Clustered PHD domains in mixed lineage leukaemia proteins are attracted by acetylation-rich active promoters and enhancers

Anna Maria Stroynowska-Czerwinska[1*], Magdalena Klimczak[1], Michal Pastor[1,2], Asgar Abbas Kazrani[1,#], Matthias Bochtler[1,2,3]*

[1]*International Institute of Molecular and Cell Biology, Warsaw, Poland*
[2]*Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland*
[3]*Lead Contact*

[#]Present address: Institut de Génétique et de Biologie Moléculaire et Cellulaire, *Illkirch-Graffenstaden, France*

[*]Correspondence:
mbochtler@iimcb.gov.pl
asczerwinska@iimcb.gov.pl

Keywords:

Histone lysine methyltransferase 2, mixed lineage leukemia, PHD domain, chromatin reader domain, histone acetylation, promoter, enhancer

SUMMARY:

Histone lysine methyltransferase (KMT2) proteins form the core of COMPASS and COMPASS-like complexes that mediate transcriptional memory by methylating H3K4 at promoters and enhancers. KMT2A-D proteins, alternatively called mixed lineage leukaemia proteins (MLL1-4), contain highly conserved unique triplet and quartet of plant homeodomains (PHDs). Here, we show that clustered PHDs, expressed in isolation in HeLa cells, localize to well-defined loci of acetylation-rich active promoters and enhancers. Binding sites overlap with targets of full-length KMT2A (MLL1) and the COMPASS-like subunit WDR5, RbBP5 and with cell cycle and cancer-related genes. COSMIC data identify frequent variations in the PHDs of KMT2 proteins, particularly KMT2C, in a wide spectrum of malignancies. Changes are enriched at conserved positions within the PHDs, indicating that they cause loss-of-function mutations. Taken together, the biochemical and cancer data suggest that the PHDs contribute to KMT2A-D targeting to active promoters and enhancers.

INTRODUCTION

Histone H3 lysine 4 (H3K4) methylation is an important chromatin mark that is typically found at promoters (Soares et al., 2017) and enhancers (Calo and Wysocka, 2013). In humans, H3K4 methylation is deposited by a family of histone lysine methyltransferases (KMT2) (Shilatifard, 2012). This protein family consists of six active KMT2 paralogues (KMT2A-D/MLL1-4, KMT2F-G/SET1A-B) (Bochyńska et al., 2018; Shilatifard, 2012) and KMT2E/MLL-5, which is phylogenetically related but believed to lack catalytic activity (Mas-Y-Mas et al., 2016). As the MLL1-4 nomenclature suffers from conflicting use of the MLL2 and MLL4 paralogue identifiers, we will use the KMT2A-D nomenclature hereafter.

All KMT2 proteins act in the context of large multiprotein complexes (Takahashi et al., 2011). In vertebrates, the complexes are classified as COMPASS or COMPASS-like complexes, with KMT2F-G/SET1A-B or KMT2A-D, respectively, acting as the enzymatically active methyltransferase subunit (Cenik and Shilatifard, 2021; Shilatifard, 2012). Other integral subunits necessary for the full methylation activity of both COMPASS and COMPASS-like complexes are WD repeat-containing protein 5 (WDR5), Retinoblastoma-binding protein 5 (RbBP5), Absent small homeotic-2 like (ASH2L), and Dumpy-30 (DPY-30). These proteins are collectively termed the WRAD (Ali and Tyagi, 2017). COMPASS-like complexes with KMT2A-B at the core (orthologous to Trithorax Trx in *D. melanogaster*) catalyse H3K4 trimethylation and are believed to act primarily at promoters (Denissov et al., 2014; Wang et al., 2009). Complexes with KMT2C-D (orthologous to Trithorax-related Trr in *D. melanogaster*) can only monomethylate H3K4 and are believed to keep enhancers active (Herz et al., 2012; Hu et al., 2013). Deletion experiments demonstrated that the methyltransferases in COMPASS complexes (KMT2E-F/SET1A-B), which are conserved from yeast to human, are responsible for the bulk of H3K4 methylation in cells (Lee and Skalnik, 2005; Lee et al., 2007). On the other hand, the COMPASS-like complexes, which are found only in animals, are more involved in specialized functions such as transcriptional memory at HOX and other loci (Milne et al., 2002). Accordingly, they can affect cell fate determination (Dhar et al., 2012; Liedtke et al., 2010) and the maintenance of segmental identity (Yu et al., 1995).

Human COMPASS-like (hCOMPASS-like) complexes are also of interest from a medical perspective (Rao and Dou, 2015). KMT2A/MLL1 was originally identified in patients with mixed lineage leukaemia (MLL), since the coding gene is a frequent breakpoint (Ziemin-van der Poel et al., 1991). Interestingly, the dominant effect of KMT2A fusion events stems from ectopic targeting of the fusion partners rather than a loss of hCOMPASS-like activity (Liu et al., 2009). Moreover, KMT2C and KMT2D are frequently mutated in other malignancies (Fagan and Dingwall, 2019; Herz et al., 2014). In fact, KMT2C appears among the most prominent cancer driver mutations in many studies across different types of malignancies (Bailey et al., 2018). In contrast to the KMT2A variants, which typically lead to fusion proteins, KMT2C and KMT2D variants in cancer are often missense mutations, nonsense mutations, small insertions, or deletions.

The mechanisms through which COMPASS and COMPASS-like complexes are recruited to their chromosomal targets or mediate transcriptional memory are poorly understood. However, some clues can be derived from the architecture of the protein complexes and the KMT2 proteins (Fig. 1A). All KMT2 proteins contain SET and postSET domains (Bochyńska et al., 2018; Zhang et al., 2002), which are the catalytically active domains in methyltransferases, except for KMT2E (Mas-Y-Mas et al., 2016). The compact KMT2F and KMT2G proteins (~180 kDa) also have RNA recognition motifs (RRMs), but no chromatin binding domains (Maris et al., 2005). This observation suggests a reliance on the other COMPASS subunits for chromatin targeting. By contrast, the much larger KMT2A-D proteins (~270-550 kDa) have multiple chromatin binding domains, suggesting an intrinsic ability to localize to chromatin targets without guidance from other interacting proteins. This view is supported by the persistent localization of at least one KMT2 protein (KMT2A) to physiologically relevant targets when RbBP5, WDR5, or ASH2L of the WRAD subunits are knocked down (Dou et al., 2006). However, intrinsic targeting of KMT2A is also complemented by guidance from COMPASS complex subunits (Chen et al., 2011; Lorton et al., 2020; Wysocka et al., 2005).

KMT2A-D contain multiple chromatin reader domains, like CXXC, bromodomains, AT hooks, and plant homeodomains (PHDs). To understand how KMT2A-D implement transcription memory through chromatin reading and targeting, we investigated PHDs as they are present within all KMT2A-D except KMT2F-G (Ali et al., 2014; Bochyńska et al., 2018). PHDs, despite their name, are widely present in transcription regulating proteins from plants to animals, and even in yeast (Bienz, 2006). They are compact domains that consist of around 65 amino acids, containing the Cys4HisCys2Cys/His (C4HC2C/H) motif which chelates two structural Zn^{2+} ions (Bienz, 2006). The stand-alone PHDs are known to be readers of H3K4me3 (Musselman and Kutateladze, 2009; Shi et al., 2006). However, some PHDs also recognize trimethylated lysine residues in other histone tail positions or non-methylated H3K4 (Mansfield et al., 2011; Musselman and Kutateladze, 2011; Sanchez and Zhou, 2011). In addition, doublet PHDs can bind to acetylated lysine residues within a histone tail (Ali et al., 2012; Dreveny et al., 2014; Lange et al., 2008; Zeng et al., 2010).

PHDs in KMT2 proteins can be found as singlets or clustered triplets or quartets. Hereafter, we designate individual PHDs by their paralogue identifier (2A-D) and by their position. For example, the third PHD in KMT2A is abbreviated as 2A3. Similarly, clusters of PHDs are designated by the first and last domains, i.e. 2A13 (comprising PHD1, PHD2 and PHD3 of KMT2A). According to this nomenclature, KMT2A and KMT2B have two triplets (2A13 and 2B13), KMT2C has a quartet (2C14) and a triplet (2C57), and KMT2D has two triplets (2D13 and 2D46) (Fig. 1A). KMT2A-D also contain one isolated PHD each: 2A4, 2B4, 2C8, and 2D7. PHDs containing an extension containing the C4 zinc finger are termed “extended” PHDs (Liu et al., 2019) and are abbreviated with an extra “e”: 2Ae3, 2Be3, 2Ce4, 2Ce7, and 2De6.

The binding properties of PHDs in KMT2 proteins have mainly been investigated in single PHDs. For instance, 2A3 with bromodomain binds H3K4me1-3 with low affinity (Wang et

al., 2010). 2D6 and 2C7 show a preference for H4K16ac and, to a lesser extent, H4K20ac (Zhang et al., 2019). 2Ce7 and 2De6 bind non-modified H4 (Liu et al., 2019). It should be noted that the alternative numbering of KMT2C PHDs in Liu *et al.* starts at 0 instead of 1. So far, only one PHD cluster in KMT2 proteins, 2D46, has been characterised biochemically. 2D46 was found to associate with non-modified H4 *in vitro*, specifically H4R3me2a and H4K20me1-3 (Dhar et al., 2012). Furthermore, other roles for PHDs in KMT2 proteins have been proposed. For example, 2A3 may act as a platform for the protein-protein interactions with binding partners such as cyclophilin peptidyl-prolyl isomerase (Cyp33) via its RRM (Fair et al., 2001; Wang et al., 2010). In another set of experiments, 2A2 and 2B2 were shown to facilitate ubiquitin transfer onto H3/H4 histones and KMT2A protein (Wang et al., 2012).

To our knowledge, the arrangement of PHDs in triplets or quartets in the KMT2 proteins is unique in the human proteome but has nevertheless received only a little attention. Prior studies have typically focused on single PHD. Moreover, they were designed to test the binding of PHDs to a pre-selected chromatin mark that is chosen from a small set of candidate marks. Here, we address this limitation by performing an unbiased characterisation of all clustered PHDs in the KMT2 protein family. We use biochemical chromatin affinity assays and mapping experiments in cultured stable cell lines using greenCUT&RUN. Our results show that clustered PHDs bind preferentially to well-defined loci within active promoter and enhancer regions that mainly correspond to genes involved in the cell cycle. Moreover, we observed a significant overlap between the targets of clustered PHDs and the full-length hCOMPASS-like subunits KMT2A, WDR5 and RbBP5. In parallel, analysis of cancer data revealed loss-of-function mutations in the PHDs of KMT2A-D, with clear enrichment in the PHDs in KMT2C. Taken together, we conclude that clustered PHDs in KMT2A-D contribute to the targeting of hCOMPASS-like complexes to active promoters and enhancers.

RESULTS

PHDs are highly conserved in KMT2 proteins

Amino acid conservation between PHDs of some mammals is surprisingly high (~99% identity) and prompted us to analyse conservation more broadly along the entire protein (Suppl. Fig. 1A). Shannon analysis (amino acid diversity) for each paralogue across different species indicated that the PHDs are among the most conserved regions within KMT2A-D proteins (Fig. 1B). In the case of 2D46, the regions of the PHD are similarly or even slightly more conserved than the catalytic SET domain. On the other hand, the differences between the clustered PHDs in human KMT2 proteins are considerable, with typical values for amino acid identity and similarity of below 50% (Suppl. Fig. 1B). The high conservation between species and variability among different modules suggests that PHDs play an important role in KMT2 biology.

Interaction between PHDs and mononucleosomes strengthens with increasing number of PHDs

For biochemical characterisation, we cloned and purified all clustered PHDs of KMT2 proteins (2A13, 2B13, 2C14, 2C57, 2D13, and 2D46), as well as some singlets or doublets as a fusion protein with GST (2A3, 2C7, 2Ce7, 2C67, 2D6, and 2De6) (Fig.2A-C). We subsequently excluded 2B13 from further studies, as it showed minimal expression and solubility.

First, we investigated the interactions between purified PHDs and recombinant histones or isolated mononucleosomes from HeLa cells by Far-Western blot (Fig. 2C-D, Suppl. Fig. 2). The analysis indicated that all KMT2 GST-PHDs, but not GST alone, bind to both recombinant and endogenous H3, whereas other nucleosome subunits were either weakly bound or not at all. In addition, the strength of the histone interactions grew with the number of fused PHDs (e.g., 2C7<2Ce7<2C67<2C57). For the triplets, the interaction was strongest for 2D46 and 2C57 and weaker for 2A13, 2C14, and 2D13. Of note, histone modifications do not show a clear impact on binding of recombinant and endogenous H3 histone, as both showed similar binding strength.

Next, to test the interaction of PHDs with endogenous mononucleosomes wrapped with DNA in solution, we performed GST pull-down assays, taking advantage of the N-terminal GST-tag of all PHD constructs (Fig. 2E). Similar to the Far-Western results, we observed stronger interactions with PHD doublets and triplets than with singlets. As before, 2D13 and 2C14 showed the weakest interactions. Deviating from the rule of increased affinity for longer constructs, the 2C67 doublet bound mononucleosomes better than the 2C57 triplet in the pull-down assay, but not in the Far-Western experiments (Fig. 2D-2E).

