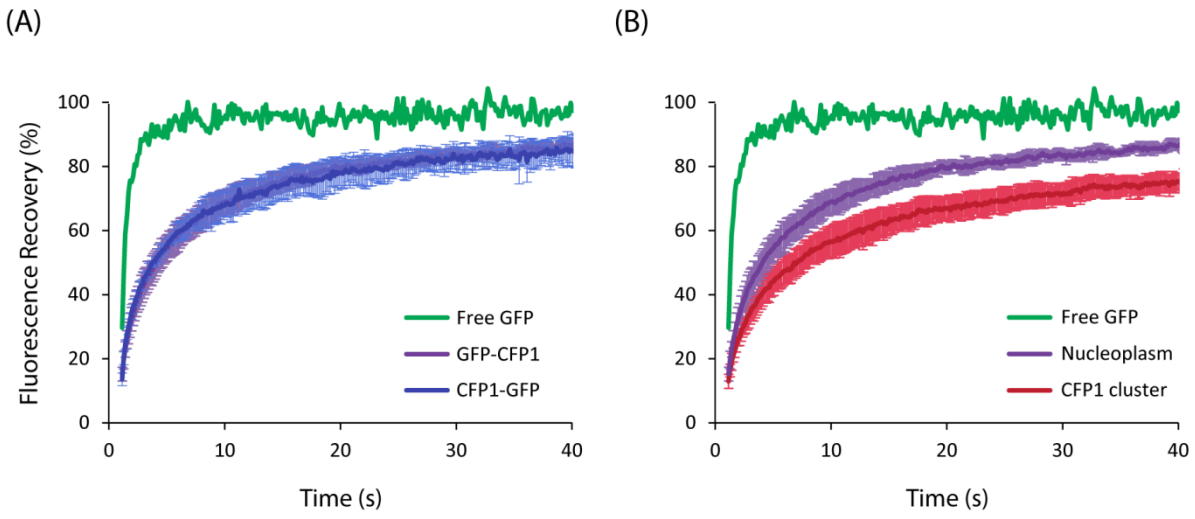


**Figure 45 – CFP1 is less dynamic than free GFP, reflecting its interactions with chromatin**
(A) The dynamics of N-terminal (GFP-CFP1) and C-terminal (CFP1-GFP) fusions were compared by measuring fluorescence recovery after photobleaching a small circular spot (radius = 2 µm). Mean recoveries and standard error bars are shown for GFP-CFP1 (n=5) or CFP1-GFP, (n=3). Close agreement is seen between N- and C-terminal fusions. The recovery of free GFP is shown for comparison (n=1). (B) The dynamics of GFP-CFP1in regions containing a bright CFP1 cluster (n=4), was then compared to the surrounding region. There was a clear difference between recovery of the CFP1 clusters, and recovery elsewhere in the nucleus.

## 6.8 The zf-CXXC and PHD domains both contribute to the nuclear dynamics of CFP1

Fluorescence microscopy and ChIP-sequencing experiments suggested that the zf-CXXC and PHD domain make different contributions to the nuclear distribution of CFP1, and to its interactions with NMI chromatin. To determine the contribution of each domain to CFP1 dynamics, FRAP experiments were conducted on cell lines expressing wild type or mutant GFP-CFP1. Four changes were made to the earlier FRAP protocol to optimise quantitation of CFP1 dynamics; 1) the acquisition rate was increased from 8 to 12 fps reducing the lag between bleaching and acquisition and more appropriately sampling the rapid initial recovery, 2) monitoring for longer (1,000 frames) after bleaching allowed more confident estimation of plateaued recovery, 3) bleaching a fixed spot instead of raster scanning a circular region drastically reduced the bleach time, and allowed synchronous bleaching of the target region. Spot bleaching also minimised the bleached volume reducing the contribution of diffusion to the recovery.

Using the optimised FRAP protocol, nuclear GFP-CFP1 recovered to >99% of its initial intensity within one minute, recovering to >90% within 10 seconds (Figure 46A). CXXC*, PHD* or both mutations in combination (DB*), accelerated the recovery of GFP-CFP1 during this rapid initial phase consistent with reduced binding to chromatin. Mutation of the zf-CXXC domain had a slightly larger effect than mutation of the PHD domain (Figure 46B and C), while no further acceleration was observed in the DB* mutant (Figure 46D). This implies that while the zf-CXXC domain can contribute to CFP1

dynamics independently of the PHD domain, the PHD domain has no effect on dynamics in the absence of the zf-CXXC domain. This result is consistent with ChIP-sequencing results which demonstrated that CXXC* is unable to bind NMIs genome-wide, while binding of PHD* was partially retained.

To quantify differences between wild type and mutant CFP1, individual recovery curves were fit using a biexponential function which provides a good fit to the average data (Figure 46E). By fitting this function to individual recovery curves the half recovery time (t½) can be estimated from a single FRAP measurement. The distributions of t½ estimates from wild type or mutant GFP-CFP1 were then compared (Figure 46F). To determine whether the dynamics of CFP1 were significantly affected by the point mutations a two-tailed Student's t-test was used to compare the distribution of t½ times from individual FRAP measurements. Point mutations in the zf-CXXC or PHD domains were found to have a small but significant effect on the nuclear dynamics of CFP1. On average mutant GFP-CFP1 recovered ~20% faster than the wild type.

Surprisingly, despite its inability to bind to NMIs recovery of DB* was still substantially slower than predicted by free diffusion. This discrepancy may represent anomalous diffusion whereby the shape of the GFP-CFP1 molecule constrains its diffusion due to molecular sieving effects (Bancaud et al. 2009; Wachsmuth et al. 2000). Alternatively, diffusion of GFP-CFP1 may be delayed by interactions which are not dependent on the zf-CXXC or PHD domains. The possibilities are explored further in Chapter 7.

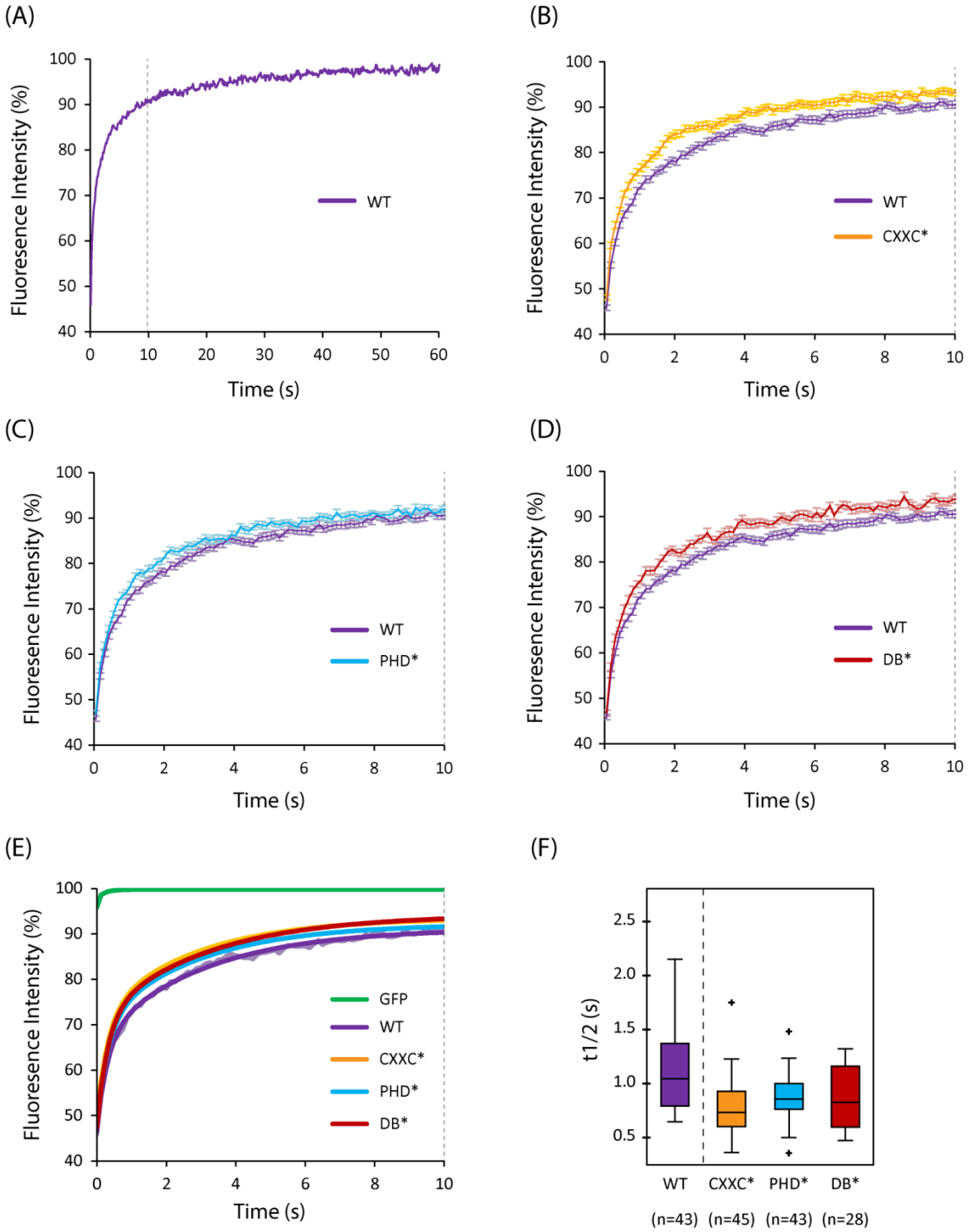



**Figure 46 – Mutation of the zf-CXXC or PHD domain increases the nuclear dynamics of CFP1**

Analysis of CFP1 binding kinetics by fluorescence recovery after photobleaching. The fluorescence recovery after photobleaching of wild type or mutant GFP-CFP1 was measured in a small region of the nucleus. (A) The average fluorescence recovery of wild type GFP-CFP1 is shown over 1 minute. Fluorescence recovers to 90% of its initial intensity within the first 10 s after photobleaching

(indicated by the hatched line). (B-D) The average recoveries of mutant GFP-CFP1 were compared the wild type over this initial period of rapid recovery. Bars indicate standard error. (E) FRAP of wild type or mutant GFP-CFP1 was modelled using a biexponential function. The average recovery curve of wild type GFP-CFP1 is shown in light purple to illustrate the close fit of this function ($R^2$=0.996). The predicted recovery of GFP based on free diffusion is shown in green to aid interpretation of CFP1 dynamics. (F) The distribution of half recovery times (t½) calculated from individual FRAP measurements are depicted as a box and whisker plot. Pairwise two-tailed Student's t-tests indicated a significant difference ($p<0.005$) between wild type and mutant GFP-CFP1, but no significant difference between the CXXC*, PHD* and DB* mutations. This distinction is indicated by a hatched line.

## 6.9 Discussion

To understand how CFP1 interacts with chromatin in live cells its nuclear dynamics were examined by fluorescence recovery after photobleaching. This analysis revealed that CFP1 is highly dynamic in the nucleus (t½ ~1 s), with no appreciable immobile fraction. Rapid turnover of CFP1 is consistent with previous FRAP studies of the DNA methyltransferase enzyme DNMT1 which also contains a type I zf-CXXC domain (Schermelleh et al. 2007; Frauer et al. 2011). Frauer et al. also demonstrated that a fragment of DNMT1 containing just the zf-CXXC domain showed similar dynamics, suggesting that the zf-CXXC domain may play a dominant role in the dynamics of the full length protein.

Despite its rapid recovery, CFP1 was less dynamic than predicted by free diffusion, suggesting its equilibration may be slowed by interactions with a relatively static structure such as chromatin. Importantly, the recovery of GFP-CFP1 is considerably faster than reported for core histone proteins (Kimura & Cook 2001) and slightly faster than the linker histone H1 (Lever et al. 2000), implying that GFP-CFP1 interacts with chromatin less stably, or less frequently. Crucially, its rapid dynamics suggest CFP1 does not provide a long-residency landmark on chromatin. Transient interactions with chromatin may allow CFP1 to recognise specific regions of the genome by rapidly sampling different binding sites and remaining longer in more favourable regions.

To dissect the contribution of the zf-CXXC domain to CFP1 dynamics, FRAP analysis was performed on the CXXC* cell line. The recovery of CFP1 was accelerated by the CXXC* mutation supporting the suggestion that the mobility of the wild type protein is reduced due to interactions with chromatin. Surprisingly, while statistically significant, the effect of the CXXC* mutation was very subtle, suggesting that the recognition of non-methylated CpGs is not solely responsible for CFP1 dynamics. This may indicate that recognition of a non-methylated CpG is a rare event with the majority of CFP1 molecules remaining unbound, or that binding to non-methylated CpGs is relatively short lived presenting a minor impediment to equilibration. Alternatively, mutation of the zf-CXXC domain may have a subtle effect in this assay because the dynamics of CFP1 are already so fast. Although the t½ is decreased by just 0.2 s in the CXXC* mutant this represents an acceleration of 20%. Nevertheless mutation of the zf-CXXC domain does not result in free diffusion of CFP1 suggesting that additional factors contribute its dynamics.

As for CXXC*, the PHD* mutation had a subtle but significant effect on the mobility of CFP1, however mutation of the PHD domain was shown to have a striking effect on the spatial distribution of CFP1, which no longer formed the bright clusters observed for wild type CPF1. These clusters have also been observed by another group immunostaining Flag-tagged CFP1 in HEK293 cells (Lee & Skalnik 2005), and have also recently been observed in mouse ES cells (Supplementary Figure 2). Although the molecular nature of these structures is not clear, recent evidence from a collaborating group (Koseki Lab, unpublished observation) suggests they co-localise with Pol II foci, implying they may be related to transcriptional activity. PHD-dependent clustering may contribute to the enrichment of CFP1 at specific regions of the genome by elevating its local concentration. Conversely, by increasing the local concentration of a zf-CXXC protein, CFP1 clusters may promote a local aggregation of genomic regions containing non-methylated CpGs. A similar function has recently been demonstrated for Spp1, the yeast homologue of CFP1, which tethers distal genomic elements in a H3K4me3 dependent manner during meiosis (Sommermeyer et al. 2013). Complementary analysis

using fluorescence in situ hybridisation may help to determine whether CFP1 clusters are associated with specific regions of the genome.

# Chapter 7 - Investigating the interaction between CFP1 and the SET1 complex

## 7.1 Introduction

The methylation state of lysine 4 on the histone H3 tail (H3K4) varies across the genome, with punctate peaks of H3K4 trimethylation (H3K4me3) corresponding to the majority of transcription start sites (Guenther et al. 2007). It is not yet clear whether H3K4 methylation represents a cause or a consequence of transcription (Henikoff & Shilatifard 2011). In yeast, H3K4A and H3K4R mutations are non-lethal indicating that H3K4me3 is not essential for transcription to occur (Briggs et al. 2001). Nevertheless, H3K4 methylation appears to be essential in metazoan organisms (Wysocka et al. 2005; Hollmann et al. 2002). In mammals, H3K4 methylation is catalysed by six different enzymes (Liu et al. 2009; Xiao et al. 2011; Aravind et al. 2011); SET1A, SET1B, and MLL proteins 1-4. Although diversification among H3K4 methyltransferases implies non redundant functions, the division of labour between this family of SET1-like enzymes remains controversial (Wang et al. 2009; Bledau et al. 2014; Cheng et al. 2014; Denissov et al. 2014).

Unlike G9a, Suv39h and ASH1L which methylate lysine residues elsewhere in the H3 tail (An et al. 2011; Patnaik et al. 2004; Tachibana et al. 2001; Strahl et al. 2002), methylation of H3K4 by the SET1-like enzymes depends upon four additional subunits (Ernst & Vakoc 2012; Dou et al. 2006). These four proteins WDR5, RBBP5, ASH2L and DPY-30 (WRAD) interact with all six H3K4 methyltransferases in mammals and their homologs purify with the SET1 enzyme in yeast indicating that the WRAD complex has been conserved during evolution. As well as stimulating catalytic activity, WDR5 and ASH2L each contain chromatin binding activities which may contribute to the localisation of SET1-like methyltransferases genome-wide (Chen et al. 2011; Wysocka et al. 2005; Han et al. 2006). Importantly, SET1A and -B differ from the MLL enzymes through their interaction with CFP1 and Wdr82 (J.-H. Lee et al. 2007; Lee & Skalnik 2005). These unique subunits may establish a division of

labour by modulating the genome-wide localisation of SET1A /B complex (SET1), directing their activities towards different regions.

Wdr82 has been proposed to couple SET1 to active transcription though its interaction with the C-terminal domain of RNA polymerase II (Lee & Skalnik 2008), however the contribution of Wdr82 to SET1 localisation genome-wide is not yet known. Additionally, the zf-CXXC protein CFP1 has been proposed to influence H3K4 methylation independently of transcription, based on the observation that exogenous CpG-rich sequences can support the generation of a novel H3K4me3 peak coincident with overlapping CFP1 enrichment (Thomson et al. 2010). As previous attempts to study SET1 localisation using ChIP-sequencing have been unsuccessful, the role of CFP1 in SET1 localisation remains unknown. To understand how CFP1 interacts with SET1 *in vivo*, this chapter compares their nuclear dynamics and localisation genome-wide.

## 7.2 A double point mutation in the CFP1 SET1 interaction domain abrogates binding to SET1A

Previous chapters have explored the roles of the zf-CXXC domain, and the PHD domain in the interaction between CFP1 and chromatin. Although significant contributions from each domain were observed, their combined effects were insufficient to fully explain the genome-wide localisation of CFP1 or its dynamics within the nucleus. Some CpG islands were not recognised by CFP1 despite the presence of trimethylated H3K4 and non-methylated CpGs, perhaps indicating an additional selectivity beyond the recognition of non-methylated CpGs by the zf-CXXC domain and H3K4me3 by the PHD domain. Similarly, attenuation of both domains using specific point mutations did not cause the mutant protein to diffuse freely, suggesting that the mobility of CFP1 is additionally reduced by zf-CXXC- and PHD-independent interactions.

To understand the impact of this interaction on CFP1 localisation and dynamics, an *in vivo* system was required in which CFP1 was isolated from the SET1 complex. A previous study suggested that the residues responsible for interaction with the SET1 enzymes mapped to a region towards the C-terminus of CFP1 (Figure 47A) (Butler et al. 2008). To identify residues critical for interaction with SET1, a multiple sequence alignment was performed using CFP1 homologues from a diverse range of eukaryotic species. Alignment of the SET1 interaction domain (SID) from these species reveals the conservation of specific residues from yeast to humans (Figure 47B).

Interestingly, the SID domain contains a conserved CXXXXC…CXXX[C/H] motif reminiscent of the zf- CXXC domain, PHD domain and other protein folds which coordinate zinc. In the original study characterising the CFP1 SID domain, mutation of the first cysteine (C375) abolished interaction with SET1A and B, as did a triple point mutation including the third cysteine (C391). It is possible that mutation of these residues disrupts interaction with the SET1 enzymes, as the mutant SID domains are no longer able to coordinate zinc, causing a local unfolding of the protein. To try to abolish interaction with SET1 with minimal effects on the folding of CFP1, single point mutations were designed to convert S392, K393 or Y394 (SKY) to alanine. This SKY motif was identical across the nine species compared, with the exception of K393 in goat (Figure 47B).

