Published in final edited form as:

Biochemistry. 2018 February 06; 57(5): 525–539. doi:10.1021/acs.biochem.7b00705.

The Characteristics of a PHD Finger Subtype

Daniel Boamah[†], Tao Lin[†], Franchesca A. Poppinga[†], Shraddha Basu[†], Shahariar Rahman[†], Francisca Essel[†], and Suvabrata Chakravarty^{*,†,§}

[†]Chemistry & Biochemistry, South Dakota State University, Brookings, SD, USA, 57007.

[§]BioSNTR, Brookings, SD, USA, 57007.

Abstract

Although the *plant homeodomain* (PHD) finger superfamily is known for its site-specific *readouts* of histone tails, the origins of the mechanistic differences in histone H3 *readout* by different PHD subtypes remain less clear. We show that sequences containing the xCDxCDx motif in the PHD treble clef (xCDxCDx-PHD) constitute a distinct subtype, based on the following observations: (i) the amino acid composition of the binding site is strikingly different from other subtypes due to position-specific enrichment of negatively charged and bulky nonpolar residues, (ii) the binding site positions are mutually and positively correlated, and this correlation is absent in other subtypes, and (iii) there are only small structural deviations, despite low sequence similarity. The xCDxCDx-PHD constitutes ~20% of the PHD family, and the *double PHD fingers* (DPFs) are 10% of the total number of xCDxCDx-PHDs. This subtype originated early in the evolution of eukaryotes but has diversified within the metazoan lineage. Despite sequence diversification, the positions of the enriched nonpolar residues, in particular, show very small structural deviations, suggesting critical contributions of nonpolar residues in the binding mechanism of this subtype. Using mutagenesis, we probed the contributions of the binding-site positions enriched in nonpolar residues in four xCDxCDx-PHD proteins and found that they contribute to the tight packing of the H3 residues. This effect may potentially be exploited, as we observed affinity enhancement upon substituting a bulky nonpolar residue at the same binding site in another histone *reader*. Overall, we present a detailed characterization of PHD subtypes.

TOC GRAPHICS

^{*}**Corresponding Author.** suvabrata.chakravarty@sdstate.edu, Tel: 1-605-688-5694; Fax: +1-605-688-6364, Box-2202 SAV367, Department of Chemistry & Biochemistry, South Dakota State University, Brookings, SD, USA, 57007.

ASSOCIATED CONTENT

Supporting Information (SI)

The supporting information (SI) attached with the text contains: (i) Supplemental Figures S1 – S10, (ii) Table S1 and S2, and (iii) Supporting Results.

The authors declare no competing financial interest.

Notes

Any opinions, findings, and conclusions or recommendations expressed in this material are those of authors and do not necessarily reflect the views of the National Science Foundation and National Institutes of Health.

INTRODUCTION

Protein–protein interactions (PPIs) on the chromosomes decisively contribute to the interpretation of information encoded in the genome, and the tethering of proteins onto the intrinsically disordered tails of histones is a critical component of chromosomal PPIs. Regulatory protein complexes tethered to histone tails are known to orchestrate chromatin signaling pathways, which collectively regulate chromatin structure and thereby exert ultimate control over all the diverse biological outcomes associated with chromatin^{1, 2}. Eukaryotes have evolved peptide-anchoring scaffolds (known as *readers*³) to specifically recognize histone peptide segments^{4–7}. In general, the disordered peptide segments of histone tails provide the binding surface for the *readers* (present within the regulatory protein complexes) to facilitate the anchoring of regulatory complexes onto the chromatin. Over the past decade, fascinating discoveries about the mechanism of histone peptide recognition, revealed by the high-resolution structures of *readers* in complex with cognate histone peptides^{4–7}, have helped us to understand chromatin regulatory mechanisms in much greater detail. For example, the structures of histone recognition are providing important starting points to further probe detailed pathway-specific molecular mechanisms (e.g., reengineering *readers* to switch their specificity in order to probe and alter DNA methylation landscapes in living stem cells^{8, 9} or probing the pH dependent chromosomal association/dissociation of *readers*¹⁰).

Unlike *de novo* design, *reader* reengineering involves manipulation of a few binding-site residues of an existing *reader* while leaving the rest of the *reader* sequence unaltered^{8, 9}. Therefore, structure-guided dissection of the key determinants of histone peptide recognition mechanisms will continue to play an important role, not just in probing chromatin signaling mechanisms but also in the design of reagents for diagnostic applications^{11, 12}. To dissect the key determinants, we previously probed the energetics for recognition of the histone H3 N-terminal peptide by *readers* belonging to the same structural scaffold, the *plant homeodomain* (PHD) finger, that recognize the same peptide sequence¹³. In that study, we found that the *readout* of the same peptide sequence could differ between *readers*, and the energetic contributions of the peptide amino acids correlated with the sequence features of the *readers*. For example, H3 Lys4 makes negligibly small energetic contributions to N-terminal H3 recognition by PHD *readers* featuring a treble clef xCDxCDx motif¹³ (Figure S1A), suggesting that characteristic sequence features within large and diverse protein families have functional consequences. In general, as zinc finger folds depend predominantly on zinc coordination (typically by position-specific conservation of cysteine

and histidine residues), the rest of the sequence positions undergo mutations without disrupting the overall fold, leading to sequence diversification. Despite this sequence diversification, distinct sequence features in zinc fingers (e.g., RBR-RING1¹⁴ distinct from the canonical RING; the subtypes of zf-CxxC¹⁵) are associated with functional consequences. In the present study, we therefore further probed the characteristics of the xCDxCDx motif featuring PHD fingers (henceforth indicated as the xCDxCDx-PHD or PHD_nW_DD) to better understand the adaptations and functional consequence of the presence of a motif in a diverse superfamily.

The mechanism of N-terminal H3 recognition by the xCDxCDx-PHD was deduced much earlier in seminal reports on the structures of UHRF1^{16–22} PHD and the *double PHD fingers* (DPFs, i.e., PHD1–PHD2 tandem fingers) of the DPF3²³ and KAT6A²⁴ proteins. The PHD fingers of the UHRF1^{16–22} and KAT6A²⁴ proteins were recognized as histone H3-Arg2 (H3R2) *readers*, and the Asp residues of the treble clef xCDxCDx anchor the H3R2 residue. In addition, DPF3²³ and KAT6A²⁴ structures in complex with histone H3 also revealed that the PHD1 of these DPFs recognizes the histone H3K14ac mark, while the PHD2 of the DPFs is responsible for xCDxCDx-mediated H3R2 anchoring. Following the above reports, Dreveny et al.²⁵, showed, for the first time, that histone H3 residues 4–10 adopt a helical conformation when complexed with KAT6A²⁵, while the anchorage of the H3R2 and K14ac residues with helical H3 was found to be the same as in earlier reports^{23, 24}. Since then, DPF structural reports have not only shown a helical H3 conformation (e.g., DPF3²⁶ or MORF²⁷) but also revealed that DPFs (e.g., in KAT6A⁷, DPF2⁷) recognize crotonyl while others (e.g., in MORF²⁷) recognize acyl marks on the H3K14 residue. H3R2 anchorage by the xCDxCDx motif remains unaltered in all of these histone H3 *readout* mechanisms^{7, 25–27}. Two important additional observations have further contributed to the perceived versatility of the xCDxCDx-PHD: (i) they also include H3K4 *readers*, revealed by structural studies on BAZ2A²⁸, BAZ2B²⁸, KDM5B^{29, 30}, KDM5A³¹, and plant ATXR5³²; and (ii) a 3₁₀-helix *acidic wall*³³ (AW) in BAZ2A/BAZ2B, as elegantly shown by Bortoluzzi et al.³³, is responsible for the helical turn of H3 when complexing with BAZ2A³³ and BAZ2B³³ PHD fingers. The sequence feature of the AW consists of two consecutive negatively charged residues, and like the xCDxCDx motif, the AW is likely another negatively charged-residue sequence feature. In short, two negatively charged-residue patterns important for the peptide-binding function of PHD fingers were observed in the above structural studies. In a large and diverse superfamily, the co-occurrence of characteristic features in a sequence raises the possibility that the features are correlated, particularly if they evolved simultaneously for a common purpose. This possibility therefore calls for a detailed analysis of PHD sequences.

While recognized for utilizing its canonical binding site for different functions, the PHD finger module is also well recognized as a versatile *reader*^{34, 35}; for example, it is well known as both an H3K4me3^{36, 37} (modified histone) and an H3K4me0^{38, 39} (unmodified histone) *reader*. To indicate the presence or absence of a characteristic tryptophan (W) residue found in the H3K4me3 and H3K4me0 *reader* subtypes, respectively, the sequences of these subtypes were, for convenience, previously referred to as PHD_W and PHD_nW¹³. In all of the structures of xCDxCDx-PHD complexed with H3, no H3 residue other than

H3K14 carries any modification marks^{7, 16–22, 25–29, 32, 33}. In other words, xCDxCDx-PHD also recognizes unmodified H3 residues 1–13, with the canonical binding site accommodating the N-terminal portion of H3. The xCDxCDx-PHD also lacks the characteristic tryptophan (W), and for the Asp (D) residues of the xCDxCDx motif, the xCDxCDx-PHD sequences are alternatively referred to here as PHD_nW_DD for convenience. The recognition of unmodified H3 by the PHD_nW_DD subtype raises the question of whether the well-known H3K4me0 *reader* subtype PHD_nW is distinct from the PHD_nW_DD subtype. In the present analysis, we therefore addressed the following questions: (i) Is PHD_nW_DD (or xCDxCDx-PHD) a subtype distinct from other typical PHD finger subtypes? (ii) What are the sequence features of PHD_nW_DD? (iii) When did PHD_nW_DD originate and how is it distributed among the eukaryotes?