KMT2 clustered PHDs recognize active promoters and enhancers

To verify the *in vitro* findings on PHDs, we examined isolated clustered PHDs in the context of chromatin by greenCUT&RUN (gC&R) (Fig. 3A). gC&R relies on an MNase-anti-GFP nanobody to cut chromatin at regions that are bound by a GFP-tagged fusion protein of interest. Therefore, we generated stable HeLa cell lines expressing NLS-EYFP-tagged triplet and quartet PHDs (NLS-EYFP-2A13, NLS-EYFP-2C14, NLS-EYFP-2C57, NLS-EYFP-2D13, and NLS-EYFP-2D46), as well as NLS-EYFP alone as a negative control. The strength of the EYFP signals measured by flow cytometry (Suppl. Fig. 3A) recapitulated the strength of the interaction seen previously between the histones and mononucleosomes (Fig. 2D-E). 2D46 interacted the most strongly and was the least expressed cluster PHD in HeLa cells, even after sorting a subpopulation of cells with higher expression. In general, the generated stable cell lines were characterised by a diffuse fluorescence pattern in the nucleus without aggregation foci, very similar to that observed for endogenous full-length KMT2D protein and histone modifications (Fig. 3B). This observation confirms that we generated stable cell lines with modest overexpression without creating aggregates in the nucleus.

In general, the pattern of peaks obtained from the gC&R data seemed similar across all clustered PHDs, and very different from the NLS-EYFP negative control which was more comparable to the HeLa cell line baseline (Fig. 3C-D). The similarity across the PHDs was

unexpected due to the various biological roles of KMT2A and KMT2C/D proteins. This similarity was also seen when comparing genomic heatmaps that show enrichment of all PHDs, particularly around transcriptional start sites (TSSs) (Fig. 3E). However, genomic distribution showed that only about 35% of all sites that bind clustered PHDs are located within promoters, with around 35% of remaining binding sites within distal intergenic regions and around 20% within introns (Fig. 3F). In addition, the comparison of peak binding sites showed high similarity (~5,500 common peaks for NLS-EYFP-2A13, NLS-EYFP-2D13, and NLS-EYFP-2D46) (Fig. 3G). In general, we obtained good-quality sequencing data for NLS-EYFP-2A13 (~12,000 peaks), NLS-EYFP-2D13 (~20,500 peaks) NLS-EYFP-2D46 (~18,500 peaks). However, we retrieved a low number of peaks for NLS-EYFP-2C14 (~2.800 peaks) and NLS-EYFP-2C57 (~1.100 peaks). Importantly, negative control of NLS-EYFP showed only around 70 peaks.

To classify binding sites of PHDs more precisely, we assigned them to imputed chromatin states for HeLa cells according to a 25-state of hidden Markov model (HMM) (Ernst and Kellis, 2015). This analysis indicated that PHDs were bound almost exclusively at active promoters (TssA, PromD1, PromU) and enhancers (EnhA1, EnhA2) (Fig. 3H, Suppl. Fig. 3B). In line with the expected role of enhancer-specific KMT2D, NLS-EYFP-2D13 preferred binding to enhancers over promoters. Similarly, NLS-EYFP-2A13 of promoter-specific KMT2A bound more selectively to promoters over enhancers. Surprisingly, NLS-EYFP-2C14, NLS-EYFP-2C57, and NLS-EYFP-2D46 from enhancer-specific KMT2C/D proteins also showed preference for promoters over enhancers. In general, the number of peaks indicated that cluster PHDs bind a relatively small subpopulation of promoters and enhancers.

Chromatin sites that bind clustered PHDs show enriched acetylation and do not strongly correlate with single H4K16ac

We also linked the binding sites for PHDs to experimentally-derived chromatin marks from the ENCODE database (Davis et al., 2018) and to HMM imputed chromatin marks (Ernst and Kellis, 2015). First, we confirmed a high similarity between imputed data, experimental ENCODE data, and our data from H3K4me1 and H3K4me3 in HeLa cells (Suppl. Fig. 3C). Second, we observed that clustered PHDs are mainly recruited to regions enriched with H3K4me1-3/ac, H3K9ac, and H3K27ac (but not H3K27me3 or H3K9me3), which is in line with binding at active enhancers and promoters (Suppl. Fig. 3D). However, by looking at a larger set of HMM imputed chromatin marks at sites bound by clustered PHDs, we observed enrichment of H2B (H2BK15ac, 20ac), H3 (H3K23ac), and to a lesser extent H4 acetylation (Fig. 3I, Suppl. Fig. 3E). Notably, we did not see such preferences with the NLS-EYFP negative control, which only showed a weak signal for H3K9me3. The level of acetylation also reflects the differences between bound and unbound chromatin states (Fig. 3H, Suppl. Fig. 3F-G). For instance, bound PromD1 is highly acetylated while unbound PromD2 is not (Suppl. Fig 3F). Similarly, EnhA1 which shows higher acetylation is also bound stronger than EnhA2 (Suppl. Fig. 3G).

H4K16ac, which was identified in previous research as a 2D6 target (Zhang et al., 2019), is not mapped in HeLa cells in ENCODE and is also not included in the HMM imputed chromatin marks. We therefore wanted to characterise their distribution in HeLa S3 cells using standard C&R. We found a broad distribution of H4K16ac and a specific enrichment within actively-transcribed gene bodies, which makes H4K16ac an unlikely candidate as the main recruiting factor for clustered PHDs (Suppl. Fig. 3H). We also mapped the distribution of multiple acetylations of H4 instead of single H4K16ac, namely H4K5acK8acK12ac (H4 3ac) and H4K5acK8acK16ac (H4 4ac). Here, we observed more specific patterns that correlated better with clustered PHDs (Suppl. Fig. 3H). In addition, we performed C&R experiments for H2A.Zac – another acetylation mark – and noticed a strong correlation with the binding sites of clustered PHDs (Suppl. Fig. 3H). Taken together, our gC&R data for clustered PHDs points towards their genome-wide affinity for well-defined active and highly acetylated promoters and enhancers.

Chromatin sites that bind clustered PHDs correlate partially with the localization of hCOMPASS-like subunits

After an initial characterisation of PHDs, we wanted to gain a better understanding of their role in recruiting full-length KMT2 proteins and hCOMPASS-like subunits. Thus, we tested multiple commercial antibodies in C&R (KMT2A, KMT2C, KMT2D, WDR5, Ash2L, RbBP5) and gathered useful mapping data for KMT2A, WDR5 and RBBP5. We observed relatively few (2,500-5,800) but very well-defined peaks for the full-length proteins, and these highly overlapped (~1,200 common peaks, Suppl. Fig. 3I) and partially co-localized with clustered PHDs (Fig. 3C-D, G ~ 1,100 common peaks with NLS-EYFP-2A13, KMT2A, and WDR5, Suppl. Fig. 3J ~800 common peaks with NLS-EYFP-2A13, 2D13, 2D46, KMT2A, and WDR5). We also found that the heatmaps for the investigated full-length hCOMPASS-like subunits were very similar to those for clustered PHDs, with a sharp peak around TSSs (Fig. 3E). However, contrary to what was seen with clustered PHDs, the distribution of all full-length proteins throughout the genome showed a much stronger preference for promoter regions (~85% of peaks for full-length proteins versus ~35% of peaks for PHD triplets in promoters) (Fig. 3F). In line with this observation, the full-length subunits mapped to active promoters with high specificity (TssA and to a lesser extent, PromD1, and even less to PromU) (Fig. 3H). However, the chromatin marks that were found to be enriched in hCOMPASS-like subunits were very consistent with those identified for clustered PHDs, such as H2BK15ac and H3K23ac (Fig. 3I). Taken together, these findings support the hypothesis that clustered PHDs take part in the stabilization of correct targeting of the hCOMPASS-like complexes.

PHDs and full-length KMT2A bind chromatin regions involved in the cell cycle and cancer-related pathways

With gC&R and C&R data, we performed KEGG and gene ontology (GO) analysis for the genes in which the peaks of NLS-EYFP-2A13, NLS-EYFP-2D13, NLS-EYFP-2D46, and full-length KMT2A localized. Due to a low amount of identified peaks, we excluded 2C14

and 2C57 from this analysis. KEGG pathway analysis of bound genomic regions showed very high consistency for all PHDs and to a lesser extent for full-length KMT2A protein (Fig. 4A, Suppl. Fig. 4A). Most of the identified pathways were highly significant and related to oncogenesis, which is in agreement with the known role of KMT2A-D and particularly KMT2C as a tumour suppressor (Larsson et al., 2020). False discovery rate (FDR) was lowest for cell cycle, hepatocellular carcinoma, and small cell lung cancer. Results of the GO analysis of biological processes showed that most targets of clustered PHDs were associated with cell cycle, DNA repair, or cellular division processes (Fig. 4B, Suppl. Fig. 4B). In addition, Multiple Expectation maximizations for Motif Elicitation (MEME) analysis of regions bound by all PHD clusters and full-length KMT2A protein identified a GGAA motif among the hits (Fig. 4C, Suppl. Fig. 4C). This motif is known to be present within promoters of genes related to DNA repair and apoptosis (Uchiumi et al., 2011). Furthermore, 2A13, 2D13, and 2D46 binding sites were also enriched for a TGAnTCA motif that is associated with cellular proliferation (Warner et al., 2008).

KMT2 PHDs are mutated in cancer patients within crucial amino acids involved in zinc finger folding

KMT2 proteins are frequently mutated in cancer (Rao and Dou, 2015). Among the KMT2 proteins, mutations in KMT2C are the most strongly associated with cancer according to the COSMIC database (~4,050 cases of cancer patients), followed by KMT2D (~2,600 cases), KMT2A (~1,580 cases), and finally KMT2B (~1,250 cases). The most frequent type of mutations affecting KMT2A are translocations that result in mixed lineage leukaemia. In contrast, missense point mutations in KMT2A-D proteins are most frequent in patients diagnosed with lung, liver, skin, or large intestine cancers (Fig. 5A). In agreement with COSMIC data, KEGG pathways indicated a very significant association between genomic regions bound by clustered PHDs and hepatocellular carcinoma and small cell lung cancer (Fig. 4A).

Furthermore, our analysis of the COSMIC mutation dataset for KMT2A-D showed that most of the mutations in cancer patients were found within the intrinsically-disordered regions (IDRs) (Fig. 5B). The only exception to this trend was mutations in KMT2C, which are enriched within the clustered PHDs, particularly the first cluster 2C14. Interestingly, some of the mutations impact crucial amino acids of the PHDs (especially within KMT2A and KMT2D proteins), specifically Cys or His residues (within the Cys4HisCys2Cys/His motif) that are needed for chelating Zn^{2+} ions which are essential for domain folding (Fig. 5C-D).

DISCUSSION:

Clustering of PHDs in KMT2 proteins is conserved in animals

To our knowledge, clustered PHDs (the triplets and the quartet) are present uniquely in the animal orthologues of human KMT2 proteins and are among the most highly conserved regions (Fig. 1B). Their presence is not limited to vertebrate proteins. The *Drosophila*

melangoster orthologue of KMT2A/B, Trx, has one PHD triplet. In contrast, the *Drosophila* orthologue of KMT2C/D, Trr, lacks clustered PHDs but interacts with separate proteins known as lost PHDs (Lpt or Cmi) that contain two clusters of PHDs (quartet and triplet). In *Drosophila*, ablation of *cmi* results in larval lethality (Chauhan et al., 2012), at least in part due to defects in dpp/BMP signalling and pattern formation (Chauhan et al., 2013). This observation highlights the crucial role of clustered PHDs in organism development.

Clustered PHDs interact primarily with acetylation-rich nucleosomes

Initially, we were surprised by a preference of clustered PHD for H3 histone as seen in Far-Western blot (Fig. 2D), contrary to previous observations of KMT2C-D PHD binding to the H4 tail (Liu et al., 2019; Zhang et al., 2019). However, a similar interaction was observed with the *Drosophila* Cmi protein that was also found to preferentially bind H3 (Chauhan et al., 2012). On the other hand, H3 and H4 tail binding may not necessarily be mutually exclusive. Based on the nucleosome core particle structures (Luger et al., 1997), one can deduce that the disordered tail of H4 can come close enough to H3 for PHDs to make contact with both. Simultaneous interaction, albeit for a protein without PHDs, has already been shown in the cryoEM structure of a Dot1L nucleosome complex (Worden et al., 2019).

Our findings from the gC&R experiments (Fig. 3I) support previous observations that clustered PHDs are primarily attracted by multiple histone acetylations (Ali et al., 2012; Zeng et al., 2010). However, the previously reported binding of singlet 2D6 to H4K16ac (Zhang et al., 2019) did not stand out as a single modification in our data and was also not supported by our gC&R and C&R profiles (Suppl. Fig. 3H). Instead, our gC&R characterisation (Fig. 3I) showed that PHDs bind in regions with multiple acetylations, particularly in regions containing acetylated H2B histones, namely H2BK15ac. However, not all H2BK15ac-rich regions are bound by PHDs. Interestingly, all investigated clustered PHDs seem to discriminate between promoters (strongest binding to TssA, weaker binding to PromU and PromD1, and no binding to PromD2 ad PromP) and enhancer types (strongest binding to EnhA1 and EnhA2) (Fig. 3H), based on their acetylation levels (Suppl. Fig. 3F-G). This observation suggests that the interactions of clustered PHDs are dependent on a pattern of modifications rather than a single modification.