To determine whether mutation of the SKY motif was sufficient to disrupt interaction with SET1, CFP1 with a double point mutation converting both K393 and Y394 to alanine (SID*) was first stably expressed in C127 cells as a GFP-fusion. Expression of full length SID* GFP-CFP1 was verified by western blot, which revealed that the wild type and SID* mutant fusions were expressed at comparable levels in their respective cell lines (Figure 48A). The ability of each fusion to interact with SET1A was assayed by immunoprecipitation (Figure 48B).

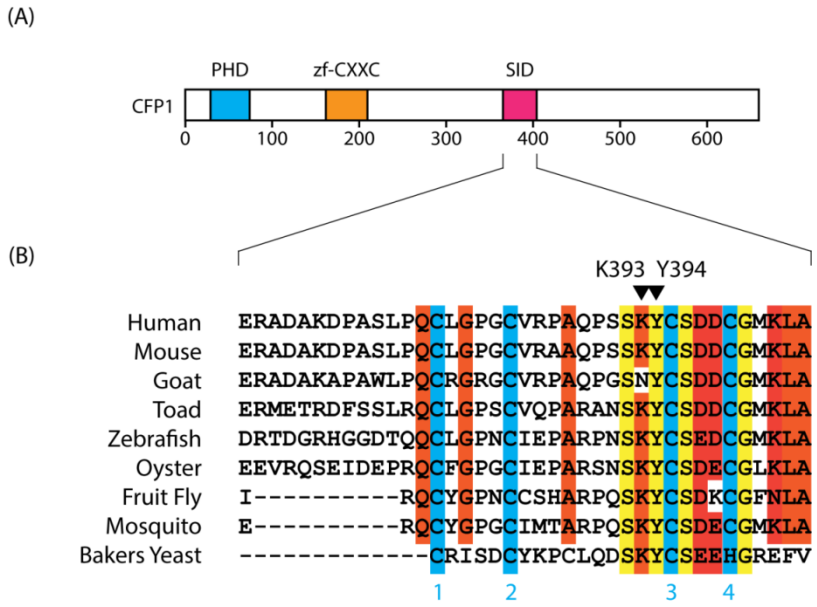



**Figure 47 – The CFP1 SID domain resembles a zinc finger, and contains a conserved SKY motif**
(A) A schematic depicting PHD, zf-CXXC and the SET1 interaction domain (SID) locations in the CFP1 sequence. (B) Multiple sequence alignment. CFP1 SID domains from; *Homo sapiens*, *Mus musculus*, *Capra aegagrus hircus* (Goat), *Xenopus laevis*, *Danio rerio*, *Crassostrea gigas* (Pacific Oyster), *Drosophila melanogaster*, *Anopheles gambiae* (Mosquito) and the PHD domain from *Saccharomyces cerevisiae* Spp1, were aligned using ClustalW. Identities are highlighted yellow, and orange, similarities in red. Putative zinc-coordinating residues are highlighted in cyan, and numbered below the alignment. K393, and Y394 residues, mutated to alanine in the SID* mutant, are indicated above the alignment.

Wild type or SID* GFP-CFP1 was immunoprecipitated by incubating nuclear extract from each cell line with α-GFP. Nuclear extract containing free GFP was immunoprecipitated in parallel as a negative control. After incubation with the extract α-GFP was captured by Protein G beads and washed to remove unbound nuclear extract. Immunoprecipitated proteins were then analysed by western blot. While SET1A was clearly immunoprecipitated in extract from C127s expressing wild type GFP-CFP1, no SET1A was detected in the SID* IP. Western blot analysis indicated that the nuclear extracts initially contained equal concentrations of SET1A suggesting that the SID* mutation abrogates interaction between CFP1 and SET1A (Figure 48D).

To determine whether CFP1 still interacts with chromatin despite isolation from the SET1 complex by the SID* mutation, a triple mutant cell line (TRP*) was generated (Figure 48C). These cells expressed GFP-CFP1 with point mutations in the PHD and zf-CXXC domain and a double point mutation in the SID domain (W49A, R200A, K393A, and Y394A). In addition, cloning of the SET1A cDNA into the plasmid backbone generated in Chapter 5, allowed the generation of another C127 cell line stably expressing GFP-SET1A at endogenous levels (Figure 48D). This cell line enabled the dynamics of SET1A to be investigated in parallel.

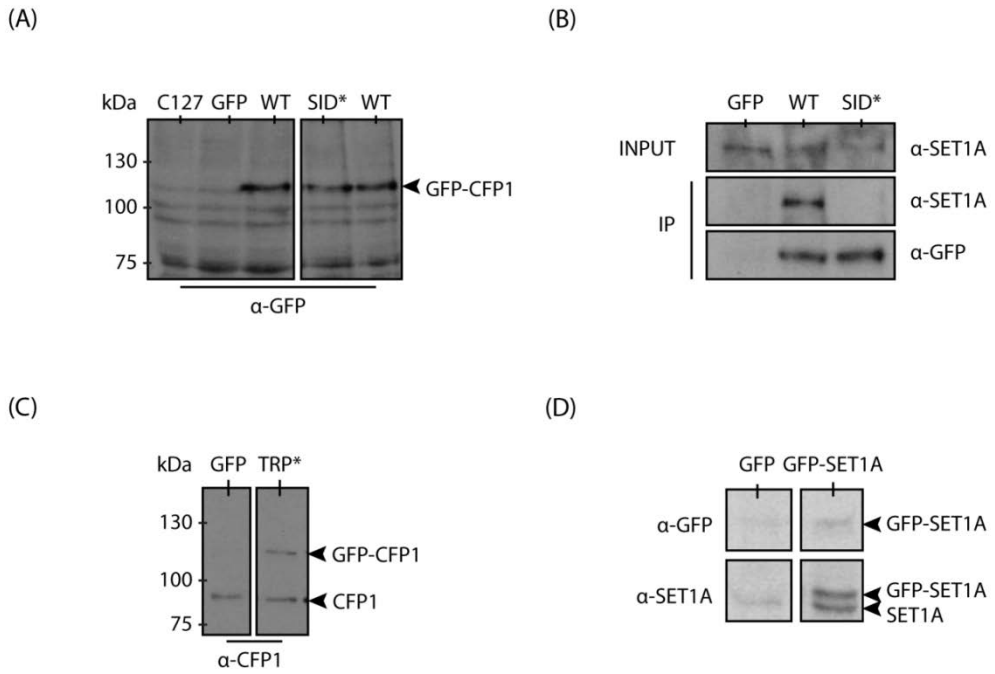



**Figure 48 - A double point mutation in the conserved SET1 interaction domain isolates CFP1 from the SET1 complex**

(A) Western blot analysis verifying the expression of full length SID* GFP-CFP1 in a stable C127 cell line. (B) The interaction between GFP-CFP1 and SET1A was analysed by α-GFP immunoprecipitation. Nuclear extracts were tested from C127 cells expressing free GFP, wild type GFP-CFP1, or SID* mutant GFP-CFP1. Immunoprecipitated proteins were analysed by western blot. While SET1A co-precipitated with the wild type GFP-CFP1, no SET1A precipitated with the SID* mutant, despite equal amounts of SET1A in the input extract, and equal precipitation of the GFP fusions. (C and D) Western blot analysis verifying the expression of full length TRP* GFP-CFP1 (E) and GFP-SET1A (F) in stable C127 cell lines.

### 7.3 The SET1 interaction domain plays a dominant role in the nuclear dynamics of CFP1

In the previous chapter, abrogation of binding to non-methylated CpGs by a point mutation in the zf-CXXC domain (CXXC*) was demonstrated to accelerate turnover of CFP1. Similarly, a mutation in the PHD domain (PHD*) which prevented binding to H3K4me3 accelerated CFP1 turnover to a similar degree. Importantly, the combination of both mutations (DB*) had no further effect on CFP1 turnover, which was still slower than predicted by free diffusion. This discrepancy may be related to the interaction of CFP1 with SET1. To determine whether interaction with SET1 contributes to the nuclear dynamics of CFP1, turnover of GFP-CFP1 containing the SID* mutation was investigated by FRAP (Figure 49). Interestingly, fluorescence imaging of the SID* cell line revealed the presence of bright CFP1 clusters suggesting that these structures do not depend on the interaction of CFP1 with SET1.

FRAP analysis revealed that mutation of the SID domain accelerated the recovery of CFP1 (Figure 49), suggesting that the mobility of wild type CFP1 is reduced due to interactions with SET1. Unexpectedly, the SID* mutation had a much larger effect on CFP1 dynamics than the combination of CXXC* and PHD* mutations (DB*). This suggested that the interaction with SET1 may play a dominant role in the interaction between CFP1 and chromatin. To determine whether binding through the zf-CXXC or PHD domains is dependent on the interaction with SET1, FRAP analysis was performed on the triple mutant (TRP*) which combines CXXC*, PHD*domains and SID* mutations (Figure 49C, E and F). Crucially, recovery of the TRP* mutant was even faster than the SID*, suggesting that the mobility of SID* CFP1 is reduced due to interactions with chromatin mediated by the zf-CXXC and PHD domains. This implies that while the effects of these domains on the mobility of wild type CFP1 is masked by the dominant effect of its interaction with SET1, they have a large and statistically significant ($p=0.00001$) effect on the mobility of isolated CFP1 (Figure 49F).

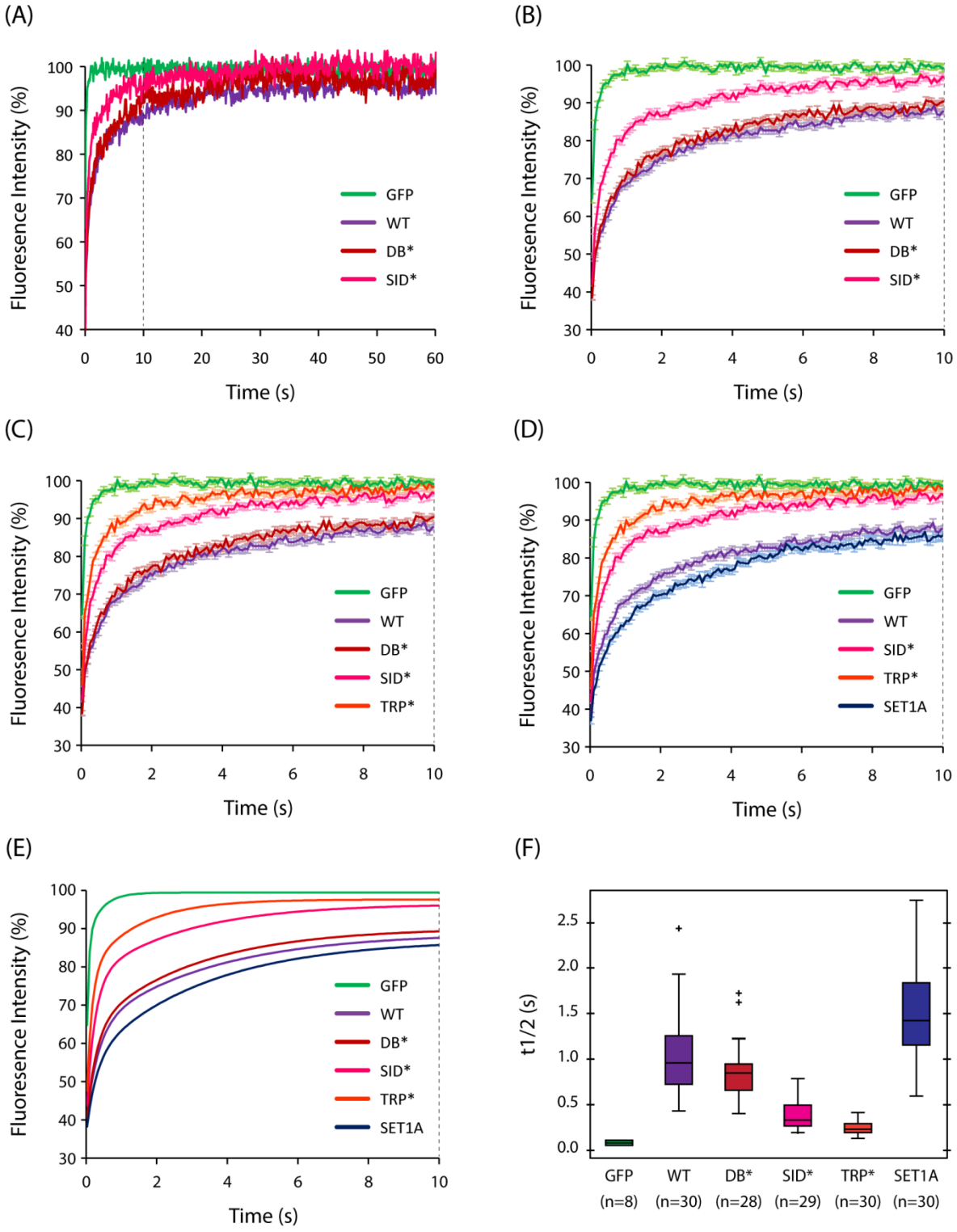



**Figure 49 – The nuclear dynamics of CFP1 reveal its interactions with SET1 as well as with chromatin**

Analysis of CFP1 binding kinetics by fluorescence recovery after photobleaching. (A) Average fluorescence recoveries are shown over 1 minute. FRAP measurements were carried out on C127 cells expressing GFP, wild type GFP-CFP1 (WT), or mutant versions containing mutations in the zf-

CXXC and PHD domains (DB*) or in the SET1 interaction domain (SID*). WT fluorescence recovers to 90% of its initial intensity within the first 10 s after photobleaching (indicated by the hatched line). (B) There is a clear difference between the average recovery of WT or SID* mutant GFP-CFP1 over this initial period. Bars indicate standard error. (C) The combination of DB* and SID* mutations (TRP*) further accelerates CFP1 recovery, but quite to the rate of free GFP. (D) GFP-SET1A recovers slightly slower than WT GFP-CFP1. (E) FRAP curves were modelled using a biexponential function, which fits to the data extremely well ($R^2$>0.96). (F) The distribution of half recovery times (t½) calculated from individual FRAP measurements are depicted as a box and whisker plot. Pairwise two-tailed Student's t-tests indicated that the observed differences were all statistically significant difference ($p<0.001$) with the exception of DB* GFP-CFP1 which was not significantly different ($p=0.08$) from WT GFP-CFP1.

To understand how interaction with SET1 might reduce the mobility of CFP1 inside the nucleus, FRAP analysis of GFP-SET1A was performed in parallel (Figure 49D, E and F). SET1A recovered at a similar rate to wild type CFP1, suggesting that like CFP1, SET1A is considerably more mobile than the core histones, but not as mobile as freely diffusing GFP. Importantly, SET1A recovered significantly slower than CFP1, suggesting that interactions with SET1A may reduce the mobility of CFP1. Although CFP1 recovers more quickly than SET1A, the similarity between them is consistent with a dominant role of SET1 interactions in the nuclear dynamics of CFP1. As the recovery of wild type CFP1 is more similar to SET1A, than to the SID* suggesting that the wild type protein may interact frequently or stably with the SET1A complex.

Interestingly, Wild type CFP1 and SET1 exhibit the greatest variation in the rate of recovery between experiments (Figure 49F). As potential interactions are stripped away by the progressive introduction of mutations, this variation is decreased. One explanation for this trend is that the wild type CFP1 is more sensitive to heterogeneity between the regions selected for FRAP analysis. For example variation in the numbers of non-methylated CpGs within the bleached region may have less of an effect on the CXXC* mutant than on wild type CFP1. Building on these experiments, complementary studies might exploit the recent advances in live cell fluorescence imaging, and

single molecule tracking to examine the dynamics of individual zf-CXXC proteins *in vivo* (Chen et al. 2014; Gebhardt et al. 2013; Mazza et al. 2012).

### 7.4 CFP1 binds to non-methylated islands independently of SET1

FRAP analysis demonstrated that mutation of the SET1 interaction domain had a dominant effect on the nuclear dynamics of CFP1. To determine whether interaction with SET1 affects the genome-wide localisation of CFP1, localisation of the SID* mutant was profiled by ChIP-sequencing. Chromatin was prepared from SID* and TRP* cell lines in parallel, and immunoprecipitated with the α-GFP antibody. Immunoprecipitated DNA from biological duplicates was submitted to the WTCHG for library preparation followed by Illumina sequencing. Sequencing reads were then aligned to the mouse genome to produce genome-wide profiles.

Visual inspection of the SID* profile revealed punctate peaks of enrichment corresponding to islands of non-methylated CpG dinucleotides (NMIs) (Figure 50A). This indicated that the SID* mutation does not abrogate CFP1 binding to chromatin, and suggested CFP1 can recognise these sites independently of interactions with SET1. Peaks of enrichment were not detected in the TRP* profile, consistent with the loss of binding in the CXXC* and DB* mutants. Interestingly enrichment of SID* at many NMIs was more pronounced than the enrichment of wild type GFP-CFP1. Peaks of SID* enrichment were also observed at many NMIs where wild type peaks were not observed. To investigate the altered localisation of SID* genome-wide, peaks of enrichment were identified using the MACS algorithm(Zhang et al. 2008).

Compared to wild type GFP-CFP1, a greater number of SID* peaks were detected by the MACS algorithm (9,581 compared to 6,946 wild type peaks). Analysis of their locations genome-wide revealed that 89% of SID* peaks (9,581 /10,767) corresponded to an experimentally determined NMI (Figure 50B). A visual inspection of SID* peaks distinct from NMI regions, revealed peaks of

BioCAP signal which were below the threshold for NMI detection (Figure 50C). This suggested that like the wild type protein, SID* mutant CFP1 binds exclusively to NMIs, consistent with the dominant role of the zf-CXXC domain in CFP1 localisation. SID* peaks were detected at 70% of NMIs (9,581/13,527) compared to just 40% in the wild type. To understand why SID* appears to bind a greater number of NMIs, the read densities of wild type and SID* mutant GFP-CFP1 were compared across all NMIs (Figure 50D). Importantly, a positive linear relationship was observed between SID* and wild type densities, suggesting that the increased number of peaks reflect a general increase in signal, rather than a redistribution of the SID* mutant across the genome. This suggestion is supported by the observation that most peaks unique to the SID* correspond to smaller NMIs. Given the increased mobility of the SID* in FRAP experiments, the increased ChIP-signal is somewhat surprising, however as the library preparation and sequencing steps for the WT and SID* mutant were not performed in parallel further experiments are required to allow a quantitative comparison. The ability of CFP1 to bind NMIs independently of CFP1 indicates that the increased mobility of SID* CFP1 observed by FRAP analysis is not caused by a failure to bind chromatin. This is consistent with the further increase in mobility upon mutation of the zf-CXXC and PHD domains in the TRP* mutant.

