2017

Dissecting the Histone-binding Mechanism of a PHD Finger Subtype

Daniel Boamah
South Dakota State University

DISSECTING THE HISTONE-BINDING MECHANISM OF A PHD FINGER SUBTYPE

BY

DANIEL BOAMAH

A dissertation submitted in partial fulfillment of the requirement for the

Doctor of Philosophy

Major in Biochemistry

South Dakota State University

2017

DISSECTING THE HISTONE-BINDING MECHANISM OF A PHD FINGER SUBTYPE

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy in Biochemistry degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

Suvobrata Chakravarty, Ph.D. Date
Dissertation Advisor

Douglas E. Raynie Date
Head, Department of Chemistry and Biochemistry

Dean, Graduate School Date

DEDICATION

I dedicate this work to the Boamah family (Yvonne, Danielle, Michelle and Gabrielle) for all the sacrifices to see this dream come through. I also dedicate it to my parents, in-laws and siblings for all the help along the way. Thank you for what I am today! To God be the Glory for all He has done.

ACKNOWLEDGEMENTS

I thank the Almighty God for his faithfulness and favor towards me throughout this journey. I thank my entire family (Yvonne, Danielle, Michelle and Gabrielle) for all the prayers, encouragement, support and their patients for making this a reality. I am highly indebted to my dissertation advisor; Dr. Suvobrata Chakravarty for his persistent measured criticisms, mentorship and uncompromising meticulousness, which has enriched my life tremendously. I thank the Department of Chemistry and Biochemistry, the Biochemical Spatio-temporal NeTwork Resource (BioSNTR) and the National Institute of Health for their financial support at different stages of this career. A special thank you goes to my dear committee members (Dr. Adam Hoppe, Dr. Fathi Halaweish, Dr. Feng Li and Dr. Alireza Salehnia) for their tremendous help and encouragement at every step of the way.

I would like to thank the leadership and all members of my church (Church of Pentecost, PIWC, Worcester, MA and Holy Life Tabernacle, Brookings, SD) for your prayers and encouragement. Finally, I thank all my lab members both past and present, my colleagues at the department and the Ghanaian Community in Brookings for making this journey an enjoyable one. God bless you all.

TABLES OF CONTENTS

ABBREVIATIONS

ASA: Accessible Surface Area

ATRX: Alpha Thalassemia/Mental Retardation Syndrome X-Linked

AIRE: Autoimmune regulator

BAZ1A: Bromodomain adjacent to zinc finger domain protein 1A

BAZ2A: Bromodomain adjacent to zinc finger domain protein 2A

BCL9: B-Cell Lymphomas 9

BPTF: Bromodomain PHD finger transcription factor

CBX: Chromobox

CMI: Cumulative Mutual Information

DLBCL: Diffuse Large B-cell lymphoma

DNMT3: DNA methyltransferase 3

DPF: Double PHD Finger

EZH2: Enhancer of Zeste Homolog 2

FPLC: Fast Protein Liquid Chromatography

GST: Glutathione S-Transferase

ING4: Inhibitor of growth protein 4

IPTG: Isopropyl β-D-1-thiogalactopyranoside

ITC: Isothermal Titration Calorimeter

KAT6A: Lysine (K) Acetylatransferase 6A

KDM5B: Lysine (K) specific demethylase 5B

MI: Mutual Information

MLL2: Mixed Lineage Leukemia 2

MLL3: Mixed Lineage Leukemia 3

NoRC: Nucleolar Remodeling Complex

NSCLC: Non–Small Cell Lung Cancer

MSA: Multiple Sequence Alignment

PCR: Polymerase Chain Reaction

PDZ: Postsynaptic density, Drosophila disc large tumor suppressor, and Zonula occludens-1

PHD: Plant Homeodomain

PHFR1: PHD And Ring Finger Domains 1

RMSD: Root Mean Squared Deviation

SDS-PAGE: Sodium dodecyl sulfate polyacrylamide gel electrophoresis

UHRF1: Ubiquitin-like PHD and RING finger domains 1

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

DISSECTING THE HISTONE-BINDING MECHANISM OF A PHD FINGER SUBTYPE

DANIEL BOAMAH

2017

Disordered tails of histones are critical information retrieval hub and thus, aberrations in the flow of information through these hubs are associated with a number of pathological consequences in human. Mechanism for retrieval of information from these hubs is achieved by protein-protein interaction, i.e. proteins dock onto histone tails to initiate chromatin signaling. Eukaryotes have a number of small peptide binding domains that have evolved to specifically interact with histone tails, and these domains called histone readers as they read the information encoded on histone tails. Plant homeodomain (hereafter PHD) finger, a binucleated zinc finger, family is one such histone readers. Next-generation sequencing efforts on diagnosed patient's genomes or cancer tissues show that mutations in PHD finger, particularly a subgroup of PHD fingers, are associated with number of pathological consequences. Therefore, for future understanding of the possible mechanisms for the pathological consequences, as an initial step, detailed characterization of the binding mechanism of the PHD subtype, the PHD_nW_DD, was undertaken here. Starting with human BAZ2A (bromodomain adjacent to zinc-finger 2A), one member of the PHD_nW_DD subtype that is associated with prostate cancer was utilized to probe the effect of mutations on histone tail binding.

We experimentally discovered two categories of mutations that disrupt peptide binding: (1) Type-A: positions that are in contact with the peptide and (2) Type-B:

positions that are remote from the peptide-binding site (distal site). For my dissertation, I focused on understanding the biochemical basis of the effects of Type-A mutations using recombinant protein chemistry and biophysical chemistry.

The peptide-anchoring residue positions of BAZ2A-PHD, interestingly, are enriched in specific type of residues in a subtype specific manner. The energetics revealed that, two non-polar amino acid residues and an Aspartate residue in the treble clef knuckle make significant contributions to the formation of the hBAZ2A-histone peptide complex as mutations at these three positions completely aborted peptide binding. The energetic contributions of the identified positions were further confirmed by mutagenesis in three members of the subtype (UHRF1-PHD, KDM5B-PHD and KAT6A-PHD) that included pairs sharing even less than 40% sequence identity with each other. Despite low sequence similarity, mutations cause similar consequences in histone H3 binding suggesting a strong similarity in the binding mechanism, and thus justifying the subtype classification.

CHAPTER I

INTRODUCTION

1.1. General Introduction

Structure dictates functions, and thus, the protein fold (the precise arrangement of the secondary structures in space) is intricately related to the function a protein performs. A protein fold is often also associated with a large sequence space meaning that several different sequences including diverse ones, that retain a few common sequence features, can adopt the same structural fold[1-3]. However, the precise biochemical function of diverse sequences, despite retaining the same overall fold can significantly differ from one another. Many protein folds are referred to as superfamilies for the diverse sequences they contain and the diverse biochemical functions they perform.

*P*lant *H*omeo*D*omain (PHD), a binucleated Zn finger[4-6], is one such superfamily where the median sequence identity (in 85% non-redundant set)[7] is only ~27%, i.e. diverse sequences populate the PHD fold. Although the primary biochemical function of PHD finger is to bind histone peptides, they can recognize different histone peptides. A correlation between the sequences of the histone peptide and sequences of PHD finger has been noted earlier[7, 8] meaning that there are different subtypes of sequences within the PHD superfamily with each subtype capable of recognizing a distinct histone sequence. Thus, correlating the energetic contributions of the binding-site residues of PHD subtypes in the sequence specific interactions with histone peptides would be valuable for sequence-based classification of PHD fingers. The classification will be very useful in the genome-wide functional annotation of PHD finger sequences. In addition, detailed information of the energetic contributions of PHD residues will be useful for

manipulating the binding behavior of PHD by residue substitution, and this in turn can be useful for manipulating gene regulatory mechanisms, and then enable probing pathway specific flow of epigenetic information.

For this dissertation, we thus focus on PHD, particularly a specific subtype of the PHD finger. In general, PHD finger contains approximately 50-80 amino acid residue[4-6] with a characteristic cross-braced topology having two zinc atoms coordinated by one Histidine and seven cysteine residues within a Cys4-His-Cys3 sequence motif[8] (Figure 1.1)[7, 9]. PHD finger was discovered in the protein HAT3.1 in *Arabidopsis thaliana* in 1993[6, 10]. Most of the proteins with PHD fingers are found in the nucleus[11, 12] because they are mainly involved in the modification of chromatin as well as the mediation of molecular interactions in gene transcription[6, 8, 11, 13]. They may occur as a single finger but can often be seen in cluster of two (Figure 1.2)[7, 9] or more, and can also occur in tandem with other domains, such as bromodomain and chromodomain[6, 8, 12, 14].

Earlier discovery of their function in chromatin regulation showed that they bind methylated-lysine histone H3[14-17]. However, recent studies have shown them to have a sophisticated ability to recognize other sequences of histones, including histone peptides lacking post-translational modification (e.g., unmodified histones)[7, 8, 18, 19]. These studies reinforce that they are versatile in their ability to recognize histone sequences (modified and unmodified) in their role as epigenome readers that control gene expression by recruiting transcription factors and chromatin regulators[7-9, 18, 19].

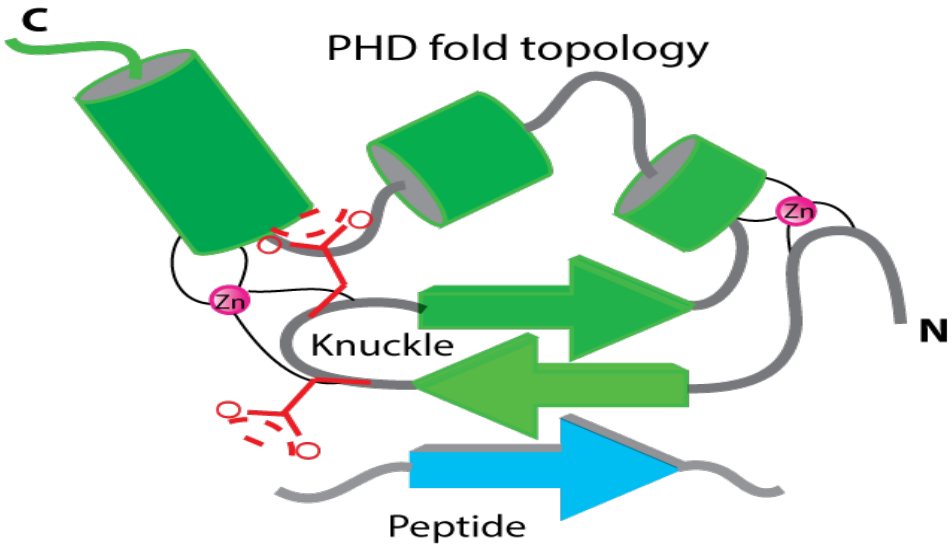


Figure 1.1. A topological diagram of the PHD fold: The Cys and His residues hold the binucleated Zn atoms. A subtype that we study here has two Aspartate residues in the treble clef knuckle. The pair of Asp residues is noted to anchor the Arg2 of Histone H3 peptide.