The binding of acetylated lysine residues appears to be a common property of clustered PHDs. Besides KMT2A-D proteins, other examples are the PHD doublet in DPF3b (Lange et al., 2008) and the double PHD finger in MOZ (Dreveny et al., 2014). Surprisingly, the binding mode for histone tails to DPF3b and MOZ differs, making a prediction of a binding mode to clustered PHDs of KMT2A-D proteins difficult. Thus, we attempted to crystallize the clustered PHDs, but without success.

Clustered PHDs are recruited to well-defined loci within active promoters and enhancers

Our data from the gC&R experiment indicates that clustered PHDs localize to loci within active promoters and enhancers (Fig. 3H). Except for 2D13, the signal for promoter binding is stronger than the signal for enhancer binding. Surprisingly, the preferences of all tested PHDs,

except for 2D13, were rather similar and did not segregate into two groups for PHDs from promoter-specific KMT2A on the one hand and PHDs from enhancer-specific KMT2C/D on the other hand. This suggests that the traditional division between promoter and enhancer maintaining KMT2 proteins may not be as clear-cut as is generally believed. Alternatively, domains other than the clustered PHDs (e.g., CXXC) in promoter-specific KMT2A/B proteins or subunits of hCOMPASS-like complexes could be responsible for the distinction between promoters and enhancers.

The clustered PHDs in KMT2A-D proteins are frequently mutated in cancer

Missense and nonsense mutations, as well as small deletions and insertions, are frequent in KMT2A-D (Fig. 5B). Some of these mutations map to the PHDs, particularly clearly for KMT2C, where the number of patients with characterised variants is very large. The distribution of mutations to many sites, rather than a few selected sites, and the enrichment of alterations in conserved residues of the PHDs suggest that the mutations are loss-of-function mutations (Vogelstein et al., 2013). Hence, it is likely that PHD dysfunction is a major contributor in KMT2A-D driven malignancies. However, in leukaemia, particularly in mixed lineage leukaemia, the KMT2A alterations are not small local changes, but chromosomal translocations generating novel fusion proteins. Importantly, translocation breakpoints are not randomly distributed. They cluster into a major Breakpoint Cluster Region (BCR) that extends from exon 8 to 14, with most breakpoints in the core region from exon 9 to 11 (Meyer et al., 2013, 2018), and a minor breakpoint cluster region from exons 19 to 26 (Meyer et al., 2019). Reciprocal translocations create two fusion proteins. Typically, the N-terminal region of KMT2A is fused in the frame, whereas the C-terminal portion is fused out of the frame. This observation suggests that the fusion protein with the KMT2A N-terminal region is selected for and therefore acts as the oncogenic driver (Meyer et al., 2013). As the coding region for clustered PHDs of KMT2A (2A13) starts in exon 11 and major BCR is localized in the region between exons 9 and 11, 2A13 are either completely absent or only present in a truncated form of the oncogenic fusion proteins. We conclude that biology of PHD is relevant for KMT2A-D loss-of-function malignancies, but not for translocation-driven KMT2A gain-of-function leukaemias.

Multiple mechanisms contribute to the targeting of KMT2A-D

There is strong evidence that COMPASS and COMPASS-like complexes are directed to their chromatin targets not only by the KMT2 proteins but also by other complex subunits. A core component of COMPASS and COMPASS-like complexes is the WRAD core (Ernst and Vakoc, 2012). For WDR5, a preference for methylated H3K4, particularly dimethylated H3K4 (Couture et al., 2006; Ruthenburg et al., 2006; Wysocka et al., 2005), with modulation by methylation of H3R2 (Lorton et al., 2020) and serotonylation of H3Q5, has been demonstrated biochemically (Zhao et al., 2021). For Ash2L, strong evidence for DNA binding was shown (Chen et al., 2011; Sarvan et al., 2011). However, it is unclear how this binding could contribute to directing COMPASS or COMPASS-like complexes to promoters or enhancers, or help to establish a positive feedback loop for transcriptional memory. Menin, a component of COMPASS-like complexes with KMT2A/B catalytic core binds to a plethora

of transcription factors (Dreijerink et al., 2017). PTIP, a component of COMPASS-like complexes with KMT2C/D catalytic core recruits them to Pax proteins (Lechner et al., 2000). Moreover, only promoter-specific KMT2A-B contain CXXC domains, which localize the proteins to non-modified CpGs and single modification of CpG was shown to abolish this interaction (Cierpicki et al., 2010; Stroynowska-Czerwinska et al., 2018). These CpGs are enriched in CpG islands (CGIs) of active promoters, but not elsewhere in the genome (Bird, 1986). The presence of CXXC domains in KMT2A-B and absence from KMT2C-D may account, fully or partially, for protein preference for promoters over enhancers. KMT2A-B, but not KMT2C-D, also contain bromodomains which typically bind acetylated lysine residues (Mujtaba et al., 2007). As lysine acetylation is generally associated with a permissive chromatin state (Eberharter and Becker, 2002), their presence could help to target methyltransferases to active chromatin, thus implementing a positive feedback loop that could contribute to positive transcriptional memory. However, this has not been confirmed for the KMT2A bromodomain. In conclusion, we believe KMT2C/D only contain PHDs because mono-methyltransferases may target both promoters and enhancers. In contrast, KMT2A contains additional domains, such as CXXC, to increase the affinity of clustered PHDs to acetylation-rich active promoters but not enhancers.

Limitations of the study

As Far-Western assays lack the definition of histone tail modifications, we attempted to probe direct affinities to modified histone tails using histone peptide arrays, but could not use the data because of inconsistencies between repeats. In contrast, the gC&R assay generated highly reproducible data. However, a limitation of gC&R is that we cannot exclude the possibility that PHDs could be attracted to chromatin marks because of their interactions with other proteins. The assay also does not allow us to distinguish bound and unbound chromatin marks, where the latter strongly correlates with the bound marks. Our data suggests that clustered PHDs recognize a combination of chromatin marks, rather than a single mark. However, the limitation is that the set of experimentally-determined chromatin marks in ENCODE, or even the larger set of HMM imputed chromatin marks, is incomplete. Hence, it remains possible that there could be a single chromatin mark not considered in our analysis that directs PHDs, rather than the combinatorial code suggested by our data. Despite these limitations, our findings clearly suggest that PHDs play an important role in targeting KMT2 proteins, and more generally hCOMPASS-like complexes, to a well-defined subset of active promoters and enhancers.

MATERIALS AND METHODS:

Cloning and expression of PHD domains

E. coli codon-optimised synthetic genes (GeneArt, ThermoScientific; Table S1) encoding the PHD domains of human KMT2A (UniProt: Q03164, 2A3, aa 1565–1631; 2A13, aa 1416–1640), KMT2B (UniProt: Q9UMN6, 2B13, aa 1182–1414), KMT2C (UniProt: Q8NEZ4, 2C14, aa 266–532; 2C7, aa 1083–1143; 2Ce7, aa 1055–1144; 2C67, aa 1008–1144; 2C57, aa 934–1149) and KMT2D (UniProt: O14686, 2D13, aa 156–341; 2D6, aa 1503–1561; 2De6, aa

1475–1564; 2D46, aa 1351–1572) were cloned using BamHI and XhoI restriction sites into pGEX-P6-2 plasmids encoding a N-terminal GST-tag. For GST overexpression, an empty pGEX-P6-2 plasmid was used. In addition, human codon-optimised synthetic genes (GeneArt, ThermoScientific; Table S1) encoding the clustered PHD domains of KMT2A (UniProt: Q03164, 2A13, aa 1565–1631), KMT2C (UniProt: Q8NEZ4, 2C14, aa 266–532; 2C57, aa 934–1149) and KMT2D (UniProt: O14686, 2D13, aa 156–341; 2D46, aa 1351–1572) were cloned using the SLIC method (Jeong et al., 2012; Li and Elledge, 2007), first to pEYFP plasmids and then into a customised pLJM1-EGF1-based plasmid encoding N-terminal EYFP for the generation of a stable cell line by lentiviral transduction. Briefly, NLS-EYFP-PHD fusion gene fragments were amplified with the primers specified in Table S2 using Q5 DNA Polymerase. Plasmid pLJM1 was linearised using AgeI and EcoRI restriction enzymes. DNA fragments were incubated for 30 minutes with T4 DNA polymerase at 22 °C to generate overhangs. The reaction was stopped by the addition of dCTP to a 2 mM final concentration. Inserts and vector were mixed in a 1:1 molar ratio and incubated for 30 minutes at 37 °C before being transformed into NEB stable *E. coli* competent cells.

All protein overexpression was performed in *E. coli* BL21(DE3) pLys cells. The cells were grown in LB media supplemented with 100 ng/ml ampicillin at 37 °C until OD600 reached 0.6/cm–0.8/cm. Cells were then transferred to 4 °C for 30 minutes. After induction with 1 mM IPTG and in the presence of 150 µM $ZnSO_4$, proteins were expressed overnight at 22 °C. The cells were then harvested and stored at -80 °C in pellets from 1L of culture batches.

Protein purification

Frozen cells overexpressing GST or GST-PHD domains were resuspended in 40 ml ice-cold lysis buffer (50 mM Tris-HCl pH 7.5, 400 mM NaCl, 5% glycerol, 150 µM $ZnSO_4$, 1mM DTT) supplemented with 0.25 mM PMSF. After sonication, the lysate was cleared by ultracentrifugation for 30 minutes. Next, the supernatant was applied to Glutathione Sepharose™ 4B beads (GE Healthcare) equilibrated with lysis buffer. After 2 hours of incubation, the beads were extensively washed with washing buffer (50 mM Tris-HCl pH 8.0, 400 mM NaCl, 5% glycerol, 150 µM $ZnSO_4$, 1 mM DTT). Protein was eluted by incubating the beads with elution buffer (200 mM Tris-HCl pH 8.0, 150 mM NaCl, 5% glycerol, 20 mM reduced L-glutathione, 150 µM $ZnSO_4$, 1 mM DTT) for 15 minutes. The purified protein was concentrated using Amicon Ultra centrifugal filters with 10 kDa cut-off (Millipore) and dialyzed overnight against storage buffer (50 mM Tris pH 8.0, 150 mM NaCl, 150 µM $ZnSO_4$, 1 mM DTT, 50% glycerol). Protein concentrations were determined spectrophotometrically (Nanodrop, Thermo Fisher Scientific). Aliquoted samples were frozen in liquid nitrogen and stored at -80 °C.

Human cell lines

HeLa S3 cells, HeLa S3-derived stable cell lines, and HEK293T cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, Biowest) supplemented with 10% fetal bovine serum (FBS, Biowest) and penicillin-streptomycin solution (Biowest) at 37 °C in an

atmosphere of 5% CO_2. Cells were passaged using trypsin and stored in 5% DMSO and DMEM containing 10% FBS and antibiotics at -80 °C.

Isolation of mononucleosomes

Mononucleosomes were isolated from HeLa S3 cells. A pellet of 20 million HeLa S3 cells was resuspended in 5 ml of ice-cold TM2 buffer (10 mM Tris pH 7.5, 2 mM $MgCl_2$, 0.5 mM PMSF). Next, NP-40 was added to a final concentration of 0.6%. After incubation on ice for 5 minutes, the pellet was washed with TM2 buffer, resuspended in 400 µl of TM2 buffer containing micrococcal nuclease (New England BioLabs) and 1mM $CaCl_2$, and incubated in a Thermomixer for 10 minutes at 37 °C. Next, 2 mM EGTA, 300 mM NaCl, and 0.1% Triton X-100 were added, and the mixture was transferred into Bioruptor tubes (Diagenode). After sonication for 5 cycles of 30 seconds ON/OFF at 4 °C (Bioruptor Pico, Diagenode), the mixture was centrifuged (13000 g) for 10 minutes at 4 °C, and the supernatant was transferred to a fresh tube. For quality control, DNA was isolated from a small aliquot (PCR clean up, A&A Biotechnology) and checked on 1.5% agarose gel for DNA length (~150nt). Finally, mononucleosomes were divided into aliquots, fast-frozen with liquid nitrogen, and stored at -80 °C.

Generation of stable cell lines

Stable HeLa cell lines expressing NLS-EYFP-tagged proteins were generated using lentiviral transduction. Lentivirus stocks were prepared using plasmids pMD2.G and psPAX2, and a pLJM1-based transfer plasmid, by standard transfection with PEI (Sigma Aldrich #408727). HEK293T cells were grown in a 15 cm dish following transfection. When cells reached ~90% confluency, the medium was aspirated and replaced with fresh medium without antibiotics. Transfer plasmid pLJM1 and plasmids psPAX2 and pMD2.G were mixed in 4:2:1 ratio in Opti-MEM (20 µg, final volume 0.7 ml). Next, PEI in Opti-MEM was added (60 µg in a total volume 0.7 ml). The transfection mixture was briefly vortexed, incubated for 20 minutes at room temperature, and slowly added to the cell culture. 17 hours post-transfection, the medium was exchanged for fresh DMEM supplemented with 25 mM HEPES, 4 mM caffeine, and 1 mM sodium butyrate. The medium was harvested 48 hours post-transfection and centrifuged for 3 minutes at 3000 g. The supernatant was then filtered through a 0.45 µm PES filter, aliquoted, and stored at -80°C.