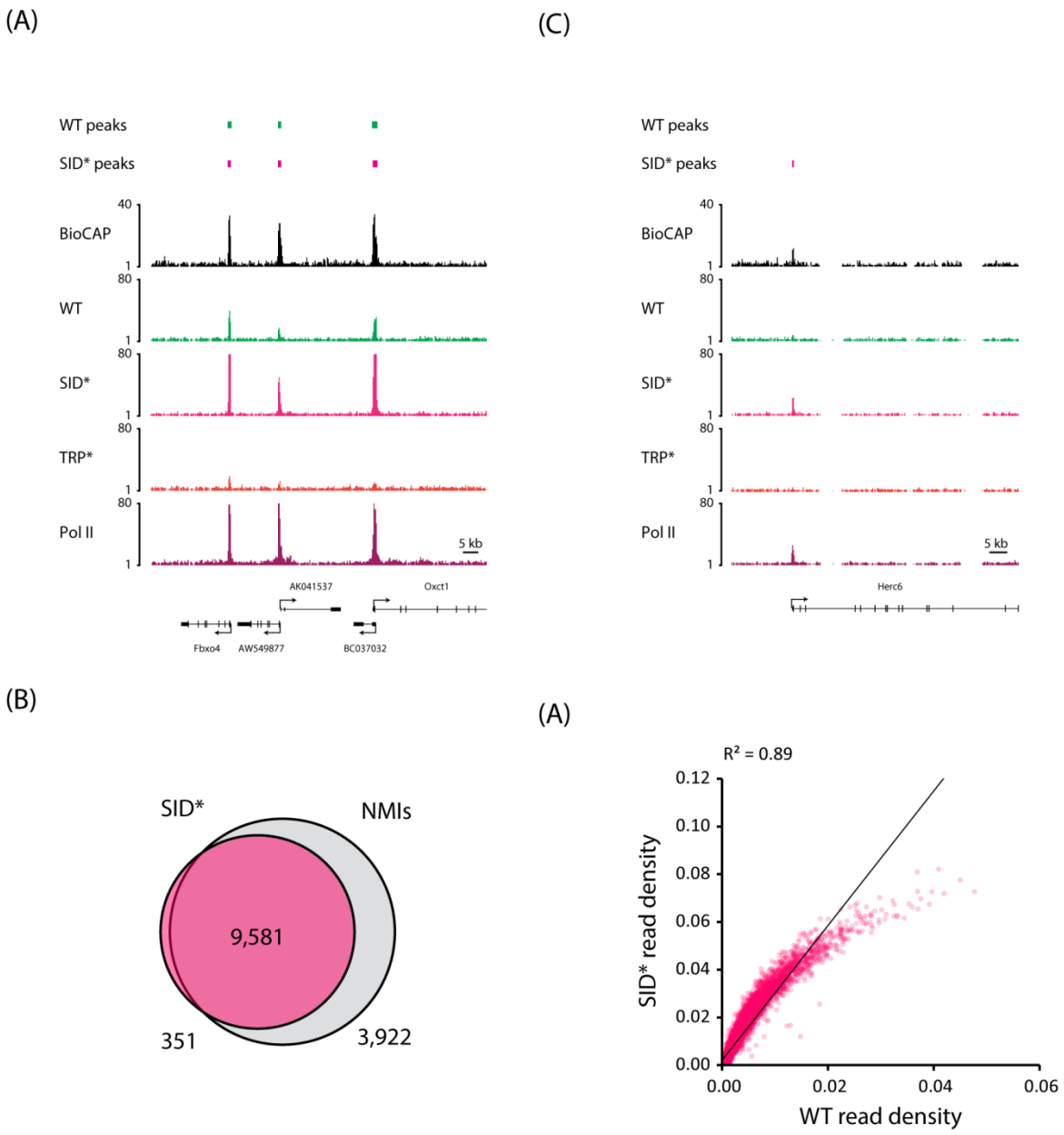



**Figure 50 – CFP1 binds to NMIs independently of its interaction with SET1**

(A) Profiles of non-methylated DNA (BioCAP), and enrichment of wild type or mutant GFP-CFP1 are shown across a region of the mouse genome containing three CFP1 peaks. Pol II enrichment is shown below for comparison. Mutation of the SET1 interaction domain (SID*) did not prevent GFP-CFP1 from binding to chromatin. When point mutations in the zf-CXXC, PHD and SID domains were combined (TRP*) binding was completely ablated. Genes are indicated below the profiles with arrows indicating the direction of transcription. The scale is given in the bottom right corner. (B) Genome-wide overlap between SID* peaks and non-methylated islands (NMIs) is depicted as a proportional Venn diagram. (C) SID* peaks which were not detected in the wild type profile, typically

corresponded to local enrichment of non-methylated DNA and Pol II. (D) Comparison of wild type and SID* mutant read densities across all NMIs reveals a positive linear relationship.

### 7.5 ChIP sequencing of GFP-SET1A reveals preferential localisation to a subset of NMIs

CFP1 may contribute to the unique chromatin architecture of NMIs by directing the placement of H3K4me3 by SET1 (Thomson et al. 2010; Clouaire et al. 2012). Importantly, comparison of wild type and SID* GFP-CFP1 suggested that CFP1 is able to bind to NMIs independently of SET1. However the contribution of CFP1 to SET1 localisation is not yet known. To understand whether CFP1 directs SET1 to specific regions of the, ChIP-sequencing was performed in C127 cells stably expressing GFP-SET1A. Chromatin was prepared from two biological replicates, immunoprecipitated using the α-GFP antibody, and sequenced using the Ion Proton from Ion Torrent. Unfortunately, the level of enrichment in one of the replicates was insufficient for ChIP-sequencing, so further experiments are necessary to verify the reproducibility of these results. However, a putative profile was generated by aligning singlicate SET1A reads to the mouse genome.

Visual inspection of the SET1A profile revealed punctate peaks of enrichment corresponding to transcription start sites (TSSs) (Figure 51A). This indicated that the N-terminal GFP tag did not prevent SET1A from binding to chromatin associated with specific regions of the genome. SET1A peaks that were not associated with a TSS were generally of lower magnitude, with the exception of a few distal NMIs that were also occupied by Pol II, possibly representing unannotated TSSs (Figure 51B). To investigate its localisation genome-wide, SET1A peaks identified using the MACS algorithm were compared to the locations of mouse TSSs genome-wide. The MACs algorithm identified 5,000 peaks of SET1A enrichment, of which 76% (3,816) were within 1 kb of an annotated TSS (Figure 51C). Preferential binding of SET1A to promoter regions is consistent with its proposed role in H3K4 trimethylation. Critically, SET1A peaks were only observed at 16% of TSSs (3,816/23,203), suggesting SET1A may not drive the placement of H3K4me3 at all genes.

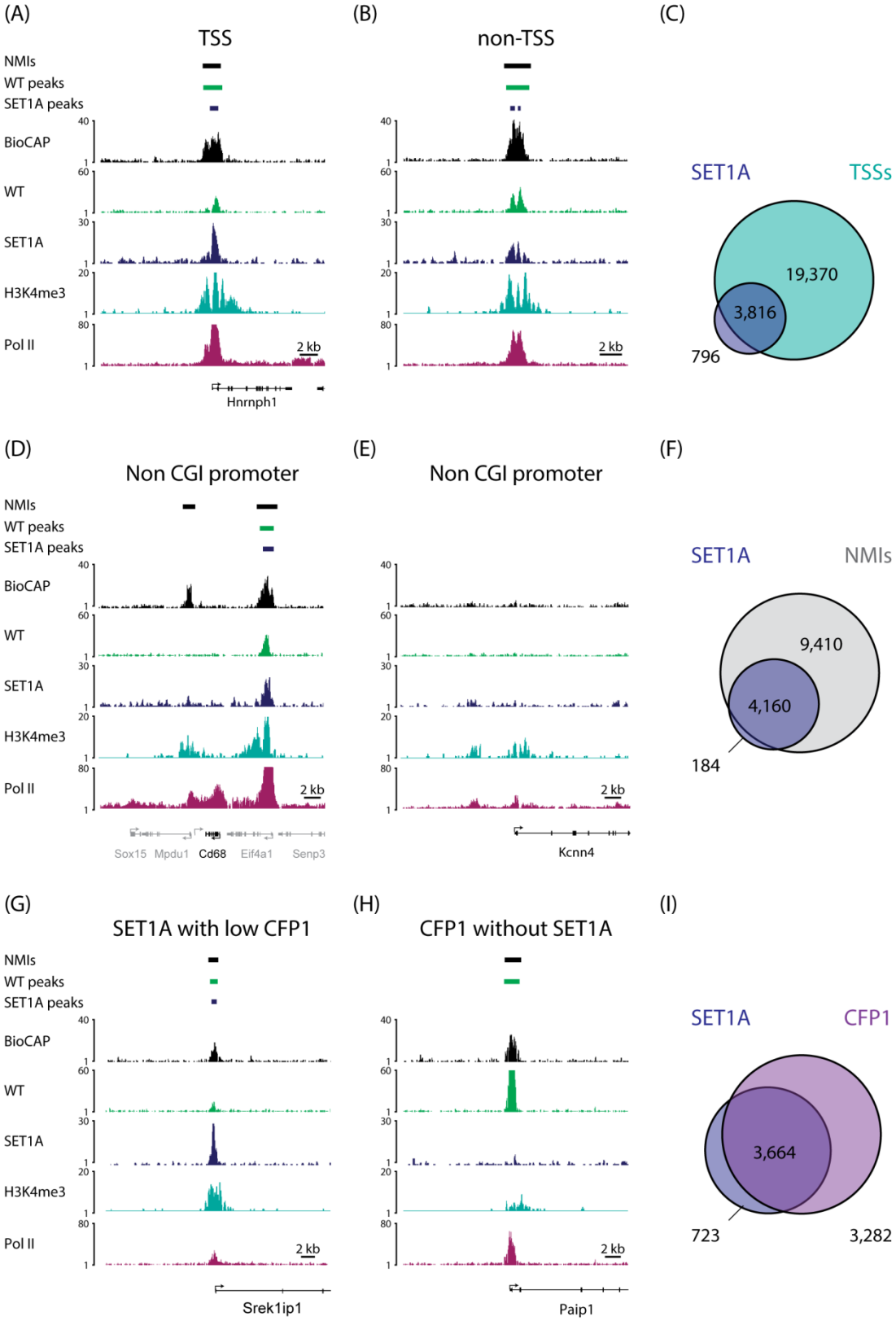

(A)
TSS
NMIs
WT peaks
SET1A peaks
BioCAP
WT
SET1A
H3K4me3
Pol II
2 kb
Hnrnph1
(B)
non-TSS
2 kb
(C)
SET1A
TSSs
19,370
3,816
796
(D)
Non CGI promoter
Sox15 Mpdu1 Cd68 Eif4a1 Senp3
(E)
Non CGI promoter
Kcnn4
(F)
SET1A
NMIs
9,410
4,160
184
(G)
SET1A with low CFP1
Srek1ip1
(H)
CFP1 without SET1A
Paip1
(I)
SET1A
CFP1
3,664
723
3,282

**Figure 51 – SET1A interacts with a subset of NMIs**

Snap shots depict the ChIP-sequencing profiles of non-methylated DNA (BioCAP), GFP-CFP1 (WT), GFP-SET1A (SET1A), H3K4me3 and RNA polymerase II (Pol II) across ~15 kb regions of the genome. Proportional Venn diagrams illustrate the overlap of SET1A peaks with other genomic features. SET1A peaks were generally associated with transcriptional start sites (TSSs) (A), however a minority of were seen at regions of non-methylated DNA without an annotated TSS (B). Peaks of SET1A were not observed at all TSSs (C), and low levels of SET1A were observed at the cd68 (D) and Kcnn4 (E) non-CGI promoters, despite enrichment of Pol II at cd68. Almost all SET1A peaks were associated with non-methylated CpG islands (NMIs) (F). High levels of SET1A were not always associated with high levels of CFP1 (G), and vice versa (H), as SET1A peaks were only observed at 53% of CFP1 peaks (I).

Consistent with specific recognition of non-methylated CpG islands through its interaction with CFP1, low levels of SET1A were observed at non-CGI promoters (Figure 51D and E). Genome-wide analysis of the overlap between SET1A peaks and NMIs revealed that over 80% (4,160/5000) of SET1A peaks corresponded to NMIs (Figure 51F). This suggested that non-methylated DNA is an important attribute for recognition by SET1A. Conversely SET1A peaks were only observed at 30% of NMIs (4,160/13,572), indicating that the presence of non-methylated CpGs is not sufficient for enrichment of SET1A. Like CFP1, SET1A appears to bind preferentially, to a subset of NMIs. Visual inspection of their genome-wide profiles revealed that high levels of SET1A are sometimes observed despite low levels of CFP1 (Figure 51G). Furthermore, some promoters exhibited low levels of SET1A despite high levels of CFP1 (Figure 51H). Intersection of the CFP1 and SET1A peak sets revealed enrichment of CFP1 at 66% of SET1A peaks (3,282/5000) (Figure 51I), suggesting that CFP1 may play an important role in the interaction between SET1A and chromatin. Conversely, SET1A peaks were only detected at 53% of CFP1 peaks (3,664/6,946), consistent with the ability of SID* mutant CFP1 to bind to chromatin independently of SET1. The incomplete overlap between CFP1 and SET1A may suggest that a fraction of CFP1 remains dissociated from SET1 inside the nucleus. These experiments demonstrate that GFP-SET1A can be used to profile interactions between SET1A and chromatin, however duplicate experiments are required to support a more detailed analysis.

## 7.6 Inducible knockout of CFP1 in embryonic stem cells

ChIP-sequencing experiments suggest that CFP1 is important for the enrichment of SET1A at a subset of promoter NMIs. To understand the role of CFP1 in the SET1 complex, and its contribution to the deposition of H3K4me3, an *in vivo* system was required for the acute removal of CFP1. To investigate the function of CFP1, a mouse ES cell line was generated to allow drug-inducible deletion of the CFP1 gene (These cells were a kind gift from Haruhiko Koseki, Riken, Japan). In this cell line two LoxP sites flank exons 4 and 5 of the endogenous CFP1 gene (Figure 52A). In the inducible knock out line (CFP1fl/fl) line LoxP sites have been introduced on both alleles. CFP1fl/fl cells also express Cre, a site specific bacteriophage recombinase, fused to a modified oestrogen receptor (ERT). ERT normally interacts with cytoplasmic chaperone proteins such as Hsp70 which prevents Cre-ERT from entering the nucleus. The small molecule inhibitor 4-hydroxytamoxifen (tamoxifen) binds to ERT preventing its interaction with chaperone proteins, and allowing Cre-ERT to enter the nucleus. Genetic knockout of CFP1 is catalysed by Cre which excises the DNA between the two LoxP sites. Removal of exons 4 and 5 causes a frame shift in leading to a nonsense CFP1 transcript containing premature stop codons.

Using this system, further experiments will attempt to address the role of CFP1 in the interaction between SET1 and chromatin. To confirm that CFP1 is eliminated from the CFP1fl/fl cell line after tamoxifen treatment, nuclear extract from treated and untreated cells was analysed by western blot (Figure 52B). CFP1 appears significantly reduced after 48 h and completely eliminated within 72 h. Previous experiments in the Klose lab revealed that acute removal of CFP1 in the CFP1fl/fl ES cell line had no effect on the global levels of H3K4me3 (Supplementary Figure 3). Maintenance of global H3K4me3 levels in the absence of CFP1 suggests that CFP1 may not be required for H3K4me3 placement, or may only be required at a minority of sites. The latter possibility is consistent with the preferential binding of CFP1 to a minority of promoter NMIs, which was observed in the profile generated for C127 cells. Similarly, a recently published profile of GFP-CFP1 in mouse ES cells

suggests that CFP1 is preferentially enriched at the most highly transcribed genes (Denissov et al. 2014). The contribution of CFP1 to H3K4me3 at these sites may have a small effect on the global level of H3K4me3. Therefore, experiments have now begun to investigate the potential effects of deleting CFP1 on the genome-wide distribution of H3K4me3.

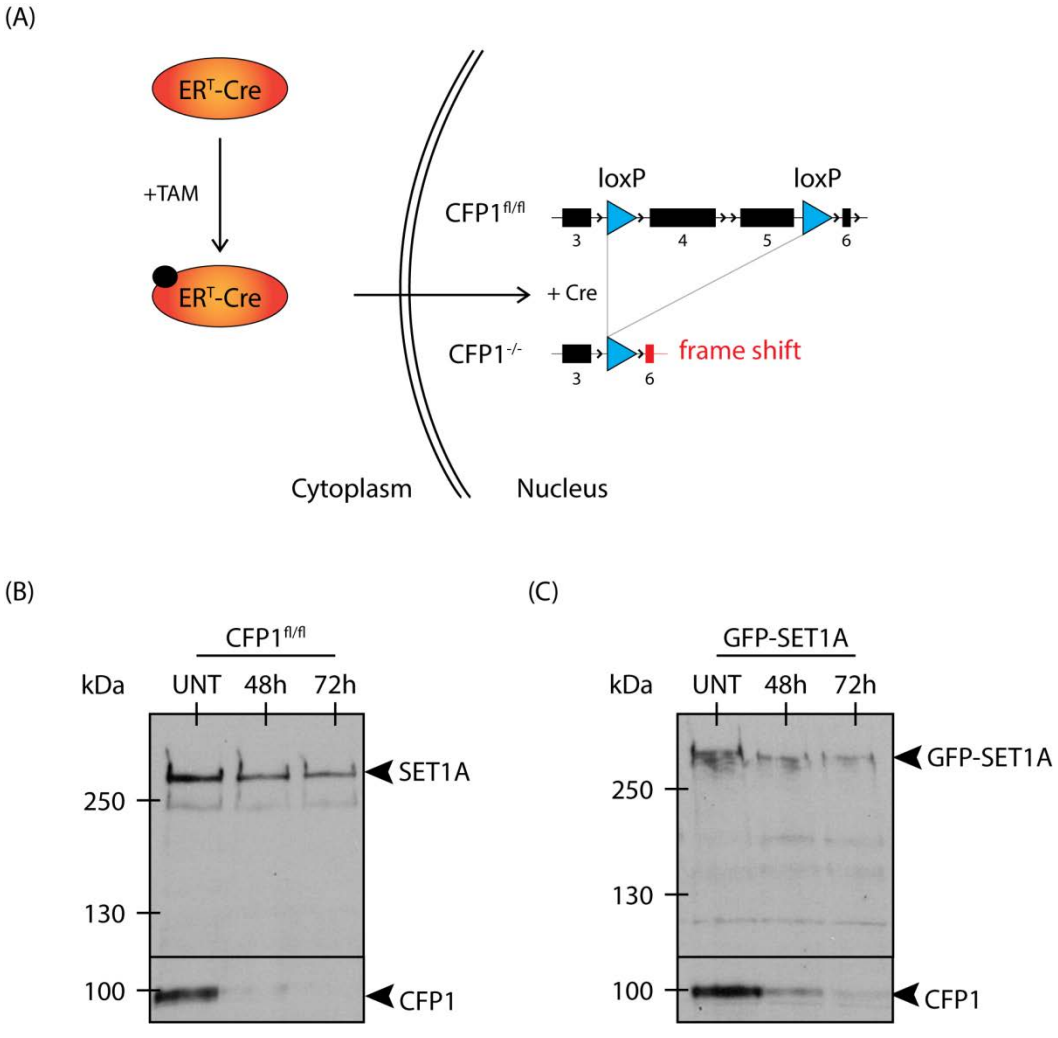



**Figure 52 - Inducible knockout of CFP1 in embryonic stem cells**

(A) Schematic Illustration of tamoxifen (TAM) inducible CFP1 knockout. Addition of TAM allows ERT-CRE to enter the nucleus. Cre then catalyses site specific recombination between the LoxP sites (blue) excising CFP1 exons 4 and 5 which creates a frame shift to knock out CFP1 expression. Deletion of CFP1 in CFP1fl/fl (B) or GFP-SET1A CFP1fl/fl (C) cells was analysed by western blot. Nuclear extract was prepared from untreated cells (UNT) and after 48 h or 72 h of treatment with

tamoxifen and separated on a 6% SDS-PAGE gel. CFP1 staining (lower panel) revealed a clear deletion of CFP1 after treatment with tamoxifen. Staining of endogenous SET1A (B) or GFP-SET1A (C) revealed a slight reduction after deletion of CFP1.