The Zn-chelating Cys and His residues in many Zn-fingers (e.g., PHD finger, RING finger, FYVE finger etc.) appear with a characteristic 3D pattern called treble clef. The treble clef 3D motif [20] refers to a combination of a knuckle and a helix (Figure 1.1). In the treble clef motif, the knuckle provides two cysteine residues and the N-terminal of the helix provides the other two Zn-chelating cysteine residues for chelating the Zn atom.

Many domains or folds contain treble cleft motif and PHD finger is one of them. Many of these domains or folds (e.g., PHD finger, RING finger, FYVE finger etc.), in addition to possessing the treble clef motif, can have other secondary structural element that distinguish them from one another. For the discussion of this dissertation we often refer to treble cleft portion of the PHD finger and our entire study is about PHD finger (and not any other treble clef containing domains or folds).

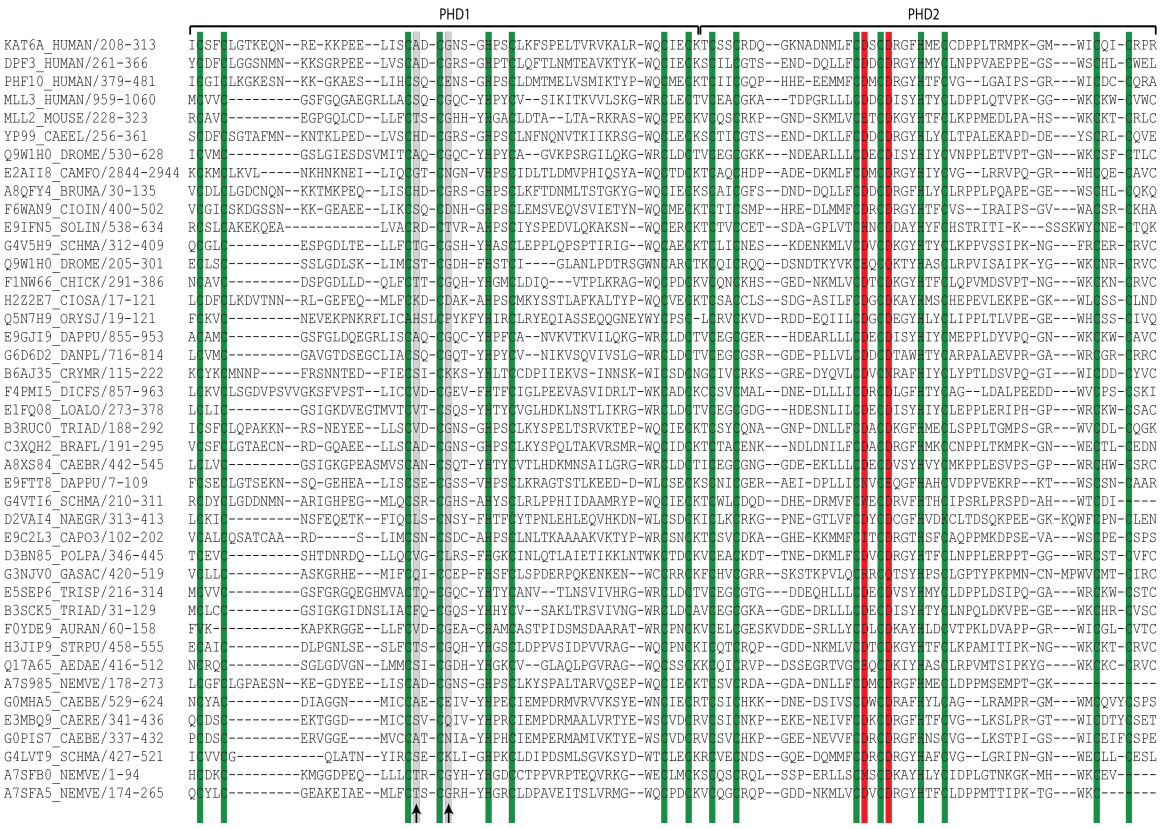


Figure 1.2. Alignment of tandem PHD-PHD sequences as reported in Chakravarty et al., 2015. The zinc chelating positions are shown in green, red bar shows the presence of Aspartate residues and the gray bar shows the absence of Aspartate residues.

The treble clef portion of some of the PHD fingers have been reported to possess a pair of Asp residues[7] that can interact with the Arg2 of the Histone H3 peptide. This characteristic pattern of Aspartic acid residue within the treble clef knuckle has been shown to be a subtype-specific property[7] and the pair of Asp residues make a significant energetic contribution to the formation of the protein-peptide complex[7]. However, it is not clear if the presence of the treble clef Asp residues in the PHD sequences is enough for histone peptide binding. It is likely that other amino acid residues in the PHD sequence also make large contributions to the energetics of the protein-peptide complex

formation. In this, I wanted to identify the characteristics of those residues that make large energetic contributions for anchoring the H3 peptide. In addition, I wanted to determine if there are residues remote from the binding site that also contribute to peptide binding as this provides an opportunity to modulate binding by manipulations without altering the binding site. For this dissertation, therefore, I discuss histone-binding mechanism of PHD finger subtype in the context of energetic contributions of PHD subtype specific residues.

1.2. Introduction to research

Protein-protein interaction on the chromosomes decisively contributes to the interpretation of the information encoded in the genome, and the tethering of proteins onto the intrinsically disordered tails of histones is a critical component of the chromosomal protein-protein interactions. Regulatory protein complexes, bound onto 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[21, 22]. Eukaryotes have evolved peptide-anchoring scaffolds (typically called readers)[23] to specifically recognize histone peptide segments[24-27]. In general, the disordered peptide segments of histone tails provide the binding surface for readers (present within the regulatory protein complexes) to facilitate the anchorage of regulatory complexes onto the chromatin.

Over the past decade, fascinating discoveries on the mechanisms of histone peptide recognition, revealed by high-resolution structures of readers in complex with cognate peptides[24-27] are helping us to understand chromatin regulatory mechanisms in much

greater detail. For example, 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 for probing and altering DNA methylation landscapes in living stem cells)[28, 29].

Opposed to *de novo* design, *reader* reengineering involves manipulation of a few binding site residues of an existing *reader* leaving the rest of the *reader* sequence unaltered[28, 29]. Therefore, structure guided dissection of the key determinants of histone peptide recognition mechanisms will continue to play important role not just in probing chromatin-signaling mechanisms but also in the design of reagents for diagnostic applications[30, 31]. Motivated by the application outcomes, our lab had earlier probed the energetics of the recognition of histone H3 N-terminal peptide by *readers* belonging to the same structural scaffold, the *p*lant homeo*d*omain (PHD)-finger, that recognize the same peptide sequence[7].

The study revealed that recognition mechanisms could differ in the *readout* of the same peptide sequence by *readers* of even the same scaffold. It was also noted that 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 towards N-terminal H3 recognition by PHD *readers* featuring a treble clef xCDxCDx motif, while H3 Lys4 makes a substantial energetic contribution for readers lacking the motif[7]. This lead to speculate that subtype specific sequence patterns for *readers* evolved for distinctly interacting with even the same peptide sequence, and a comprehension of the roles of all subtype specific sequence features would be invaluable to manipulate a *reader's* substrate preferences for the above applications. Therefore, here, we investigate

the contributions of the subtype specific features of *readers* using the treble clef xCDxCDx motif featuring PHD subtype as a model to better understand histone-anchoring adaptations.

PHD-finger of BAZ2A (*b*romodomain *a*djacent to *z*inc finger domain protein 2A), one of the xCDxCDx motif featuring *readers* was chosen as the base member for a detailed investigation of the peptide anchoring mechanism for the following reasons. BAZ2A, a subunit of the chromatin remodeling complex NoRC (*n*ucleolar *r*emodeling *c*omplex)[32], possesses the PHD-Bromodomain cassette for facilitating chromatin anchorage of NoRC[33, 34] for rDNA silencing[35, 36]. Likely for its critical role, BAZ2A overexpression is tightly associated with prostate cancer [37], where BAZ2A, along with EZH2 (*e*nhancer of *zeste h*omolog 2), coordinates epigenetic silencing in prostate cancer cells[37]. Thus, the possibility of therapeutically targeting BAZ2A chromatin anchoring scaffold has also been suggested[37], and we, therefore, dissect the binding characteristics of BAZ2A-PHD.

The recent structure of the BAZ2A-PHD in complex with histone H3 N-terminal peptide highlights the role of acidic patch, another subtype specific feature, in peptide recognition[38], suggesting the importance of recognizable sequence features in the functional characterization of *readers*. This work thus complements the structural report as we investigate a number of other sequences features, in addition to the acidic patch, not only in BAZ2A but also in several other *readers*. Initiated with BAZ2A-PHD, we follow up the probing of position specific contributions of residues using four additional *readers* (sharing low sequence similarity with BAZ2A-PHD). Based on an experimental analysis of a set consisting of a total of 29 proteins (5 wild type *readers* and their 24

mutants), we provide a comprehensive view of histone peptide recognition by a *reader* subtype.