To generate stable cell lines, 0.1 million HeLa cells were seeded per well in a 6 well plate 24 hours before infection. Infection was carried out for 72 hours. Every 24 hours, media was exchanged for a new virus aliquot and fresh media in a ratio of 2.5:0.5, supplemented with DEAE-dextran (6 µg/ml). The negative control NLS-EYFP HeLa stable cell line was infected once after 24 hours. 72 hours after the first virus addition, the cells were transferred to a 10 cm plate and maintained in media supplemented with 2 µg/ml puromycin. Non-infected parental HeLa cells were also treated with puromycin to confirm successful selection. After selection, stable cell lines were sorted using a BD FACSAriaII (BD BioSciences). As a negative control, the original HeLa S3 cells were used. Next, cells expressing EYFP above

background levels were collected and transferred back to a culture plate. Finally, the cells were stored in 5% DMSO and DMEM containing 10% FBS and antibiotics at -80 °C.

GST pull-down

For GST pull-down, purified GST-tagged PHD domains (1 μM) were incubated with 30 μg of isolated mononucleosomes in 500 μl of pull-down buffer (10 mM Tris pH 7.5, 150 mM NaCl, 0.5% TritonX-100, 1 mM DTT, 150 μM $ZnSO_4$, 1 mM PMSF, 1xEDTA-free protease inhibitors [Roche]) overnight on a rotating wheel at 4 °C. 10% of the mixtures were put aside as input. Next, 40 μl of Glutathione Sepharose™ 4B beads (GE Healthcare) were washed three times with pull-down buffer and added to the sample containing the PHD mononucleosome complexes. The mixtures were then incubated for 2 hours on a rotating wheel at 4 °C. Following this, the supernatants were collected as flow through and the beads were washed three times with washing buffer (10 mM Tris pH 7.5, 150 mM NaCl, 0.5% Triton X-100). Finally, the beads were incubated at 95 °C with SDS-loading buffer (50 mM Tris pH 6.8, 2% SDS, 5% β-mercaptoethanol, 10% glycerol, bromophenol blue) for 10 min for Western blot.

SDS-PAGE, Western blots, and Far-Western blots

Proteins were diluted in SDS-loading buffer and heated at 95 °C for 5 minutes. The samples were then run on Tris/glycine/SDS-polyacrylamide gel (the 18% separating gel). For the Western blots, proteins were wet-transferred onto a nitrocellulose membrane (100 V, 1 hour, 4 °C). The membrane was blocked with 5% non-fat milk, incubated with a specific primary antibody for anti-histone H3 (Cell Signaling Technologies) overnight at 4 °C, washed 3 times with TBS-T, and incubated at room temperature with anti-rabbit HRP-conjugated secondary antibodies (Sigma) for 1 hour. For detection of GST-tagged proteins, the membrane was incubated at room temperature with an anti-GST primary antibody conjugated with HRP (Abcam).

For Far-Western blotting, 5 μg of mononucleosomes were resolved using an 18% SDS-PAGE gel, wet-transferred onto a nitrocellulose membrane (1 hour, 100V, 4 °C), and blocked for 1 hour at room temperature or overnight at 4 °C with 5% non-fat milk. The membranes were then washed three times with TBS-T and once with interaction buffer (20 mM HEPES 7.5, 100 mM KCl, 10% glycerol, 1 mM DTT, 150 μM $ZnSO_4$), incubated with 100 nM PHD domains in interaction buffer for 1 hour at room temperature, washed three times with TBS-T, and incubated with an anti-GST antibody conjugated with HRP (Abcam) for 1 hour at room temperature.

Immunodetection was performed using self-made ECL (0.2 mM coumaric acid, 1.25 mM luminol, 100 mM Tris 8.5, 0.03% H_2O_2) or Femto ECL (ThermoScientific). The chemiluminescent signals were scanned from the membranes using a BioRad documentation system. All antibodies used are listed in the Table S3**.**

Immunofluorescence

Cells for immunofluorescence were seeded onto cover slides at the bottom of 24 well plates (50.000 cells per well). After 24 hours, the cells were washed twice with PBS, fixed for 10 minutes at room temperature (4% paraformaldehyde in PBS), and washed twice with PBS again. The cover slides were either used immediately or stored sealed with parafilm for up to two weeks at 4 °C. For immunostaining, cells were incubated with permeabilisation buffer (0.25% TritonX100 in PBS) at room temperature for 20 minutes, washed once with PBS, placed in blocking buffer (3% bovine serum albumin, BSA in PBS) for 30 minutes, incubated with primary antibody diluted in Ab buffer (1% BSA in PBS) in a humid box for 1 hour or overnight at 4 °C, washed two times with PBS, incubated for 1 hour with secondary antibody and DAPI in Ab buffer, and washed two times with PBS. The cover slides were then briefly washed with MilliQ water and put on ethanol-cleaned glass slides containing Mowiol. The specimens were kept protected from light and analysed by confocal microscopy (LSM 800, Zeiss). All antibodies used are listed in the Table S3.

Standard CUR&RUN (C&R) and greenCUT&RUN (gC&R)

For standard C&R (Skene and Henikoff, 2017), the protocol published by Janssens and Henikoff (dx.doi.org/10.17504/protocols.io.zcpf2vn) was used. For gC&R, the protocol described by Nizamuddin and colleagues (Nizamuddin et al., 2021), with minor modifications, was used. Briefly, for both protocols, 0.5 million HeLa S3 cells were collected by trypsinisation and washed twice with wash buffer (20 mM HEPES–KOH pH 7.5, 150 mM NaCl, 0.5 mM spermidine and EDTA-free complete protease inhibitor cocktail). In parallel, BioMag®Plus Concanavalin A (conA) magnetic beads (10 µl per 0.5 million HeLa S3 cells) were incubated twice with binding buffer (20 mM HEPES–KOH pH 7.5, 10 mM KCl, 1 mM $CaCl_2$ and 1 mM $MnCl_2$). The washed cells were then mixed with the conA beads and incubated on a vortexer for 8 minutes. All steps up to this point were performed at room temperature to minimise cellular stress and DNA fragmentation.

For C&R, conA immobilised cells were incubated with 100 µl antibody buffer (20 mM HEPES–KOH pH 7.5, 150 mM NaCl, 0.5 mM spermidine and EDTA-free protease inhibitor cocktail, 0.01% digitonin, 2 mM EDTA) supplemented with the antibody overnight on a rotating wheel at 4 °C. All antibodies used are listed in the Table S3**.** After the overnight incubation, the cells were washed twice with ice-cold digitonin buffer and incubated on a nutator for 1 hour with 50 µl digitonin buffer containing 1.5 µl C&R MNase (Cell Signaling Technologies). For gC&R, conA immobilised cells were incubated for 4 minutes at room temperature with antibody buffer and then washed twice with digitonin buffer. Next, the immobilised cells were incubated in 100 µl digitonin buffer with 0.4 ng of the fusion protein consisting of anti-GFP nanobody and MNase domains (a kind gift from Marc Timmers) for 1 hour at 4 °C on a rotating wheel.

For both protocols, cells were then washed two times with ice-cold digitonin buffer, resuspended in 100 µl digitonin buffer, and kept on a wet-ice bath for at least 5 minutes. Next, MNase digestion was activated by adding $CaCl_2$ to a 2 mM final concentration. After 30

minutes, digestion was stopped by adding 100 µl 2xSTOP buffer (340 mM NaCl, 20 mM EDTA, 10 mM EGTA, 0.01% digitonin, 100 µg/ml of RNase A [ThermoScientific]). Next, the mixtures were incubated in a thermoblock at 37 °C to release chromatin-MNase complexes. Supernatants (200 µl) were supplemented with 2 µl of 10% (w/v) SDS and 2.5 µl proteinase K (20 mg/ml; ThermoScientific), and incubated for 1 hour at 50 °C in a thermoblock. DNA was extracted using a ChIP DNA Clean & Concentrator kit (ZymoResearch). For background correction in the C&R experiments, rabbit normal IgG (Sigma-Aldrich) was used, and in the gC&R experiments, the parental HeLa S3 cells were used.

Library preparation

For library preparation, a NEBNext II Ultra DNA Prep Kit was used according to Liu's protocol (dx.doi.org/10.17504/protocols.io.bagaibse). Briefly, DNA isolated after C&R or gC&R was mixed with END Prep Enzyme Mix and Reaction Buffer, and incubated in a thermocycler at 20 °C for 30 minutes, and at 50 °C for 1 hour. Next, NEB Next Adaptor (25x diluted, 0.6 µM) was ligated using Ligation Master Mix and Ligation Enhancer for 15 minutes at 20 °C, followed by hairpin cleavage by USER enzyme for 15 minutes at 37 °C. For the cleaning of adaptor-ligated DNA, left-side selection using 1.7x AMPure XP beads (Beckman Coulter) was used. The purified DNA was mixed with Q5 MasterMix, Universal PCR primer, appropriate index primer (NEB Single Index, NEBNext® Multiplex Oligos for Illumina), and used for PCR in a thermocycler (Biorad, Eppendorf). Finally, the amplified DNA libraries were cleaned with AMPure XP beads (first 0.55x right-sided, then left-sided in total 1.15x). The quality of libraries was verified using TapeStation (Perlan, Agilent).

NGS sequencing and data analysis

The concentration of the libraries was measured by qPCR using a Kapa Library Quantification kit (Kapa Biosciences, KK4824), according to the manufacturer's protocol. Paired-end (2x100nt) sequencing (10 MR histone modifications, 20-30MR full length hCOMPASS-like subunits and gC&R) was performed on an Illumina NovaSeq 6000 instrument, using a NovaSeq 6000 S1 Reagent Kit (200 cycles, Illumina) with the addition of 0.5% control library Phix (Illumina).

The upstream analysis was performed using the Galaxy server (Afgan et al., 2018). Original fastq files were trimmed using both the TrimGalore! and FastQTrimmer algorithms. The quality of the reads was confirmed by FastQC. Next, the paired-end reads were aligned to the human genome hg38 using Bowtie2 (Langmead and Salzberg, 2012), filtered for minimum MAPQ20 and against a blacklist, and deduplicated. Peak calling was done using MACS2 (Feng et al., 2012; Zhang et al., 2008). Heatmaps were created using plotHeatmap from deeptools in Galaxy (Ramírez et al., 2016). Peaks were first analysed in IGV (Robinson et al., 2011), then genome-wide using KEGG (Kanehisa and Goto, 2000) and GO (The Gene Ontology Consortium, 2019) analysis on the Cistrome-GO webserver (Li et al., 2019), and MEME-ChIP (Bailey et al., 2009; Machanick and Bailey, 2011) on the Galaxy server. For the comparison of the bed files with ENCODE or HMM imputed data, a measure

of enrichment that quantifies the excess over the random overlap was used. For bed files A and B, the size of the overlap region (A and B) was multiplied with the genome length, and divided by the sizes of A and B. Except for scale, the results were qualitatively similar to the standard Jaccard statistic from bedtools (Quinlan and Hall, 2010) that measures overlap as (A and B)/(A or B) for clustered PHD domains A13, 2D13, 2D46 with comparatively many binding sites. The non-standard overlap measure was chosen because, unlike the Jaccard statistic, it made it possible to meaningfully extend the analysis to 2C14 and 2C57 PHD domain clusters, which appeared to be bound at far fewer sites (presumably due to technical difficulties).

Diversity analysis (Shannon analysis)

For sequence conservation analysis, UNIPROT (UniProt Consortium, 2021) was queried for proteins annotated as KMT2A, KMT2B, KMT2C, or KMT2D. The program cd-hit (Li and Godzik, 2006) was used to cluster sequences at 80% sequence similarity. Fusion proteins or truncated proteins (any proteins with a length less than 90% of the length of the reference human protein) were excluded from the analysis. To avoid bias from multiple annotated splice variants, only one protein per paralogue and species was retained. Proteins were aligned using Clustal Omega (Sievers et al., 2011). Inspired by the program BALCONY (Płuciennik et al., 2018), sequence divergence in each position was quantified by a Shannon entropy measure of amino acid divergence (sum of $p_i log_2 p_i$, where p_i is the frequency of amino acid "i" in a given position). As scores were needed regarding the human protein and not to a gapped alignment, BALCONY was not used directly, and the calculation of Shannon scores was re-implemented in Python.

COSMIC data analysis

The data for the analysis of the distribution of mutations throughout the length of the KMT2 proteins and in the patient population was obtained from the Catalogue of Somatic Mutations in Cancer (COSMIC) release v.92. The COSMIC Mutation Data (Genome Screens) dataset containing data on mutations found in whole genome sequencing analysis was downloaded from the database as a filtered file to comprise the changes found in the KMT2A–D genes in all the malignancies described in COSMIC. The data from targeted screens (non-WGS) was not used in the distribution analysis of mutations as too many (~92%) data points are missing in relation to the type and location of mutations for the non-WGS dataset to be informative.

For the analysis of the distribution of mutations among the patients, the COSMIC Sample Features non-filtered dataset was downloaded to calculate the number of patients diagnosed with a given malignancy and to assign the mutations described in the Mutation Data file to individuals. The data points from the WGS Mutation Data dataset with 'Mutation description' assigned as 'Unknown' (33% of the whole dataset) were removed for the analysis.

AUTHOR CONTRIBUTION

Conceptualization of the study: ASC, MB. Manuscript preparation: ASC and MB with the input from the other authors. Preparation of figures: ASC, MB, MK. Shannon analysis: MB. Protein purification, Far-Western, GST pull-down, immunofluorescence, CUT&RUN, data deposition: ASC. Generation of HeLa stable cell lines: MP. Optimisation of PHD expression constructs: AAK, MB, ASC. Bioinformatics analysis of CUT&RUN data: MB and ASC. Analysis of COSMIC data: MK, ASC, MB. Supervision of the study and funding: MB.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY

The standard and greenCUT&RUN data have been deposited to Gene Expression Omnibus (GEO) under the accession number: GSE185921.