Our analysis suggests that PHD_nW_DD is a distinct subtype that originated early in the evolution of eukaryotes and that enriched site-specific nonpolar peptide contacts are notable sequence features of this subtype. Therefore, in addition to addressing the above questions, we also probed the energetic contributions of the enriched contacts by mutagenesis in four representative members of the PHD_nW_DD subtype (the PHD fingers of the proteins KDM5B, UHRF1, BAZ2A, and KAT6A) to validate the roles of the evolved specific contacts. Even though the structural mechanisms of H3 recognition by UHRF1^{16–22}, KAT6A^{7, 24, 25}, KDM5B^{29, 30}, and BAZ2A^{28, 33} have all been very well analyzed by point mutations (*see* Table S1), particularly by the replacement of Asp residues in the xCDxCDx motif, here we probed the contributions of enriched contacts (e.g., nonpolar contacts, Table S1). Based on computational analysis and follow-up experimental analysis of a set consisting of five *readers* (their 14 mutants and one H3 mutant), here we report on the detailed characterization of a *reader* subtype. As distinct sequence features in other zinc fingers (e.g., the RBR-RING1¹⁴ and zf-CxxC¹⁵ subtypes) are now emerging, our analysis will be useful in the functional characterization of subtypes of zinc finger domains and other protein families in general.

METHODS

The PFAM Alignment, Alignment Positions, and The PHD subtypes

10,858 PFAM-aligned sequences of PHD finger superfamily (PFAM⁴⁰ accession number PF00628.26) in STOCKHOLM 1.0 format were manually downloaded from the PFAM ftp server. This alignment was first *condensed* by eliminating the alignment positions having the symbol “.” in the original PFAM-alignment. The eliminated positions in this step have the symbol “.” greater than 85% of times in the alignment column. Sequences are extracted from the *condensed* alignment for creating a non-redundant set. Using CD-Hit⁴¹, redundant sequences at 85% sequence identity were eliminated. The final non-redundant set contained 3,469 sequences. A part of the *condensed* master PFAM-alignment of these 3,469 sequences is shown in Figure S2. In this master alignment, the correspondence between residues in the original PFAM alignment is not altered even though redundant sequences were eliminated. Without altering the master alignment, PHD sequences are further segregated into subtypes (Figure S2). A breakdown of these 3,469 sequences into subtypes: (i) PHD_W subtype (1,425 sequences), (ii) PHD_nW subtype (1,362), and (iii) PHD_nW_DD (692). As a

control, PHD_W subtype sequences were also segregated into: (a) PHD_W_DD (234 sequences) and (b) the remaining 1191 of the PHD_W sequences. The PHD_W_DD control subtype refers to PHD_W sequences having a pair of Asp residues in the treble-clef knuckle of the PHD scaffold. As all of these groups are extracted from the same *condensed* master PFAM-alignment, the alignment positions in each subtype remain the same (Figure S2), and therefore, enable comparison of position specific statistics among the subtypes. For example, Occurrence frequencies of residues at **L1** and **L2** positions are calculated based on the 85% non-redundant master PFAM alignment and the same alignment when partitioned into subtypes (Figure S2). Segregating the sequences into subtypes was carried out by simply extracting sequences that met the residue-type selection criterion for an alignment position, e.g., selecting sequences with residue of the 26th position = W in the PFAM alignment (Figure S2) resulted in PHD_W subtype. Similarly, selecting sequences with residues of the 24th position = D and that of the 26th position = W in the PFAM alignment (Figure S2) resulted in PHD_nW_DD subtype. For the convenience of discussion, we have named the alignment positions (columns) as **AW1**, **AW2** (i.e. *acidic wall 1/2*), **L1**, **L2**, **D1**, **D2**, **W/G**, **V** (Figure S1B).

Sequence Perturbation—The multiple sequence alignment (MSA) of a PFAM domain consisting of 85% non-redundant sequences is considered as the *master* MSA of the family (*see above*). We consider a position/column of a multiple sequence alignment (MSA) to be *perturbed* when we permit only a specific type of residues to appear in that column. For example, **W** position of PFAM PHD alignments (Figure S2) is perturbed when we permit only W, Y and F to appear at the **W** position. The set of sequences that meets the condition is considered here to be *perturbed*. Thus, PHD_W is considered to consist of *perturbed* sequences. Mentioned above, the *perturbed* sequences retain the same alignment as that of the *master* MSA. The occurrence frequency of amino acids in a column of the *master* alignment is referred to as *probability* while that of a *perturbed* alignment as *conditional probability*, i.e. a condition is imposed on the observed probability. The *conditional probability* of an amino acid in a column is compared with its *probability* in the corresponding column as, $p = (\text{conditional probability}) - (\text{probability})$ (*see* Figure 4B–D, Figure S3). For p , a positive value is considered as *positive enrichment* ($\uparrow$) and a negative value is considered as *negative enrichment* ($\downarrow$) (Figure 4B–D, Figure S3). For convenience, p is calculated for the most frequent amino acid in a column. To avoid confusion, **W** and **G** positions refer to the same position, and a *positive enrichment* of W is always associated with *negative enrichment* of G or *vice versa*. As PHD_nW_DD do not have W in the **W**-position, but a G, we use both symbols in the discussion.

Residue Occurrence Frequency, Sequence Identity, and Sequence Logos

The *probability* and the *conditional probability* are calculated as the frequency of occurrence (p_{ji}) of the j th amino acid (or residue type) in the i th alignment position (column), i.e., p_{ji} is the ratio between the number of times the j th residue type appear in the i th alignment position and the total number of sequences in the alignment (e.g., subtype alignment). The sequence logo (e.g., Figure 1, 2, and 3) was created using the online version of the Skyline program⁴² by uploading PHD subtype specific alignments created in Figure S2. The relative entropy, X-axis of sequence logo (Figure 1, 2, and 3), is the default one returned by the

Skyalign program⁴². For the sequence identity distribution (Figure 5B), sequence identity between the sequences, i and j , was calculated by comparing residue identity between the aligned position of *master* PFAM alignment. The observed median of the sequence identity distributions (Figure 5B, dashed line) of a subtype (e.g., PHD_nW_DD) was used to set the sequence identity cutoff (40%) for structural analysis. Sequence identity cutoff (e.g., 30% or 40%) by blastclust⁴³ is also used for clustering PHD finger sequences.

Other Alignments

In addition to PFAM alignments, two other alignments (MUSTANG⁴⁴ and MUSCLE⁴⁵) were also used here. Structural alignments (Figure S1B, S10A) and structural superposition (e.g., Figure 3, Figure 7D) were generated using MUSTANG⁴⁴. MUSCLE⁴⁵ sequence alignment was used for aligning the DPF sequences (Figure S1A) with the remaining PHD_nW_DD sequences. The sequences of the DPFs were first compiled as described earlier¹³, and the sequences with 85% identical residues were eliminated after MUSCLE alignment. However, the PFAM alignment of the PHD superfamily is the key alignment used here for computing the position specific frequencies of amino acids for subtype characterization (see above).

ASA_{SC}, Packing Density, and Secondary Structure—NACCESS⁴⁶ program was used to compute the accessible surface area (ASA). The amount of surface area lost by the sidechain atoms of a protein residue upon formation of a peptide–protein complex is measured by ASA_{SC}. ASA_{SC} of the i th residue is, $ASA_{SC_res_i} = (\text{residue_i_}ASA_{SC_protein}) - (\text{residue_i_}ASA_{SC_complex})$ where residue_i_ASA_{SC}_protein and residue_i_ASA_{SC}_complex are respectively the sidechain atom ASA of the i th residue in the free and bound form of the protein. Protein residues with $ASA_{SC} > 10\text{\AA}^2$ are considered to make contact with the peptide (Figure 1B). We used the Voronoia program⁴⁷ to compute packing density of atoms as described in our earlier study¹³. A set of 15 high-resolution (2.0Å) PHD–peptide complexes belonging to different subtypes (Figure S10A) and having a low sequence similarity with one another (sequence identity < 40%) were chosen. We refer to the structures of this set for the packing density (PD) discussion of **L1**, **L2** and **V** positions. Assignment of the secondary structural elements (SSE) of coordinate files was carried out using the SEC-STR program⁴⁸, and the PISCES⁴⁹ high-resolution non-redundant set of 3699 single chain structures was used for estimating amino acid occurrence frequencies within a selected SSE pattern.

Taxonomic Affiliations—UniProt accession numbers in the 85% non-redundant PFAM alignment master alignment were used to retrieve the text file for each protein ID from the UniProt database. The ‘OC’ symbol in the first column of this text file provided the taxonomic information of a protein.

Peptide Swapping and Peptide Backbone Clash—Using the MUSTANG⁴⁴ structural alignment (Figure S10A), the coordinate set of one PHD–peptide complex was superposed onto another PHD–peptide complex as described as *peptide swapping* in our earlier study¹³. For convenience, we use one PHD–peptide complex as the *reference complex*, and the remaining complexes are superposed onto the *reference complex*. Only

structures with resolution $2.0\text{\AA}$ were used. This is a simple but a crude approach, as it does not optimize backbone hydrogen bonding between the peptide and the protein. The extended conformations of peptides in the PHD_nW subtype (2puy, 4gne, 4qbs and 3qla) were placed onto the PHD_nW_DD subtype (3asl, 4q6f, 5b75 and 5i3l). PHD-peptide inter-molecular backbone carbon-carbon contact distances for H3 residues 4–7 (Figure 8A) were determined using LIGPLOT⁵⁰. As a positive control, peptides in the PHD_nW subtype (2puy, 4gne, 4qbs and 3qla) were swapped among themselves for a rough estimate of the error (~16% in Figure 8A) contributed by the crude swapping approach. As PHD sequences of these complexes share < 40% sequence identity, comparison of results obtained by swapping between subtypes (e.g., PHD_nW and PHD_nW_DD) and those within a subtype (e.g., PHD_nW) was acceptable.