The versatile canonical PHD binding-site[8], like the versatile PDZ domain binding-site[39, 40], binds terminal peptides (e.g., the N-terminal H3) predominantly by beta augmentation[41], and therefore , comparison of the binding mechanisms between the two is natural. The extensive work on PDZ binding mechanisms probed by large-scale mutagenesis[40], sequence randomization[42], and proteomic scale binding analysis[43, 44] have been complemented by computational studies to extract correlation between PDZ binding-site positions[45] for an evolutionary view on the origin of PDZ subtypes and binding versatility. Although the number of mutations considered in this study is comparatively much smaller than the hundreds of mutations used in the PDZ[40] binding mechanism study, the results obtained using a reasonable size of the PHD proteins studied here, encouraged us to similarly probe the relationship between PHD binding-site positions by computational approaches here for a view on nature's design principles of versatile binding-sites. In addition, the experimental binding thermodynamics observed here also encourages us to computationally analyze natural adaptations for anchoring histone tails, as these peptides lack the common peptide hotspot residues (W, Y, F, L and I)[46]. Overall, we provide a case-specific study of histone-*reader* interactions in the context of: (a) superfamily subtype characterization, (b) evolutionary relationship between positions of versatile binding-sites and (c) general adaptations for anchoring onto unmodified histone tails.

CHAPTER II

ENERGETICS OF THE INTERACTIONS BETWEEN HUMAN BAZ2A-PHD FINGER-HISTONE H3

ABSTRACT

PHD-finger of BAZ2A (*b*romodomain *a*djacent to *z*inc-finger 2A) protein is a histone H3 N-terminal peptide *reader* that uses the treble clef xCDxCDx motif for the recognition of the peptide. It is overexpressed in prostate cancer. The possibility of therapeutically targeting BAZ2A chromatin-anchoring scaffold for the metastatic potential of prostate cancer has recently been suggested[37]. Though some work has been done on the molecular mechanism as to how this protein interacts with the histone H3, the energetics involved in the interaction is lacking. As an initial step for the understanding of the possible mechanisms and the energetics, detailed characterization of the binding mechanism of the BAZ2A-PHD was undertaken.

Here, we have performed systematic site-specific mutagenesis in the recombinant version of this PHD finger, purified mutant proteins and measured interaction affinity between the mutant proteins and histone H3 using ITC (Isothermal Titration Calorimetry). We show that, two categories of mutations disrupt peptide binding: (1) Type-A: positions that are in contact with the peptide and (2) Type-B: positions that are remote from the peptide-binding site (distal site). For this dissertation, we focused on understanding the biochemical basis of the effects of Type-A mutations.

The peptide-anchoring residue positions of BAZ2A-PHD making large energetic contributions are enriched in specific amino acids (bulky non-polar and acidic). The bulky non-polar amino acid residues identified here are Leu-1692 and Leu-1693 and are

tightly packed with the small side chains of the histone peptide. An Aspartate residue in the treble clef knuckle, Asp 1695, also makes significant energetic contribution to the formation of the hBAZ2A-histone peptide complex. Ala substitutions at these positions mentioned above completely aborted peptide binding.

The recognition features identified here could be exploited in the PHD-finger subtype possessing the same scaffold and possibly utilized to develop peptide diagnostics for epigenome.

INTRODUCTION

The human Bromodomain Adjacent to Zinc finger 2A (BAZ2A) also called TIP5, is a part of the chromatin remodeling complex NoRC (nucleolar remodeling complex) known to be involved in epigenetic rRNA gene silencing[37, 47]. BAZ2A regulates numerous protein-coding genes and directly interacts with EZH2 to maintain epigenetic silencing at genes repressed in metastasis[37]. BAZ2A is a very large multidomain protein of a molecular weight of 211kDa and consists of a number of DNA and protein interaction domains[33, 47]. The protein contains a DNA binding homeobox and different transcription factors, methyl-CpG-binding domain, AT-hook DNA –binding domains, and a PHD (plant homeodomain) zinc finger adjacent to a bromodomain at their C terminus[47](Figure 2.1). The PHD zinc finger adjacent to the bromodomain has been shown to have a significant role in the NoRC function, binding to HDAC1 (histone deacetylase 1) DNMTs (DNA methyltransferases), and ATPase subunit SNF2h[36]. It has also been shown to belong to a PHD zinc finger family proteins that harbor PHD_nW_DD subtype[7]. Through next-generation sequencing efforts on diagnosed

patients' genomes or cancer tissues, it has come to light that aberrations (mutations, overexpression) in proteins harboring PHD_nW_DD subtype are associated with different pathological consequences[37, 48-53](Figure 4.1). For example, mutations in MLL2 PHD has been reported to result in different types of cancers[54-60]. BAZ1A and MLL3 have also been reported in rare tumors of gynaecological origin[61].Therefore for future comprehension of the mechanisms for pathological consequences, as a first step, detailed characterization of the binding mechanism of the PHD_nW_DD was undertaken here, starting with one member (BAZ2A-PHD) of the subtype.

Systematic site-specific mutagenesis in the recombinant version of this BAZ2A-PHD was performed, purified mutant proteins and measured interaction affinity between the mutant proteins and histone H3 using ITC (Isothermal Titration Calorimetry) to infer the molecular basis of the spatial connectivity or coupling between binding site amino acids and those away from the functional site.

Figure 2.1. Domain organization of human BAZ2A containing PHD *z*inc finger *a*djacent to *b*romodomain: The MBD (methyl-CpG-binding domain) is the first domain, followed by the DDT (DNA binding homeobox and different transcription factors) domain, the PHD zinc finger is colored indigo which is adjacent to the bromodomain.

MATERIALS AND METHODS

DNA constructs

The recombinant BAZ2A-PHD finger (residues 1673-1728 UniProt ID Q9UIF9/BAZ2A_HUMAN) used here was taken from an earlier work in the lab[7]. Briefly, BAZ2A-PHD was originally cloned into pETite N-His SUMO Kanamycin expression vector. 14 BAZ2A-PHD mutants (*see* mutant list in Table 2.8) were created using site directed mutagenesis. DNA sequences of wild type and all mutants were confirmed by DNA sequencing (Genscript). In total, 15 proteins (1 wild type and 14 mutants) created in this study are used to probe the peptide binding mechanisms of PHD finger subtype (BAZ2A).

Molecular Cloning

Synthetic DNA molecule of hBAZ2A-PHD of UniProt ID Q9UIF9/BAZ2A_HUMAN, residues 1673-1728 was obtained from GenScript and PCR (polymerase chain reaction) amplified using Thermo Scientific Phusion High Fidelity DNA Polymerase. It was sub-cloned into pETite N-His SUMO Kanamycin expression vector (Lucigen). This was a ligation independent (enzyme-free) cloning and expression system in which the gene of interest was cloned under the control of the L-rhamnose-inducible rhaP$_{BAD}$ promoter harbored on one of the pRham^{TM} vectors[62, 63], which replaces the T7 promoter. The synthetic BAZ2A was PCR amplified with primers that contained short homology to the ends of the pETite N-His SUMO Kan vector (18 nucleotides of overlap with the ends of the vector)[62-65]. The PCR product was then mixed with the pre-processed vector in a 1:1 ratio and transformed directly into the high-efficiency

chemically competent cells (E. cloni® 10G SOLOs Chemically Competent Cells, Lucigen). It was incubated overnight at 37°C for about 14 hours on kanamycin LB plate (see transformation). Colony PCR was done to confirm clones, and individual colonies were picked, cultured overnight at 37°C (for about 18 hours on kanamycin LB media). The recombinant DNA (pETite_BAZ2A) was isolated using QIAprep spin miniprep kit (QIAGEN). Agarose gel electrophoresis was run to confirm the presence and size of the miniprep sample before they were sent to GenScript for DNA sequencing.

After DNA sequence was verified, plasmid was transformed into HI-Control® BL21 (DE3) SOLOs Chemically Competent cells (see transformation), protein expression was done and purified for the biophysical measurements with the ITC. Fourteen mutations were made in pETite_BAZ2A to change the amino acid residue positions to Alanine with Site-Directed Mutagenesis. They were then expressed and purified for ITC measurements.