ACKNOWLEDGEMENT

We thank Marc Timmers (Freiburg University, Germany) for the kind gift of MNase-anti-GFP nanobody, Agnieszka Rawluszko-Wieczorek (Poznan University of Medical Sciences, Poland) for her help in cloning some plasmids, Dorota Adamska (CeNT, Warsaw University, Poland) for Illumina sequencing, Derek Janssens (Fred Hutchinson Cancer Research Center, USA) for his advice regarding CUT&RUN protocols, Bartlomiej Czerwinski who helped with informatics and Katarzyna Misztal, Karim Abu Nahia and Kamil Jastrzębski (IIMCB core facility) for their help with cell sorting and TapeStation analysis. We are also grateful to Albert Jeltsch (Stuttgart University, Germany), Agnieszka Rawluszko-Wieczorek (Poznan University of Medical Sciences, Poland) and Aleksandra Pekowska (Nencki Institute of Experimental Biology, Warsaw, Poland) for the fruitful discussions and to Katarzyna Szafran for comments on figures. NGS was performed at the Genomics Core Facility CeNT UW, using the NovaSeq 6000 platform financed by the Polish Ministry of Science and Higher Education (6817/IA/2018 2018-04-10). This work was supported by a HARMONIA grant from the Polish National Science Centre (NCN, UMO-2014/14/M/NZ5/00558) to MB and a grant from the Polish National Agency for Academic Exchange (NAWA, PPI/APM/2018/1/00034) to IIMCB.

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Grüning, B.A., et al. (2018). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res *46*, W537–W544.

Ali, A., and Tyagi, S. (2017). Diverse roles of WDR5-RbBP5-ASH2L-DPY30 (WRAD) complex in the functions of the SET1 histone methyltransferase family. J Biosci *42*, 155–159.

Ali, M., Yan, K., Lalonde, M.-E., Degerny, C., Rothbart, S.B., Strahl, B.D., Côté, J., Yang, X.-J., and Kutateladze, T.G. (2012). Tandem PHD fingers of MORF/MOZ acetyltransferases display selectivity for acetylated histone H3 and are required for the association with chromatin. J Mol Biol *424*, 328–338.

Ali, M., Hom, R.A., Blakeslee, W., Ikenouye, L., and Kutateladze, T.G. (2014). Diverse functions of PHD fingers of the MLL/KMT2 subfamily. Biochim Biophys Acta *1843*, 366–371.

Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell *173*, 371-385.e18.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208.

Bienz, M. (2006). The PHD finger, a nuclear protein-interaction domain. Trends Biochem Sci *31*, 35–40.

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

Bochyńska, A., Lüscher-Firzlaff, J., and Lüscher, B. (2018). Modes of Interaction of KMT2 Histone H3 Lysine 4 Methyltransferase/COMPASS Complexes with Chromatin. Cells *7*, 17.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol Cell *49*, 825–837.

Cenik, B.K., and Shilatifard, A. (2021). COMPASS and SWI/SNF complexes in development and disease. Nat Rev Genet *22*, 38–58.

Chauhan, C., Zraly, C.B., Parilla, M., Diaz, M.O., and Dingwall, A.K. (2012). Histone recognition and nuclear receptor co-activator functions of Drosophila cara mitad, a homolog of the N-terminal portion of mammalian MLL2 and MLL3. Development *139*, 1997–2008.

Chauhan, C., Zraly, C.B., and Dingwall, A.K. (2013). The Drosophila COMPASS-like Cmi-Trr coactivator complex regulates dpp/BMP signaling in pattern formation. Developmental Biology *380*, 185–198.

Chen, Y., Wan, B., Wang, K.C., Cao, F., Yang, Y., Protacio, A., Dou, Y., Chang, H.Y., and Lei, M. (2011). Crystal structure of the N-terminal region of human Ash2L shows a winged-helix motif involved in DNA binding. EMBO Rep *12*, 797–803.

Cierpicki, T., Risner, L.E., Grembecka, J., Lukasik, S.M., Popovic, R., Omonkowska, M., Shultis, D.D., Zeleznik-Le, N.J., and Bushweller, J.H. (2010). Structure of the MLL CXXC domain-DNA complex and its functional role in MLL-AF9 leukemia. Nat Struct Mol Biol *17*, 62–68.

Couture, J.-F., Collazo, E., and Trievel, R.C. (2006). Molecular recognition of histone H3 by the WD40 protein WDR5. Nat Struct Mol Biol *13*, 698–703.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res *46*, D794–D801.

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

Dhar, S.S., Lee, S.-H., Kan, P.-Y., Voigt, P., Ma, L., Shi, X., Reinberg, D., and Lee, M.G. (2012). Trans-tail regulation of MLL4-catalyzed H3K4 methylation by H4R3 symmetric dimethylation is mediated by a tandem PHD of MLL4. Genes Dev *26*, 2749–2762.

Dou, Y., Milne, T.A., Ruthenburg, A.J., Lee, S., Lee, J.W., Verdine, G.L., Allis, C.D., and Roeder, R.G. (2006). Regulation of MLL1 H3K4 methyltransferase activity by its core components. Nat Struct Mol Biol *13*, 713–719.

Dreijerink, K.M.A., Timmers, H.T.M., and Brown, M. (2017). Twenty years of menin: emerging opportunities for restoration of transcriptional regulation in MEN1. Endocr Relat Cancer *24*, T135–T145.

Dreveny, I., Deeves, S.E., Fulton, J., Yue, B., Messmer, M., Bhattacharya, A., Collins, H.M., and Heery, D.M. (2014). The double PHD finger domain of MOZ/MYST3 induces α-helical structure of the histone H3 tail to facilitate acetylation and methylation sampling and modification. Nucleic Acids Research *42*, 822–835.

Eberharter, A., and Becker, P.B. (2002). Histone acetylation: a switch between repressive and permissive chromatin. EMBO Rep *3*, 224–229.

Ernst, J., and Kellis, M. (2015). Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol *33*, 364–376.

Ernst, P., and Vakoc, C.R. (2012). WRAD: enabler of the SET1-family of H3K4 methyltransferases. Briefings in Functional Genomics *11*, 217–226.

Fagan, R.J., and Dingwall, A.K. (2019). COMPASS Ascending: Emerging clues regarding the roles of MLL3/KMT2C and MLL2/KMT2D proteins in cancer. Cancer Letters *458*, 56–65.

Fair, K., Anderson, M., Bulanova, E., Mi, H., Tropschug, M., and Diaz, M.O. (2001). Protein Interactions of the MLL PHD Fingers Modulate MLL Target Gene Regulation in Human Cells. Mol Cell Biol *21*, 3589–3597.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nat Protoc *7*, 1728–1740.

The Gene Ontology Consortium (2019). The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res *47*, D330–D338.

Herz, H.-M., Mohan, M., Garruss, A.S., Liang, K., Takahashi, Y., Mickey, K., Voets, O., Verrijzer, C.P., and Shilatifard, A. (2012). Enhancer-associated H3K4 monomethylation by Trithorax-related, the Drosophila homolog of mammalian Mll3/Mll4. Genes Dev *26*, 2604–2620.

Herz, H.-M., Hu, D., and Shilatifard, A. (2014). Enhancer Malfunction in Cancer. Mol Cell *53*, 859–866.

Hu, D., Gao, X., Morgan, M.A., Herz, H.-M., Smith, E.R., and Shilatifard, A. (2013). The MLL3/MLL4 branches of the COMPASS family function as major histone H3K4 monomethylases at enhancers. Mol Cell Biol *33*, 4745–4754.

Jeong, J.-Y., Yim, H.-S., Ryu, J.-Y., Lee, H.S., Lee, J.-H., Seen, D.-S., and Kang, S.G. (2012). One-step sequence- and ligation-independent cloning as a rapid and versatile cloning method for functional genomics studies. Appl Environ Microbiol *78*, 5440–5443.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28*, 27–30.

Lange, M., Kaynak, B., Forster, U.B., Tönjes, M., Fischer, J.J., Grimm, C., Schlesinger, J., Just, S., Dunkel, I., Krueger, T., et al. (2008). Regulation of muscle development by DPF3, a novel histone acetylation and methylation reader of the BAF chromatin remodeling complex. Genes Dev *22*, 2370–2384.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357–359.

Larsson, C., Cordeddu, L., Siggens, L., Pandzic, T., Kundu, S., He, L., Ali, M.A., Pristovšek, N., Hartman, K., Ekwall, K., et al. (2020). Restoration of KMT2C/MLL3 in human colorectal cancer cells reinforces genome-wide H3K4me1 profiles and influences cell growth and gene expression. Clin Epigenetics *12*, 74.

Lechner, M.S., Levitan, I., and Dressler, G.R. (2000). PTIP, a novel BRCT domain-containing protein interacts with Pax2 and is associated with active chromatin. Nucleic Acids Res *28*, 2741–2751.

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

Lee, J.-H., Tate, C.M., You, J.-S., and Skalnik, D.G. (2007). Identification and characterization of the human Set1B histone H3-Lys4 methyltransferase complex. J Biol Chem *282*, 13419–13428.

Li, M.Z., and Elledge, S.J. (2007). Harnessing homologous recombination in vitro to generate recombinant DNA via SLIC. Nat Methods *4*, 251–256.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Li, S., Wan, C., Zheng, R., Fan, J., Dong, X., Meyer, C.A., and Liu, X.S. (2019). Cistrome-GO: a web server for functional enrichment analysis of transcription factor ChIP-seq peaks. Nucleic Acids Res *47*, W206–W211.

Liedtke, S., Buchheiser, A., Bosch, J., Bosse, F., Kruse, F., Zhao, X., Santourlidis, S., and Kögler, G. (2010). The HOX Code as a "biological fingerprint" to distinguish functionally distinct stem cell populations derived from cord blood. Stem Cell Research *5*, 40–50.

Liu, H., Cheng, E.H., and Hsieh, J.J. (2009). MLL fusions. Cancer Biol Ther *8*, 1204–1211.

Liu, Y., Qin, S., Chen, T.-Y., Lei, M., Dhar, S.S., Ho, J.C., Dong, A., Loppnau, P., Li, Y., Lee, M.G., et al. (2019). Structural insights into trans-histone regulation of H3K4 methylation by unique histone H4 binding of MLL3/4. Nat Commun *10*, 36.

Lorton, B.M., Harijan, R.K., Burgos, E.S., Bonanno, J.B., Almo, S.C., and Shechter, D. (2020). A Binary Arginine Methylation Switch on Histone H3 Arginine 2 Regulates Its Interaction with WDR5. Biochemistry *59*, 3696–3708.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251–260.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696–1697.

Mansfield, R.E., Musselman, C.A., Kwan, A.H., Oliver, S.S., Garske, A.L., Davrazou, F., Denu, J.M., Kutateladze, T.G., and Mackay, J.P. (2011). Plant Homeodomain (PHD) Fingers of CHD4 Are Histone H3-binding Modules with Preference for Unmodified H3K4 and Methylated H3K9. J Biol Chem *286*, 11779–11791.

Maris, C., Dominguez, C., and Allain, F.H.-T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J *272*, 2118–2131.

Mas-Y-Mas, S., Barbon, M., Teyssier, C., Déméné, H., Carvalho, J.E., Bird, L.E., Lebedev, A., Fattori, J., Schubert, M., Dumas, C., et al. (2016). The Human Mixed Lineage Leukemia 5 (MLL5), a Sequentially and Structurally Divergent SET Domain-Containing Protein with No Intrinsic Catalytic Activity. PLoS One *11*, e0165139.

Meyer, C., Hofmann, J., Burmeister, T., Gröger, D., Park, T.S., Emerenciano, M., Pombo de Oliveira, M., Renneville, A., Villarese, P., Macintyre, E., et al. (2013). The MLL recombinome of acute leukemias in 2013. Leukemia *27*, 2165–2176.

Meyer, C., Burmeister, T., Gröger, D., Tsaur, G., Fechina, L., Renneville, A., Sutton, R., Venn, N.C., Emerenciano, M., Pombo-de-Oliveira, M.S., et al. (2018). The MLL recombinome of acute leukemias in 2017. Leukemia *32*, 273–284.

Meyer, C., Lopes, B.A., Caye-Eude, A., Cavé, H., Arfeuille, C., Cuccuini, W., Sutton, R., Venn, N.C., Oh, S.H., Tsaur, G., et al. (2019). Human MLL/KMT2A gene exhibits a second breakpoint cluster region for recurrent MLL-USP2 fusions. Leukemia *33*, 2306–2340.

Milne, T.A., Briggs, S.D., Brock, H.W., Martin, M.E., Gibbs, D., Allis, C.D., and Hess, J.L. (2002). MLL targets SET domain methyltransferase activity to Hox gene promoters. Mol Cell *10*, 1107–1117.

Mujtaba, S., Zeng, L., and Zhou, M.-M. (2007). Structure and acetyl-lysine recognition of the bromodomain. Oncogene *26*, 5521–5527.

Musselman, C.A., and Kutateladze, T.G. (2009). PHD fingers: epigenetic effectors and potential drug targets. Mol Interv *9*, 314–323.