DNA constructs and Mutagenesis—The recombinant GST fusion hKDM5B PHD (residues 305–366 of UniProt ID Q9UGL1/KDM5B_Human), hKDM5B PHD D331A mutant and hAIRE PHD (residues of 294–343 UniProt ID O43918/AIRE_HUMAN) used here was taken from our earlier study¹³. Recombinant His-tag fusion of hUHRF1 PHD (residues 299–367 UniProt ID Q96T88/UHRF1_HUMAN) and hKAT6A DPF (residues 196–319 UniProt ID Q92794/KAT6A_HUMAN) were cloned into a modified pET28A vector using *Bam*HI and *Xho*I restriction sites. The recombinant BAZ2A PHD finger (residues 1673–1728 UniProt ID Q9UIF9/BAZ2A_HUMAN) used was cloned into pETite N-His SUMO Kan vector. PCR (polymerase chain reaction) amplification of synthetic DNA representing the sequences of selected *readers* was first carried out for creating the desired constructs. Synthetic DNA molecules were obtained from GeneScript. Mutants of hKDM5B PHD (L325A, L326A, and V346A), hUHRF1 PHD (L344A, M345A, D350A, and V365A), hBAZ2A PHD (L1692A, L1693A, D1695A, D1698A, and V1713A), hKAT6A DPF (L279A, F280A, and D285A) and hAIRE PHD (C310L) were created using site directed mutagenesis.

Expression and purification of proteins—Recombinant hKDM5B PHD and its mutants (L325A, L326A, D331A and V346A), recombinant hAIRE PHD and its mutant (C310L), were purified as described in Chakravarty *et al.*¹³ using the pGEX-4T3 vector constructs. Briefly, the fusion proteins were isolated using GST affinity chromatography (Pierce glutathione agarose, Thermo scientific). GST tag was eliminated by overnight human alpha thrombin (Haematologic Technologies) digestion at 4°C prior to gel filtration chromatography. Recombinant hUHRF1 PHD and its mutants (L344A, M345A, D350A and V365A), hBAZ2A and its mutants (L1692A, L1693A, D1695A, D1698A, and V1713A), hKAT6A DPF and its mutants (L279A, F280A, and D285A) were all purified following the His-tag affinity chromatography step, and the further purified by gel filtration chromatography.

Synthetic Peptides and Isothermal Titration Calorimetry (ITC)—We used three synthetic peptides for this study. They are: (i) Histone H3 residues 1–11 (H3-1-11W), (ii) Histone H3T3S residues 1–11 (H3T3S-1-11W), and (iii) Histone H3 residue 1–16 with acetylation at K14 (H3K14ac-1-16W). The W in the synthetic peptide sequence represents C-terminal Tryptophan residue, used here for determining molar extinction coefficient⁵¹.

H3-1-11W was used for the binding studies of the wild type and the mutants of hKDM5B, hUHRF1, hBAZ2A, and hAIRE PHD finger proteins. H3K14ac-1-16W peptide was used for the binding studies of the wild type and the mutants of hKAT6A DPF. The H3T3S-1-11W peptide was used in the binding studies of hKDM5B and hUHRF1 PHD finger wild type proteins. 98% pure (confirmed with mass-spec) synthetic peptides were obtained from GeneScript. These peptides were used for ITC (Isothermal Titration Calorimetry) experiments. The thermodynamics of the interactions between the histone H3 peptide and all the *readers* were probed by a single method (ITC), and all the ITC titration experiments were carried out under identical experimental conditions, buffer (25 mM phosphate, 150 mM sodium chloride at pH 7.6). The thermodynamics of binding between a protein and the synthetic peptides were studied using MicroCal ITC200 (GE Healthcare) with protein (0.05 – 0.15 mM) and peptide (10 fold higher) respectively loaded in the cell and syringe at 25°C. In this study, we utilize the values of K_d or K_a (i.e., ΔG) for comparing a mutant to its wildtype, as we are cautious that interpretation of ΔH when K_d is higher (weakened affinity) for low protein concentration can be less accurate⁵² if “c” values ($[\text{protein concentration}]/K_d$) are close to or lower than its limiting range where interpretation of ΔG is still reliable⁵². Despite its report in our earlier study¹³, KDM5B D331A (*see* Table S1) is included because the buffer conditions (150 mM sodium chloride) used here is different from that used earlier (250 mM sodium chloride).

RESULTS

Binding-site Residue Frequencies

Although diverse sequences populate this family, based on the canonical binding-site, PHD finger sequences can be grouped into two major functional subtypes^{34, 35}: (i) the PHD_W subtype and (ii) the PHD_nW subtype (Figure S2). The justification for this grouping is that functionally related positions show enrichment in specific residue types among the subtype sequences; for example, aromatic cage-forming positions are enriched with aromatic residues in the PHD_W subtype¹³. Therefore, to probe the characteristic features of the PHD_nW_DD (xCDxCDx-PHD) subtype, we relied on enrichment analysis of residue types at specific alignment positions, particularly for positions **L1**, **L2**, **D1**, **D2**, **AW1**, **AW2**, and **V** (*see* Figure S1B). In the PHD_nW_DD subtype, these positions undergo a change in accessible surface area (ASA) of sidechain atoms upon peptide binding, with $\Delta ASA_{SC} \approx 10 \text{ \AA}^2$ (Figure S1B). The enrichment analysis in the following was performed on a set of PHD sequences with an Asp residue at the **D2** position while lacking a Trp residue at the **W** position (i.e., PHD_nW_DD).

The nonpolar L2 position—The overall result for occurrence frequency and residue enrichment at the **L1** and **L2** positions of the PHD superfamily is represented as sequence logos in Figure 1. It is clear that bulky nonpolar residues (Ile, Val, and Leu) are always present at the **L1** position irrespective of the subtype; that is, $p = 0.95$ is observed for the entire superfamily (Figure 1A–D). However, this is not the case for the **L2** position (Figure 1A–D). The occurrence frequency of bulky nonpolar residues (Leu, Phe, and Ile) at the **L2** position is 0.32 for the family, and its values in the PHD_W, PHD_nW, and PHD_nW_DD subtypes are 0.05, 0.36, and 0.82, respectively (Figure 1A–D). In other words, the preference

for bulky nonpolar residues at **L2** varies substantially from one subtype to the other, and thus a preference for a bulky nonpolar residue at **L2** seems to be a key feature of the PHD_nW_DD subtype. The increase in occurrence frequency of bulky nonpolar residues for this subtype at **L2** (0.32 $\rightarrow$ 0.82) relative to the average for the PHD superfamily is considered positive enrichment ($p > 0$). Similarly, a drop in the occurrence frequency is considered negative enrichment ($p < 0$); for example, the PHD_W subtype showed negative enrichment at **L2** (0.32 $\rightarrow$ 0.05). The most frequent residue at the **L2** position in the PHD_W, PHD_nW, and PHD_nW_DD subtypes is Gln ($p = 0.32$), Cys ($p = 0.26$), and Leu ($p = 0.60$), respectively. As a control, we also checked the xCDxCDx motif-containing sequences (e.g., TAF3 PHD) in the PHD_W subtype, and we observed that they never include bulky nonpolar residues ($p = 0.03$) at **L2** (Figure 1C, right). This suggests that PHD_nW_DD is likely to be a distinct subtype, and in addition to having the pair of Asp residues in a treble clef knuckle (xCDxCDx) (Figure S1), this subtype likely has additional features. We identified all such features associated with the xCDxCDx motif (see below).

The acidic wall (AW) residues—The AW positions (**AW1** and **AW2**) in the PHD_nW_DD subtype also showed positive enrichment for negatively charged residues, and the frequencies of Asp/Glu residues at these positions in this same subtype were observed to be high, with $p(D,E) > 0.50$ (Figure 2). In other subtypes, the occurrence frequency of Asp/Glu residues at the **AW1** and **AW2** positions were observed to be significantly lower than that of PHD_nW_DD ($p(D,E) \approx 0.3$), especially PHD_nW ($p(D,E) < 0.2$). Another notable difference for the AW positions of the PHD_nW_DD subtype is that the occurrence frequency of Gly residues at these positions is negligibly low ($p(G) \approx 0.09$). Gly is usually the most frequent residue at the AW positions in other subtypes, and therefore the value of $p(G)$ at these positions in other sequences is higher (Figure 2). In other words, the preferences for Asp/Glu residues at the AW positions in the PHD_nW_DD subtype likely are in exchange for a decreased preference for Gly residues.

Other nonpolar positions and structural variations—Side chains at the **V** position are physically in contact with the histone H3 peptide, while those in the **P** position are in close proximity to the peptide ligand. An enrichment of nonpolar residues at the **V** and the **P** positions in the PHD_nW_DD subtype (Figure 3A, B) is also consistent with the preferred nonpolar contact with the peptide (see below). The occurrence frequency of nonpolar residues at the **V** and the **P** positions in other sequences are low (Figure 3C, D). The segment harboring the **V** and the **P** positions in the PHD superfamily, however, shows a large structural deviation due to insertions/deletions (Figure 3B–D). This can be observed, for example, in PHD_W structures, in which this segment has helices of varying lengths (Figure 3D). For the PHD_nW_DD sequences, even those with a sequence identity $< 30\%$ (e.g., BAZ2A, UHRF1, and DPF3-PHD2), the structural deviations for this segment are negligibly small (Figure 3A) in comparison with other subtypes (Figure 3C, D). In the PHD_W and PHD_nW structures, for sequences with sequence-identity $< 40\%$ this region typically shows large structural deviations (Figure 3C, D). For the observed small structural deviation, we anticipate that the structure of this segment, which is enriched in proline residues, is a characteristic feature of the PHD_nW_DD subtype and typically lacks insertions/deletions (Figure 3A). A lack of structural deviation suggests that the quality of

the alignment for this segment in the PHD_nW_DD subtype is more accurate and thus more reliable than that of the other subtypes. Therefore, the estimate of enrichment for nonpolar residues ($p(V,I) = 0.64$) at the **V** position is likely to be accurate in the PHD_nW_DD subtype. At the same time, the estimate at the **V** position for this region in the PHD_W subtype ($p(V,I) = 0.23$) is likely to be less reliable.