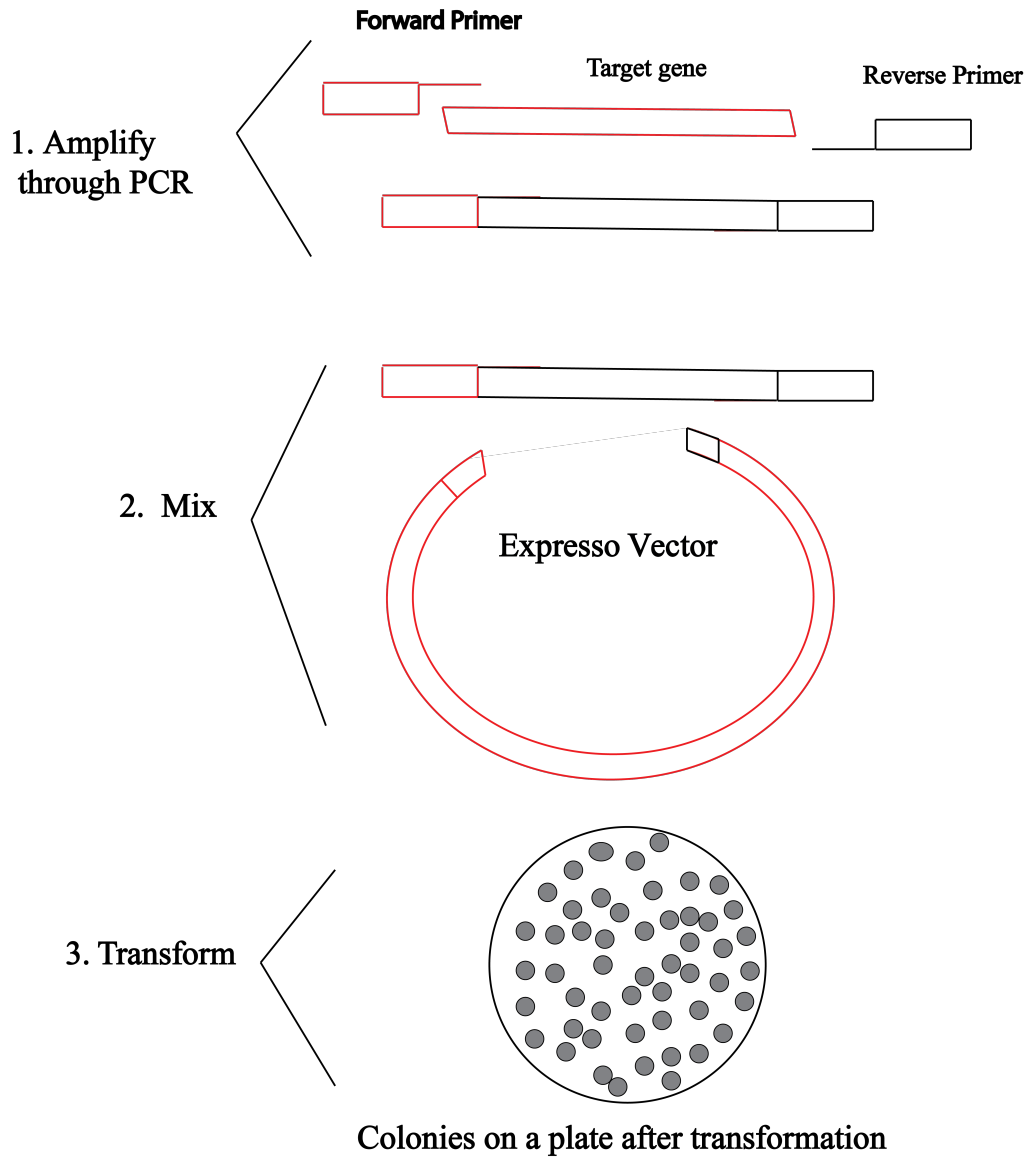


Figure 2.2. The cloning design strategy of pETite_BAZ2A-PHD using the enzyme-free technology by Lucigen Simplifying genomics, where both forward and reverse primers are designed with 18 nucleotide sequences overlapping that of the pETite N-His SUMO Kan vector.

Table 2.1. The nucleotide and translated protein sequence of hBAZ2A for the cloning. The red colored portion of the primer sequence represents the short homology to the ends of the pETite N-His SUMO Kan vector (18 nucleotides of overlap with the ends of the vector).

	Nucleotide Sequence
hBAZ2A-PHD	TCTGTCAACAAAGTGACATGTCTAGTCTGCCGGAAG GGTGACAATGATGAGTTTCTTCTGCTTTGTGATGGG TGTGACCGTGGCTGCCACATTTACTGCCATCGTCCC AAGATGGAGGCTGTCCCAGAAGGAGATTGGTTCTG TACTGTCTGTTTGGCTCAGCAGGTG
	Protein Sequence
	S V N K V T C L V C R K G D N D E F L L L C D G C D R G C H I Y C H R P K M E A V P E G D W F C T V C L A Q Q V
Primers	**Primer Sequence**
Forward Primer (5' to 3')	CATCATCACCACCATCACTCTGTCAACAAAGTGACA TGT
Reverse Primer (5'to 3')	GTGGCGGCCGCTCTATTACACCTGCTGAGCCAAACA GAC

Table 2.2. Reagents for PCR: The reaction was set up on ice and the reagents were added in the order they appear in the table.

Component	**Volume in 50µL Reaction/µL**
Distilled water	35.5
5X Phusion buffer	10
10mM dNTPs	1
Forward Primer	1
Reverse Primer	1

Template DNA	1
Phusion DNA Polymerase	0.5

Table 2.3. PCR reaction cycle: The entire reaction took about 2 hours

Cycle step	**Temperature**	**Time**	**Cycles**
Initial Denaturation	98 °C	30sec	1
Denaturation Annealing Extension	98 °C 62 °C 72 °C	10sec 30sec 1min30sec	30
Final Extension Final Hold	72 °C 4 °C	10min ∞	1

Transformation

These steps were applicable to E. coli DH5α, BL21 cells or any other E. coli cells used for the cloning. The materials needed were 42 °C water bath, Ice, LB Plates with appropriate antibiotics (kanamycin for hBAZ2A-PHD), Sterile SOC or LB (no antibiotics)

Procedure

- DH5α or BL21 competent cells from -80 °C freezer was put on ice for 30min. The competent cells were usually in eppendorf tube
- Cells were gently mixed with the pipet tip (was not pipet up and down) and 30µL was aliquot into 1.5ml eppendorf tube that had been pre-chilled on ice.
- 1µL of DNA was added to the cells and mixed gently
- The tube was then incubated on ice for 30 minutes.
- The SOC was Pre-warmed at 42°C
- The cells were heat shocked for 45 seconds in a 42°C water bath without shaking.
- The tube was then placed on ice for 2 minutes.
- 250µL of SOC was added to the cells and incubated at 37°C for 1 hour at 225rpm.
- The Kanamycin agar plate was pre-warmed in the incubator at 37°C.
- 100µL of cells was inoculated and incubated at 37°C for overnight (12-16hours).

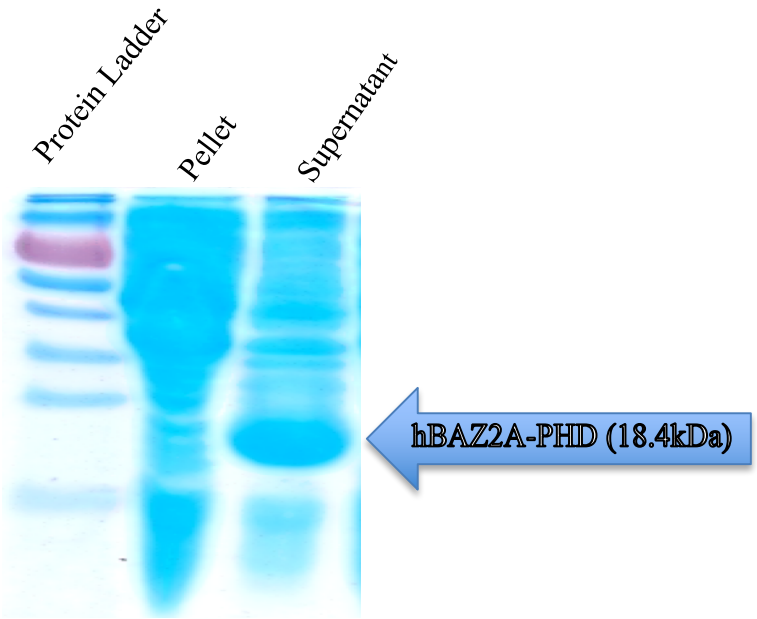
Protein Expression and Solubility Test

The recombinant hBAZ2A-PHD was transformed into HI-Control® BL21 (DE3) SOLOs Chemically Competent cells (Lucigen) after the sequence was verified from GenScript. Three separate colonies on the Kanamycin agar plate were selected for a small-scale protein expression and solubility testing. This was done by inoculating them in a 4ml LB/kanamycin media and incubated at 225rpm, 37°C till the optical density at 600nm (OD_{600}) of each of the culture was between 0.2-0.4 (This was usually reached in 4hours). Half of the volume of each culture (2ml) was transferred into a new eppendorf tube and induced with 1mM isopropyl β-D-1-thiogalactopyranoside (IPTG). They were

then allowed to grow for another 3hours at the same conditions as above. The rest of the test was carried out by the procedure below:

- The cell cultures were pelleted in pre-weighed eppendorf tubes by centrifuging at 13000rpm for 10mins
- 0.5g of the pelleted cells were completely resuspended in a 1ml solution of 1x bug buster.
- The suspension was incubated on a shaker at 4°C for 20mins and then centrifuged at 13000rpm for 10mins.
- The supernatant, which now contains the soluble protein, was separated from the pellet and SDS-PAGE gel run to determine the solubility of the protein
- The remaining pellet (both induced and uninduced) was also used to run SDS-PAGE gel to determine the expression of each colony picked.

A



B

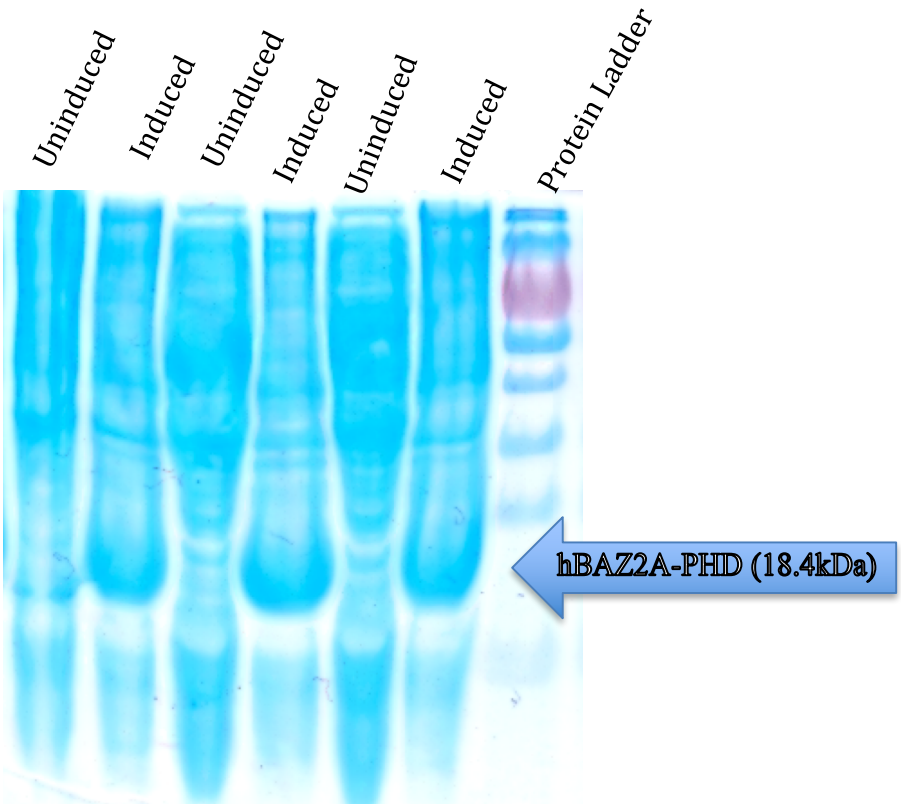


Figure 2.3. (A) SDS-PAGE gel of hBAZ2A-PHD Solubility test showing that the protein is soluble hence can be seen in the supernatant with the size of 18.4KDa. Lane 2, which is the pellet, does not show such a band, which means the protein, is clearly not in the pellet hence it is a soluble protein.