Musselman, C.A., and Kutateladze, T.G. (2011). Handpicking epigenetic marks with PHD fingers. Nucleic Acids Res *39*, 9061–9071.

Nizamuddin, S., Koidl, S., Bhuiyan, T., Werner, T.V., Biniossek, M.L., Bonvin, A.M.J.J., Lassmann, S., and Timmers, Ht. (2021). Integrating quantitative proteomics with accurate genome profiling of transcription factors by greenCUT&RUN. Nucleic Acids Res *49*, e49.

Płuciennik, A., Stolarczyk, M., Bzówka, M., Raczyńska, A., Magdziarz, T., and Góra, A. (2018). BALCONY: an R package for MSA and functional compartments of protein variability analysis. BMC Bioinformatics *19*, 300.

Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

Ramírez, F., Ryan, D.P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A.S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res *44*, W160-165.

Rao, R.C., and Dou, Y. (2015). Hijacked in cancer: the MLL/KMT2 family of methyltransferases. Nat Rev Cancer *15*, 334–346.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24–26.

Ruthenburg, A.J., Wang, W., Graybosch, D.M., Li, H., Allis, C.D., Patel, D.J., and Verdine, G.L. (2006). Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. Nat Struct Mol Biol *13*, 704–712.

Sanchez, R., and Zhou, M.-M. (2011). The PHD Finger: A Versatile Epigenome Reader. Trends Biochem Sci *36*, 364–372.

Sarvan, S., Avdic, V., Tremblay, V., Chaturvedi, C.-P., Zhang, P., Lanouette, S., Blais, A., Brunzelle, J.S., Brand, M., and Couture, J.-F. (2011). Crystal structure of the trithorax group protein ASH2L reveals a forkhead-like DNA binding domain. Nat Struct Mol Biol *18*, 857–859.

Shi, X., Hong, T., Walter, K.L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Peña, P., Lan, F., Kaadige, M.R., et al. (2006). ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. Nature *442*, 96–99.

Shilatifard, A. (2012). The COMPASS family of histone H3K4 methylases: mechanisms of regulation in development and disease pathogenesis. Annu Rev Biochem *81*, 65–95.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol *7*, 539.

Skene, P.J., and Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife *6*, e21856.

Soares, L.M., He, P.C., Chun, Y., Suh, H., Kim, T., and Buratowski, S. (2017). Determinants of Histone H3K4 Methylation Patterns. Mol Cell *68*, 773-785.e6.

Stroynowska-Czerwinska, A., Piasecka, A., and Bochtler, M. (2018). Specificity of MLL1 and TET3 CXXC domains towards naturally occurring cytosine modifications. Biochim Biophys Acta Gene Regul Mech *1861*, 1093–1101.

Takahashi, Y., Westfield, G.H., Oleskie, A.N., Trievel, R.C., Shilatifard, A., and Skiniotis, G. (2011). Structural analysis of the core COMPASS family of histone H3K4 methylases from yeast to human. Proc Natl Acad Sci U S A *108*, 20526–20531.

Uchiumi, F., Miyazaki, S., and Tanuma, S.-I. (2011). Biological functions of the duplicated GGAA-motifs in various human promoter regions. Yakugaku Zasshi *131*, 1787–1800.

UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res *49*, D480–D489.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Wang, J., Muntean, A.G., Wu, L., and Hess, J.L. (2012). A Subset of Mixed Lineage Leukemia Proteins Has Plant Homeodomain (PHD)-mediated E3 Ligase Activity. J Biol Chem *287*, 43410–43416.

Wang, P., Lin, C., Smith, E.R., Guo, H., Sanderson, B.W., Wu, M., Gogol, M., Alexander, T., Seidel, C., Wiedemann, L.M., et al. (2009). Global analysis of H3K4 methylation defines MLL family member targets and points to a role for MLL1-mediated H3K4 methylation in the regulation of transcriptional initiation by RNA polymerase II. Mol Cell Biol *29*, 6074–6085.

Wang, Z., Song, J., Milne, T.A., Wang, G.G., Li, H., Allis, C.D., and Patel, D.J. (2010). Pro isomerization in MLL1 PHD3-bromo cassette connects H3K4me readout to CyP33 and HDAC-mediated repression. Cell *141*, 1183–1194.

Warner, J.B., Philippakis, A.A., Jaeger, S.A., He, F.S., Lin, J., and Bulyk, M.L. (2008). Systematic identification of mammalian regulatory motifs' target genes and functions. Nat Methods *5*, 347–353.

Worden, E.J., Hoffmann, N.A., Hicks, C.W., and Wolberger, C. (2019). Mechanism of Cross-talk between H2B Ubiquitination and H3 Methylation by Dot1L. Cell *176*, 1490-1501.e12.

Wysocka, J., Swigut, T., Milne, T.A., Dou, Y., Zhang, X., Burlingame, A.L., Roeder, R.G., Brivanlou, A.H., and Allis, C.D. (2005). WDR5 associates with histone H3 methylated at K4 and is essential for H3 K4 methylation and vertebrate development. Cell *121*, 859–872.

Yu, B.D., Hess, J.L., Horning, S.E., Brown, G.A., and Korsmeyer, S.J. (1995). Altered Hox expression and segmental identity in Mll-mutant mice. Nature *378*, 505–508.

Zeng, L., Zhang, Q., Li, S., Plotnikov, A.N., Walsh, M.J., and Zhou, M.-M. (2010). Mechanism and regulation of acetylated histone binding by the tandem PHD finger of DPF3b. Nature *466*, 258–262.

Zhang, X., Tamaru, H., Khan, S.I., Horton, J.R., Keefe, L.J., Selker, E.U., and Cheng, X. (2002). Structure of the Neurospora SET domain protein DIM-5, a histone H3 lysine methyltransferase. Cell *111*, 117–127.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhang, Y., Jang, Y., Lee, J.-E., Ahn, J., Xu, L., Holden, M.R., Cornett, E.M., Krajewski, K., Klein, B.J., Wang, S.-P., et al. (2019). Selective binding of the PHD6 finger of MLL4 to histone H4K16ac links MLL4 and MOF. Nat Commun *10*, 2314.

Zhao, J., Chen, W., Pan, Y., Zhang, Y., Sun, H., Wang, H., Yang, F., Liu, Y., Shen, N., Zhang, X., et al. (2021). Structural insights into the recognition of histone H3Q5 serotonylation by WDR5. Sci Adv *7*, eabf4291.

Ziemin-van der Poel, S., McCabe, N.R., Gill, H.J., Espinosa, R., Patel, Y., Harden, A., Rubinelli, P., Smith, S.D., LeBeau, M.M., and Rowley, J.D. (1991). Identification of a gene, MLL, that spans the breakpoint in 11q23 translocations associated with human leukemias. Proc Natl Acad Sci U S A *88*, 10735–10739.

FIGURES:

Figure 1. PHDs are highly conserved in KMT2A-D proteins.
(A) Scheme of KMT2A-G protein family architecture. Yellow: CXXC, orange: PHD, light green: bromodomain, violet: FY, green: SET, grey: AT hook. Taspase I cleavage sites in KMT2A and KMT2B proteins are labelled with a black line. **(B)** Amino acid diversity analysis (Shannon analysis) of KMT2A-D proteins across species. The scheme of full-length proteins with indicated domains is labelled below the plot for each paralogue. All pictures are shown with the same scale. See also Figure S1.

Figure 2. Interaction between PHDs and histones strengthens with an increasing number of PHDs.
(A) Graphical scheme of purified isolated PHDs. **(B)** Table with protein name and borders, based on the UniProt sequence. **(C)** *In vitro* approaches used with purified GST-tagged proteins and scheme of Far-Western blot methodology. **(D)** Far-Western blot performed with 100 nM solution of singlets (2C7, 2Ce7), doublet (2C67), and triplet (2C57) PHDs with membrane containing recombinant (H3/H4) and endogenous histones isolated from HeLa

cells. All membranes were exposed in the same session. The experiment was performed with at least two replicates. **(E)** Western blot analysis of the samples after GST-pull down with singlets (2A3, 2C7, 2D6), enhanced singlets (2Ce7, 2De6), doublet (2C67), and triplet (2A13, 2C14, 2C57, 2D13, 2D46) GST-tagged PHDs. Bound mononucleosomes were identified using an anti-H3 antibody. The experiment was performed at the same time. See also Figure S2.

Figure 3. Isolated PHDs, expressed in the HeLa cell lines, recognize specifically acetylation-enriched regions of active promoters and enhancers.

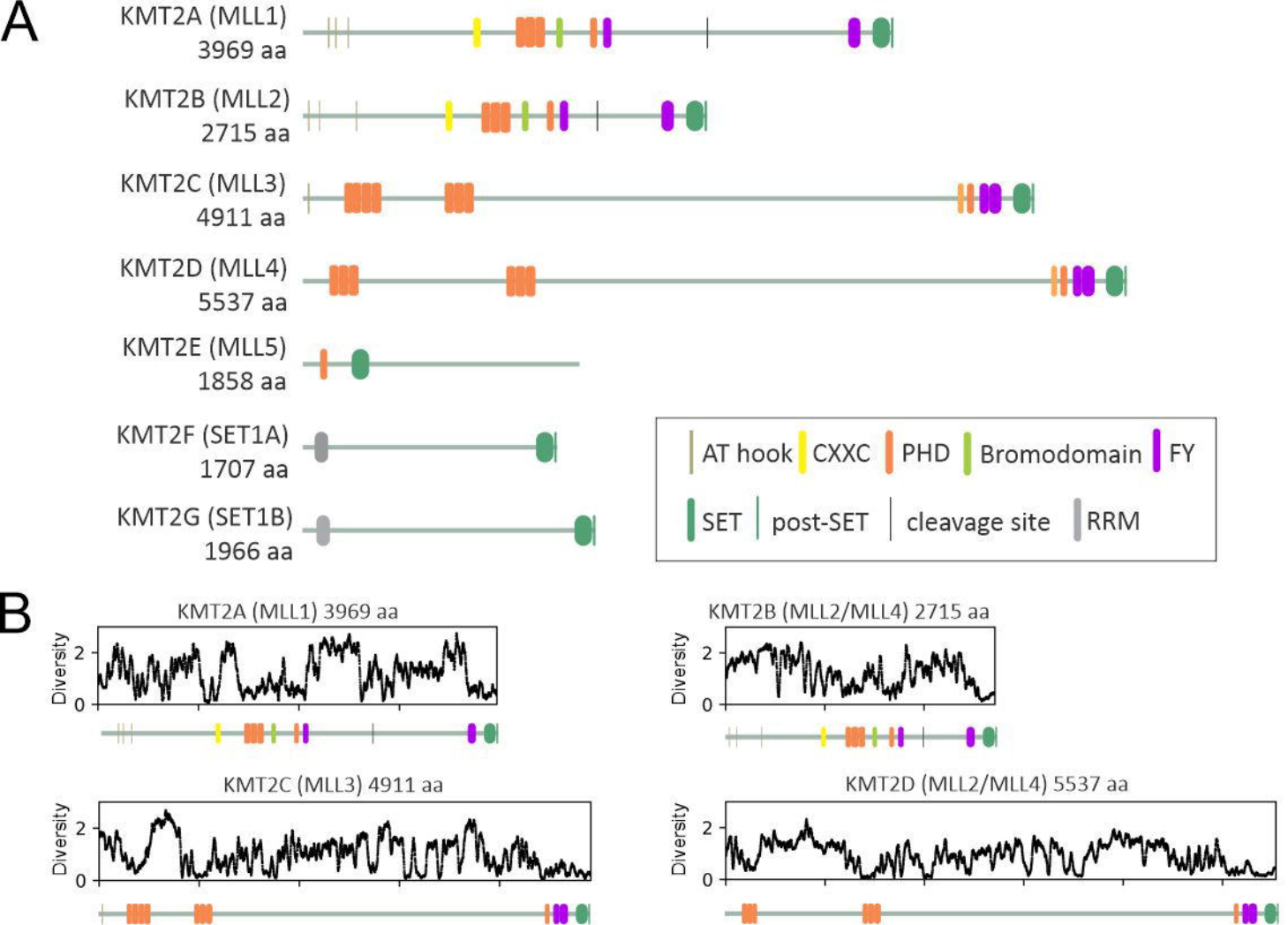
(A) Scheme of greenCUT&RUN (gC&R) experiments, performed in stable HeLa cell lines, expressing NLS-EYFP-tagged clustered PHDs of KMT2A-D proteins. **(B)** Fluorescent images of NLS-EYFP-tagged clustered PHDs, expressed in stable HeLa cell lines as well as immunofluorescent images with antibodies against full-length KMT2D protein (anti-KMT2D) and selected histone modification (anti-H3K4me1, anti-H3K4me3). **(C-D)** Pictures from IGV genome browser, showing **(C)** common and **(D)** various reads patterns of C&R selected hCOMPASS-like core subunits (darker grey WDR5, RbBP5 and brighter grey KMT2A) as well as gC&R of clustered PHDs (orange 2A13, light violet 2C14, dark violet 2C57, light blue 2D13, darker blue 2D46). The genomic localization is shown at the top and a scale of a genome at the bottom right corner. HMM is the lane with imputed chromatin states; red is an indicator of the active promoter region, yellow and orange of active enhancers and green of transcribed regions. **(E)** Heatmaps around TSS region (+- 2kb) of clustered PHDs (2A13, 2C14, 2C57, 2D13, 2D46), negative control (NLS-EYFP) gC&R signals in comparison to the background of parental HeLa cells and hCOMPASS-like core subunits (WDR5, RbBP5, KMT2A) or histone marks (H3K4me1/3, H4K16ac) in comparison to IgG negative control. **(F)** Genomic distribution of identified peaks (n, number of peaks) for the indicated proteins – clustered PHDs (2A13, 2C14, 2C57, 2D13, 2D46) and selected hCOMPASS-like core subunits (WDR5, RbBP5 and KMT2A). **(G)** Venn diagrams, based on Intervene results. Left panel – comparison of the unique and common peaks of selected clustered PHDs (2A13, 2D13, 2D46). Right panel – comparison of the unique and common peaks of isolated PHDs of KMT2A (2A13) with KMT2A full-length protein and hCOMPASS-like core subunit, WDR5. **(H-I)** Enrichment analysis comparing the genome-wide distribution of the peaks of selected protein (PHDs, WDR5, RbBP5 and KMT2A) and **(H)** HMM imputed states or **(I)** HMM imputed chromatin marks. The darker signal, the stronger correlation is observed between the two samples. **(H)** In bright grey enhancer regions, in orange active promoters and red repressed or poised regions are labelled. Detailed characterisation of imputed chromatin promoter and enhancer states and marks is shown in Suppl Fig. 3F-G. **(I)** Acetylations labelled in green and repressive methylations labelled in red. See also Figure S3.