Subtype Status for PHD_nW_DD

The overall result of the enrichment analysis of binding-site residues upon perturbation of the PHD finger master alignment at the **D2** position is in Figure 4. Although the PHD_nW and PHD_nW_DD sequences recognize the unmodified N-terminal histone H3, the amino acid compositions at the PHD finger canonical binding site are strikingly different (Figure 4A). The compositional differences are not only due to the presence of an Asp residue at the **D2** position but also to the presence of specific amino acids at a number of other positions. This suggests specific adaptations for function (*see below*), and PHD_nW_DD sequences could thus be distinct from those of the PHD_nW or PHD_W subtypes. Using enrichment analysis, we further probed the correlations between the binding-site positions (**L2**, **D2**, **AW1**, **AW2**, and **V**) as a justification for considering PHD_nW_DD as a distinct subtype. In all the discussions above, the observed enrichments were with a single perturbing condition, namely, restricting the **D2** position to have an Asp residue. Here we checked whether the **D2** position would similarly show *positive enrichment* ($p > 0$) if other binding-site positions were perturbed (*reciprocal enrichment*, Figure 4B). A reciprocal enrichment would suggest correlations between the positions, which would indicate that the types of amino acids appearing at the correlating positions were selected for a common function (*see below*). In the reciprocal enrichment analysis, Asp was enriched at the **D2** position when other peptide-anchoring positions (e.g., **L2**, **V**, or **AW1**) were perturbed (Figure 4B). Figure S3 is an example of detailed perturbation analysis. Like the **D2** positions, other peptide-anchoring positions were also checked to see whether there is a mutual relationship among the set of binding-site positions (Figures 4C). All of the binding-site positions (**L2**, **D2**, **AW1**, **AW2**, and **V**) show reciprocal enrichment with each other (Figure 4C). The observed correlations suggest that the PHD_nW_DD subtype binding-site residues likely coevolved for a common purpose (function).

In the reciprocal enrichment analysis, the following two positions were included as controls (Figure 4D): (i) the **W** position, the characteristic feature of the PHD_W subtype, and (ii) the **D1** position, as an Asp residue at the **D1** position is present in the majority of PHD sequences; that is, $p(D,E) \sim 0.7$ in the PHD superfamily. In the controls, we observed that a positive enrichment ($p > 0$) at the selected positions in the PHD_nW_DD subtype is correlated with a negative enrichment ($p < 0$) at the **W** position, suggesting a lack of correlation between these positions in the PHD_W subtype. In other words, the results indicate that the features characteristic of the PHD_nW_DD subtype were not selected for in the PHD_nW subtype. The enrichment (p) observed at the **D1** position, when perturbed at the selected positions, was relatively small compared with the other perturbations (Figure 4B–C). This observation is also consistent with the notion that a family-specific trait, unlike a subtype-specific trait, is less likely to undergo a change upon sequence perturbation. Therefore, based on the following observed criteria, we consider PHD_nW_DD to be a

distinct subtype because of: (i) binding-site compositional differences (Figure 4A), (ii) the mutual relationship between the binding-site residues in PHD_nW_DD (Figure 4B–C), (iii) a lack of correlation with other subtypes for the binding-site positions (Figure 4D), and (iv) a small structural deviation despite low sequence identity (Figure 3A).

PHD Subtypes and Taxonomic Affiliations

In general, in the 85% non-redundant PFAM alignment (Figure S2), the proportions of the PHD subtype sequences are: PHD_W (41.10%), PHD_nW (38.95%), and PHD_nW_DD (19.95%), while the DPFs constitute 10% of the PHD_nW_DD subtype (Figure 5A). In the 85% non-redundant PFAM dataset, the distribution of sequence identity between any two PHD finger sequences (Figure 5B) suggests that PHD_nW_DD sequences, with a median sequence identity of 38.78%, are diverse. As the sequences are grouped by the binding-site residue preferences, the procedure of subtyping is not biased towards grouping highly similar sequences. The median of the PHD_nW_DD subtype is, however, higher than that of the PHD superfamily (27.08%) or the PHD_nW (30.77%) or PHD_W (29.41%) subtypes (Figure 5B). The enriched residues at several binding-site positions in a small protein and the lack of insertions/deletions likely contributes to the observed higher value of the median sequence identity for the PHD_nW_DD subtype.

We then checked the taxonomic affiliations of the PHD finger sequences present in the 85% non-redundant PFAM alignment to better understand the origins of the PHD finger subtypes (Figure 5C). At atomic resolution, details of the recognition mechanisms of histone peptides by *readers* have come predominantly from the structures of human proteins (i.e., mammals, chordates, metazoans), and the taxonomic affiliations of the ancestors of human proteins provide a view of the origins of such mechanisms. All the PHD subtype sequences were observed to be present earlier than the metazoan lineage, as we found their presence in fungal, plants, stramenopiles, and other lineages (Figure 5C). These findings indicate that the subtypes likely existed early in the eukaryotic tree of life (Figure 5C). However, there are differences in the proportions of these lineages among the subtypes. For example, the fungal and metazoan lineages dominate the sequence space of the PHD_W subtype, while the affiliations for the other subtypes are quite different (Figure 5C). The majority of the PHD_nW_DD and PHD_nW sequences belong to the metazoan lineage (Figure 5C). The proportions are even more striking for DPFs, as 80% of DPF sequences are found in the metazoan lineage (Figure 5C). This suggests that the PHD_nW_DD and PHD_nW subtype sequences expanded in the metazoan lineage, and thus it is tempting to believe that the readout of unmodified N-terminal histone H3 by either mechanism (PHD_nW_DD or PHD_nW) has contributed in the evolution of metazoans. To better understand how the enriched residues, particularly in the metazoan lineage, enhance peptide binding, we probed the contributions of these positions.

Enriched Residue Contributions

In the following section, we probed the energetic contributions of the enriched residues at the binding-site position of four PHD_nW_DD fingers (hKDM5B, hUHRF1, hBAZ2A, and hKAT6A) by mutagenesis, that is, by replacing the wild type residue with alanine. The PHD fingers of hKDM5B, hUHRF1, and hKAT6A were chosen, as they share less than 30%

sequence identity (Figure S1B). Isothermal titration calorimetry (ITC) binding thermodynamics was used to probe the effect of the mutations, and the thermodynamic parameters of the ITC experiments are in Table S2 (SI page S13).

The nonpolar L2 position—The bulky nonpolar residues at **L2** in the hKDM5B PHD (L326A), hUHRF1 PHD (M345A), hBAZ2A PHD (L1693A), and hKAT6A DPF (F280A) were replaced with Ala (Table S1, S2; SI pages S13–S14). The **L2** position mutants of the hUHRF1 PHD, hKDM5B PHD, hBAZ2A PHD, and hKAT6A DPF did not show any detectable H3 peptide binding according to ITC (Figure 6A–C, S4A), suggesting that substitution of the bulky nonpolar residues completely disrupted peptide binding. This suggests the importance of the energetic contributions of bulky nonpolar residues at the **L2** position in the PHD_{nW_DD} subtype. The energetic effect of **L2** substitution in the hUHRF1, hBAZ2A, and hKAT6A domains has not been previously reported (Table S1). The effect of the **L2** position substitution (L326A) in the KDM5B PHD reported earlier²⁹ was a ~2-fold drop in affinity. We expect the fold drop in affinity for the **L2** substitution of KDM5B to be higher than that as we note a complete loss of peptide binding upon **L2** substitution. In the study of PHD finger complexes, the disruption of peptide binding by introducing a Trp residue at the L2 position^{38, 53} (e.g., L326W in the KDM5B PHD³⁰) has also been reported earlier. Although the **L2**→W substitution is important for demonstrating contact between the **L2** position and the peptide^{30, 38, 53}, the binding disruption in **L2**→W is likely to be the result of overpacking of the interface with the bulky Trp residue. The **L2**→A substitution further captures the contribution of the loss of packing in a tightly packed *reader*–peptide interface. As a control, the highly conserved **L1** position (conserved among all PHD sequences) was also subjected to alanine mutagenesis in these four proteins. Consistent with this conservation, **L1** replacements also completely disrupted peptide binding in all four proteins (Figure 6A–C). The energetic effect on peptide binding by **L1** mutation is discussed separately (see SI, page S15 – S16). As **L2** is highly conserved among PHD_{nW_DD} sequences, as is **L1** among all PHD sequences, the complete disruption of peptide binding is consistent with the conservation pattern, and we expect the ΔG for the **L2** position to be very large (>1.5 kcal/mol).

The nonpolar V position—The bulky nonpolar residues at the **V** position in the hKDM5B PHD (V346A) and hUHRF1 PHD (V365A) were similarly replaced with Ala. Mutation at the **V** position weakened binding significantly (Figure 6D, S4B). However, unlike the **L2** position, mutations at the **V** position did not disrupt H3 peptide binding completely (Figure 6A–C). We therefore expect the energetic contribution for the **V** mutations in the PHD_{nW_DD} subtype to be lower than that of the **L2** position but be large enough to cause significant weakening of peptide binding. The observed weakened binding upon Ala substitution at the **V** position suggested that the nonpolar contacts made with the side chains of residues at this position likely play an important role. It was also interesting to see that a position with higher conservation (**L2**, $p = 0.82$) had a larger energetic contribution than one with a lower level of conservation (**V**, $p = 0.64$) in the PHD_{nW_DD} subtype. To probe the van der Waals' contact between the side chain of the **V** position and the CH₃– group of the H3 peptide Thr3 residue (H3-Thr3), we replaced this group with Ser.