(B) SDS-PAGE gel of the expression test of 3 colonies of hBAZ2A-PHD. Lanes 1, 3 and 5 represents the uninduced colonies. Lanes 2, 3 and 6 represents the colonies induced with 1mM IPTG. It can clearly be seen that the induced colonies have bands at the 18.4KDa mark whereas the uninduced do not have that band in the gel. Lane 7 represents the protein ladder that was used to determine the size of the protein.

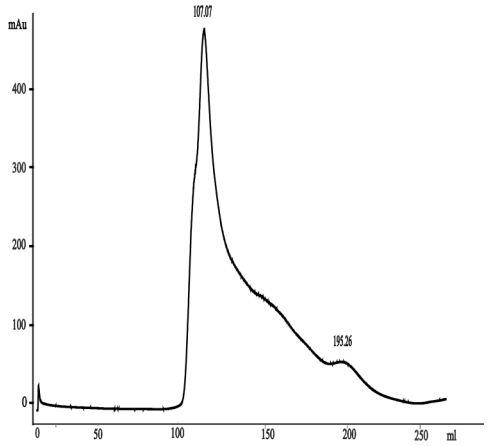
Protein Expression and Purification

After it was determined that the recombinant protein was soluble and could be expressed, a 4litre culture was prepared by growing 30ml culture in an LB/kanamycin media overnight at 37°C, 225rpm. The overnight grown culture was transferred to an autoclaved terrific broth/kanamycin media and allowed to grow in a 37°C incubator at the same rotation speed as above till the OD_{600} was between 0.6 and 1. The culture was then induced with 1mM IPTG and transferred to a 16°C shaker (N-Biotek) at rotating speed of 225rpm and allowed to grow overnight (approximately 18hrs). The protein was isolated and purified by His-tag affinity chromatography.

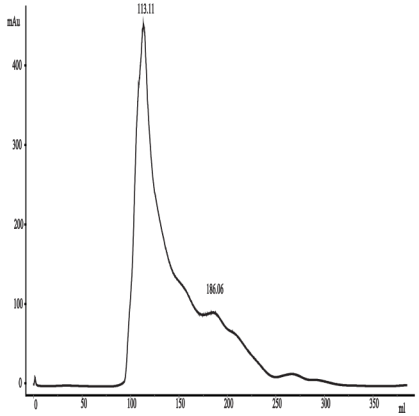
The overexpressed cells were lysed in a phosphate buffer (25mM phosphate buffer, 300mM NaCl at pH 7.6) with 10mg/ml of lysozyme, which facilitates in breaking the cell wall of the bacteria to release the soluble protein in the buffer and 0.1mg/ml of IGEPAL-CA 600 (a nonionic, non-denaturing detergent which prevents surface-induced protein aggregation in the buffer). The number of bacteria cells in the resuspension phosphate buffer should be 30%. The resuspension was allowed to stand on ice for 30minutes before it was treated with a Misonix ultrasonic liquid processor (Sonicator S-4000) for 5 minutes that further lysed the cells. The lysates were cleared by centrifugation (18000rpm for 30min at 4°C, Beckman J2-MI centrifuge). After centrifugation, the supernatant that contained the soluble protein was loaded on a phosphate buffer equilibrated nickel sepharose high performance bead (GE Healthcare Life Sciences) overnight. The protein was then eluted with an imidazole (200-250mM) solution at pH 7.6. It was finally polished or purified to homogeneity by size exclusion gel filtration chromatography using a HiLoad 26/600 superdex column using AKTA prime FPLC (GE Healthcare Life

Sciences). The phosphate buffer at this step had 150mM NaCl. The peak corresponding to the size of the protein was concentrated to the desired volume and concentration determined by UV-Vis Nanodrop Spectrophotometer and Bradford Method. The purity of the protein was confirmed by SDS-PAGE (Bio-Rad).

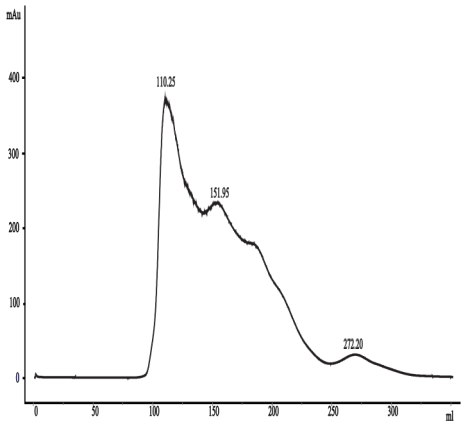
It must be stated that different conditions at the elution step (in terms of the amount of salt in the buffer, the concentration of imidazole needed for the elution of the protein) were tried in order to arrive at the optimized condition used for the purification (Figure 2.3 and 2.4).



300mM Salt

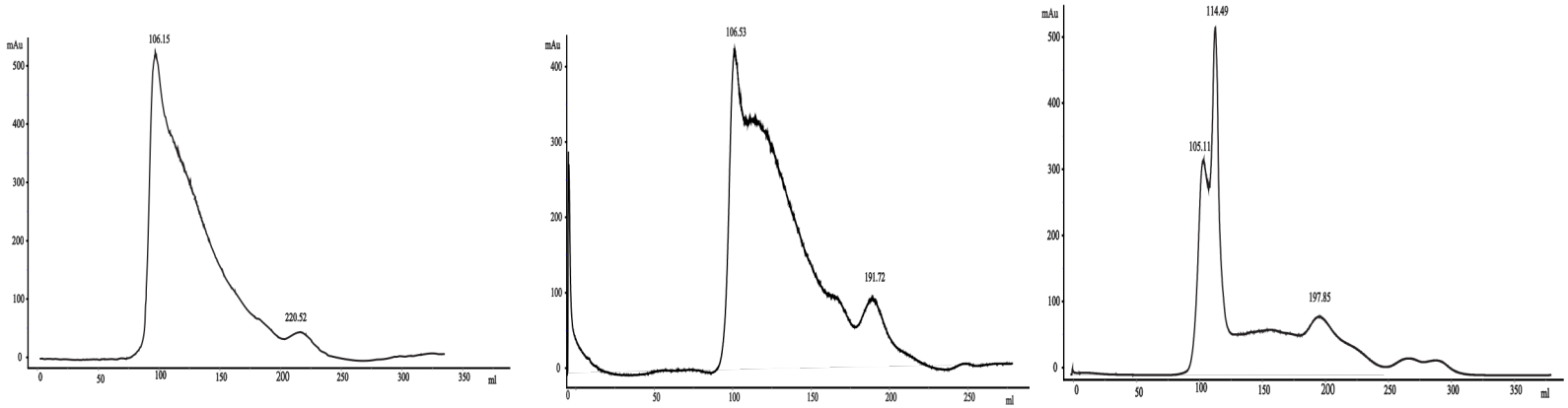


250mM Salt



150mM Salt

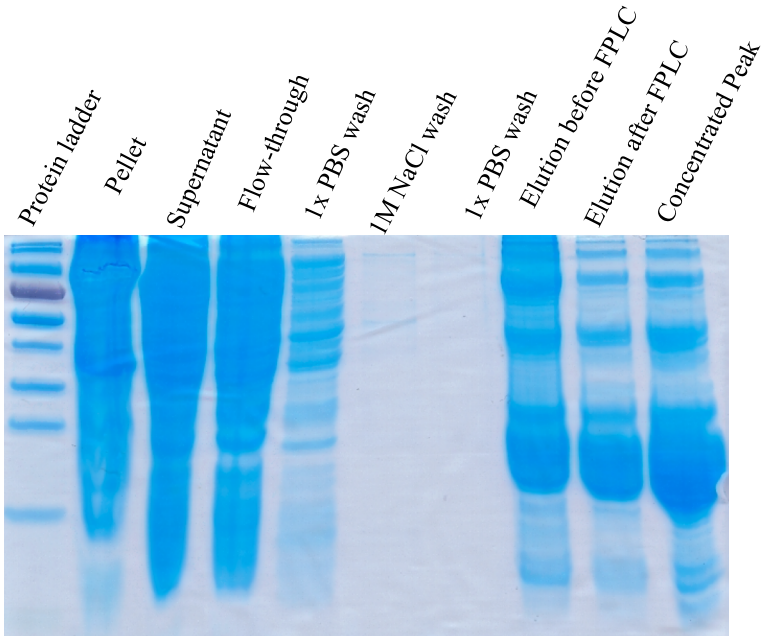
Figure 2.4. FPLC profile with different NaCl salt concentrations in 25mM phosphate buffer to determine the optimum salt concentration needed for the elution of BAZ2A-PHD. The Y-axis is the absorbance in mAu and the X-axis is the volume in ml. The various concentrations of the salt were used at the elution step before loading the protein into the AKTA prime FPLC



250-300mM Imidazole 150-200mM Imidazole 50-100mM Imidazole

Figure 2.5. FPLC profile with different imidazole concentrations in 25mM phosphate buffer to determine the range of imidazole concentration required for the elution of BAZ2A-PHD. The Y-axis is the absorbance in mAu and the X-axis is the volume in ml. The various concentrations of the imidazole were used at the elution step before loading the protein into the AKTA prime FPLC

(A)


Protein ladder
Pellet
Supernatant
Flow-through
1x PBS wash
1M NaCl wash
1x PBS wash
Elution before FPLC
Elution after FPLC
Concentrated Peak

(B)