As the H3K4me3 modification is relatively stable (Radman-Livaja et al. 2010), compared to the dynamics of SET1A observed by FRAP analysis, the contribution of CFP1 to H3K4me3 deposition may be masked in measurements of the steady state levels. Furthermore, redundancy between the SET1-like H3K4 methyltransferase enzymes may obscure the contribution of CFP1 to SET1 dependent methylation. To directly examine the role of CFP1 in the SET1A complex, a new stable cell line was generated by transfecting CFP1fl/fl mouse ES cells with the GFP-SET1A expression plasmid. As demonstrated in C127 cells, the GFP-SET1A fusion should allow the genome-wide localisation of SET1A to be assessed by ChIP-seq using the α-GFP antibody.

In yeast, the CFP1 homologue Spp1 is reported to bind to nascent SET1 during translation (Halbach et al. 2009) and deletion of Spp1 results in reduced levels of the SET1 protein, despite equivalent levels of the SET1 transcript (Nedea et al. 2008). To understand the role of CFP1 in the SET1 complex, it was therefore important to determine whether SET1 levels are affected by conditional deletion of CFP1 in mouse ES cells. Nuclear extracts were prepared from parental CFP1fl/fl and GFP-SET1A CFP1fl/fl cell lines with or without tamoxifen treatment. Western blot analysis of these extracts suggests that endogenous SET1A and the GFP fusion are both reduced by the deletion CFP1 upon tamoxifen treatment based on their depletion relative to non-specific bands (Figure 52B and C). This implies that like the WRAD subunits, CFP1 may contribute to the stability of the SET1 complex in mouse ES cells. To confirm this result, future experiments will compare SET1 and CFP1 levels to an unrelated nuclear protein as a gel loading control, to rule out differences due to variation in total protein concentration. Using the panel of CFP1 mutants generated and characterised in this study, the contribution of CFP1 to the stability, genome-wide localisation and nuclear dynamics of SET1 can

now be mechanistically dissected by examining the ability of each mutant to rescue acute removal of wild type CFP1.

## 7.7 Discussion

Interaction with CFP1 distinguishes the SET1 enzymes from MLL1-4; however the role of CFP1 in the SET1A complex is unclear. In particular, the contribution of CFP1 to the stability, activity and genome-wide localisation of the SET1 complex remains controversial. Consistent with previous studies, co-precipitation of SET1A with wild type CFP1 suggests that the interaction between them is relatively stable (J.-H. Lee et al. 2007; Lee & Skalnik 2005). Nevertheless deletion of CFP1 or mutation of the SID domain demonstrated that both proteins can exist independently *in vivo*. Importantly, SET1A protein levels were reduced upon conditional deletion of CFP1 in ES cells, consistent with previous experiments in constitutive CFP1 null ES cells (Tate et al. 2010), and experiments in yeast deleting Spp1, the yeast homologue of CFP1 (Nedea et al. 2008). This suggests either reduced expression, or accelerated degradation of SET1 in the absence of CFP1. However, previous studies report that SET1A mRNA levels are increased in the absence of CFP1. (Tate et al. 2010). In yeast, the SET1 complex is thought to be assembled co-transcriptionally (Halbach et al. 2009), with Spp1 binding to the nascent SET1 protein prior to homologues of the WRAD complex. A recent study demonstrated that SET1 protein levels are tightly regulated in response to transcriptional activity and H3K4me3 levels, suggesting that degradation of unbound SET1 may suppress promiscuous methylation and restrict H3K4 methylation to actively transcribed genes (Soares et al. 2014). Surprisingly, Tate et al. found that expression of an N-terminal fragment of CFP1 which contained the PHD and CXXC domains, but did not interact with SET1 was sufficient to rescue normal levels of SET1A. Rescue of conditionally deleted CFP1, using the CFP1 mutants characterised in this study, will allow the contribution of CFP1 domains to SET1A stability to be dissected.

Although the contribution of CFP1 to SET1 catalytic activity was not investigated in this study, work by other groups suggests that CFP1 may modulate the specificity of SET1, facilitating trimethylation H3K4me3 (Schneider et al. 2005; Takahashi et al. 2009). A recent study examining SET1 methyltransferase activity *in vitro* proposed that CFP1 mediates interactions between the n-SET domain and the catalytic core which are required to methylate H3K4 on nucleosomes containing ubiquityl H2B (Kim et al. 2013). To understand how CFP1 interacts with SET1 *in vivo*, their interaction was disrupted to determine effect of isolating CFP1 from the complex. In isolation, CFP1 retained its ability to bind to NMIs, but was considerably more mobile in the nucleus. Interactions with SET1 appear not to influence the genome-wide localisation of CFP1. Inclusion of CFP1 in the SET1 complex may reduce its mobility due to non-specific interactions between the complex and the surrounding chromatin, or due to molecular sieving effects (Sprague & McNally 2005) which impede the SET1 complex, but do not affect isolated CFP1. Rescue of conditionally deleted CFP1 using CXXC* or DB* CFP1, should allow the contribution of CFP1 to the catalytic activity of SET1A to be separated from its proposed role in mediating the interaction between SET1A and chromatin. Experiments are currently underway to dissect these functions.

# Chapter 8 - Discussion

CpG islands are a prominent feature of vertebrate genomes, and are associated with the majority of gene promoters, suggesting they may be involved in the regulation of transcription. As CpG islands are generally resistant to DNA methylation, they serve as a binding site for zf-CXXC proteins, which bind specifically to non-methylated CpGs. Many zf-CXXC proteins are associated with chromatin modifying activities, suggesting that CpG islands may impinge upon transcriptional potential, by providing a platform for activating or repressive modifications (Blackledge & Klose 2011).

The zf-CXXC protein CFP1 is essential for mammalian embryogenesis (Carlone & Skalnik 2001), and has been proposed to contribute to normal development by influencing chromatin modification and gene regulation. Based on its inclusion in the SET1A and SET1B methyltransferase complexes (Lee & Skalnik 2005; J.-H. Lee et al. 2007), CFP1 was proposed to direct H3K4 methylation to CpG island regions (CGIs), which contain a high density of non-methylated CpGs (Tate et al. 2010; Thomson et al. 2010; Clouaire et al. 2012). Although CFP1 also contains a PHD domain proposed to recognise H3K4me3, the function of this domain was not clear. To understand how CFP1 contributes to the function of CpG islands, the primary aims of this thesis were to determine how CFP1 interacts with the genome, and to dissect the contribution of the zf-CXXC and PHD domains.

## 8.1 How does CFP1 interact with the genome?

In this study, ChIP-sequencing of CFP1 in a mammalian cell line confirmed that like other zf-CXXC proteins, CFP1 interacts exclusively with non-methylated regions of the genome. Although the CFP1 zf-CXXC domain is able to interact with non-methylated DNA throughout the genome, preferential enrichment was observed at a subset of non-methylated promoters, suggesting more frequent or more stable interactions at these sites. A complementary analysis using fluorescence microscopy revealed that CFP1 is heterogeneously distributed in the nucleus, where it is excluded from nucleoli

and pericentromeric domains. Based on FRAP analysis, CFP1 is highly dynamic and interacts transiently with chromatin in living cells. Given its small size and rapid rate of diffusion it is unlikely that the nucleoli or pericentromeric domains physically block the entry of CFP1, however they may present a less favourable environment, due to a lack of potential binding sites (Kang et al. 2010), or an unfavourable environment due to a higher density of competing proteins (Bancaud et al. 2009). CFP1 is thought to play a role in H3K4 methylation through its interaction with the SET1 complex (Tate et al. 2010; Thomson et al. 2010). As CFP1 appears to be excluded from heterochromatic regions, this may restrict the activity of SET1A to euchromatin regions, protecting heterochromatin from H3K4 methylation. By sequestering CFP1, euchromatic regions may facilitate H3K4 methylation by SET1 reinforcing the euchromatic state. By spatially partitioning enzyme complexes associated with their maintenance euchromatic and heterochromatic states may be self-reinforcing at a global level.

### 8.2 What use is a PHD when you can bind to CpG islands?

The recent proposal that CFP1 could bind to H3K4me3 via its PHD domain was particularly interesting because it implies a capacity for feedback. Theoretically, positive feedback loops result in an amplification, while negative feedback loops may be oscillatory or self-stabilising. It is possible that H3K4me3 binding by the CFP1 PHD domain has an inhibitory effect on H3K4 methylation, consistent with increased levels of H3K4 methylation in CFP1 null ES cells (Carlone et al. 2005). However, based on the punctate and varied profiles of H3K4me2 and 3, this possibility seems unlikely to be the case. *In vitro* assays confirmed that the CFP1 PHD is able to bind H3K4me2/3, but not to H3K4me0/1, which has since been reported by another group (Eberl et al. 2012). To understand the role of the PHD domain, this study investigated its contribution to the genome-wide localisation, dynamics, and nuclear distribution of CFP1.

Unlike the zf-CXXC domain, the PHD domain is not required for the interaction between CFP1 and chromatin and does not appear to contribute to the genome-wide localisation of CFP1. Interestingly, the PHD domain is required for normal levels of CFP1 across all non-methylated islands. Suggesting it contributes to stability but not localisation. This suggestion was supported by FRAP analysis, which detected increased mobility in PHD mutant CFP1, indicating less stable or less frequent interactions with chromatin. Interestingly, mutation of the PHD domain alters the nuclear distribution of CFP1, abolishing the bright foci which result from a dense clustering of CFP1 interaction sites. Initial experiments using the H3K4 demethylase, suggest that this effect may be H3K4me3 dependent, however as clustering is not affected by the zf-CXXC mutation, which abolishes CFP1 binding genome-wide, the effect may be chromatin-independent.

### 8.3 What is the role of CFP1 in the SET1 complex?

Based on its specificity for non-methylated DNA, CFP1 was previously proposed to 'restrict' SET1A/B to euchromatin (Tate et al. 2010), in keeping with increased H3K4 methylation in CFP1 null ES cells (Carlone et al. 2005). Several studies have suggested that CFP1 'recruits' SET1 to CpG islands (Thomson et al. 2010; Orlando et al. 2012; Long, Blackledge, et al. 2013), implying that CFP1 is required for the interaction to occur. Given that the CFP1 PHD domain is required for SET1A pull-down with an H3K4me3 peptide, CFP1 was proposed to mediate H3K4me3-dependent recruitment of SET1A (Eberl et al. 2012). This does not appear to be the case *in vivo*, as the PHD domain is unable to support recruitment of CFP1 when the zf-CXXC domain is mutated. The work presented here demonstrates that disruption of the interaction between CFP1 and SET1 has a dramatic effect on the nuclear mobility of CFP1, but not its genomic localisation. This implies that the SET1 complex interacts non-specifically with chromatin independently of CFP1, suggesting that CFP1 may restrict the placement of H3K4me3, by limiting the activity of SET1A/B to specific regions of the genome. This conclusion is supported by the depletion of H3K4me3 at active genes in CFP1 null ES cells (Clouaire et al. 2012), despite the overall increase in H3K4me3 levels (Carlone et al. 2005). Rescue of

the CFP1 null cells using zf-CXXC mutant CFP1, results in ectopic peaks of H3K4me3 (Clouaire et al. 2012), implying that the zf-CXXC domain of CFP1 restricts the activity of SET1A/B to CpG islands.

Direct investigation of the SET1A /B methyltransferases has been hampered by the lack of an effective antibody for ChIP-sequencing or high resolution microscopy. By developing a stable cell line expressing GFP-SET1A, this study detected enrichment of SET1A at a subset of non-methylated promoters. Building on this work, future experiments profiling GFP-SET1A in the presence or absence of CFP1 should help to elucidate the contribution of CFP1 to the genome-wide localisation of SET1A.

Importantly, conditional deletion of CFP1 results in a reduced level of SET1A, consistent with previous proposals that CFP1 contributes to the stability of the SET1 protein (Carlone et al. 2005; Tate et al. 2009; Tate et al. 2010). A recent study of the yeast SET1 complex suggested that SET1 protein levels are regulated by proteolytic degradation, depending on; the interaction of the complex with chromatin, the global level of active transcription, and the global level of H3K4me3 (Soares et al. 2014). Homologues of CFP1 and Wdr82 were both found to be important regulators of this degradation. CFP1 exerts separable positive and negative effects on the activity of SET1A/B, by providing resistance to proteolytic degradation while restricting its activity to CpG islands (Tate et al. 2009). By rescuing conditional deletion of CFP1 with the PHD point mutants described in this study, it would be exciting to determine how interaction with H3K4me3 contributes to the protective and restrictive roles of CFP1. Association with chromatin may protect SET1 from degradation, by occluding a degradation site within the protein, or by sequestering it away from proteolytic enzymes.

## 8.4 How do SET1A and SET1B contribute to H3K4me3?

Like the SET1A and SET1B complexes, MLL1/2 both possess zf-CXXC and PHD domains. As MLL3/4 lack these domains, they may potentially perform a distinct role in methylation of H3K4. Recent studies support this division of labour, suggesting that MLL3/4 are responsible for monomethylation of H3K4 rather than di- or trimethylation (Hallson et al. 2012; Herz et al. 2012; Cheng et al. 2014; Tie et al. 2014). RNAi knock downs of SET1A/B, or dSET1 in *Drosophila*, result in a global reduction in H3K4me3 (Wu et al. 2008; Mohan et al. 2011; Ardehali et al. 2011). Therefore these enzymes were thought to be responsible for the majority of H3K4me3 genome-wide (Cheng et al. 2014). Although SET1A/B may contribute to the level of H3K4me3 across all non-methylated islands, the genome-wide profiles presented in this study suggest that SET1A, and CFP1 are preferentially enriched at a subset of non-methylated promoters. The localisation of CFP1 to a minority of promoters has since been reported by another group, who profiled GFP-CFP1 in mouse ES cells (Denissov et al. 2014). The same study suggests that MLL2 is primarily required to maintain H3K4me3 at bivalent promoters, and evokes SET1A/B as the predominant H3K4 methyltransferases at active genes. Refining this picture of distinct but overlapping functions among the H3K4 methyltransferase family, distinct phenotypes were recently described for conditional knockouts of SET1A, which is essential at the epiblast stage, or SET1B, which is not required until gastrulation (Bledau et al. 2014). Concordantly, SET1A is essential in embryonic stem cells, while SET1B is dispensable. Although SET1A null mice die at the peri-implantation stage similar to CFP1 null phenotype, CFP1 null ES cells are viable, perhaps supporting a role for CFP1 in restricting rather than facilitating SET1A activity.

## 8.5 How does CFP1 contribute to H3K4 methylation?

CFP1 appears to be involved in several feedback loops, which may play an important role in the regulation of H3K4 methylation with consequences for gene expression. On a macro scale, CFP1 is heterogeneously distributed, favouring euchromatin and particularly the interface between chromatin domains and the inter-chromatin compartment. As mentioned above, this localisation

may produce a self-reinforcing loop, where the spatial distribution of CFP1 restricts H3K4 methylation by SET1, and in turn H3K4me2/3 restricts the spatial distribution of CFP1. Based on its rapid nuclear dynamics, this global segregation effect may not depend on the stable association of CFP1 with a given site on the genome, but results from an increased local concentration of CFP1 due to favourable interactions. As this effect is highly dependent on the spatial organisation of chromatin within the nucleus, it may be overlooked by sequencing approaches. Although global segregation is inherently difficult to perturb, a first step might be to understand if H3K4me3, or DNA methylation are required to maintain the heterogeneous distribution of CFP1. If DNA methylation is required to prevent association of CFP1 with heterochromatic domain, CFP1 should re-distribute to these regions in DNMT triple knock outs which lack DNA methylation. If H3K4 methylation is required to sequester CFP1 to euchromatic domains, redistribution of CFP1 should be observed in cells transfected with the H3K4 demethylase RBP2.

A second important feedback loop may exist between H3K4me3 and DNA methylation. H3K4me3 is thought to prevent *de novo* DNA methylation by blocking DNMT3a/b (Zhang et al. 2010), while non-methylated DNA may attract H3K4me3, through interactions with the CFP1-SET1 complex (Tate et al. 2010; Thomson et al. 2010). This loop is particularly important during DNA replication, when CpG methylation in the parental DNA is re-established on the daughter strands. By trimethylating H3K4, the CFP1-SET1 complex may protect CpG islands from *de novo* methylation prior to DNA replication. Once replicated, recognition of the non-methylated CpG island by CFP1 may then facilitate H3K4 methylation by SET1. If H3K4me3 is required *in vivo*, to protect CpG islands from DNA methylation, then removal of H3K4me3 should lead to the acquisition of DNA methylation.

Failure to protect CpG island promoters from DNA methylation, may explain why CFP1 null ES cells are prone to apoptosis, and fail to differentiate, while CFP1 null mice die at the implantation stage (Carlone et al. 2005). Like the global segregation effect, this is a difficult system to perturb. In yeast,

H3K4 mutants are viable however this species lacks DNA methylation. Knock down of WDR5, a core subunit of all six H3K4 methyltransferase complexes prevents stem cell renewal (Ang et al. 2011), while deletion is most likely lethal. Conflicting with the proposal that H3K4me3 protects CpG islands from methylation, a recent study suggests that CpG islands which are non-methylated in wild type ES cells do not become methylated in the absence of CFP1 despite reduced levels of H3K4me3 (Clouaire et al. 2012). This could suggest that H3K4me3 does not block DNA methylation, or that DNA methylation is blocked by additional factors, however interpretation of this result is confounded by the global reduction of DNMT1 and DNA methylation in the CFP1 null cell line. The constitutive CFP1 null cells may possess secondary adaptations critical to their viability. By using a conditional deletion of CFP1 the complications may be avoided, as the cells have not been selected to survival without CFP1. To directly determine whether H3K4me3 is required to protect CpG islands from DNA methylation, it may be possible to target an H3K4 demethylase to a specific CpG island, and monitor the loss of H3K4me3 and DNA methylation at that target. With recent improvements in site specific targeting strategies (Chen et al. 2013), this may now be feasible approach.