The H3(Thr→Ser)₃ substitution significantly affected peptide binding (Figure 6E), and this is also consistent with the contribution of the nonpolar contacts with the **V** position.

The xCDxCDx acidic residues—As the effects of the substitution (Asp→Ala) at the **D1** position in the hUHRF1^{16–19}, hKDM5B^{13, 29, 30}, and hKAT6A²⁴ have thoroughly been analysed by earlier studies, here we checked the effect of the Asp→Ala substitution at **D1** only for the BAZ2A (D1695A) domain (Figure 7C). This substitution severely compromised binding, consistent with the observed effects of **D1** substitution in the hUHRF1^{16–19}, hKAT6A²⁴, and hKDM5B proteins^{13, 29, 30}. Due to the substitution of the Asp residues at the **D2** position in UHRF1 and KDM5B, the respective affinities were lowered by ~9 fold (Figure 7A–B). The 9-fold drop for the **D2** mutant in UHRF1 is consistent with that reported by Lallous et al.¹⁸ and is within the range 3–20 fold reported in other studies^{16, 17, 19}. For the **D2** substitution in the BAZ2A PHD, the affinity was lowered by ~3 fold (Figure 7C). It is possible that the structural deviation of the treble clef knuckle in the BAZ2A PHD may contribute to the observed smaller drop in the affinity (Figure 7D) for the **D2** substitution in BAZ2A. For the KAT6A domain, we expect the energetic contribution of **D2** to be as large as we observed for complete disruption of binding upon **D2** substitution (Figure S4A). In short, the **D2** position also made a substantial contribution to anchoring the H3R2 residue in the PHD_nW_DD subtype. This finding also justifies PHD finger sequence grouping by the treble clef knuckle xCDxCDx pattern for subtype-specific analysis, in which the Asp residues are expected to serve as binding hotspots. Among the PHD_nW_DD subtype binding-site positions examined for enrichment analysis, the **D2** position had the largest structural deviation (1.58 ± 0.57 Å, Figure S1B), and this is due to the structural deviation of the xCDxCDx segment, like that of the BAZ2A (Figure 7D). Due to this deviation, the ASA of **D2** in BAZ2A is 0.0 Å², while the other proteins have a ASA value of 17.14 ± 2.55 Å² for the **D2** position. The BAZ2A structural deviation places D1698 farther away from the H3Arg2 residue (Figure 7D). The larger energetic contribution of the **D1** position in comparison with the **D2** position is also consistent with the larger ASA (33.09 ± 2.28 Å²) for the **D1** position (Figure S1B).

The AW acidic residues—Recently, the consequences of the replacement of negatively charged residues at the **AW** positions in the BAZ2A (D1688, E1689) and BAZ2B (E1943, E1944) domains were thoroughly discussed by Bortoluzzi et al.³³. We have therefore focused on sequence analysis of the acidic wall. Using a small structural dataset, early conformational analysis⁵⁴ showed that the occurrence frequency of Asp residues within the first three positions of the ₃₁₀-helix or turns are higher than for other amino acids. ₃₁₀-helical conformations are commonly observed as N- and C-terminal extensions of regular α -helices, in which the backbone amide-NH of the *i*th residue forms a hydrogen bond with the backbone carbonyl >C=O of the (*i*-3)th residue within the ₃₁₀-helical segment. In all the structures (< 2.0 Å) of the PHD_nW_DD subtype, we observed at least one “(*i*-3) ← *i*” backbone hydrogen bond involving the **AW** region (Figure S5). However, the **AW** ₃₁₀-helix neither precedes nor follows an α -helix and is retained within a secondary structural pattern, **llhhhee** (l, loop; h, ₃₁₀-helix; e, strand; Figure S5). In a non-redundant ($< 25\%$ sequence identity) set of 3699 structures (< 1.6 Å) we observed 340 segments featuring this pattern, and the occurrence frequency of amino acids at the positions of each of the h-symbols is: **h₁**

[p(P+D+E+S) = 0.42], **h₂** [p(D+E+S+N) = 0.60], and **h₃** [p(D+N) = 0.32] (Figure S5). Although no single amino acid emerged as having a clear predominance at any of the h-positions, negatively charged residues along with proline and non-charged polar residues were observed to be more frequent in such ₃₁₀-helical segments. In the plant ATXR5 protein, the **AW1** and **AW2** positions contain proline and serine, respectively; that is, ATXR5 lacks negatively charged residues within the ₃₁₀-helix (Figure S1B). Similarly, the UHRF1 **AW1** position also features a proline residue (Figure S1B). For occurrence frequency estimation, we therefore accounted for the presence of proline and uncharged polar serine and asparagine residues. Interestingly, the occurrence frequencies for **l₂** [p(P+D+S+N) = 0.52], **e₁** [p(I+V+L+F) = 0.69], and **e₂** [p(I+V+L+F) = 0.41] (Figure S5) also suggested preferences for amino acids at these positions. In the PHD_nW_DD subtype, the ₃₁₀-helical segment precedes a strand rich in bulky nonpolar leucine residues (Figure S5), and the observed preferences for the **e₁** and **e₂** positions suggested that the **llhhhee** segments are similar in amino acid composition to the PHD_nW_DD subtype, and the amino acid composition of these segments likely contributes to the structure of the **AW** segment. In the **llhhhee** structural pattern, the sidechains of Asp (D), Glu (E), Ser (S), and Asn (N) were, however, not observed to participate in hydrogen bonds with the polypeptide backbone. Similarly, the **llhhhee** structural pattern lacking Asp, Glu, Pro, Ser, and Asn residues was also observed, suggesting that further study of sequence–structure relationships⁵⁵ will be needed to understand whether the structure of the **AW** region is due to a local effect of the sequence alone or has contributions from distal regions of the protein.

The sequences of other subtypes prefer Gly for the corresponding segment (*see above*), and the conformation of this segment in other subtypes is distinctly different from a ₃₁₀-helix (Figure 8A). As discussed above, the preference for Gly at the **AW1** and **AW2** positions in the PHD_nW_DD subtype is low. Interestingly, within the **llhhhee** structural segments the occurrence frequency of Gly for **h₁₋₃** is also low, [p(G) = 0.05]. The ₃₁₀-helical conformation in the PHD_nW_DD subtype has consequences for the peptide conformation. In general, the cognate peptide in the binding sites of the PHD_W and PHD_nW subtypes is typically accommodated by β -augmentation. However, the strand conformation, beyond residue number 4 of H3, will likely sterically interfere with the ₃₁₀-helical conformation of the **AW** segment (Figure 8A). For an estimate of this interference, we take peptides (in the β -augmented conformation) from the PHD_nW and PHD_W complexes to insert at the PHD_nW_DD binding site (peptide swapping). The distribution of the interfacial contact distances in the swapped complexes of the PHD_nW_DD subtype shows that more than 40% of the interfacial contact distances are ≥ 3.0 Å (Figure 8A). As a control, peptides swapped between PHD_nW complexes (sharing low sequence identity) have less than 16% of the interfacial distances as ≥ 3.0 Å. Therefore, it is likely that the H3 peptide adopts a helical conformation in order to evade steric interference with the **AW** segment. Histone H3 thus adopts a context-dependent conformation consistent with earlier structural observations (Figure S6). A few sequence features of H3 likely contribute to adopting a helical conformation. They are: (a) the N-cap H3T3-OH⁵⁶, (b) H3G11, a helix-terminating Gly⁵⁷, and (c) the K9 positive charge for accommodating the helix dipole⁵⁸ (Figure 8C). H3 is likely to be structurally malleable due to these factors and thus is capable of *one-to-many* interactions⁵⁹, like other network hubs of disordered segments. The N-cap H3T3-OH

hydrogen bonds with the H3 backbone are observed in the peptide complexes of BAZ2A³³, KAT6A^{7, 25}, DPF3²⁶, MORF²⁷ and UHRF1²⁰. Tallant et al.²⁸ had earlier shown that peptide binding in BAZ2A and BAZ2B PHD fingers is disrupted upon the introduction of the H3T3ph mark²⁸. As suggested by Bortoluzzi et al.³³, this disruption is likely due to the interference of the H3T3ph mark with the N-cap role of H3T3-OH in impeding H3 helix initiation. It is also interesting that the proportions of the negatively charged residues at the **AW1** and **AW2** positions among the PHD_nW_DD subtype sequences increases from ~40% in non-metazoan lineages to >60% in the metazoan lineage (Figure S7A). By contrast, the amino acid frequencies at other binding-site positions (in PHD_nW_DD as well as other subtypes) do not show this behavior (Figure S7). Although the binding mechanisms existed early in the evolution of eukaryotes, the contributions of the **AW1** and **AW2** positions towards peptide binding are likely to be more prominent among the metazoans.

DISCUSSION

The finding that there are patterns of negatively charged residues (xCDxCDx and **AW**) at the PHD finger binding site prompted an investigation to determine the relationship between these patterns. Here we report that not only is there a correlation between these patterns, but these negatively charged residues also correlate with nonpolar binding-site residues. The mutually correlating pattern of residues in PHD fingers had not been previously discussed, and therefore we performed a detailed analysis to determine whether the distinct composition of the correlated binding-site positions constitutes a subtype that developed early in evolution. As the site-specific histone *readers* are referred to as *Lys/Arg-readers*, negatively charged residues of *readers* generally receive more attention in the characterization of these proteins (*see* Table S1). The observed enriched nonpolar residues in the readout of highly charged histone tails therefore also encouraged us to explore the nature of the adaptation. For example, why should there be a need for selecting bulky nonpolar residues at the **L2** and **V** positions in a subtype-specific manner? We attempt to explain these adaptations in the following discussion.