Figure 4. PHDs and KMT2A bind chromatin regions involved in the cell cycle and cancer-related pathways.

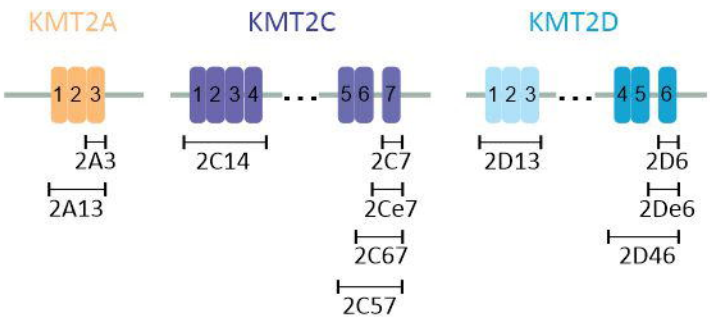
Top twelve common hits of **(A)** KEGG pathway and **(B)** GO biological processes analysis of the peaks identified for full-length KMT2A (grey) and clustered PHDs (orange 2A13, light

blue 2D13, darker blue 2D46). The top twelve hits for all proteins are shown separately in Suppl. Fig. 4A-B. 2C14 and 2C57 were not included due to the low number of identified peaks. **(C)** Common DNA motif (GGAA) enriched in C&R/gC&R peaks of full-length KMT2A, as well as isolated clustered PHDs, identified by MEME-ChIP analysis. A set of top three DNA motifs identified by MEME-ChIP for each protein are shown separately in Suppl. Fig. 4C. See also Figure S4.

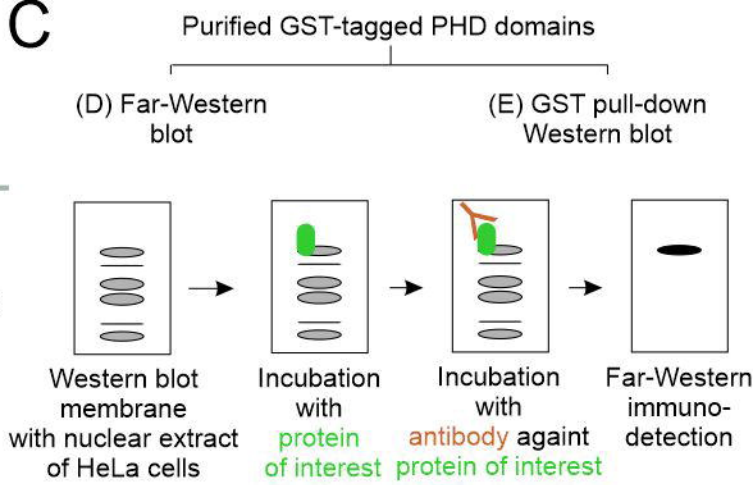
Figure 5. KMT2A-D PHDs, especially KMT2C, are mutated in cancer patients within PHD domains and amino acids involved in Zn2+ finger folding.
Data obtained from the COSMIC database. **(A)** Distribution of mutations among patients: missense, nonsense and indel within full-length KMT2A-D (left panel); missense within full-length KMT2A-D (central panel); missense in PHDs (right panel) of KMT2A-D. **(B)** Distribution of missense (violet), in-frame (pink), frameshift (light blue), and nonsense (green) mutations within KMT2A-D full-length proteins. **(C)** The frequency of missense mutations within crucial amino acids (Cys4HisCys2Cys/His, C4HC2C/H motif) of PHDs. Grey: mutations in all KMT2 proteins, orange: KMT2A, light orange: KMT2B, violet: KMT2C, blue: KMT2D. **(D)** The ratio between missense mutations and the number of residues within a particular protein region (full-length, PHD, C4HC2C/H motif, or rest of residues in PHD). Bars are collected separately for KMT2A-D.


A
KMT2A (MLL1)
3969 aa
KMT2B (MLL2)
2715 aa
KMT2C (MLL3)
4911 aa
KMT2D (MLL4)
5537 aa
KMT2E (MLL5)
1858 aa
KMT2F (SET1A)
1707 aa
KMT2G (SET1B)
1966 aa
AT hook
CXXC
PHD
Bromodomain
FY
SET
post-SET
cleavage site
RRM
B
KMT2A (MLL1) 3969 aa
KMT2B (MLL2/MLL4) 2715 aa
KMT2C (MLL3) 4911 aa
KMT2D (MLL2/MLL4) 5537 aa
Diversity
2
0

A



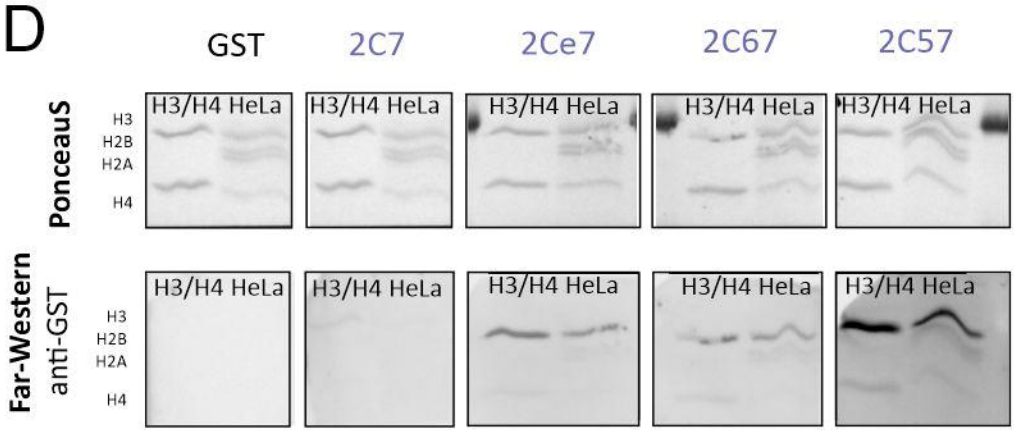
C



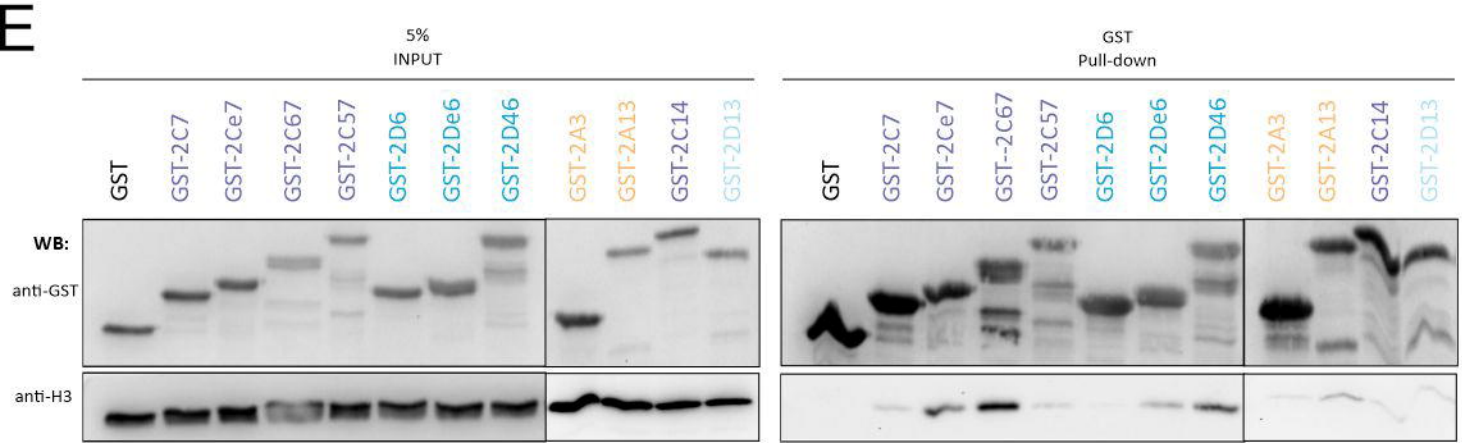
B

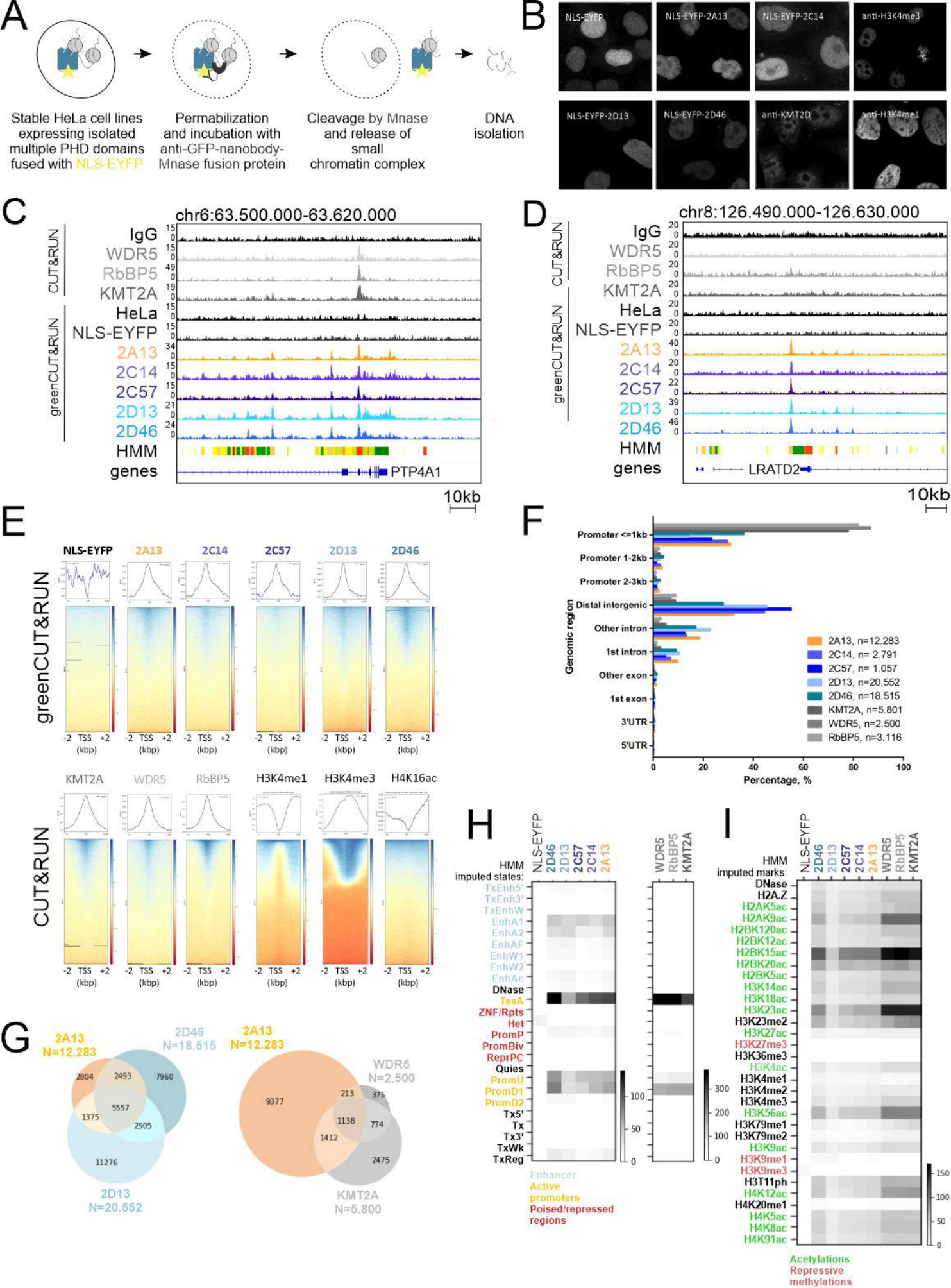
UniProt:	Name:	Residues:
Q03164	2A3	1565-1631
	2A13	1416-1640
Q8NEZ4	2C14	266-532
	2C7	1082-1143
	2Ce7	1055-1144
	2C67	1008-1144
	2C57	934-1149
O14686	2D13	156-341
	2D6	1503-1561
	2De6	1475-1564
	2D46	1351-1572