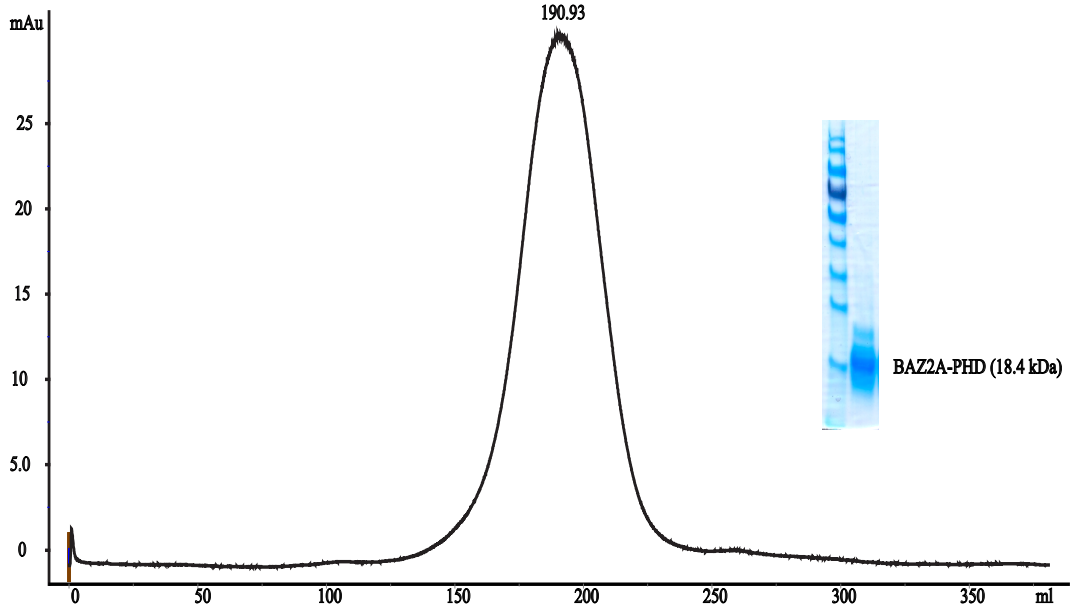

mAu
190.93
25
20
15
10
5.0
0
0
50
100
150
200
250
300
350
ml
BAZ2A-PHD (18.4 kDa)

Figure 2.6. (A) SDS-PAGE gel of hBAZ2A-PHD purification process. The concentrated peak was reloaded on the AKTA prime to get a much more purer protein as seen at B (B) FPLC profile of purified hBAZ2A-PHD in 25mM phosphate buffer, 150mM NaCl, pH of 7.6, and the SDS-PAGE gel representing the peak.

Determination of Protein Concentration by Bradford Method

Table 2.4. Absorbance measurements and concentration of Bovine Serum Albumin (BSA) used for the Bradford standard curve

Tubes	Absorbance 1	Absorbance 2	Absorbance 3	Average Absorbance at 595nm	Concentration (mg/ml)
1	1.347	1.373	1.37	1.36	2
2	1.278	1.254	1.251	1.26	1.5
3	1.118	1.117	1.112	1.12	1
4	0.953	0.938	0.948	0.95	0.75
5	0.8	0.792	0.791	0.79	0.5
6	0.616	0.602	0.606	0.61	0.25
7	0.509	0.494	0.476	0.49	0.125
8	0.366	0.386	0.362	0.37	0

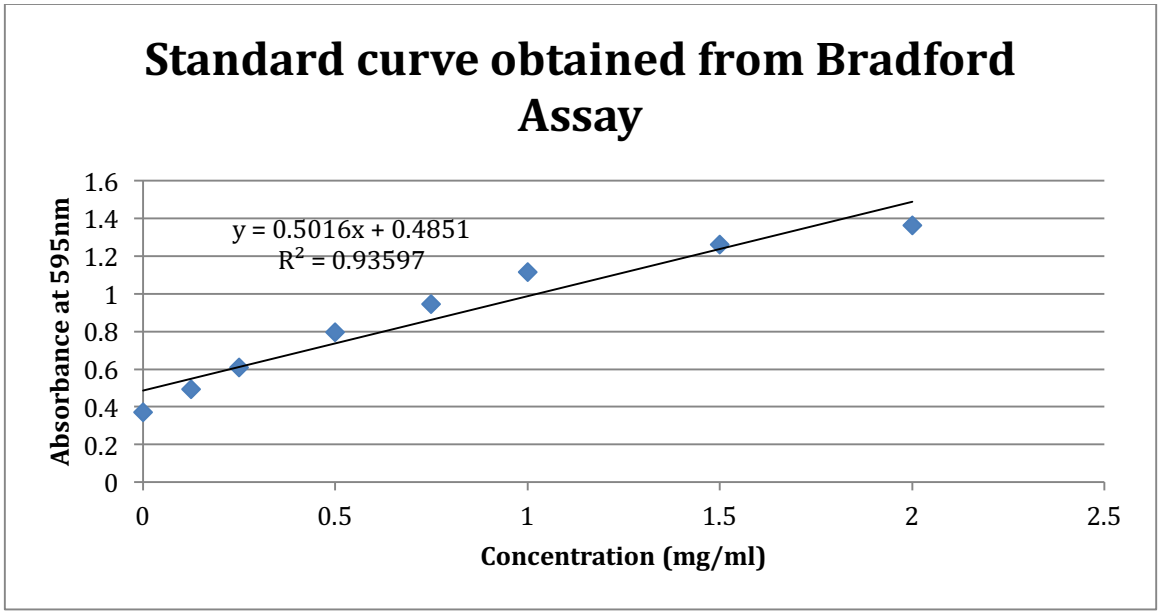


Figure 2.7. Standard curve used for the determination of the protein concentration. It was performed from the Quick Start Bradford Protein Assay protocol where 250µL of the Bradford solution and 5µL of each of

the dilutions made from the standard BSA were used for the absorbance measurements. The equation of the curve was used for the estimation of the protein concentration.

Mutagenesis

Fourteen mutants were made from the hBAZ2A-PHD to change the amino acid residue positions to Alanine with Site-Directed Mutagenesis. hBAZ2A-PHD residues that undergo a change in the accessible surface area with the peptide, $\Delta ASA_{SC}, \geq 10 Å^2$ of the side chain atoms were considered. Distal amino acid residues as well as aspartates in the PHD treble-clef knuckle were also considered for the mutagenesis. Another factor that was also taken into consideration in selecting which amino acid residue to mutate was the distal negatively charged amino acid residues since histone H3 peptide is positively charged hence it was expected that, electrostatics could contribute to the binding.

Primers were designed and ordered from Integrated DNA Technologies (IDT) and Site Directed Mutagenesis performed (Table 2.5 and Table 2.7). Most of the mutants were obtained by using gradient PCR, where varying annealing temperatures were employed to get the mutations to work.

Table 2.5. Mutagenesis reaction cycle: The entire reaction took about 1hour 45minutes

Cycle step	**Temperature**	**Time**	**Cycles**
Initial Denaturation	98 °C	30sec	1
Denaturation Annealing	98 °C Varying	10sec 30sec	 18

Extension	72 °C	1min30sec	
Final Extension Final Hold	72 °C 4 °C	10min ∞	1

Table 2.6. Reagents for Mutagenesis: The reaction was set up on ice and the reagents were added in the order they appear in the table.

Component	**Volume in 50µL Reaction/µL**
Distilled water	35.5
5X Phusion buffer	10
10mM dNTPs	1
Forward Primer	1
Reverse Primer	1
Template DNA (pETite_BAZ2A-PHD)	1
Phusion DNA Polymerase	0.5

Table 2.7. Primers used for the Site Directed Mutagenesis of hBAZ2A-PHD. The letter colored red is the site of mutation

Mutant	**Forward Primer (5'-3')**	**Reverse Primer (5'-3')**	**Protein Sequence**
V1674A	CAGATTGGAGGT TCTGCCAACAAA GTGACCTGC	GCAGGTCACTT TGTTGGCAGAA CCTCCAATCTG	AHREQIGGSANK VTCLVCRKGDND EFLLLCDGCDRGC HIYCHRPKMEAVPE GDWFCTVCLAQQV

D1688A	CGGAAGGGTGAC AATGCTGAGTTT CTTCTGCTT	AAGCAGAAGAA ACTCAGCATTGT CACCCTTCCG	SVNKVTCLVCRKG DNAEFLLLCDGCDR GCHIYCHRPKMEAV PEGDWFCTVCLAQQV
E1689A	AAGGGTGACA ATGATGCGTTT CTTCTGCTTTGT	ACAAAGCAGAA GAAACGCATCA TTGTCACCCTT	SVNKVTCLVCRKG DNDAFLLLCDGCDR GCHIYCHRPKMEAV PEGDWFCTVCLAQQV
L1692A	ACAATGATGAG TTTCTTGCGCTTT GTGATGGGTGTGA	TCACACCCATCA CAAAGCGCAAGA AACTCATCATTGT	SVNKVTCLVCRKG DNDEFLALCDGCDR GCHIYCHRPKMEAV PEGDWFCTVCLAQQV
L1693A	ATGATGAGTTTC TTCTGGCTTGTGA TGGGTGTGACCG	CGGTCACACCCA TCACAAGCCAGA AGAAACTCATCAT	SVNKVTCLVCRKG DNDEFLLACDGCDR GCHIYCHRPKMEAV PEGDWFCTVCLAQQV
D1695A	TTTCTTCTGCTT TGTGCTGGGTGT GACCGTGGC	GCCACGGTCACA CCCAGCACAAAG CAGAAGAAA	SVNKVTCLVCRKG DNDEFLLLCAGCDR GCHIYCHRPKMEAV PEGDWFCTVCLAQQV
D1698A	CTTTGTGATGGG TGTGCCCGTGGC TGCCACATT	AATGTGGCAGCC ACGGGCACACCC ATCACAAAG	SVNKVTCLVCRKG DNDEFLLLCDGCAR GCHIYCHRPKMEAV PEGDWFCTVCLAQQV
V1713A	CCCAAGATGGA GGCTGCCCCAGA AGGAGATTGG	CCAATCTCCTT CTGGGGCAGCC TCCATCTTGGG	SVNKVTCLVCRKG DNDEFLLLCDGCDR GCHIYCHRPKMEAA PEGDWFCTVCLAQQV
E1711A	CATCGTCCCAA GATGGCGGCTG TCCCAGAAGGA	TCCTTCTGGGA CAGCCGCCATCT TGGGACGATG	SVNKVTCLVCRKG DNDEFLLLCDGCDR GCHIYCHRPKMAAV PEGDWFCTVCLAQQV
E1715A	ATGGAGGCTGTC CCAGCAGGAGAT TGGTTCTGT	ACAGAACCAAT CTCCTGCTGGGA CAGCCTCCAT	SVNKVTCLVCRKG DNDEFLLLCDGCDR GCHIYCHRPKMEAV PAGDWFCTVCLAQQV
V1728A	GTCTGTTTGGCTC AGCAGGCGTAAT AGAGCGGCCGCCAC	GTGGCGGCCGCTC TATTACGCCTGCT GAGCCAAACAGAC	SVNKVTCLVCRKG DNDEFLLLCDGCDR GCHIYCHRPKMEAV PEGDWFCTVCLAQQA SSSGRHR
D1686A	GTCTGCCGGAAG GGTGCCAATGAT GAGTTTCTT	AAGAAACTCATC ATTGGCACCCTTC CGGCAGAC	SVNKVTCLVCRKG ANDEFLLLCDGCDR GCHIYCHRPKMEAV PEGDWFCTVCLAQQV