Anchorage of the peptide side chain H3Lys4 quaternary ammonium ion ($-\text{CH}_2\text{N}^+(\text{CH}_3)_3$) by the aromatic cage likely dominates the peptide-binding energetics in the PHD_W subtype (Figure S8C), while salt bridge-mediated H3Lys4 likely dominates the binding energetics in the PHD_nW subtype (Figure S8C). For the PHD_nW_DD subtype, we previously found that peptide binding might not require an energetic contribution from the H3Lys4 residue¹³. The lack of energetic contribution from this residue must be compensated by additional contacts, and **L2** likely provides the compensating contact. For example, in the BAZ2A **L2** position, the CD1 and CG atoms of L1693 (**L2**) are in van der Waals contact with the histone H3 Thr3 backbone C atom (Figure 9A), and these contacts are maintained in a well-packed environment. For example, the packing density of the L1693-CD1 and -CG atoms in the BAZ2A complex were found to be higher than typically observed for other Leu CD1 and CG atoms (Figure 9A). In general, in the PHD_nW_DD subtype, the **L2** residue side chain atoms contribute to tight packing of the peptide backbone atoms, which is in contrast to the contacts made by the **L2** residues in other subtypes. In the PHD_W subtype, **L2** residues, with smaller side chains (e.g., Gly, *see* Figure 9B) are typically not in contact with H3 peptide backbone atoms (Figure 9B). Based on the side chain size, the PHD_nW subtype

members may or may not retain contacts with the peptide backbone (Figure S8B). Thus, the **L2** contribution is likely to be an essential feature of peptide binding in the PHD_nW_DD subtype, as we found that substitutions of bulky **L2** residues with Ala in the BAZ2A, KDM5B, UHRF1, and KAT6A proteins (Figure 6A–C) completely disrupted peptide binding in each case. Other subtypes may not require **L2** contacts (e.g., PHD_W with p(**L2**) $\cong 0.05$), as they have other strong, specific, peptide-anchoring contacts and thus can afford to have residues with smaller side chains (e.g., Gly) at the **L2** position. Inspecting the alignments (Figures S10A) and structures of the PHD_nW subtype (Figures S8D), one observes that the DNMT3A ATRX–DNMT3–DNMT3L (ADD) domain retains a bulky nonpolar **L2** residue, but the ATRX and AIRE proteins do not. It is interesting that the **D1** Asp residue forming a salt bridge with H3R2 is absent in the DNMT3A ADD domain but is present in the ATRX and AIRE proteins (Figure S8D, S10A). The alteration in position-specific contacts for the same peptide for *readers* of the same fold suggests that compensatory adjustments likely account for the bindingsite changes (Figure S8D). Therefore, for AIRE, we introduced a Leu residue at the **L2** position by replacing the wild type Cys residue (C310L). Interestingly, this substitution enhances AIRE's affinity for the unmodified H3 peptide by ~15 fold (Figure 9C). A gain in free energy (ΔG) of ~1.5 kcal/mol could compensate for the loss of another interaction of similar strength. It is thus tempting to speculate that the loss of one type of contact may be compensated by another, such as the **L2** contact, and the compensatory adjustments likely contributed to the origin of distinct subtypes for the PHD fingers.

Like the **L2** position, the contribution of the **V** position also suggests that nonpolar contacts at the *reader*–H3 interface play a significant role. The enriched nonpolar residues at the **V** position (e.g., UHRF1 V365, KDM5B V346, BAZ2A V1713, or KAT6A M300) is a subtype-specific adaptation to strengthen the anchorage of H3T3–CH₃ (Figure 6D). Therefore, the earlier observation that substitution of H3Thr3 with Ala severely compromises peptide binding in the BAZ2A³³ and KDM5B¹³ domains is consistent with the observed nonpolar contacts. The contribution of H3Thr3–CH₃ (CG2 atom) is anticipated to be different from that of the –OH group (OG1 atom) in the observed effects of the H3(Thr→Ala)3 substitution (i.e., in the N-cap role of –OH). To distinguish between the contributions of the H3Thr3 functional groups (–OH and –CH₃), here we utilized the H3Thr3Ser mutant peptide, assuming that the N-cap role of the Thr3 would remain unaltered, while the nonpolar contacts of Thr3 would be compromised if indeed the Thr3–CH₃ anchorage is an adaptation for this subtype. Similar to the effect of H3Ala1Gly¹³, we observed a substantial loss of binding by the H3Thr3Ser mutant peptide to the KDM5B and UHRF1 domains (Figure 6D). This suggests that the H3Thr3–CH₃ functional group, like the H3Ala1–CH₃ group, also makes a significant energetic contribution to H3 recognition. We therefore believe that *reader*–peptide nonpolar contacts have contributed significantly to the evolution of this binding mechanism, particularly for making contacts with the first three H3 residues (H3Ala1–Thr3), including the anchorage of the peptide side chain –CH₃ groups of small-volume residues. It is also interesting that the structural deviation of C α atoms at the **L1**, **L2**, and **V** positions in the PHD_nW_DD structures are small (≤ 0.84 Å, Figure S1B). By comparison, the structural deviation of C α atoms of the negatively charged positions (**AW1**, **AW2**, **D1**, and **D2**) are all >1.0 Å (Figure S1B). In comparison with the negatively

charged residues, the positions of nonpolar residues are likely to have been structurally restrained to a larger extent to preserve function.

The observed role of nonpolar residues in the recognition of histone peptides by the PHD_{nW_DD} subtype also encouraged us to look at the role of nonpolar binding-site residues of CBX chromodomains, for example, the histone H3K9me3 and H3K27me3 readout by the chromodomains of human CBX paralogues⁶⁰. CBX2, -4, -6, -7, and -8 retain a pair of position-specific *clasping* nonpolar residues at the peptide binding site (Figure S8) in place of a pair of *clasping* negatively charged residues in the corresponding positions in CBX1, -3, and -5⁶⁰. CBX1, -3, and -5 bind peptides with a higher affinity and are more specific for H3K9me3. By contrast, CBX2, -4, -6, -7, and -8 bind peptides with lower affinity and poorly discriminate between H3K9me3 and H3K27me3⁶⁰. Thus, the nonpolar residues in the CBX example contribute to lowering both affinity and specificity. The packing density of the CBX2 binding-site nonpolar residue (V11 and L50) side chain atoms in complex with the H3K27me3 peptide (Figure S8) are only 0.39 and 0.43, respectively, which are much lower than at the **L1**, **L2**, and **V** positions in the PHD_{nW_DD} subtype. The PHD_{nW_DD} nonpolar binding-site residues strengthen specific interactions by tightly packing against the peptide, a feature distinct from that of the CBX nonpolar binding-site residues.

Continued interest in understanding the origins of protein functional diversity is evident from a number of recent reports^{61–63}. In this context, the PHD finger is a useful example: (i) the interactions of PHD fingers are predominantly dedicated to a single interacting partner, the histone H3 molecule at its N-terminus, (ii) histone sequences vary little, as they are 90–98% identical⁶⁴ among eukaryotes. With such an invariant interacting partner, how a superfamily has diversified remains an interesting question. In this study, we discovered that coevolving position-specific binding-site residues likely contributed to the origin of a unique and functionally different PHD finger subtype, and our analysis will thus be useful in the study of other protein family functional diversification.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the National Institutes of Health (NIH) grant 1R15GM116040-01A1 to SC. Additional support was provided by the National Science Foundation/EPSCoR Award No. AII-1355423 and the state of South Dakota, through BioSNTR (Biochemical Spatiotemporal NeTwork Resource) and the Department of Chemistry & Biochemistry startup funds to SC. The authors thank the East Family and the Joseph Nelson undergraduate fellowships for supporting FP.

References

1. Smith E, Shilatifard A. The chromatin signaling pathway: diverse mechanisms of recruitment of histone-modifying enzymes and varied biological outcomes. *Mol. Cell.* 2010; 40:689–701. [PubMed: 21145479]
2. Allis CD, Jenuwein T. The molecular hallmarks of epigenetic control. *Nat. Rev., Genet.* 2016; 17:487–500. [PubMed: 27346641]