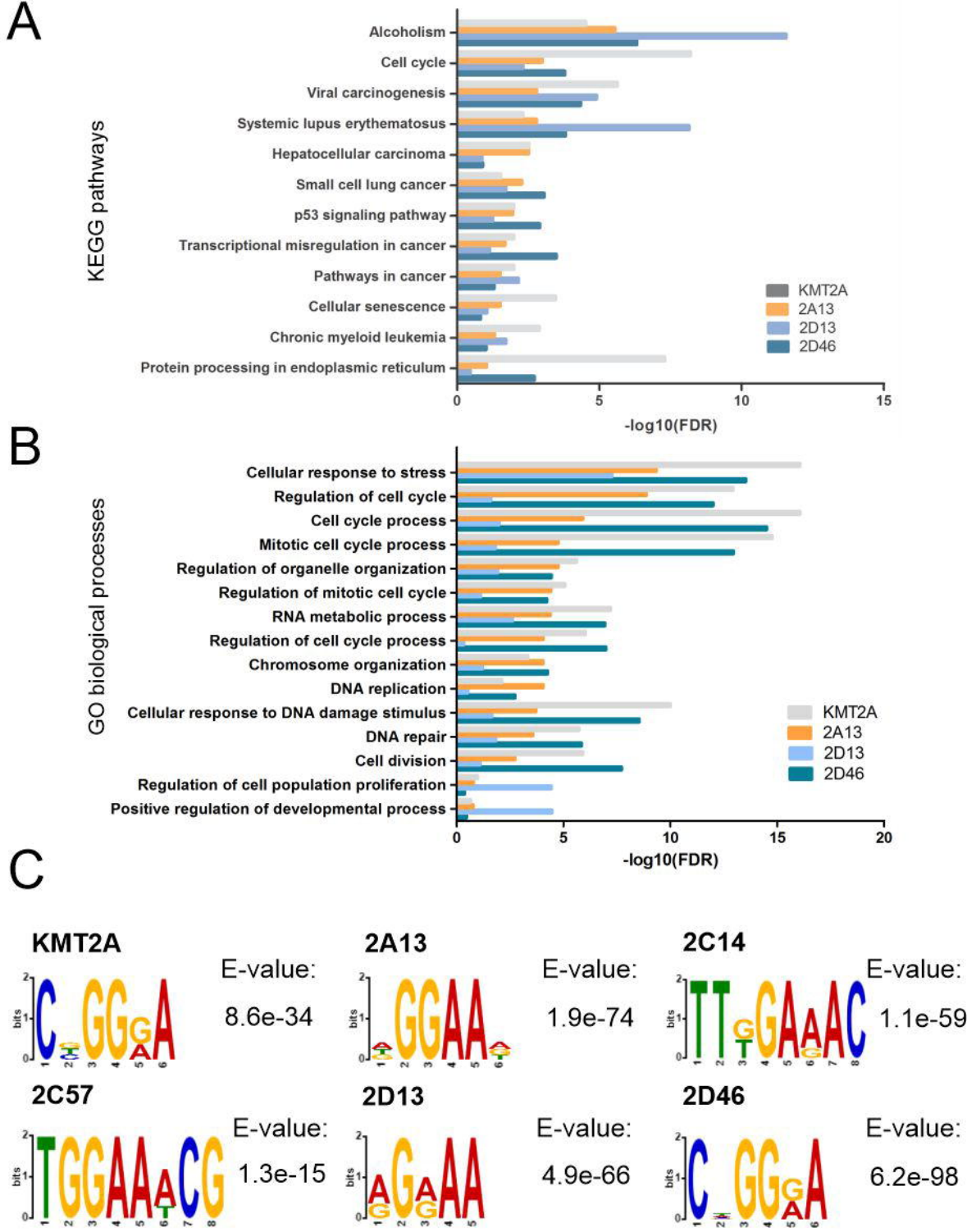
D

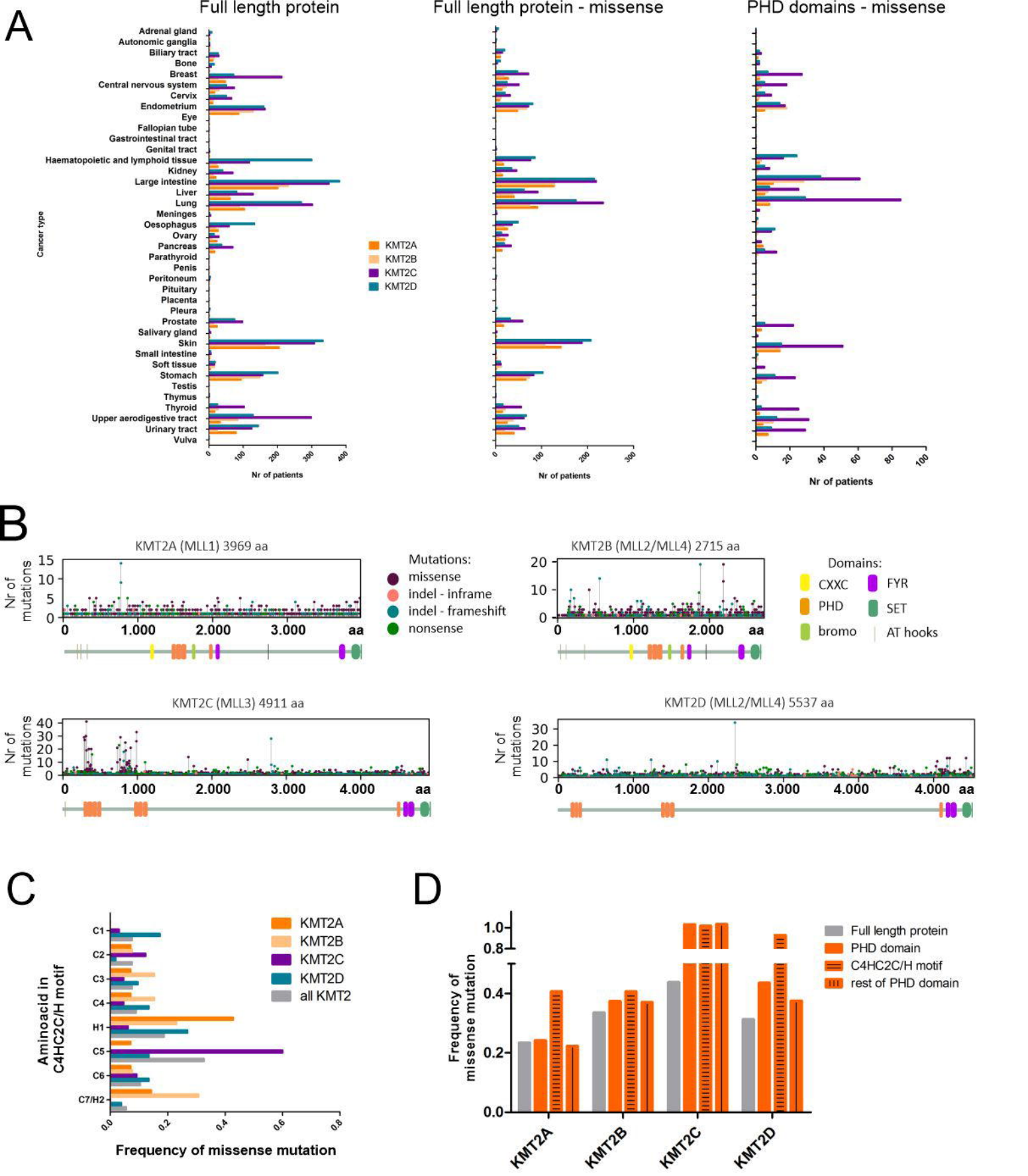


E




A
Stable HeLa cell lines expressing isolated multiple PHD domains fused with NLS-EYFP
Permabilization and incubation with anti-GFP-nanobody-Mnase fusion protein
Cleavage by Mnase and release of small chromatin complex
DNA isolation
B
NLS-EYFP
NLS-EYFP-2A13
NLS-EYFP-2C14
anti-H3K4me3
NLS-EYFP-2D13
NLS-EYFP-2D46
anti-KMT2D
anti-H3K4me1
C
chr6:63.500.000-63.620.000
CUT&RUN
IgG
WDR5
RbBP5
KMT2A
greenCUT&RUN
HeLa
NLS-EYFP
2A13
2C14
2C57
2D13
2D46
HMM
genes
PTP4A1
10kb
D
chr8:126.490.000-126.630.000
LRATD2
10kb
E
NLS-EYFP
2A13
2C14
2C57
2D13
2D46
greenCUT&RUN
KMT2A
WDR5
RbBP5
H3K4me1
H3K4me3
H4K16ac
CUT&RUN
-2 TSS +2
(kbp)
F
Genomic region
Promoter <=1kb
Promoter 1-2kb
Promoter 2-3kb
Distal intergenic
Other intron
1st intron
Other exon
1st exon
3'UTR
5'UTR
2A13, n=12.283
2C14, n= 2.791
2C57, n= 1.057
2D13, n=20.552
2D46, n=18.515
KMT2A, n=5.801
WDR5, n=2.500
RbBP5, n=3.116
Percentage, %
G
2A13 N=12.283
2D46 N=18.515
2D13 N=20.552
2804
2493
7960
5557
1375
2505
11276
2A13 N=12.283
WDR5 N=2.500
KMT2A N=5.800
9377
213
375
1138
774
1412
2475
H
HMM imputed states:
NLS-EYFP
2D46
2D13
2C57
2C14
2A13
WDR5
RbBP5
KMT2A
TxEnh5'
TxEnh3'
TxEnhW
EnhA1
EnhA2
EnhAF
EnhW1
EnhW2
EnhAc
DNase
TssA
ZNF/Rpts
Het
PromP
PromBiv
ReprPC
Quies
PromU
PromD1
PromD2
Tx5'
Tx
Tx3'
TxWk
TxReg
Enhancer
Active promoters
Poised/repressed regions
I
HMM imputed marks:
DNase
H2A.Z
H2AK5ac
H2AK9ac
H2BK120ac
H2BK12ac
H2BK15ac
H2BK20ac
H2BK5ac
H3K14ac
H3K18ac
H3K23ac
H3K23me2
H3K27ac
H3K27me3
H3K36me3
H3K4ac
H3K4me1
H3K4me2
H3K4me3
H3K56ac
H3K79me1
H3K79me2
H3K9ac
H3K9me1
H3K9me3
H3T11ph
H4K12ac
H4K20me1
H4K5ac
H4K8ac
H4K91ac
Acetylations
Repressive methylations


A
KEGG pathways
Alcoholism
Cell cycle
Viral carcinogenesis
Systemic lupus erythematosus
Hepatocellular carcinoma
Small cell lung cancer
p53 signaling pathway
Transcriptional misregulation in cancer
Pathways in cancer
Cellular senescence
Chronic myeloid leukemia
Protein processing in endoplasmic reticulum
KMT2A
2A13
2D13
2D46
0
5
10
15
-log10(FDR)
B
GO biological processes
Cellular response to stress
Regulation of cell cycle
Cell cycle process
Mitotic cell cycle process
Regulation of organelle organization
Regulation of mitotic cell cycle
RNA metabolic process
Regulation of cell cycle process
Chromosome organization
DNA replication
Cellular response to DNA damage stimulus
DNA repair
Cell division
Regulation of cell population proliferation
Positive regulation of developmental process
KMT2A
2A13
2D13
2D46
0
5
10
15
20
-log10(FDR)
C
KMT2A
E-value:
8.6e-34
2A13
E-value:
1.9e-74
2C14
E-value:
1.1e-59
2C57
E-value:
1.3e-15
2D13
E-value:
4.9e-66
2D46
E-value:
6.2e-98


A
Full length protein
Full length protein - missense
PHD domains - missense
Cancer type
Adrenal gland
Autonomic ganglia
Biliary tract
Bone
Breast
Central nervous system
Cervix
Endometrium
Eye
Fallopian tube
Gastrointestinal tract
Genital tract
Haematopoietic and lymphoid tissue
Kidney
Large intestine
Liver
Lung
Meninges
Oesophagus
Ovary
Pancreas
Parathyroid
Penis
Peritoneum
Pituitary
Placenta
Pleura
Prostate
Salivary gland
Skin
Small intestine
Soft tissue
Stomach
Testis
Thymus
Thyroid
Upper aerodigestive tract
Urinary tract
Vulva
KMT2A
KMT2B
KMT2C
KMT2D
Nr of patients
B
KMT2A (MLL1) 3969 aa
KMT2B (MLL2/MLL4) 2715 aa
KMT2C (MLL3) 4911 aa
KMT2D (MLL2/MLL4) 5537 aa
Nr of mutations
aa
Mutations:
missense
indel - inframe
indel - frameshift
nonsense
Domains:
CXXC
PHD
bromo
FYR
SET
AT hooks
C
Aminoacid in C4HC2C/H motif
C1
C2
C3
C4
H1
C5
C6
C7/H2
KMT2A
KMT2B
KMT2C
KMT2D
all KMT2
Frequency of missense mutation
D
Frequency of missense mutation
Full length protein
PHD domain
C4HC2C/H motif
rest of PHD domain
KMT2A
KMT2B
KMT2C
KMT2D