D1717A	GCTGTCCCAGAAG GAGCTTGGTTCTG TACTGTC	GACAGTACAGAA CCAAGCTCCTTC TGGGACAGC	SVNKVTCLVCRKG DNDEFLLLCDGCDR GCHIYCHRPKMEAV PEGAWFCTVCLAQQV
N1675A	AGATTGGAGGTTC TGTCGCCAAAGTG ACCTGCCTGGT	ACCAGGCAGGTCA CTTTGGCGACAGAA CCTCCAATCT	HREQIGGSVNKVTCLV CRKGDNDEFLLLCDGC DRGCHIYCHRPKMEAV PEGAWFCTVCLAQQV

After the mutagenesis reaction, Dpn I digestion was done at 37°C for 2hours to destroy the template (hBAZ2A-PHD wild type DNA). Enzyme inactivation of the Dpn I was done at 80°C for 20mins to inactivate the enzyme after digesting the parental supercoil dsDNA. The mutants were transformed into XL 10-gold competent cells and then incubated on a kanamycin agar plate at 37°C for 14hours. Colonies obtained on the kanamycin agar plates were cultured and allowed to grow at 37°C for 18hours in a kanamycin/LB media and then miniprep performed according to the QIAprep spin miniprep kit protocol (QIAGEN). 1% DNA agarose gel was prepared in 1x TAE buffer; pH 8.0 and the samples run to determine the presence of the mutants (Figure 2.8). The samples were then packaged and sent for sequencing (Genscript) to determine the right nucleotide sequence of each of the mutants.

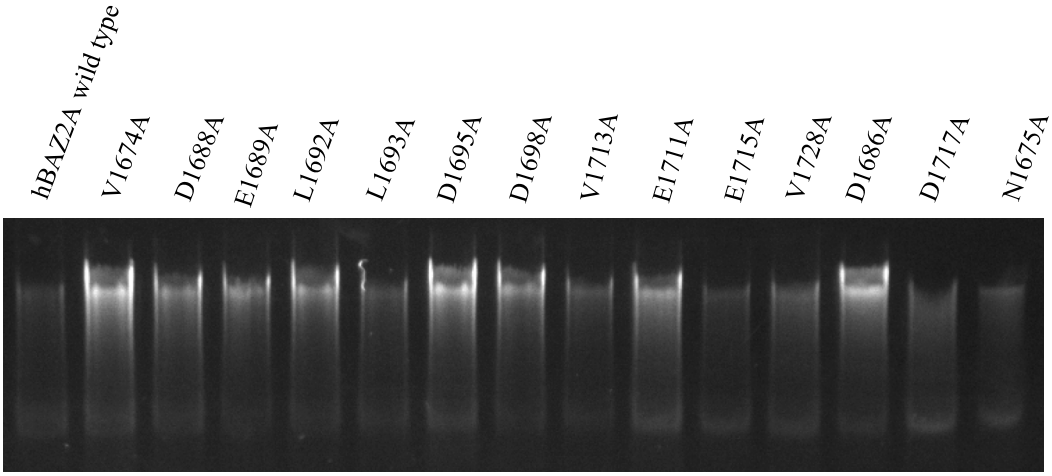


Figure 2.8. Agarose gel electrophoresis of hBAZ2A mutants before DNA samples were sent for DNA sequencing. Lane 1 (hBAZ2A wild type) was run as marker to confirm the size and behavior of the plasmid in the agarose gel.

Transformation, Expression and Purification of Mutants

After confirming the DNA sequences of all the mutants, they were transformed into HI-Control® BL21 (DE3) SOLOs Chemically Competent cells (see transformation). Solubility test and small-scale expression was done to select the best colonies for all the mutants. 4-liter culture was prepared for each mutant and purified as described above for the purification of the wild type.

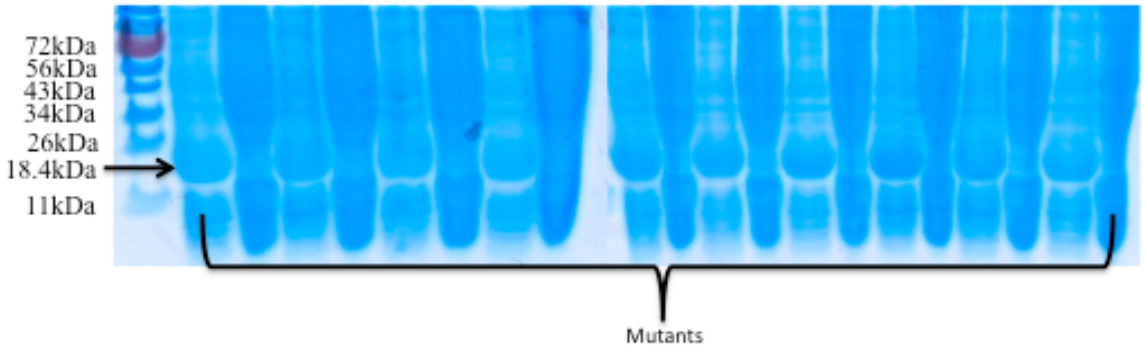


Figure 2.9. SDS-PAGE gel showing the expression of some of the mutants of hBAZ2A-PHD: For each of them, an induced and an uninduced sample was run to clearly show that proteins are being expressed. The protein was run with a 12% SDS-PAGE gel in a 1x running buffer with a 120 voltage for 60 minutes.

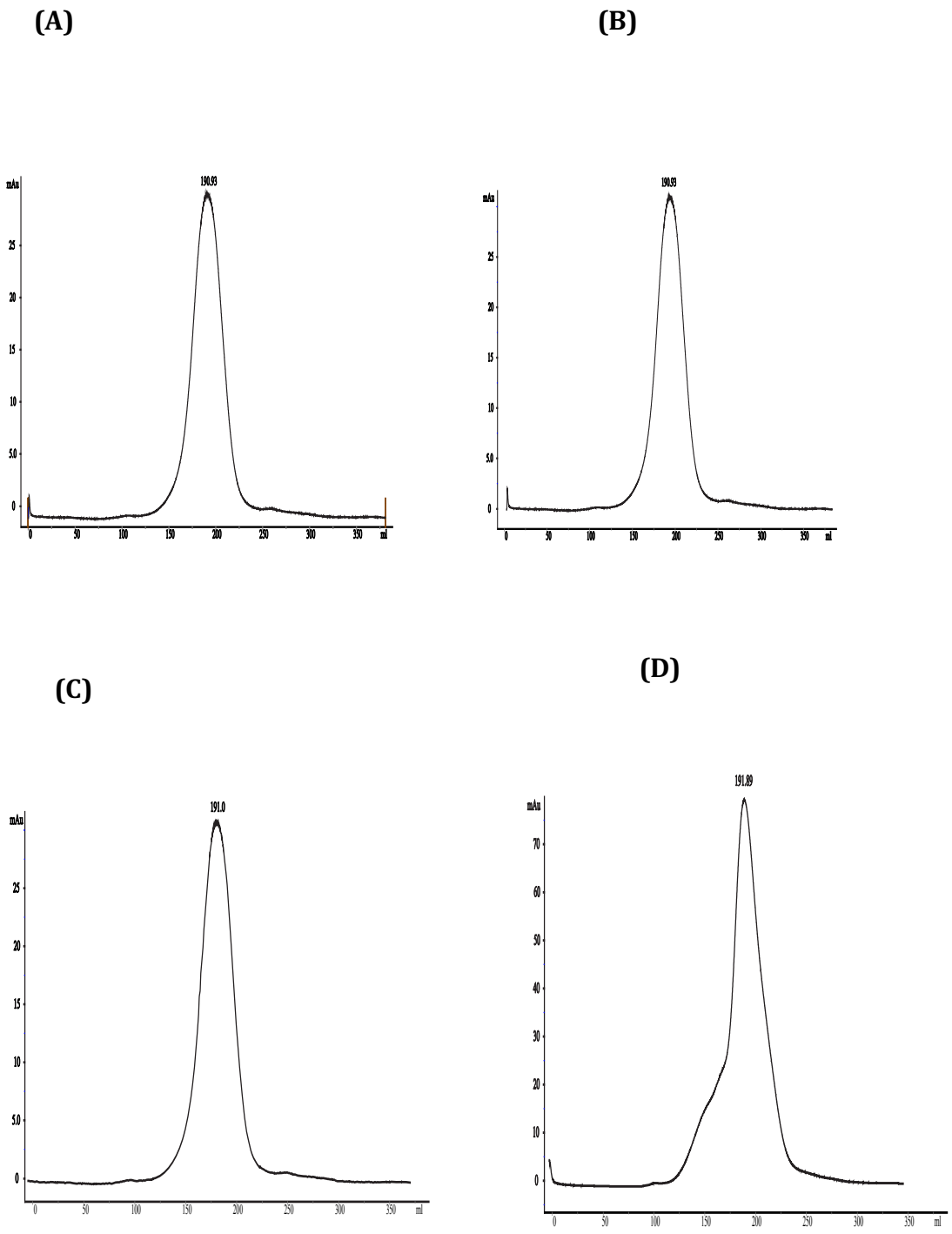


Figure 2.10 (A-D). Chromatograms showing the FPLC profile peaks of purified V1674A, D1688A, E1689A and L1692A respectively. They were all run at 4°C in 25mM phosphate buffer (150mM NaCl salt, pH of 7.6). The tubes corresponding to these peaks were concentrated using an ultra-filtration tube with a 10kDa molecular weight cut off.