A third feedback loop may stem from the presence of an H3K4me2/3-specific PHD domain in a complex which methylates H3K4. Similar feedback loops have been described for several other histone modifications (Figure 53). The histone acetyltransferase GCN5 binds acetylated histones via a bromodomain (Yu et al. 1995). Similarly, the H3K9 methyltransferase complex Suv39h-HP1 binds to H3K9me3 via the chromodomain of HP1. Importantly, EZH2-EED (analogous the SET1-WDR5 complex) both places and binds to H3K27me3 via the EED WD40 domain (Hansen et al. 2008; Margueron et al. 2009). Perhaps most similar to CFP1-SET1, the MLL1 complex which methylates H3K4, contains a PHD domain (PHD3) which binds to H3K4me3 (Chang et al. 2010). As discussed above, modification-recognition loops may result in amplification through positive reinforcement, or they may serve to limit the modification through negative feedback. Considering the molecular context of the 'CFP1 PHD-H3K4me3 loop', recognition of H3K4me3 may inhibit or promote

methylation within the same nucleosome, methylation of a neighbouring nucleosome, or trans-methylation of a nearby nucleosome.

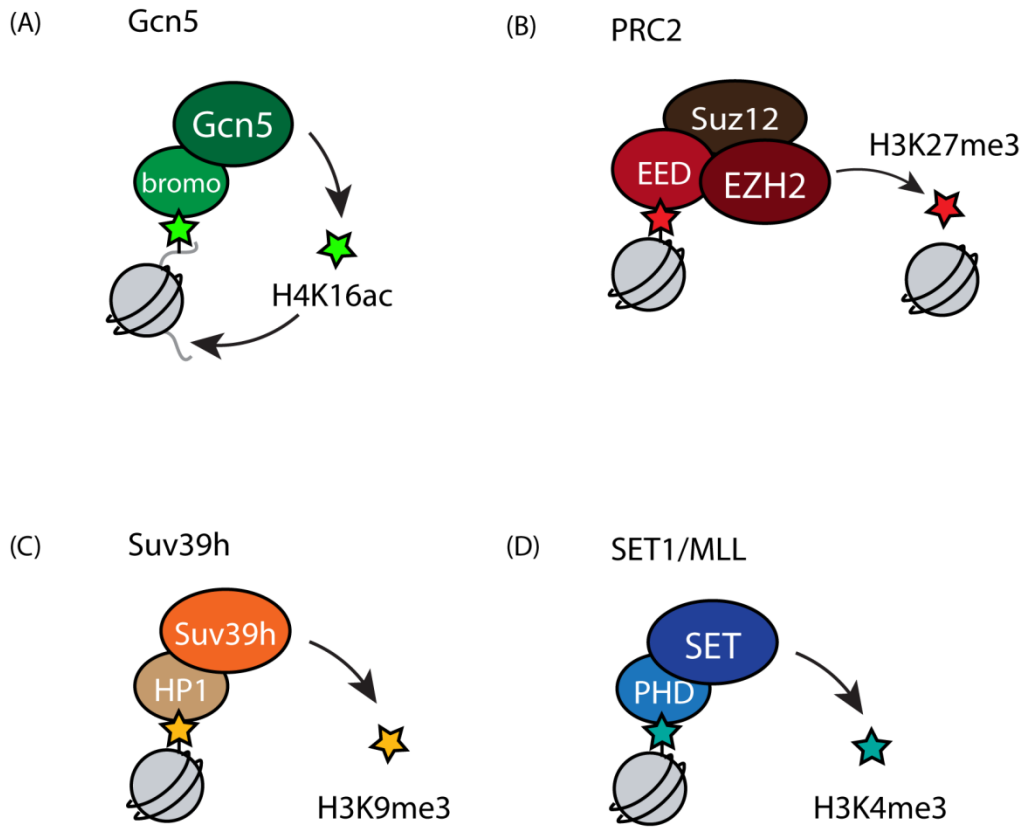



**Figure 53 - Feedback systems in histone modifying complexes**

Schematic illustrations depict four examples of complexes which interact specifically with the histone modifications they produce. (A) Recognition of H4K16 acetylation by the Gcn5 bromodomain stimulates acetylation of the opposite H4 tail within the same nucleosome. (B) Recognition of H3K27me3, by the EED subunit of PCR2, can stimulate the methylation of K27 on the H3 tail of other nucleosomes. (C) The H3K9 methyltransferase complex Suv39h-HP1 binds to H3K9me3 via the HP1 chromodomain which may help to propagate this heterochromatic mark. (D) PHD domains specific for H3K4me3 are present in SET1A/B and MLL1 methyltransferase complexes, potentially contributing to the methylation of H3K4.

In the GCN5 example, H4K16ac was shown to promote acetylation of the other H4 tail within the same nucleosome (Li & Shogren-Knaak 2009). In the EZH2-EED (PRC2 complex) example, recognition

of H3K27me3 by EED stimulates trans-methylation by EZH2 (Margueron & Reinberg 2010), facilitating the spread of the H3K27me3 along the chromatin fibre (Suganuma & Workman 2010). By analogy, a similar WD40 protein (WDR5) which recognises H3KR2me2s, may facilitate spreading of H3K4 methylation by the SET1A/B and MLL1-4 complexes (Migliori et al. 2012). This is a particularly exciting possibility with regards to regulation of broad CpG islands, including the hox genes which encode key developmental transcription factors.

Importantly, recognition of H3K4me3 by PHD3 is not required for binding of MLL1 to chromatin, or for trimethylation of H3K4 interaction (Chang et al. 2010). However, H3K4me3 recognition is essential for MLL1-dependent activation of target genes, suggesting that H3K4me3 is dispensable for MLL1 recruitment and activity, but that recognition somehow alters the protein to produce downstream effects (Chang et al. 2010). Comparatively, the CFP1 PHD mutant, used in this study remained able to bind chromatin, albeit at lower levels. It would be interesting to know how mutation of the CFP1 PHD domain affects the localisation and enrichment of SET1A and H3K4me3.

Unlike the PRC2, which appears able to spread the H3K27me3 modification over hundreds of kilobases, the punctate profile of H3K4me3 suggests that SET1A/B and MLL1-4 complexes are only able to spread this activating modification over shorter distances. More importantly the spread of this mark appears to be limited to non-methylated DNA, consistent with a restrictive role for the zf-CXXC domains of CFP1, MLL1 and MLL2. Whether the PHD domain of CFP1 plays a positive or a negative role in H3K4 methylation, it appears to operate between the tendency of the zf-CXXC domain to restrict methylation, and a tendency of WDR5 to spread methylation. Considering that the PHD domain increases the enrichment of CFP1 at non-methylated DNA it seems most likely that it amplifies H3K4me3 within these regions.

## 8.6 How does CFP1 contribute to the function of CpG islands?

Classical examples of epigenetic regulation suggest that although chromatin may potentiate or silence gene promoters, specific transcription factors are required to stimulate productive transcription. A considerable fraction of vertebrate genomes is constitutively silenced by DNA methylation and heterochromatic histone modifications. During development, additional regions of the genome may become silenced in a tissue-specific manner, and may subsequently become methylated by DNMT3a/b. By remaining non-methylated, CpG islands can be recognised by trithorax-group proteins associated with potentiation (CFP1, MLL1 and MLL2) and by polycomb-group proteins which are associated with gene silencing.

Like MLL1/2, CFP1 potentiates transcription, and resists silencing of CpG islands through the methylation of H3K4 (Thomson et al. 2010; Carlone et al. 2005). CFP1 promotes H3K4 methylation, by stabilising the SET1 methyltransferase enzymes (Soares et al. 2014; Carlone et al. 2005; Tate et al. 2009), and appears to be required for the complex to produce H3K4me3, which is specifically recognised by several chromatin remodelling complexes (Li et al. 2006; Wysocka et al. 2006; Shi et al. 2006). Evidence from this study suggests that CFP1 is highly dynamic in the nucleus, interacting transiently but specifically with chromatin through its zf-CXXC and PHD domains. This may allow a limited number of CFP1 molecules to sample an excess of non-methylated CpGs. Cooperative binding to regions with H3K4me2/3 and non-methylated CpGs may stabilise the CFP1-SET1 complex enabling restricted propagation of H3K4me3 mark, although stochastic sampling may allow limited placement of H3K4me3 at all non-methylated regions (Klose et al. 2013). Importantly, preferential enrichment of CFP1 at a subset of gene promoters suggests that additional factors are involved in localisation of the SET1 complex. Based on the positive correlation between CFP1 and Pol II, and the suggestion that CFP1 is most highly enriched at active genes (Denissov et al. 2014), it is likely that CFP1 is recruited to these sites via interactions between Pol II, and the SET1 complex. Using the

point mutations characterised in this study to disrupt specific interactions, it will be exciting to discover how enrichment of CFP1, SET1 and H3K4me3 is established during gene induction.

## 8.7 Conclusion

Far from being a simple adaptor molecule which tethers the SET1 complex to non-methylated DNA, CFP1 combines multiple interactions with distinct binding properties, and appears to be essential for tight regulation of the SET1 complex - both facilitating and restricting its activity.

# References


Allfrey, V., Faulkner, R. & Mirsky, A., 1964. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 315(1938), pp.786–794.

An, S. et al., 2011. Crystal structure of the human histone methyltransferase ASH1L catalytic domain and its implications for the regulatory mechanism. *The Journal of biological chemistry*, 286(10), pp.8369–74.

Ang, Y.-S. et al., 2011. Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network. *Cell*, 145(2), pp.183–97.

Antequera, F., 2003. Structure, function and evolution of CpG island promoters. *Cellular and Molecular Life Sciences CMLS*, 60, pp.1647–1658.

Aravind, L., Abhiman, S. & Iyer, L.M., 2011. Natural history of the eukaryotic chromatin protein methylation system. *Progress in Molecular Biology and Translational Science*, 101, pp.105–176.

Ardehali, M.B. et al., 2011. Drosophila Set1 is the major histone H3 lysine 4 trimethyltransferase with role in transcription. *The EMBO journal*, 30(14), pp.2817–28.

Aslanidis, C. & de Jong, P.J., 1990. Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic acids research*, 18(20), pp.6069–74.

Badenhorst, P. et al., 2005. The Drosophila nucleosome remodeling factor NURF is required for Ecdysteroid signaling and metamorphosis. *Genes & development*, 19(21), pp.2540–5.

Bancaud, A. et al., 2009. Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin. *The EMBO journal*, 28(24), pp.3785–98.

Bannister, a J. et al., 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*, 410(6824), pp.120–4.

Bannister, a J. & Kouzarides, T., 1996. The CBP co-activator is a histone acetyltransferase. *Nature*, 384, pp.641–643.

Barr, M.L. & Bertram, E.G., 1949. A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature*, 163, p.676.

Barski, A. et al., 2007. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), pp.823–37.

Bell, O. et al., 2007. Localized H3K36 methylation states define histone H4K16 acetylation during transcriptional elongation in Drosophila. *The EMBO journal*, 26(24), pp.4974–84.

Berger, S.L., 2007. The complex language of chromatin regulation during transcription. *Nature*, 447(7143), pp.407–412.

Bernard, P. & Allshire, R., 2002. Centromeres become unstuck without heterochromatin. *Trends in cell biology*, 12(9), pp.419–24.

Bernstein, B.E. et al., 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2), pp.315–26.

Bernstein, B.E. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Bernstein, B.E. et al., 2005. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2), pp.169–81.

Bestor, T. et al., 1988. Cloning and sequencing of a cDNA encoding DNA methyltransferase of mouse cells. The carboxyl-terminal domain of the mammalian enzymes is related to bacterial restriction methyltransferases. *Journal of molecular biology*, 203(4), pp.971–83.

Bestor, T.H. & Verdine, G.L., 1994. DNA methyltransferases. *Current opinion in cell biology*, 6(3), pp.380–9.

Bird, a P., 1986. CpG-rich islands and the function of DNA methylation. *Nature*, 321, pp.209–213.

Bird, A. et al., 1985. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell*, 40(1), pp.91–99.

Bird, A.P., 1987. CpG islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, 3(12), pp.342–347.

Birke, M. et al., 2002. The MT domain of the proto-oncoprotein MLL binds to CpG-containing DNA and discriminates against methylation. *Nucleic acids research*, 30(4), pp.958–965.

Birney, E. et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799–816.

Blackledge, N.P. et al., 2012. Bio-CAP: a versatile and highly sensitive technique to purify and characterise regions of non-methylated DNA. *Nucleic acids research*, 40(4), p.e32.

Blackledge, N.P. et al., 2010. CpG islands recruit a histone H3 lysine 36 demethylase. *Mol Cell*, 38(2), pp.179–190.

Blackledge, N.P. & Klose, R., 2011. CpG island chromatin: A platform for gene regulation. *Epigenetics*, 6(2), pp.147–152.

Bledau, A.S. et al., 2014. The H3K4 methyltransferase Setd1a is first required at the epiblast stage, whereas Setd1b becomes essential after gastrulation. *Development (Cambridge, England)*, 141(5), pp.1022–35.

Boggs, B. a et al., 2002. Differentially methylated forms of histone H3 show unique association patterns with inactive human X chromosomes. *Nature genetics*, 30(1), pp.73–6.

Boyes, J. & Bird, A., 1991. DNA methylation inhibits transcription indirectly via a methyl-CpG binding protein. *Cell*, 64, pp.1123–1134.

Briggs, S.D. et al., 2001. Histone H3 lysine 4 methylation is mediated by Set1 and required for cell growth and rDNA silencing in Saccharomyces cerevisiae. *Genes & development*, 15(24), pp.3286–95.

Brown, S.W., 1966. Heterochromatin. *Science*, 151(3709), pp.417–425.

Brownell, J.E. et al., 1996. Tetrahymena histone acetyltransferase A: A homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell*, 84(6), pp.843–851.

Brownell, J.E. & Allis, C.D., 1995. An activity gel assay detects a single, catalytically active histone acetyltransferase subunit in Tetrahymena macronuclei. *Proceedings of the National Academy of Sciences of the United States of America*, 92(14), pp.6364–8.

Brykczynska, U. et al., 2010. Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. *Nature structural & molecular biology*, 17(6), pp.679–87.

Buratowski, S. et al., 1989. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell*, 56(4), pp.549–61.

Burke, T.W. & Kadonaga, J.T., 1996. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes & Development*, 10(6), pp.711–724.

Bushnell, D. a et al., 2004. Structural basis of transcription: an RNA polymerase II-TFIIB cocrystal at 4.5 Angstroms. *Science (New York, N.Y.)*, 303(5660), pp.983–8.

Butler, J.E.F. & Kadonaga, J.T., 2002. The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes & development*, 16(20), pp.2583–92.

Butler, J.S., Lee, J.-H. & Skalnik, D.G., 2008. CFP1 interacts with DNMT1 independently of association with the Setd1 Histone H3K4 methyltransferase complexes. *DNA and cell biology*, 27(10), pp.533–43.

Butler, V.P. & Beiser, S.M., 1973. Antibodies to small molecules: biological and clinical applications. *Advances in immunology*, 17, pp.255–310.

Cao, F. et al., 2010. An Ash2L/RbBP5 heterodimer stimulates the MLL1 methyltransferase activity through coordinated substrate interactions with the MLL1 SET domain. *PloS one*, 5(11), p.e14102.

Cao, R. et al., 2002. Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science (New York, N.Y.)*, 298(5595), pp.1039–43.

Carlone, D.L. et al., 2005. Reduced Genomic Cytosine Methylation and Defective Cellular Differentiation in Embryonic Stem Cells Lacking CpG Binding Protein. *Mol Cell Biol*.

Carlone, D.L. & Skalnik, D.G., 2001. CpG Binding Protein Is Crucial for Early Embryonic Development CpG Binding Protein Is Crucial for Early Embryonic Development. *Mol Cell Biol*, 21(22).

Carlson, M. & Laurent, B.C., 1994. The SNF/SWI family of global transcriptional activators. *Current opinion in cell biology*, 6(3), pp.396–402.

Carrington, E. a & Jones, R.S., 1996. The Drosophila Enhancer of zeste gene encodes a chromosomal protein: examination of wild-type and mutant protein distribution. *Development (Cambridge, England)*, 122(12), pp.4073–83.

Carrozza, M.J. et al., 2005. Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell*, 123(4), pp.581–92.

Chang, P.-Y. et al., 2010. Binding of the MLL PHD3 finger to histone H3K4me3 is required for MLL-dependent gene transcription. *Journal of molecular biology*, 400(2), pp.137–44.

Chen, B. et al., 2013. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell*, 155(7), pp.1479–91.

Chen, F.E. et al., 1998. Crystal structure of p50/p65 heterodimer of transcription factor NF-kappaB bound to DNA. *Nature*, 391(6665), pp.410–3.

Chen, J. et al., 2014. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell*, 156(6), pp.1274–85.

Chen, Y. et al., 2011. Crystal structure of the N-terminal region of human Ash2L shows a winged-helix motif involved in DNA binding. *EMBO reports*, 12(8), pp.797–803.

Cheng, J. et al., 2014. A role for H3K4 monomethylation in gene repression and partitioning of chromatin readers. *Molecular cell*, 53(6), pp.979–92.

Cierpicki, T. et al., 2010. Structure of the MLL CXXC domain-DNA complex and its functional role in MLL-AF9 leukemia. *Nat Struct Mol Biol*, 17(1), pp.62–68.

Cléard, F., Delattre, M. & Spierer, P., 1997. SU(VAR)3-7, a Drosophila heterochromatin-associated protein and companion of HP1 in the genomic silencing of position-effect variegation. *The EMBO journal*, 16(17), pp.5280–8.

Clouaire, T. et al., 2012. Cfp1 integrates both CpG content and gene activity for accurate H3K4me3 deposition in embryonic stem cells. *Genes and Development*, 26, pp.1714–1728.

Collins, R.E. et al., 2005. In vitro and in vivo analyses of a Phe/Tyr switch controlling product specificity of histone lysine methyltransferases. *The Journal of biological chemistry*, 280(7), pp.5563–70.

Cooper, D.N., Taggart, M.H. & Bird, A.P., 1983. Unmethlated domains in vertebrate DNA. *Nucleic Acids Research*, 11(3), pp.647–658.

Côté, J. et al., 1994. Stimulation of GAL4 derivative binding to nucleosomal DNA by the yeast SWI/SNF complex. *Science (New York, N.Y.)*, 265(5168), pp.53–60.

Coulondre, C. et al., 1978. Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 274, pp.775–780.

Cross, S.H. et al., 1997. A component of the transcriptional repressor MeCP1 shares a motif with DNA methyltransferase and HRX proteins. *Nature genetics*, 16, pp.256–259.

Czermin, B. et al., 2002. Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites. *Cell*, 111(2), pp.185–96.

Deaton, A.M. et al., 2011. Cell type-specific DNA methylation at intragenic CpG islands in the immune system. *Genome research*, 21(7), pp.1074–86.

Deaton, A.M. & Bird, A., 2011. CpG islands and the regulation of transcription. *Genes & development*, 25(10), pp.1010–22.

Dehé, P.-M. et al., 2006. Protein interactions within the Set1 complex and their roles in the regulation of histone 3 lysine 4 methylation. *The Journal of biological chemistry*, 281(46), pp.35404–12.

DeLange, R.J. & Smith, E.L., 1971. The Structures of Histones. *Accounts of Chemical Research*, 1145(11), pp.368–373.