3. Ruthenburg AJ, Allis CD, Wysocka J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol. Cell.* 2007; 25:15–30. [PubMed: 17218268]
4. Patel DJ, Wang Z. Readout of epigenetic modifications. *Annu. Rev. Biochem.* 2013; 82:81–118. [PubMed: 23642229]
5. Musselman CA, Lalonde ME, Cote J, Kutateladze TG. Perceiving the epigenetic landscape through histone readers. *Nat. Struct. Mol. Biol.* 2012; 19:1218–1227. [PubMed: 23211769]
6. Andrews FH, Strahl BD, Kutateladze TG. Insights into newly discovered marks and readers of epigenetic information. *Nat. Chem. Biol.* 2016; 12:662–668. [PubMed: 27538025]
7. Xiong X, Panchenko T, Yang S, Zhao S, Yan P, Zhang W, Xie W, Li Y, Zhao Y, Allis CD, Li H. Selective recognition of histone crotonylation by double PHD fingers of MOZ and DPF2. *Nat. Chem. Biol.* 2016; 12:1111–1118. [PubMed: 27775714]
8. Noh KM, Wang H, Kim HR, Wenderski W, Fang F, Li CH, Dewell S, Hughes SH, Melnick AM, Patel DJ, Li H, Allis CD. Engineering of a Histone-Recognition Domain in Dnmt3a Alters the Epigenetic Landscape and Phenotypic Features of Mouse ESCs. *Mol. Cell.* 2015; 59:89–103. [PubMed: 26073541]
9. Noh KM, Allis CD, Li H. Reading between the Lines: "ADD"-ing Histone and DNA Methylation Marks toward a New Epigenetic "Sum". *ACS Chem. Biol.* 2016; 11:554–563. [PubMed: 26596909]
10. Tencer AH, Gatchalian J, Klein BJ, Khan A, Zhang Y, Strahl BD, van Wely KHM, Kutateladze TG. A Unique pH-Dependent Recognition of Methylated Histone H3K4 by PPS and DIDO. *Structure.* 2017; 25:1530–1539. e1533. [PubMed: 28919441]
11. Kungulovski G, Kycia I, Tamas R, Jurkowska RZ, Kudithipudi S, Henry C, Reinhardt R, Labhart P, Jeltsch A. Application of histone modification-specific interaction domains as an alternative to antibodies. *Genome Res.* 2014; 24:1842–1853. [PubMed: 25301795]
12. Moore KE, Carlson SM, Camp ND, Cheung P, James RG, Chua KF, Wolf-Yadlin A, Gozani O. A general molecular affinity strategy for global detection and proteomic analysis of lysine methylation. *Mol. Cell.* 2013; 50:444–456. [PubMed: 23583077]
13. Chakravarty S, Essel F, Lin T, Zeigler S. Histone Peptide Recognition by KDM5B-PHD1: A Case Study. *Biochemistry.* 2015; 54:5766–5780. [PubMed: 26266342]
14. Dove KK, Olszewski JL, Martino L, Duda DM, Wu XS, Miller DJ, Reiter KH, Rittinger K, Schulman BA, Klevit RE. Structural Studies of HHARI/UbcH7 approximately Ub Reveal Unique E2 approximately Ub Conformational Restriction by RBR RING1. *Structure.* 2017; 25:890–900. e895. [PubMed: 28552575]
15. Long HK, Blackledge NP, Klose RJ. ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochem. Soc. Trans.* 2013; 41:727–740. [PubMed: 23697932]
16. Rajakumara E, Wang Z, Ma H, Hu L, Chen H, Lin Y, Guo R, Wu F, Li H, Lan F, Shi YG, Xu Y, Patel DJ, Shi Y. PHD finger recognition of unmodified histone H3R2 links UHRF1 to regulation of euchromatic gene expression. *Mol. Cell.* 2011; 43:275–284. [PubMed: 21777816]
17. Hu L, Li Z, Wang P, Lin Y, Xu Y. Crystal structure of PHD domain of UHRF1 and insights into recognition of unmodified histone H3 arginine residue 2. *Cell Res.* 2011; 21:1374–1378. [PubMed: 21808300]
18. Lallous N, Legrand P, McEwen AG, Ramon-Maiques S, Samama JP, Birck C. The PHD finger of human UHRF1 reveals a new subgroup of unmethylated histone H3 tail readers. *PloS One.* 2011; 6:e27599. [PubMed: 22096602]
19. Wang C, Shen J, Yang Z, Chen P, Zhao B, Hu W, Lan W, Tong X, Wu H, Li G, Cao C. Structural basis for site-specific reading of unmodified R2 of histone H3 tail by UHRF1 PHD finger. *Cell Res.* 2011; 21:1379–1382. [PubMed: 21808299]
20. Arita K, Isogai S, Oda T, Unoki M, Sugita K, Sekiyama N, Kuwata K, Hamamoto R, Tochio H, Sato M, Ariyoshi M, Shirakawa M. Recognition of modification status on a histone H3 tail by linked histone reader modules of the epigenetic regulator UHRF1. *Proc. Natl. Acad. Sci. U. S. A.* 2012; 109:12950–12955. [PubMed: 22837395]
21. Xie S, Jakoncic J, Qian C. UHRF1 double tudor domain and the adjacent PHD finger act together to recognize K9me3-containing histone H3 tail. *J. Mol. Biol.* 2012; 415:318–328. [PubMed: 22100450]

22. Cheng J, Yang Y, Fang J, Xiao J, Zhu T, Chen F, Wang P, Li Z, Yang H, Xu Y. Structural insight into coordinated recognition of trimethylated histone H3 lysine 9 (H3K9me3) by the plant homeodomain (PHD) and tandem tudor domain (TTD) of UHRF1 (ubiquitin-like, containing PHD and RING finger domains, 1) protein. *J. Biol. Chem.* 2013; 288:1329–1339. [PubMed: 23161542]
23. Zeng L, Zhang Q, Li S, Plotnikov AN, Walsh MJ, Zhou MM. Mechanism and regulation of acetylated histone binding by the tandem PHD finger of DPF3b. *Nature.* 2010; 466:258–262. [PubMed: 20613843]
24. Qiu Y, Liu L, Zhao C, Han C, Li F, Zhang J, Wang Y, Li G, Mei Y, Wu M, Wu J, Shi Y. Combinatorial readout of unmodified H3R2 and acetylated H3K14 by the tandem PHD finger of MOZ reveals a regulatory mechanism for HOXA9 transcription. *Genes Dev.* 2012; 26:1376–1391. [PubMed: 22713874]
25. Dreveny I, Deeves SE, Fulton J, Yue B, Messmer M, Bhattacharya A, Collins HM, Heery DM. The double PHD finger domain of MOZ/MYST3 induces alpha-helical structure of the histone H3 tail to facilitate acetylation and methylation sampling and modification. *Nucleic Acids Res.* 2014; 42:822–835. [PubMed: 24150941]
26. Li W, Zhao A, Tempel W, Loppnau P, Liu Y. Crystal structure of DPF3b in complex with an acetylated histone peptide. *J. Struct. Biol.* 2016; 195:365–372. [PubMed: 27402533]
27. Klein BJ, Simithy J, Wang X, Ahn J, Andrews FH, Zhang Y, Cote J, Shi X, Garcia BA, Kutateladze TG. Recognition of Histone H3K14 Acylation by MORF. *Structure.* 2017; 25:650–654. e652. [PubMed: 28286003]
28. Tallant C, Valentini E, Fedorov O, Overvoorde L, Ferguson FM, Filippakopoulos P, Svergun DI, Knapp S, Ciulli A. Molecular basis of histone tail recognition by human TIP5 PHD finger and bromodomain of the chromatin remodeling complex NoRC. *Structure.* 2015; 23:80–92. [PubMed: 25533489]
29. Zhang Y, Yang H, Guo X, Rong N, Song Y, Xu Y, Lan W, Zhang X, Liu M, Xu Y, Cao C. The PHD1 finger of KDM5B recognizes unmodified H3K4 during the demethylation of histone H3K4me2/3 by KDM5B. *Protein Cell.* 2014; 5:837–850. [PubMed: 24952722]
30. Klein BJ, Piao L, Xi Y, Rincon-Arango H, Rothbart SB, Peng D, Wen H, Larson C, Zhang X, Zheng X, Cortazar MA, Pena PV, Mangan A, Bentley DL, Strahl BD, Groudine M, Li W, Shi X, Kutateladze TG. The histone-H3K4-specific demethylase KDM5B binds to its substrate and product through distinct PHD fingers. *Cell Rep.* 2014; 6:325–335. [PubMed: 24412361]
31. Torres IO, Kuchenbecker KM, Nnadi CI, Fletterick RJ, Kelly MJ, Fujimori DG. Histone demethylase KDM5A is regulated by its reader domain through a positive-feedback mechanism. *Nat. Commun.* 2015; 6:6204. [PubMed: 25686748]
32. Bergamin E, Sarvan S, Malette J, Eram MS, Yeung S, Mongeon V, Joshi M, Brunzelle JS, Michaels SD, Blais A, Vedadi M, Couture JF. Molecular basis for the methylation specificity of ATXR5 for histone H3. *Nucleic Acids Res.* 2017; 45:6375–6387. [PubMed: 28383693]
33. Bortoluzzi A, Amato A, Lucas X, Blank M, Ciulli A. Structural basis of molecular recognition of helical histone H3 tail by PHD finger domains. *Biochem. J.* 2017; 474:1633–1651. [PubMed: 28341809]
34. Sanchez R, Zhou MM. The PHD finger: a versatile epigenome reader. *Trends Biochem. Sci.* 2011; 36:364–372. [PubMed: 21514168]
35. Slama P, Geman D. Identification of family-determining residues in PHD fingers. *Nucleic Acids Res.* 2011; 39:1666–1679. [PubMed: 21059680]
36. Li H, Ilin S, Wang W, Duncan EM, Wysocka J, Allis CD, Patel DJ. Molecular basis for site-specific read-out of histone H3K4me3 by the BPTF PHD finger of NURF. *Nature.* 2006; 442:91–95. [PubMed: 16728978]
37. Pena PV, Davrazou F, Shi X, Walter KL, Verkhusha VV, Gozani O, Zhao R, Kutateladze TG. Molecular mechanism of histone H3K4me3 recognition by plant homeodomain of ING2. *Nature.* 2006; 442:100–103. [PubMed: 16728977]
38. Lan F, Collins RE, De Cegli R, Alpatov R, Horton JR, Shi X, Gozani O, Cheng X, Shi Y. Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature.* 2007; 448:718–722. [PubMed: 17687328]