Denissov, S. et al., 2014. Mll2 is required for H3K4 trimethylation on bivalent promoters in embryonic stem cells, whereas Mll1 is redundant. *Development (Cambridge, England)*, (January), pp.526–537.

Dhalluin, C. et al., 1999. Structure and ligand of a histone acetyltransferase bromodomain. *Nature*, 399(June), pp.491–496.

Dharmarajan, V. et al., 2012. Structural basis for WDR5 interaction (Win) motif recognition in human SET1 family histone methyltransferases. *The Journal of biological chemistry*, 287(33), pp.27275–89.

Dou, Y. et al., 2006. Regulation of MLL1 H3K4 methyltransferase activity by its core components. *Nat Struct Mol Biol*, 13(8), pp.713–719.

Dover, J. et al., 2002. Methylation of histone H3 by COMPASS requires ubiquitination of histone H2B by Rad6. *The Journal of biological chemistry*, 277(32), pp.28368–71.

Duncan, B.K. & Miller, J.H., 1980. Mutagenic deamination of cytosine residues in DNA. *Nature*, 287, pp.560–561.

Eberl, H.C. et al., 2012. A Map of General and Specialized Chromatin Readers in Mouse Tissues Generated by Label-free Interaction Proteomics. *Molecular Cell*, 3, pp.1–11.

Edmunds, J.W., Mahadevan, L.C. & Clayton, A.L., 2008. Dynamic histone H3 methylation during gene induction: HYPB/Setd2 mediates all H3K36 trimethylation. *The EMBO journal*, 27(2), pp.406–20.

Eissenberg, J.C. et al., 1990. Mutation in a heterochromatin-specific chromosomal protein is associated with suppression of position-effect variegation in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24), pp.9923–7.

Eissenberg, J.C., 2012. Structural biology of the chromodomain: form and function. *Gene*, 496(2), pp.69–78.

Eissenberg, J.C. & Shilatifard, A., 2010. Histone H3 lysine 4 (H3K4) methylation in development and differentiation. *Developmental biology*, 339(2), pp.240–9.

Elkin, S.K. et al., 2005. A PHD finger motif in the C terminus of RAG2 modulates recombination activity. *The Journal of biological chemistry*, 280(31), pp.28701–10.

Eltsov, M. et al., 2008. Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. *Proceedings of the National Academy of Sciences of the United States of America*, 105(50), pp.19732–7.

Ernst, P. & Vakoc, C.R., 2012. WRAD: enabler of the SET1-family of H3K4 methyltransferases. *Briefings in functional genomics*.

Farcas, A.M. et al., 2012. KDM2B links the Polycomb Repressive Complex 1 (PRC1) to recognition of CpG islands. *eLife*, 1, pp.1–26.

Feng, Q., Wang, H. & Ng, H., 2002. Methylation of H3-lysine 79 is mediated by a new family of HMTases without a SET domain. *Current biology*, 12(02), pp.1052–1058.

FitzGerald, K.T. & Diaz, M.O., 1999. MLL2: A new mammalian member of the trx/MLL family of genes. *Genomics*, 59(2), pp.187–92.

Flemming, W., 1882. *Zellsubstanz, kern und zelltheilung*,

Franklin, C., McCulloch, A. & Kraft, A., 1995. In vitro association between the Jun protein family and the general transcription factors, TBP and TFIIB. *Biochem. J*, 974, pp.967–974.

Frauer, C. et al., 2011. Different binding properties and function of CXXC zinc finger domains in Dnmt1 and Tet1. *PloS one*, 6(2), p.e16627.

Fudenberg, G. & Mirny, L., 2012. Higher-order chromatin structure: bridging physics and biology. *Current opinion in genetics & development*, 22(2), pp.115–124.

Fujita, N. et al., 2000. Mechanism of transcriptional regulation by methyl-CpG binding protein MBD1. *Molecular and cellular biology*, 20(14), pp.5107–18.

Fujita, N. et al., 2003. Methyl-CpG binding domain 1 (MBD1) interacts with the Suv39h1-HP1 heterochromatic complex for DNA methylation-based transcriptional repression. *The Journal of biological chemistry*, 278(26), pp.24132–8.

Gardiner-Garden, M. & Frommer, M., 1987. CpG islands in vertebrate genomes. *Journal of molecular biology*, 196(2), pp.261–82.

Gebhardt, J.C.M. et al., 2013. Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. *Nature methods*, 10(5), pp.421–6.

Gnatt, a L. et al., 2001. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science (New York, N.Y.)*, 292(5523), pp.1876–82.

Goldberg, M.L., 1979. *Sequence Analysis of Drosophila Histone Genes*, Stanford University.

Goldknopf, I.L. & Busch, H., 1977. Isopeptide linkage between nonhistone and histone 2A polypeptides of chromosomal conjugate-protein A24. *Proceedings of the National Academy of Sciences of the United States of America*, 74(3), pp.864–8.

Gori, F., Divieti, P. & Demay, M.B., 2001. Cloning and characterization of a novel WD-40 repeat protein that dramatically accelerates osteoblastic differentiation. *The Journal of biological chemistry*, 276(49), pp.46515–22.

Gozani, O. et al., 2003. The PHD finger of the chromatin-associated protein ING2 functions as a nuclear phosphoinositide receptor. *Cell*, 114(1), pp.99–111.

Gruenbaum, Y., Cedar, H. & Razin, a, 1982. Substrate and sequence specificity of a eukaryotic DNA methylase. *Nature*, 295, pp.620–622.

Guenther, M.G. et al., 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1), pp.77–88.

Halbach, A. et al., 2009. Cotranslational assembly of the yeast SET1C histone methyltransferase complex. *The EMBO journal*, 28(19), pp.2959–70.

Hallson, G. et al., 2012. dSet1 is the main H3K4 di- and tri-methyltransferase throughout Drosophila development. *Genetics*, 190(1), pp.91–100.

Han, Z. et al., 2006. Structural basis for the specific recognition of methylated histone H3 lysine 4 by the WD-40 protein WDR5. *Molecular cell*, 22(1), pp.137–44.

Han, Z. et al., 2007. Structural basis of EZH2 recognition by EED. *Structure*, 15(10), pp.1306–15.

Hansen, K.H. et al., 2008. A model for transmission of the H3K27me3 epigenetic mark. *Nature cell biology*, 10(11), pp.1291–300.

Hassan, A., Neely, K. & Workman, J., 2001. Histone acetyltransferase complexes stabilize swi/snf binding to promoter nucleosomes. *Cell*, 104, pp.817–827.

Hebbes, T.R. et al., 1994. Core histone hyperacetylation co-maps with DNase I sensitivity in the chicken beta-globin chromosomal domain. *EMBO J*, 13(8), pp.1823–1830.

Heintzman, N.D. et al., 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3), pp.311–8.

Heitz, E., 1928. *Das Heterochromatin der Moose*, Bornträger.

Henikoff, S. & Shilatifard, A., 2011. Histone modification: cause or cog? *Trends in genetics : TIG*, 27(10), pp.389–96.

Herz, H.M. et al., 2012. Enhancer-associated H3K4 monomethylation by trithorax-related, the drosophila homolog of mammalian MLL3/MLL4. *Genes and Development*, 26, pp.2604–2620.

Hewish, D.R. & Burgoyne, L. a, 1973. Chromatin sub-structure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochemical and biophysical research communications*, 52(2), pp.504–510.

Hilliker, a J., Appels, R. & Schalet, a, 1980. The genetic analysis of D. melanogaster heterochromatin. *Cell*, 21(3), pp.607–619.

Holde, K. Van, 1988. Chromatin. *Springer Series in Molecular Biology*, 12.

Holliday, R. & Pugh, J.E., 1975. DNA modification mechanisms and gene activity during development. *Science (New York, N.Y.)*, 187(4173), pp.226–32.

Hollmann, M. et al., 2002. The essential Drosophila melanogaster gene wds (will die slowly) codes for a WD-repeat protein with seven repeats. *Molecular genetics and genomics*, 268(4), pp.425–33.

Hughes, R.M. et al., 2007. Recognition of trimethyllysine by a chromodomain is not driven by the hydrophobic effect. *Proceedings of the National Academy of Sciences of the United States of America*, 104(27), pp.11184–8.

Huntsman, D.G. et al., 1999. MLL2, the second human homolog of the Drosophila trithorax gene, maps to 19q13.1 and is amplified in solid tumor cell lines. *Oncogene*, 18(August), pp.7975–7984.

Iguchi-Ariga, S.M. & Schaffner, W., 1989. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. *Genes & Development*, 3(5), pp.612–619.

Illingworth, R. et al., 2008. A novel CpG island set identifies tissue-specific methylation at developmental gene loci. *PLoS biology*, 6(1), p.e22.

Illingworth, R.S. & Bird, A.P., 2009. CpG islands--'a rough guide'. *FEBS Lett*, 583(11), pp.1713–1720.

Imbalzano, a N. et al., 1994. Facilitated binding of TATA-binding protein to nucleosomal DNA. *Nature*, 370, pp.481–485.

Ingham, P., 1983. Differential expression of bithorax complex genes in the absence of the extra sex combs and trithorax genes. *Nature*.

Ingham, P.W., 1981. Trithorax: A new homoeotic mutation of Drosophila melanogaster - II. The role of trx+ after embryogenesis. *Wilhelm Roux's Archives of Developmental Biology*, 190, pp.365–369.

Ip, Y.T., Levine, M. & Small, S.J., 1992. The bicoid and dorsal morphogens use a similar strategy to make stripes in the Drosophila embryo. *Journal of cell science. Supplement*, 16, pp.33–8.

Jacobs, S. a et al., 2002. The active site of the SET domain is constructed on a knot. *Nature structural biology*, 9(11), pp.833–8.

Jacobs, S. a & Khorasanizadeh, S., 2002. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science (New York, N.Y.)*, 295(5562), pp.2080–3.

Jenuwein, T. & Allis, C.D., 2001. Translating the histone code. *Science (New York, N.Y.)*, 293(5532), pp.1074–80.

Jiang, H. et al., 2013. Regulation of transcription by the MLL2 complex and MLL complex-associated AKAP95. *Nature structural & molecular biology*, 20(10), pp.1156–63.

Jørgensen, H.F., Ben-porath, I. & Bird, A.P., 2004. Mbd1 Is Recruited to both Methylated and Nonmethylated CpGs via Distinct DNA Binding Domains. *Society*, 24(8), pp.3387–3395.

Joseph, A., Mitchell, A.R. & Miller, O.J., 1989. The organization of the mouse satellite DNA at centromeres. *Experimental cell research*, 183, pp.494–500.

Kadonaga, J.T., 2012. Perspectives on the RNA polymerase II core promoter. *Wiley Interdisip Rev Dev Biol.*

Kadonaga, J.T., 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell*, 116(2), pp.247–57.

Kang, M. et al., 2010. A quantitative approach to analyze binding diffusion kinetics by confocal FRAP. *Biophysical journal*, 99(9), pp.2737–47.

Karmodiya, K. et al., 2012. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC genomics*, 13(1), p.424.

Kaufmann, J. & Smale, S.T., 1994. Direct recognition of initiator elements by a component of the transcription factor IID complex. *Genes & Development*, 8(7), pp.821–829.

Keshet, I., Yisraeli, J. & Cedar, H., 1985. Effect of regional DNA methylation on gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 82(9), pp.2560–4.

Kim, J. et al., 2013. The n-SET Domain of Set1 Regulates H2B Ubiquitylation-Dependent H3K4 Methylation. *Molecular Cell*, pp.1–13.

Kimura, H. & Cook, P.R., 2001. Kinetics of core histones in living human cells: little exchange of H3 and H4 and some rapid exchange of H2B. *The Journal of cell biology*, 153(7), pp.1341–53.

Kirmizis, A. et al., 2007. Arginine methylation at histone H3R2 controls deposition of H3K4 trimethylation. *Nature*, 449(7164), pp.928–32.

Klein, C. & Struhl, K., 1994. Increased recruitment of TATA-binding protein to the promoter by transcriptional activation domains in vivo. *Science*, 266, pp.280–282.

Klose, R.J. et al., 2013. Chromatin sampling--an emerging perspective on targeting polycomb repressor proteins. *PLoS genetics*, 9(8), p.e1003717.

Knezetic, J. a & Luse, D.S., 1986. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell*, 45(1), pp.95–104.

Kornberg, R.D., 2007. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 104(32), pp.12955–61.

Kornberg, R.D. & Thomas, J.O., 1974. Chromatin structure: oligomers of the histones. *Science (New York, N.Y.)*, 184(4139), pp.865–8.

Kossel, A., 1928. The Protamines and Histones. *Monographs on Biochemistry*.

Koyama-Nasu, R., David, G. & Tanese, N., 2007. The F-box protein Fbl10 is a novel transcriptional repressor of c-Jun. *Nature cell biology*, 9(9), pp.1074–80.

Kuo, M.H. et al., 1998. Histone acetyltransferase activity of yeast Gcn5p is required for the activation of target genes in vivo. *Genes & development*, 12, pp.627–639.

Kuzmichev, A. et al., 2002. Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein. *Genes & development*, 16(22), pp.2893–905.

De la Cruz, X. et al., 2005. Do protein motifs read the histone code? *BioEssays : news and reviews in molecular, cellular and developmental biology*, 27(2), pp.164–75.

Laemmli, U.K., 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, 227, pp.680–685.

Lagrange, T. et al., 1998. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes & Development*, 12(1), pp.34–44.

Lauberth, S.M. et al., 2013. H3K4me3 Interactions with TAF3 Regulate Preinitiation Complex Assembly and Selective Gene Activation. *Cell*, 152(5), pp.1021–1036.

Lee, J.-H. et al., 2010. Identification and characterization of a novel human PP1 phosphatase complex. *The Journal of biological chemistry*, 285(32), pp.24466–76.

Lee, J.-H. et al., 2007. Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. *The Journal of biological chemistry*, 282(18), pp.13419–28.

Lee, J.H. & Skalnik, D.G., 2005. CpG-binding protein (CXXC finger protein 1) is a component of the mammalian Set1 histone H3-Lys4 methyltransferase complex, the analogue of the yeast Set1/COMPASS complex. *J Biol Chem*, 280(50), pp.41725–41731.

Lee, J.H. & Skalnik, D.G., 2002. CpG-binding protein is a nuclear matrix- and euchromatin-associated protein localized to nuclear speckles containing human trithorax: Identification of nuclear matrix targeting signals. *Journal of Biological Chemistry*, 277(44), pp.42259–42267.

Lee, J.-H. & Skalnik, D.G., 2008. Wdr82 is a C-terminal domain-binding protein that recruits the Setd1A Histone H3-Lys4 methyltransferase complex to transcription start sites of transcribed human genes. *Molecular and cellular biology*, 28(2), pp.609–18.

Lee, J.H., Voo, K.S. & Skalnik, D.G., 2001. Identification and characterization of the DNA binding domain of CpG-binding protein. *J Biol Chem*, 276(48), pp.44669–44676.

Lee, J.-S. et al., 2007. Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell*, 131(6), pp.1084–96.

Letunic, I., Doerks, T. & Bork, P., 2012. SMART 7: recent updates to the protein domain annotation resource. *Nucleic acids research*, 40(Database issue), pp.D302–5.

Lever, M. a et al., 2000. Rapid exchange of histone H1.1 on chromatin in living human cells. *Nature*, 408(6814), pp.873–6.

Lewis, E., 1949. P. Pc. Polycomb Drosophila. *Information Service*.

Lewis, E.B., 1978. A gene complex controlling segmentation in Drosophila. *Nature*, 276, pp.565–570.

Lewis, J.D. et al., 1992. Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell*, 69(6), pp.905–14.

Li, H. et al., 2006. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature*, 442(7098), pp.91–5.

Li, S. & Shogren-Knaak, M. a., 2009. The Gcn5 bromodomain of the SAGA complex facilitates cooperative and cross-tail acetylation of nucleosomes. *Journal of Biological Chemistry*, 284(14), pp.9411–9417.

Lister, R. et al., 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, 462(7271), pp.315–22.

Liu, H., Westergard, T.D. & Hsieh, J.J.-D., 2009. MLL5 governs hematopoiesis: a step closer. *Blood*, 113(7), pp.1395–6.

Liu, X. et al., 2010. Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. *Science (New York, N.Y.)*, 327(5962), pp.206–9.

Lohe, a R., Hilliker, a J. & Roberts, P. a, 1993. Mapping Simple Repeated DNA-Sequences in Heterochromatin of Drosophila-Melanogaster. *Genetics*, 134, pp.1149–1174.

Long, H.K., Sims, D., et al., 2013. Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *eLife*, 2, p.e00348.

Long, H.K., Blackledge, N.P. & Klose, R.J., 2013. ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochemical Society transactions*, 41(3), pp.727–40.

Lorch, Y., LaPointe, J.W. & Kornberg, R.D., 1987. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*, 49(2), pp.203–10.

Lowy, D.R., Rands, E. & Scolnick, E.M., 1978. Helper-independent transformation by unintegrated Harvey sarcoma virus DNA. *Journal of virology*, 26(2), pp.291–8.

Luger, K. et al., 1997. Crystal structure of the nucleosome core particle at 2 . 8 Angstrom resolution. *Nature*, 7, pp.251–260.

Lyon, M.F., 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.). *Nature*, 190, pp.372–373.

M. Lienhard Schmitz et al., 1995. Interaction of the COOH-terminal Transactivation Domain of p65 NF-kB with TATA-binding Protein, Transcription factor IIB and Coactivators. *Journal of Biological Chemistry*, 270, pp.7219–7226.

Ma, Q. et al., 1993. Analysis of the murine All-1 gene reveals conserved domains with human ALL-1 and identifies a motif shared with DNA methyltransferases. *Proceedings of the National Academy of Sciences of the United States of America*, 90(13), pp.6350–4.

Margueron, R. et al., 2009. Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature*, 461(7265), pp.762–7.

Margueron, R. & Reinberg, D., 2010. Chromatin structure and the inheritance of epigenetic information. *Nature Reviews Genetics*, 11(4), pp.285–296.

Markaki, Y. et al., 2010. Functional nuclear organization of transcription and DNA replication: a topographical marriage between chromatin domains and the interchromatin compartment. *Cold Spring Harbor symposia on quantitative biology*, 75, pp.475–92.

Markaki, Y. et al., 2012. The potential of 3D-FISH and super-resolution structured illumination microscopy for studies of 3D nuclear architecture: 3D structured illumination microscopy of defined chromosomal structures visualized by 3D (immuno)-FISH opens new perspectives for stud. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 34(5), pp.412–26.

Marmorstein, R., 2003. Structure of SET domain proteins: a new twist on histone methylation. *Trends in biochemical sciences*, 28(2), pp.59–62.

Matthews, A.G.W. et al., 2007. RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. *Nature*, 450(7172), pp.1106–10.

Mazza, D. et al., 2012. A benchmark for chromatin binding measurements in live cells. *Nucleic Acids Research*, 40(15), pp.1–13.

Mellor, J., 2006. It takes a PHD to read the histone code. *Cell*, 126(1), pp.22–4.

Metz, R. et al., 1994. c-Fos-induced activation of a TATA-box-containing promoter involves direct contact with TATA-box-binding protein. *Molecular and cellular biology*, 14(9), pp.6021–9.

Migliori, V., Mapelli, M. & Guccione, E., 2012. On WD40 proteins propelling our knowledge of transcriptional control? *Epigenetics*, 7(August), pp.815–822.

Mikkelsen, T.S. et al., 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153), pp.553–60.

Miller, T. et al., 2001. COMPASS: a complex of proteins associated with a trithorax-related SET domain protein. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23), pp.12902–7.

Milne, T. et al., 2002. MLL Targets SET Domain Methyltransferase Activity to Hox Gene Promoters. *Molecular cell*, 10, pp.1107–1117.

Mizzen, C. a. et al., 1996. The TAF(II)250 subunit of TFIID has histone acetyltransferase activity. *Cell*, 87(7), pp.1261–1270.

Mohan, M. et al., 2011. The COMPASS family of H3K4 methylases in Drosophila. *Molecular and cellular biology*, 31(21), pp.4310–8.

Morse, R.H., 2007. Transcription factor access to promoter elements. *Journal of cellular biochemistry*, 102(3), pp.560–70.

Mueller, F. et al., 2012. Monitoring Dynamic Binding of Chromatin Proteins In Vivo by Fluorescence Recovery After Photobleaching R. H. Morse, ed. *Methods in Molecular Biology*, 833, pp.153–176.

Müller, F., Demény, M. a & Tora, L., 2007. New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. *The Journal of biological chemistry*, 282(20), pp.14685–9.

Muller, H.J., 1930. Types of visible variations induced by X-rays in Drosophila. *Journal of Genetics*, 22(3), pp.299–334.

Müller, J. et al., 2002. Histone methyltransferase activity of a Drosophila Polycomb group repressor complex. *Cell*, 111(2), pp.197–208.

Murray, K., 1964. the Occurrence of Epsilon-N-Methyl Lysine in Histones. *Biochemistry*, 3, pp.10–15.

Murton, B.L. et al., 2010. Characterising the binding specificities of the subunits associated with the KMT2/Set1 histone lysine methyltransferase. *Journal of molecular biology*, 398(4), pp.481–8.

Musselman, C. a & Kutateladze, T.G., 2011. Handpicking epigenetic marks with PHD fingers. *Nucleic acids research*, 39(21), pp.9061–9071.

Mutskov, V. & Felsenfeld, G., 2004. Silencing of transgene transcription precedes methylation of promoter DNA and histone H3 lysine 9. *The EMBO journal*, 23(1), pp.138–49.

Nady, N. et al., 2008. A SPOT on the chromatin landscape? Histone peptide arrays as a tool for epigenetic research. *Trends in biochemical sciences*, 33(7), pp.305–13.

Nedea, E. et al., 2008. The Glc7 phosphatase subunit of the cleavage and polyadenylation factor is essential for transcription termination on snoRNA genes. *Molecular cell*, 29(5), pp.577–87.

Ng, H.H., Jeppesen, P. & Bird, A., 2000. Active repression of methylated genes by the chromosomal protein MBD1. *Molecular and cellular biology*, 20(4), pp.1394–406.

Nielsen, P.R. et al., 2002. Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature*, 416(6876), pp.103–7.

Nikolov, D.B. et al., 1992. Crystal structure of TFIID TATA-box binding protein. *Nature*, 360, pp.40–46.

Nishino, Y. et al., 2012. Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *The EMBO Journal*, 31(7), pp.1644–1653.

Van Nuland, R., Schram, A.W., et al., 2013. Multivalent engagement of TFIID to nucleosomes. *PloS one*, 8(9), p.e73495.

Van Nuland, R., Smits, A.H., et al., 2013. Quantitative dissection and stoichiometry determination of the human SET1/MLL histone methyltransferase complexes. *Molecular and cellular biology*, 33(10), pp.2067–77.

Ohki, I. et al., 2001. Solution structure of the methyl-CpG binding domain of human MBD1 in complex with methylated DNA. *Cell*, 105(4), pp.487–97.

Ohno, S., Kaplan, W. & Kinosita, R., 1959. Formation of the sex chromatin by a single X-chromosome in liver cells of Rattus norvegicus. *Experimental cell research*, pp.415–418.

Okano, M. et al., 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3), pp.247–57.

Olins, a L. & Olins, D.E., 1974. Spheroid chromatin units (v bodies). *Science (New York, N.Y.)*, 183, pp.330–332.

Orlando, D. a et al., 2012. CpG Island Structure and Trithorax/Polycomb Chromatin Domains in Human Cells. *Genomics*, pp.1–7.

Palacios, A. et al., 2008. Molecular basis of histone H3K4me3 recognition by ING4. *The Journal of biological chemistry*, 283(23), pp.15956–64.

Parker, J., Guo, D. & Hodges, R., 1986. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X. *Biochemistry*, (1978), pp.5425–5432.

Patel, A., Vought, V.E., et al., 2008. A conserved arginine-containing motif crucial for the assembly and enzymatic activity of the mixed lineage leukemia protein-1 core complex. *The Journal of biological chemistry*, 283(47), pp.32162–75.

Patel, A. et al., 2009. On the mechanism of multiple lysine methylation by the human mixed lineage leukemia protein-1 (MLL1) core complex. *The Journal of biological chemistry*, 284(36), pp.24242–56.

Patel, A., Dharmarajan, V. & Cosgrove, M.S., 2008. Structure of WDR5 bound to mixed lineage leukemia protein-1 peptide. *The Journal of biological chemistry*, 283(47), pp.32158–61.

Patnaik, D. et al., 2004. Substrate specificity and kinetic mechanism of mammalian G9a histone H3 methyltransferase. *The Journal of biological chemistry*, 279(51), pp.53248–58.

Pazin, M.J. & Kadonaga, J.T., 1997. SWI2/SNF2 and related proteins: ATP-driven motors that disrupt protein-DNA interactions? *Cell*, 88(6), pp.737–40.

Pelegri, F. & Lehmann, R., 1994. A role of polycomb group genes in the regulation of gap gene expression in Drosophila. *Genetics*, 136, pp.1341–1353.

Peña, P. V et al., 2008. Histone H3K4me3 binding is required for the DNA repair and apoptotic activities of ING1 tumor suppressor. *Journal of molecular biology*, 380(2), pp.303–12.

Peña, P. V et al., 2006. Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature*, 442(7098), pp.100–3.

Pengelly, A.R. et al., 2013. A histone mutant reproduces the phenotype caused by loss of histone-modifying factor Polycomb. *Science (New York, N.Y.)*, 339(6120), pp.698–9.

Peters, A.H. et al., 2001. Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell*, 107(3), pp.323–37.

Pirrotta, V., 1998. Polycombing the genome: PcG, trxG, and chromatin silencing. *Cell*, 93(3), pp.333–6.

Pokholok, D.K. et al., 2005. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122(4), pp.517–27.

Pradhan, M. et al., 2008. CXXC domain of human DNMT1 is essential for enzymatic activity. *Biochemistry*, 47(38), pp.10000–9.

Radman-Livaja, M. et al., 2010. Replication and active demethylation represent partially overlapping mechanisms for erasure of H3K4me3 in budding yeast. *PLoS genetics*, 6(2), p.e1000837.

Ramón-Maiques, S. et al., 2007. The plant homeodomain finger of RAG2 recognizes histone H3 methylated at both lysine-4 and arginine-2. *Proceedings of the National Academy of Sciences of the United States of America*, 104(48), pp.18993–8.

Ramsahoye, B.H. et al., 2000. Non-CpG methylation is prevalent in embryonic stem cells and may be mediated by DNA methyltransferase 3a. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10), pp.5237–42.

Rea, S. et al., 2000. Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, 406(6796), pp.593–9.

Riggs, A., 1975. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research*.

Rodda, D.J. et al., 2005. Transcriptional regulation of nanog by OCT4 and SOX2. *The Journal of biological chemistry*, 280(26), pp.24731–7.

Roguev, a et al., 2001. The Saccharomyces cerevisiae Set1 complex includes an Ash2 homologue and methylates histone 3 lysine 4. *The EMBO journal*, 20(24), pp.7137–48.

Rose, N.R. & Klose, R.J., 2014. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochimica et biophysica acta*.

Ruthenburg, A.J. et al., 2006. Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. *Nature structural & molecular biology*, 13(8), pp.704–12.

Sambrook, J. & Russell, D.W., 2001. *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press.

Sanchez, R. & Zhou, M.-M., 2011. The PHD finger: a versatile epigenome reader. *Trends in biochemical sciences*, 36(7), pp.364–72.

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

Sarma, K. & Reinberg, D., 2005. Histone variants meet their match. *Nature reviews. Molecular cell biology*, 6(2), pp.139–49.

Sarraf, S. a. & Stancheva, I., 2004. Methyl-CpG binding protein MBD1 couples histone H3 methylation at lysine 9 by SETDB1 to DNA replication and chromatin assembly. *Molecular Cell*, 15, pp.595–605.

Sato, M.H. et al., 1999. Atomic force microscopy sees nucleosome positioning and histone H1-induced compaction in reconstituted chromatin. *FEBS letters*, 452(3), pp.267–71.

Sawadogo, M. & Roeder, R.G., 1985. Factors involved in specific transcription by human RNA polymerase II: analysis by a rapid and quantitative in vitro assay. *Proceedings of the National Academy of Sciences of the United States of America*, 82(July), pp.4394–8.

Saxonov, S., Berg, P. & Brutlag, D.L., 2006. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5), pp.1412–7.

Schermelleh, L. et al., 2007. Dynamics of Dnmt1 interaction with the replication machinery and its role in postreplicative maintenance of DNA methylation. *Nucleic acids research*, 35(13), pp.4301–12.

Schindler, U., Beckmann, H. & Cashmore, a R., 1993. HAT3.1, a novel Arabidopsis homeodomain protein containing a conserved cysteine-rich region. *The Plant journal : for cell and molecular biology*, 4, pp.137–150.

Schneider, J. et al., 2005. Molecular regulation of histone H3 trimethylation by COMPASS and the regulation of gene expression. *Molecular cell*, 19(6), pp.849–56.

Schuettengruber, B. et al., 2007. Genome regulation by polycomb and trithorax proteins. *Cell*, 128(4), pp.735–45.

Schultz, J. et al., 1998. SMART, a simple modular architecture research tool: identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), pp.5857–64.

Schwartz, Y.B. et al., 2006. Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nature genetics*, 38(6), pp.700–5.

Shi, X. et al., 2006. ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. *Nature*, 442(7098), pp.96–9.

Shi, X. et al., 2007. 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. *The Journal of biological chemistry*, 282(4), pp.2450–5.

Shogren-Knaak, M. et al., 2006. Histone H4-K16 acetylation controls chromatin structure and protein interactions. *Science (New York, N.Y.)*, 311(5762), pp.844–847.

Simon, M.D. et al., 2007. The site-specific installation of methyl-lysine analogs into recombinant histones. *Cell*, 128(5), pp.1003–12.

Sims, R.J. et al., 2005. Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *The Journal of biological chemistry*, 280(51), pp.41789–92.

Sims, R.J. et al., 2007. Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing. *Molecular cell*, 28(4), pp.665–76.

Smeets, D. et al., 2014. Three-dimensional super-resolution microscopy of the inactive X chromosome territory reveals a collapse of its active nuclear compartment harboring distinct Xist RNA foci. *Epigenetics & chromatin*, 7(1), p.8.

Smith, S.T. et al., 2004. Modulation of heat shock gene expression by the TAC1 chromatin-modifying complex. *Nature cell biology*, 6(2), pp.162–7.

Soares, L.M. et al., 2014. Feedback Control of Set1 Protein Levels Is Important for Proper H3K4 Methylation Patterns. *Cell Reports*, 6(6), pp.961–972.

Sommermeyer, V. et al., 2013. Spp1, a Member of the Set1 Complex, Promotes Meiotic DSB Formation in Promoters by Tethering Histone H3K4 Methylation Sites to Chromosome Axes. *Molecular Cell*, 49, pp.43–54.

Song, J. et al., 2011. Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation. *Science (New York, N.Y.)*, 331(6020), pp.1036–40.

Song, J.-J. & Kingston, R.E., 2008. WDR5 interacts with mixed lineage leukemia (MLL) protein via the histone H3-binding pocket. *The Journal of biological chemistry*, 283(50), pp.35258–64.

Soumpasis, D.M., 1983. Theoretical analysis of fluorescence photobleaching recovery experiments. *Biophysical journal*, 41(January), pp.95–97.

Spencer, T.E. et al., 1997. Steroid receptor coactivator-1 is a histone acetyltransferase. *Nature*, 389(6647), pp.194–198.

Spitz, F. & Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. *Nature reviews. Genetics*, 13(9), pp.613–26.

Sprague, B.L. et al., 2004. Analysis of binding reactions by fluorescence recovery after photobleaching. *Biophysical journal*, 86(6), pp.3473–95.

Sprague, B.L. & McNally, J.G., 2005. FRAP analysis of binding: proper and fitting. *Trends in cell biology*, 15(2), pp.84–91.

Stadler, M.B. et al., 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480(7378), pp.490–5.

Steward, M.M. et al., 2006. Molecular regulation of H3K4 trimethylation by ASH2L, a shared subunit of MLL complexes. *Nature structural & molecular biology*, 13(9), pp.852–4.

Strahl, B.D. et al., 1999. Methylation of histone H3 at lysine 4 is highly conserved and correlates with transcriptionally active nuclei in Tetrahymena. *Proceedings of the National Academy of Sciences of the United States of America*, 96(26), pp.14967–72.

Strahl, B.D. et al., 2002. Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression. *Molecular and cellular biology*, 22, pp.1298–1306.

Strunnikova, M. et al., 2005. Chromatin inactivation precedes de novo DNA methylation during the progressive epigenetic silencing of the RASSF1A promoter. *Molecular and cellular biology*, 25, pp.3923–3933.

Suganuma, T. & Workman, J.L., 2010. WD40 repeats arrange histone tails for spreading of silencing. *Journal of molecular cell biology*, 2(2), pp.81–3.

Sun, Z.-W. & Allis, C.D., 2002. Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature*, 418(6893), pp.104–8.

Sundin, O. & Varshavsky, a, 1981. Arrest of segregation leads to accumulation of highly intertwined catenated dimers: dissection of the final stages of SV40 DNA replication. *Cell*, 25(3), pp.659–69.

Tachibana, M. et al., 2001. Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3. *The Journal of biological chemistry*, 276(27), pp.25309–17.

Takahashi, K. & Yamanaka, S., 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4), pp.663–76.

Takahashi, Y.H. et al., 2009. Regulation of H3K4 trimethylation via Cps40 (Spp1) of COMPASS is monoubiquitination independent: implication for a Phe/Tyr switch by the catalytic domain of Set1. *Molecular and cellular biology*, 29(13), pp.3478–86.

Tate, C.M., Lee, J.H. & Skalnik, D.G., 2009. CXXC finger protein 1 contains redundant functional domains that support embryonic stem cell cytosine methylation, histone methylation, and differentiation. *Mol Cell Biol*, 29(14), pp.3817–3831.

Tate, C.M., Lee, J.-H.H. & Skalnik, D.G., 2010. CXXC finger protein 1 restricts the Setd1A histone H3K4 methyltransferase complex to euchromatin. *FEBS J*, 277(1), pp.210–223.

Thomas, J.O. & Kornberg, R.D., 1975. An octamer of histones in chromatin and free in solution. *Proceedings of the National Academy of Sciences of the United States of America*, 72(7), pp.2626–2630.

Thomson, J.P. et al., 2010. CpG islands influence chromatin structure via the CpG-binding protein Cfp1- Supplementary Information. *Nature*, 464(7291), pp.1082–6.

Tie, F. et al., 2014. Trithorax monomethylates histone H3K4 and interacts directly with CBP to promote H3K27 acetylation and antagonize Polycomb silencing. *Development (Cambridge, England)*, 141(5), pp.1129–39.

Tirode, F. et al., 1999. Reconstitution of the transcription factor TFIIH: assignment of functions for the three enzymatic subunits, XPB, XPD, and cdk7. *Molecular cell*, 3(1), pp.87–95.

Trievel, R.C. et al., 2002. Structure and catalytic mechanism of a SET domain protein methyltransferase. *Cell*, 111(1), pp.91–103.

Tschiersch, B. et al., 1994. The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes. *The EMBO journal*, 13(16), pp.3822–3831.

Tsukada, Y. et al., 2006. Histone demethylation by a family of JmjC domain-containing proteins. *Nature*, 439(7078), pp.811–816.

Vastenhouw, N.L. & Schier, A.F., 2012. Bivalent histone modifications in early embryogenesis. *Current opinion in cell biology*, 24(3), pp.374–86.

Venters, B.J. & Pugh, B.F., 2013. Genomic organization of human transcription initiation complexes. Retracted! *Nature*, 502(7469), pp.53–8.

Vermeulen, M. et al., 2007. Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4. *Cell*, 131(1), pp.58–69.

Voo, K.S. et al., 2000. Cloning of a mammalian transcriptional activator that binds unmethylated CpG motifs and shares a CXXC domain with DNA methyltransferase, human trithorax, and methyl-CpG binding domain protein 1. *Molecular and cellular biology*, 20(6), pp.2108–21.

Wachsmuth, M., Waldeck, W. & Langowski, J., 2000. Anomalous diffusion of fluorescent probes inside living cell nuclei investigated by spatially-resolved fluorescence correlation spectroscopy. *Journal of molecular biology*, 298(4), pp.677–89.

Wallrath, L.L. & Elgin, S.C.R., 1995. Position effect variegation in Drosophila is associated with an altered chromatin structure. *Genes and Development*, 9(10), pp.1263–1277.

Wang, L. et al., 1997. Histone acetyltransferase activity is conserved between yeast and human GCN5 and is required for complementation of growth and transcriptional activation. *Molecular and cellular biology*, 17(1), pp.519–27.

Wang, P. et al., 2009. Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. *Molecular and cellular biology*, 29(22), pp.6074–85.

Wang, Z. et al., 2008. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, 40(7), pp.897–903.

Weaver, D.T., Fields-Berry, S.C. & DePamphilis, M.L., 1985. The termination region for SV40 DNA replication directs the mode of separation for the two sibling molecules. *Cell*, 41(2), pp.565–75.

Weil, P. a et al., 1979. Selective and accurate initiation of transcription at the Ad2 major late promotor in a soluble system dependent on purified RNA polymerase II and DNA. *Cell*, 18(2), pp.469–84.

Weiler, K.S. & Wakimoto, B.T., 1995. Heterochromatin and gene expression in Drosophila. *Annual review of genetics*, 29, pp.577–605.

Welling, G.W. et al., 1985. Prediction of sequential antigenic regions in proteins. *FEBS letters*, 188(2), pp.215–8.

Wines, D.R. et al., 1996. Introduction of a DNA methytransferase into Drosophila to probe chromatin structure in vivo. *Chromosoma*, 104, pp.332–340.

Winston, F. & Carlson, M., 1992. Yeast SNF/SWI transcriptional activators and the SPT/SIN chromatin connection. *Trends in Genetics*, 8(11), pp.387–391.