39. Chakravarty S, Zeng L, Zhou MM. Structure and site-specific recognition of histone H3 by the PHD finger of human autoimmune regulator. *Structure*. 2009; 17:670–679. [PubMed: 19446523]
40. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. *Nucleic Acids Res*. 2012; 40:D290–301. [PubMed: 22127870]
41. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*. 2002; 18:77–82. [PubMed: 11836214]
42. Wheeler TJ, Clements J, Finn RD. Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*. 2014; 15:7. [PubMed: 24410852]
43. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997; 25:3389–3402. [PubMed: 9254694]
44. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: a multiple structural alignment algorithm. *Proteins*. 2006; 64:559–574. [PubMed: 16736488]
45. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004; 32:1792–1797. [PubMed: 15034147]
46. Hubbard SJ, Thornton JM. NACCESS: Computer Program. Dept. Biochemistry and Molecular Biology, University College London. 1993
47. Rother K, Hildebrand PW, Goede A, Gruening B, Preissner R. Voronoia: analyzing packing in protein structures. *Nucleic Acids Res*. 2009; 37:D393–395. [PubMed: 18948293]
48. Fodje MN, Al-Karadaghi S. Occurrence, conformational features and amino acid propensities for the π -helix. *Protein Eng*. 2002; 15:353–358. [PubMed: 12034854]
49. Wang G, Dunbrack RL Jr. PISCES: a protein sequence culling server. *Bioinformatics*. 2003; 19:1589–1591. [PubMed: 12912846]
50. Wallace AC, Laskowski RA, Thornton JM. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng*. 1995; 8:127–134. [PubMed: 7630882]
51. Pace CN, Vajdos F, Fee L, Grimsley G, Gray T. How to measure and predict the molar absorption coefficient of a protein. *Protein Sci*. 1995; 4:2411–2423. [PubMed: 8563639]
52. Turnbull WB, Daranas AH. On the value of c : can low affinity systems be studied by isothermal titration calorimetry? *J. Am. Chem. Soc*. 2003; 125:14859–14866. [PubMed: 14640663]
53. Musselman CA, Mansfield RE, Garske AL, Davrazou F, Kwan AH, Oliver SS, O'Leary H, Denu JM, Mackay JP, Kutateladze TG. Binding of the CHD4 PHD2 finger to histone H3 is modulated by covalent modifications. *Biochem J*. 2009; 423:179–187. [PubMed: 19624289]
54. Karpen ME, de Haseth PL, Neet KE. Differences in the amino acid distributions of 3(10)-helices and α -helices. *Protein Sci*. 1992; 1:1333–1342. [PubMed: 1303752]
55. Bystroff C, Simons KT, Han KF, Baker D. Local sequence-structure correlations in proteins. *Curr. Opin. Biotech*. 1996; 7:417–421. [PubMed: 8768900]
56. Aurora R, Rose GD. Helix capping. *Protein Sci*. 1998; 7:21–38. [PubMed: 9514257]
57. Gunasekaran K, Nagarajaram HA, Ramakrishnan C, Balaram P. Stereochemical punctuation marks in protein structures: glycine and proline containing helix stop signals. *J. Mol. Biol*. 1998; 275:917–932. [PubMed: 9480777]
58. Shoemaker KR, Kim PS, York EJ, Stewart JM, Baldwin RL. Tests of the helix dipole model for stabilization of α -helices. *Nature*. 1987; 326:563–567. [PubMed: 3561498]
59. Hsu WL, Oldfield CJ, Xue B, Meng J, Huang F, Romero P, Uversky VN, Dunker AK. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Sci*. 2013; 22:258–273. [PubMed: 23233352]
60. Kaustov L, Ouyang H, Amaya M, Lemak A, Nady N, Duan S, Wasney GA, Li Z, Vedadi M, Schapira M, Min J, Arrowsmith CH. Recognition and specificity determinants of the human cbx chromodomains. *J. Biol. Chem*. 2011; 286:521–529. [PubMed: 21047797]
61. Plach MG, Semmelmann F, Busch F, Busch M, Heizinger L, Wysocki VH, Merkl R, Sterner R. Evolutionary diversification of protein-protein interactions by interface add-ons. *Proc. Natl. Acad. Sci. of U. S. A*. 2017; 114:E8333–E8342. [PubMed: 28923934]

62. Akiva E, Copp JN, Tokuriki N, Babbitt PC. Evolutionary and molecular foundations of multiple contemporary functions of the nitroreductase superfamily. *Proc. Natl. Acad. Sci. U. S. A.* 2017; 114:E9549–E9558. [PubMed: 29078300]
63. Starr TN, Picton LK, Thornton JW. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature.* 2017; 549:409–413. [PubMed: 28902834]
64. Krylov DM, Wolf YI, Rogozin IB, Koonin EV. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res.* 2003; 13:2229–2235. [PubMed: 14525925]

Figure 1. Occurrence frequency at the L1 and L2 positions in the PHD fold

The topology cartoon (left) in (A) represents the location of the **L1** and **L2** positions within the strand–knuckle–strand segment of the PHD fold. Sequence logos for this segment in (A), (B), (C), and (D) are for the PHD superfamily and the PHD_nW_DD, PHD_W, and PHD_nW subtypes, respectively. The occurrence frequencies of bulky nonpolar residues at the **L1** and **L2** positions are indicated below the logos. Side chains (pink) in the plant ATXR5 PHD (5vab) structure (left) in (B) represents the location of the **L1**, **L2**, **D1**, and **D2** positions in the PHD_nW_DD structure. The sequence logo (right) in (C) is a control group (see text). For convenience, the positions labeled a **D1**, **D2**, and **W** are indicated in the sequence logos in (B) and (C).

Figure 2. Occurrence frequency of the acidic wall (AW) positions in the PHD

(A) The BAZ2A PHD structure (5t8r, left) showing the acidic wall residues. The sequence logo (right) represents the occurrence frequency for Asp, Glu, and Gly residues in the acidic wall region of the PHD_nW_DD subtype. (B) Sequence logos of the acidic wall region in the PHD superfamily and in the subtypes PHD_W and PHD_nW.

Figure 3. Structural deviations and enrichment at the VP positions

(A) PHD_nW_DD structures (left) representing a small structural deviation for the segment (dashed lines) harboring the **V** and **P** positions. The sequence logo (right) represents nonpolar residue enrichment at these positions. The sequence logos in (B)–(D) represent the occurrence frequency of nonpolar residues at these positions in the PHD superfamily as well as in the PHD_nW and PHD_W subtypes. The PHD IDs ($\sim 2.0 \text{ \AA}$) for each of the protein structures are: 4q6f (BAZ2A), 3asl (UHRF1), 5i3l (DPF3), 2puy (PHF21A), 4gne (NSD3), 4qbs (DNMT3A), 3qla (ATRX), 4l7x (DIDO), 3o7a (PHF13), 3c6w (ING5), and 3lqj (MLL1). The sequences of the PHD_nW_DD structures in (A) share < 30% sequence identity, and those of the PHD_nW (C) and PHD_W (D) subtypes share < 40% sequence identity.

Figure 4. Binding-site residue composition, correlation, and reciprocal enrichment

(A) Binding-site residue composition represented as residue occurrence frequency. In (B)–(D) the occurrence frequency (black bar) at an indicated position is altered (gray bar) as a result of perturbation (*see Methods*). The gray bars represent a change in frequency (Δp) with respect to the horizontal line (i.e., the black bar height). The perturbed positions are indicated with circled smaller fonts, while the circled larger fonts represent positions undergoing a change in frequency. For convenience, the most frequent amino acid at each position in the 85% non-redundant PFAM set and the corresponding BAZ2A PHD (an example of the PHD_nW_DD subtype) sequences are at the top in (B).

Figure 5. Distribution, sequence diversity, and taxonomic affiliations of PHD subtypes (A) The proportions of each subtype in the PHD superfamily. (B) The distribution of sequence identity between a pair of sequences in the PHD superfamily and in the PHD_nW_DD, PHD_W, and PHD_nW subtypes. The dashed lines represent the median sequence identity. (C) Taxonomic affiliations of the sequences in each subtype. The eukaryotic tree (bottom, right) is shown for convenience.

Figure 6. The contributions of the L1, L2, and V positions in the PHD_{nW}_DD subtype
 The ITC-binding curves of histone H3-1-11 titrated into wild type or various mutants of the KDM5B, UHRF1, and BAZ2A PHD. In (A)–(C) wild type proteins (black) are compared with the L1 (red), and L2 (green) mutants. (D) The V mutant of KDM5B PHD (purple). (E). ITC binding curve of H3T3S (left) titrated with the KDM5B (purple) or UHRF1 (gray) wild type domains. The symbolic representation (right) of the nonpolar contact between the V position and peptide H3T3 in the UHRF1 complex (3asl).

Figure 7. The contribution of the D2 position in the PHD_{nW}_DD subtype
 The ITC binding curves of histone H3-1-11 titrated with wild type (black), **D2** (green), or **D1** (red) mutants of the KDM5B, UHRF1, and BAZ2A PHDs in (A)–(C), respectively. (D) The structural deviation of the BAZ2A (4q6f) treble clef knuckle. The pdb-ids are 3asl (UHRF1), 5b75 (KAT6A), and 5i3l (DPF3).

Figure 8. The AW residues and peptide redirection in the PHD_nW_DD subtype

(A) Peptide backbone steric interference in the PHD_nW_DD subtype estimated by peptide swapping. Histograms represent the contact distances between H3 and PHD fingers in natural (left) and swapped complexes (middle and right). (B) Symbolic representation of peptide redirection. (C) Factors likely contributing to the helical conformation of histone H3.

Figure 9. L2 position packing comparison among PHD subtypes

(A) The packing density of the BAZ2A L1693 CD1 (top, left) and CG (bottom, left) atoms are shown as red dashed lines among the distributions of Leu CD1 and CG atoms, respectively, in protein structures. Surface mesh representing the contact between the **L2** side chain and the peptide backbone for the BAZ2A (4qf6, top right) and UHRF1 (3asl, bottom right) domains. (B) Surface mesh representations of the **L2** contacts in the PHD_W subtype [DIDO (4i7x), ING5 (3c6w), and MLL1 (3lqj)]. (C) ITC titration (top) profiles of the wild type and **L2** (C310L) mutant of AIRE PHD (2kft) with the histone H3-1-11 peptide. Surface mesh of the L2 positions in the AIRE wildtype and modeled C310L mutant are at the bottom.