Wong, A.K. & Rattner, J.B., 1988. Sequence organization and cytological localization of the minor satellite of mouse. *Nucleic acids research*, 16, pp.11645–11661.

Wu, J. & Grunstein, M., 2000. 25 Years After the Nucleosome Model: Chromatin Modifications. *Trends in biochemical sciences*, 25(12), pp.619–23.

Wu, M. et al., 2008. Molecular regulation of H3K4 trimethylation by Wdr82, a component of human Set1/COMPASS. *Molecular and cellular biology*, 28(24), pp.7337–44.

Wysocka, J. et al., 2006. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*, 442(7098), pp.86–90.

Wysocka, J. et al., 2005. WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. *Cell*, 121(6), pp.859–72.

Xiao, Y. et al., 2011. Caenorhabditis elegans chromatin-associated proteins SET-2 and ASH-2 are differentially required for histone H3 Lys 4 methylation in embryos and adult germ cells. *Proceedings of the National Academy of Sciences of the United States of America*, 108(20), pp.8305–10.

Xu, C. et al., 2011. The structural basis for selective binding of non-methylated CpG islands by the CFP1 CXXC domain. *Nature communications*, 2, p.227.

Xu, Y. et al., 2012. Tet3 CXXC domain and dioxygenase activity cooperatively regulate key genes for Xenopus eye and neural development. *Cell*, 151(6), pp.1200–13.

Yamagishi, T., Hirose, S. & Kondo, T., 2008. Secondary DNA structure formation for Hoxb9 promoter and identification of its specific binding protein. *Nucleic acids research*, 36(6), pp.1965–75.

Yan, Q. et al., 1999. Dual Roles for Transcription Factor IIF in Promoter Escape by RNA Polymerase II. *Journal of Biological Chemistry*, 274(50), pp.35668–35675.

Yisraeli, J. et al., 1988. Effect of in vitro DNA methylation on beta-globin gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 85(13), pp.4638–42.

Yu, B.D. et al., 1995. Altered Hox expression and segmental identity in Mll-mutant mice. *Nature*, 378, pp.505–508.

Yuan, C.-C.C. et al., 2012. Histone H3R2 symmetric dimethylation and histone H3K4 trimethylation are tightly correlated in eukaryotic genomes. *Cell reports*, 1(2), pp.83–90.

Yun, M. et al., 2011. Readers of histone modifications. *Cell research*, 21(4), pp.564–78.

Zeng, L. & Zhou, M.M., 2002. Bromodomain: an acetyl-lysine binding domain. *FEBS letters*, 513(1), pp.124–8.

Zhang, G., Gurtu, V. & Kain, S.R., 1996. An enhanced green fluorescent protein allows sensitive detection of gene transfer in mammalian cells. *Biochemical and biophysical research communications*, 227(3), pp.707–11.

Zhang, Y. et al., 2010. Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. *Nucleic acids research*, 38(13), pp.4246–53.

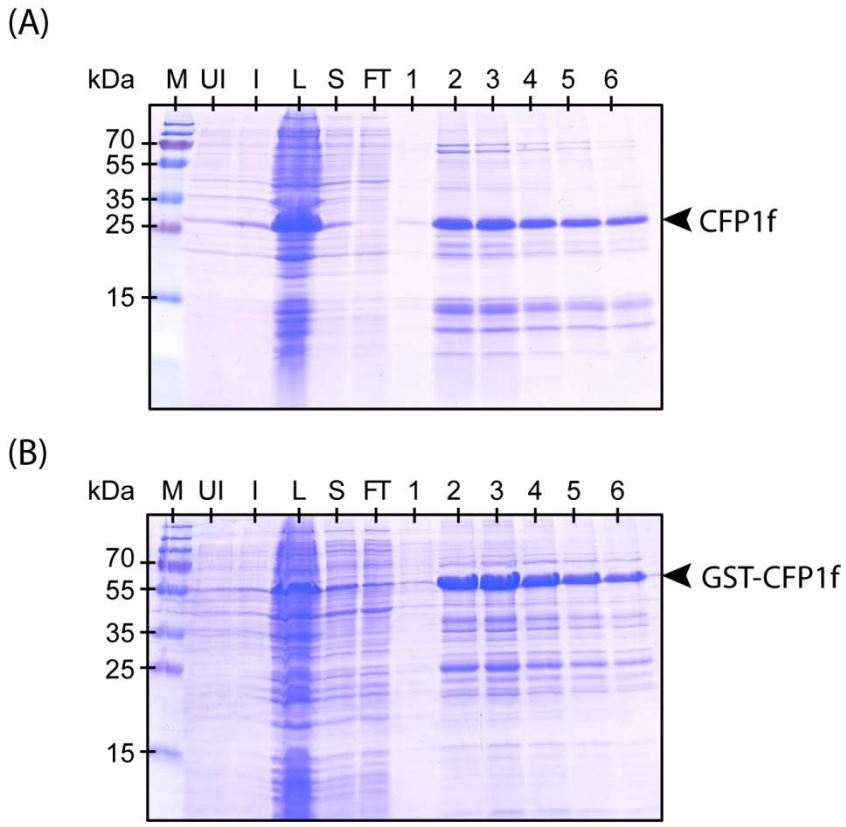
Zhang, Y., 2006. It takes a PHD to interpret histone methylation. *Nature structural & molecular biology*, 13(7), pp.572–4.

Zhang, Y. et al., 2008. Model-based analysis of ChIP-Seq (MACS). *Genome biology*, 9(9), p.R137.

Zhou, J.C. et al., 2012. Recognition of CpG island chromatin by KDM2A requires direct and specific interaction with linker DNA. *Molecular and cellular biology*, 32(2), pp.479–89.

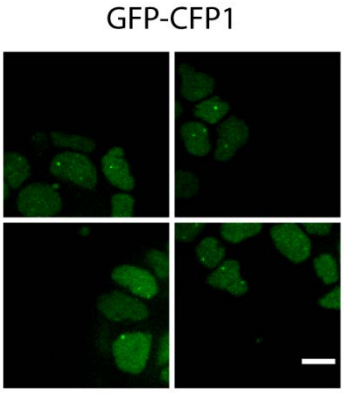
# Appendix

**Supplementary figures**



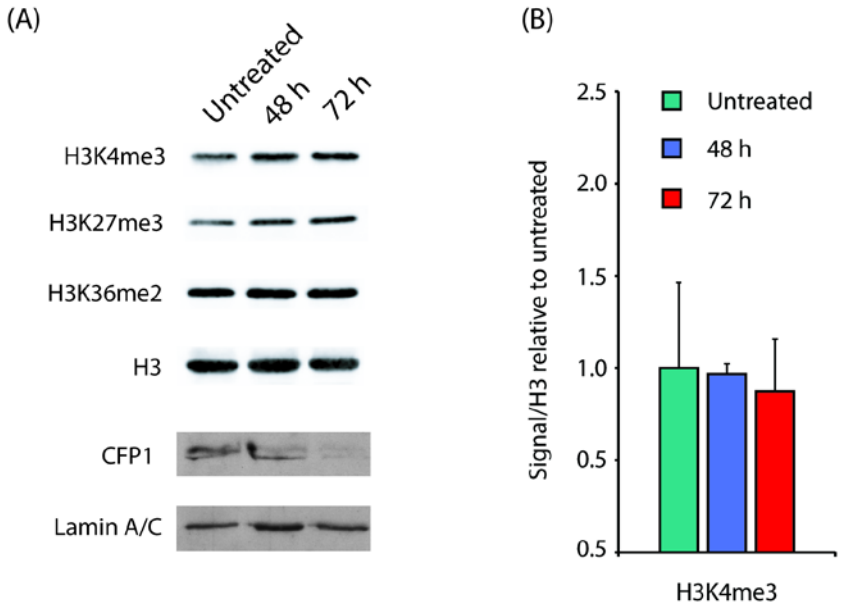

**Supplementary Figure 1 - Initial expression and purification of CFP1f**

Expression and purification of the CFP1f antigen was analysed by SDS-PAGE. (A) 6His-CFP1f, and (B) 6His-GST-CFP1f, were expressed in *E. coli* and purified using a Ni-NTA column. Samples were resolved on an 8% polyacrylamide gel and visualised by staining with Coomassie Brilliant Blue. M: Protein Markers, UI: Uninduced sample before addition of IPTG, I: IPTG induced sample, L: Lysate after sonication, S: Soluble supernatant after centrifugation of lysate, FT: Flow through from the Ni-NTA resin column, 1-6: consecutive elutions. Arrowheads indicate IPTG induced protein bands of the expected size.




**Supplementary Figure 2 – CFP1 forms bright clusters in embryonic stem cells**

Live imaging of mouse embryonic stem cells stably expressing GFP-CFP1 revealed bright foci in the majority of cells.




**Supplementary Figure 3 – Deletion of CFP1 has no effect on global levels of H3K4me3**

(A) Western blot analysis of histone modifications after deletion of CFP1. (B) ChIP-qPCR of H3K4me3 after deletion of CFP1.

**MATLAB FRAP Normalisation scripts**
**Contents**

FRAP_analysis.m is a MATLAB function for processing FRAP videos using a circular bleach according to the protocol described in F. Mueller, et al. 2012 Methods in Molecular Biology



*function [FRAP_summary, par_out]=FRAP_analysis_annotated_for_thesis(fileFolder, fileName, BleachFrame, fps, exposureTime, radius_um, bleach_duration, pixelSize)*

FRAP_analysis.m is a MATLAB function for processing FRAP videos using a circular bleach according to the protocol described in F. Mueller, et al. 2012 Methods in Molecular Biology. It requires a fileFolder and fileName, the BleachFrame (last pre bleach frame), bleach 'radius_um' and pixelSize. The capture rate 'fps', frame, 'exposureTime' and bleach_duration are also input to be saved as metadata.

It is a semi-automated process, requiring the user to select regions of interest (ROIs) manually, but automatically measuring and normalising the FRAP data according to Muller et al, while saving the raw data.

**1. Set up**

### Calculate the bleach radius in pixels 'radius_pix'

*radius_pix=radius_um/pixelSize;*

### Import the .tif file data into MATLAB

*Video = tiffread29(fullfile(fileFolder,fileName));*

### Get number of frames

*nFrames_all = length(Video);*

### Get image dimensions

*[nRow, nCol] = size(Video(1).data);*

**2. Define regions of interest to be measured**

**2.1 Automatically detect the bleached, 'ROI1'**

Section 2.1 is based on scripts written by F. Muller.

### Generate a difference image 'imDiff', by subtracting the post-bleach frame from the pre-bleach frame

*imDiff = double(Video(BleachFrame).data) - double(Video(BleachFrame+1).data);*

*imDiff = imDiff.*(imDiff>0);*

### Calculate round 'Filter' for convolution of the difference image The filter is essentially a circle with the bleach radius supplied by the user. However the filter also accounts for the pixelation of the data.

*Filter = fspecial('disk', radius_pix);*

### Process the difference image 'imDiff' with the new Filter

*Cmatrix = conv2(imDiff,Filter,'same');*

**Find maximum (or best fit) of the convolution matrix. This is reliably the centreof the bleach spot [ROI1_x, ROI1_y].**

*Cmax = max(max(Cmatrix));*

*[y, x] = find(Cmatrix==Cmax);*

*ROI1_x=x;*

*ROI1_y=y;*

**2.2 Define region for observational photobleaching correction, 'ROI2'**

This region is usually the entire nucleus (A requirement for quantitative measurements). Meaning that it contains thebleached region, and will therefore lose signal due to the photobleach immediately and in a single step.

### Open GUI window showing postbleach frame

*figure('Name',fileName);*

*imshow(Video(BleachFrame+1).data,[],'InitialMagnification', 'fit');*

### Overlay the autodetected bleach spot ROI1

*rectangle('Position',[x-radius_pix,y-radius_pix,2*radius_pix,2*radius_pix],'Curvature',[1,1],'EdgeColor','r')*

### Prompt user to select a region for observational photobleaching correction

*title('Pick the observational photobleaching correction')*

**Binary mask**

*mask_ROI2 = roipoly;*

*close(1)*

**Convert mask to polygon for saving and average nuclear radius calculations**

ROI2 = mask2poly(mask_ROI2,'MINDIST');

**2.2 Calculate nuclear area and average nuclear radius to be saved as metadata**

**Calculate nuclear area in µm**

*Nuclear_Area=(bwarea(mask_ROI2))*(pixelSize^2);*

**Find nuclear center**

*ROI2_center=regionprops(mask_ROI2, 'centroid');*

*ROI2_center_x=ROI2_center.Centroid(1);*

*ROI2_center_y=ROI2_center.Centroid(2);*

**Calculate nuclear radius**

**Define radius for each of the pixels with respect to the nuclear center [1]**

```
cMask = zeros(nRow,nCol);
warning off all
for iRow = 1:nRow
   for iCol  = 1:nCol
      rPix  = sqrt((iCol-ROI2_center_x)^2 + (iRow-ROI2_center_y)^2);
      cMask(iRow,iCol) = uint16(rPix);
   end
end
warning on all
```

**Define pixels on the nuclear perimeter [2].**

```
nPerimeterPixels=length(ROI2);
radial_list=zeros(nPerimeterPixels, 1);
for perimeter=1:nPerimeterPixels
        radial_distance=sqrt((ROI2(perimeter,1)-ROI2_center_x)^2 + (ROI2(perimeter,2)-ROI2_center_y)^2);
        radial_list(perimeter,:)=radial_distance;
end
```

**Average radial measurements to get average Nuclear_radius in µm**

*Nuclear_radius=mean(radial_list)*pixelSize;*

**2.3 Define background region, 'ROI3'**

This should be away from the cell to measure the imaging background Outline of polygon is defined by simply following the desired outline and then closing the polygon by double-clicking on the first point.

*figure('Name',fileName);*

*imshow(Video(BleachFrame+1).data,[],'InitialMagnification','fit');*

**Prompt user to select a region for background subtraction**

*title('Pick the background for subtraction')*

**This is a built in user defined roipoly, which once you have drawn it can be saved as a variable.**

*mask_ROI3 = roipoly;*

*close(1)*

**Convert mask to polygon for saving**

*ROI3 = mask2poly(mask_ROI3,'MINDIST');*

**3. Measure ROI intensities**

**3.1 Measure ROI1 intensity**

**Move the center inbetween the pixels**

*xCenterPixMod = ROI1_x+0.5;*

*yCenterPixMod = ROI1_y+0.5;*

**Generate a distance image 'dImage' in which each pixel intensity reflects its distance from the center of ROI1**

**Preallocate dImage**

*dImage = zeros(nRow,nCol);*

**Loop through the whole image calculating the distance for each pixel**

```
warning off all
for iRow = 1:nRow
   for iCol  = 1:nCol
      rPix  = sqrt((iRow-yCenterPixMod)^2 + (iCol-xCenterPixMod)^2);
      dImage(iRow,iCol) = uint16(rPix);
   end
end
warning on all
```

**Define pixels within the measurement radius 'rMeasure'. Here rMeasure is the same as the bleach spot size 'radius_pix', however this could be modulated so that a fraction of the bleach spot was measured.**

```
rMeasure=radius_pix;
[iRow, iCol] = find(dImage<=rMeasure);
linearROI1Index = sub2ind([nRow nCol],iRow,iCol);
```

**3.2 Preallocate ROI_means for speed**

```
ROI1mean=zeros(nFrames_all,1);
ROI2mean=zeros(nFrames_all,1);
ROI3mean=zeros(nFrames_all,1);
```

**Loop over images**

```
for frame = 1:nFrames_all
   fprintf('\b\b\b\b%4i',frame);
```

**Get data for current frame**

```
   data_loop = double(Video(frame).data);
```

**Extract mean ROI intensity values**

```
   ROI1_int    = mean(double(data_loop(linearROI1Index)));
   ROI2_int    = mean(double(data_loop(mask_ROI2)));
   ROI3_int    = mean(double(data_loop(mask_ROI3)));
```

**Save values**

```
   ROI1mean(frame,1) = ROI1_int;
   ROI2mean(frame,1) = ROI2_int;
   ROI3mean(frame,1) = ROI3_int;
end
```

**4. Normalise the FRAP data**

**Total cell prebleach intensity "T0" is calculated, and background "B0" subtracted.**

```
ROI2pre=ROI2mean(1:BleachFrame);
T0=mean(ROI2pre);
ROI3pre=ROI3mean(1:BleachFrame);
B0=mean(ROI3pre);
T0sub=T0-B0;
```

**Spot intensity prebleach "I0" is also calculated and background subtracted.**

```
ROI1pre=ROI1mean(1:BleachFrame);

I0=mean(ROI1pre);

I0sub=I0-B0;
```

**As well as normalising to the initial conditions, curves are normalised for monitor bleaching, giving Inorm and Tnorm.**

```
Inorm=ROI1mean-ROI3mean;

Tnorm=ROI2mean-ROI3mean;

normIsub=Inorm*T0sub;

subInorm=Tnorm*I0sub;
```

**The end result is a relative intensity curve "Irel".**

```
Irel=normIsub./subInorm;
```

**5. Compile FRAP_summary**

```
ROI1_xy=[ROI1_x, ROI1_y];

FRAP_summary=struct('ROI1mean', ROI1mean, 'ROI2mean', ROI2mean, 'ROI3mean', ROI3mean,'Irel',Irel);

par_out=struct('BleachXY', ROI1_xy, 'bleachRadius_um',  radius_um,'I0sub',I0sub,'ROI2poly',
ROI2,'NuclearArea', Nuclear_Area, 'NuclearRadius', Nuclear_radius, 'T0sub', T0sub, 'ROI3poly', ROI3,
'BleachFrame', BleachFrame, 'fps', fps, 'ExposureTime', exposureTime, 'BleachDuration', bleach_duration);
```

**Published with MATLAB® R2013a**

**Batch FRAP Analysis**

David Brown 14/05/13

**Clear workspace for new analysis**

```
clear all; close all; clc
```

**Specify parameters**

```
pixelSize=0.2778;          µm
fps=14;
exposureTime=0.04;         s
BleachFrame=50;
radius_um=1;               µm
bleach_duration=0.020;     s
Switch=0.005;              s
```

**Specify Input Folder**

```
fileFolder = fullfile(path);
```

**List .tif files**

```
fileList = dir(fullfile(fileFolder,'*.tif'));
```

**Allocate FRAP_output**

```
FRAP_output{length(fileList),3}=[];
```

**Loop through files**

```
for i=1:length(fileList);
    fileName=fileList(i).name;
    [FRAP_summary, par_out]=FRAP_analysis(fileFolder, fileName, BleachFrame, fps, exposureTime,
radius_um, bleach_duration, pixelSize);
    FRAP_output{i,1}=fileName;
    FRAP_output{i,2}=FRAP_summary;
    FRAP_output{i,3}=par_out;
end
```

**Organise frap curves for export to Excel.**

```
frap_list((length(FRAP_summary.Irel)-BleachFrame),length(fileList))=zeros;
for i=1:length(fileList)
  frap=FRAP_output{i,2}.Irel((BleachFrame+1):end);
  frap_list(:,i)=frap;
end
```

**Published with MATLAB® R2013a**