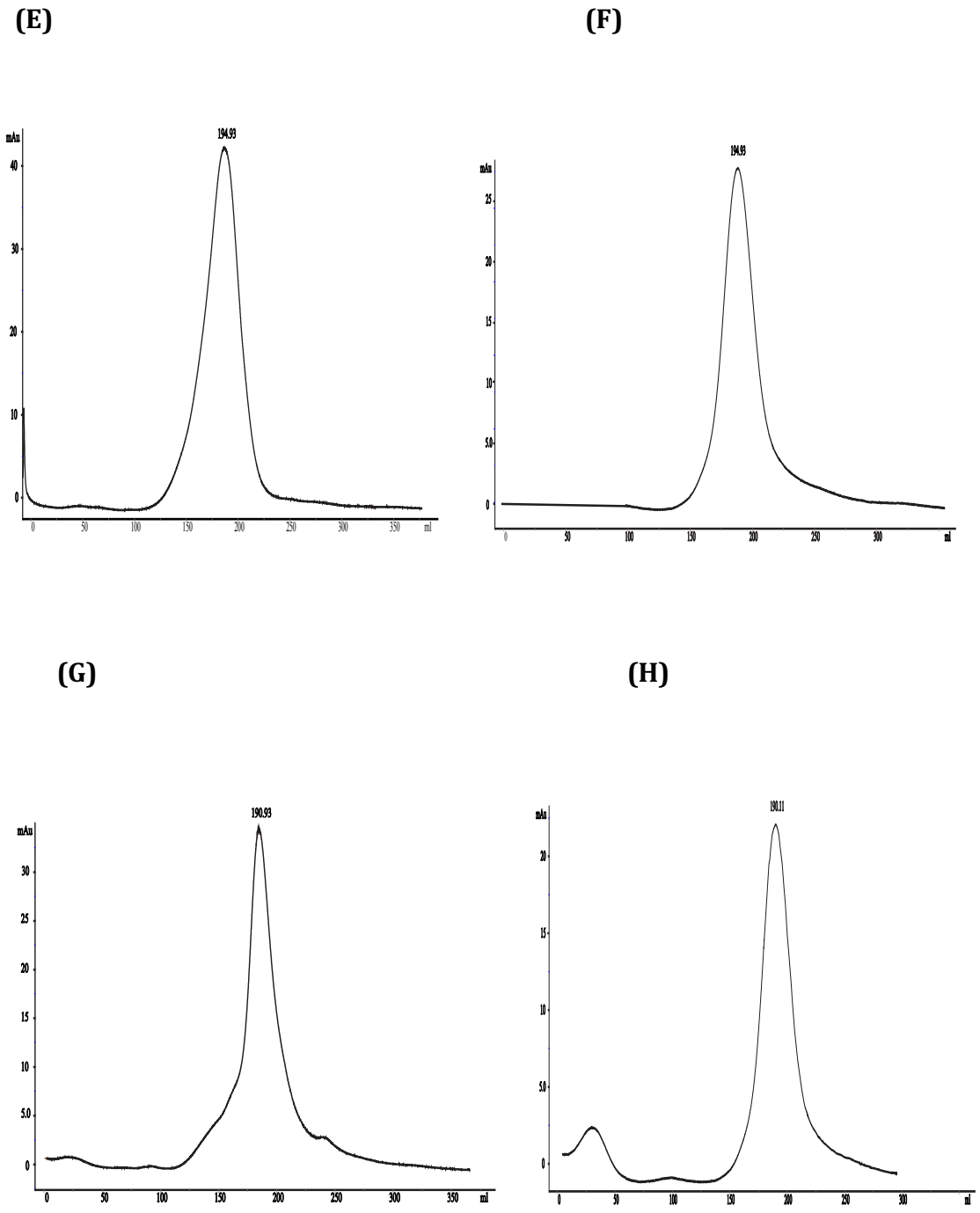
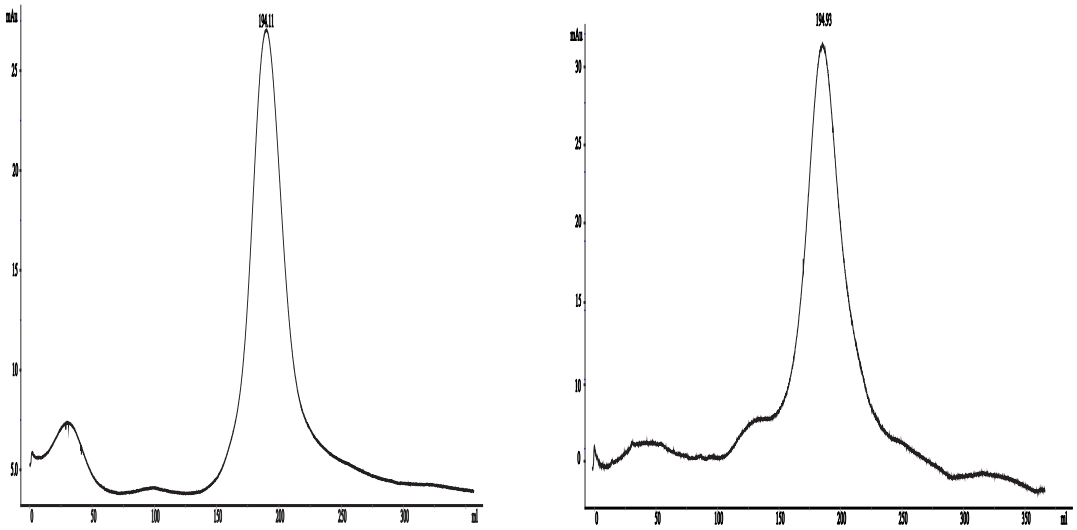


Figure 2.10 (E-H). Chromatograms showing the FPLC profile peaks of purified L1693A, D1695A, D1698A and V1713A respectively.

(I) **(J)**



(K)

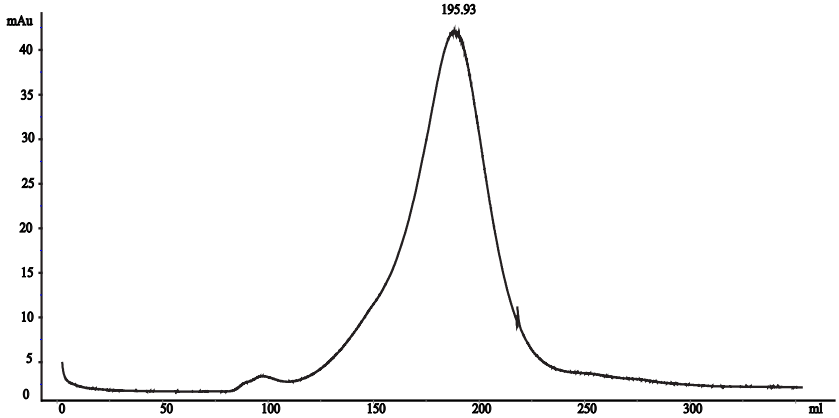
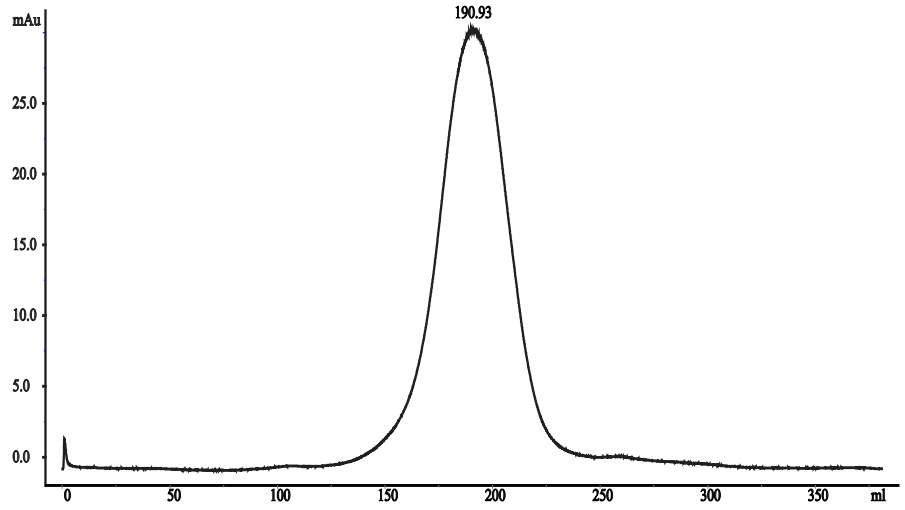


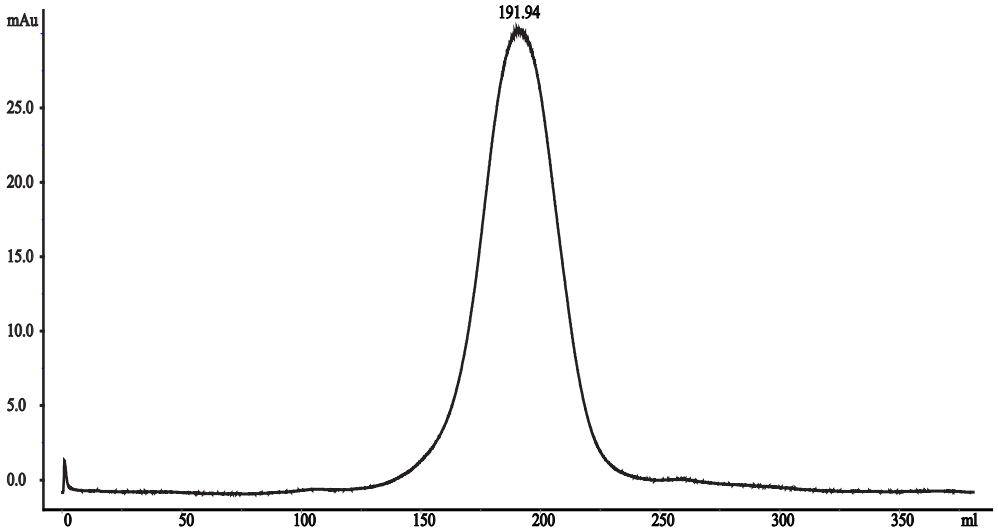
Figure 2.10 (I-K). Chromatograms showing the FPLC profile peaks of purified E1711A, E1715A, and V1728A respectively. All the proteins were eluted with 100ml of 150mM imidazole in 1x phosphate buffer

and concentrated to 4ml before loading into the AKTA prime FPLC at 4°C. The column used was the HiLoad 26/600 superdex column. The tubes corresponding to the peaks were concentrated to 0.1mM for the ITC measurement.

(L)



(M)



(N)

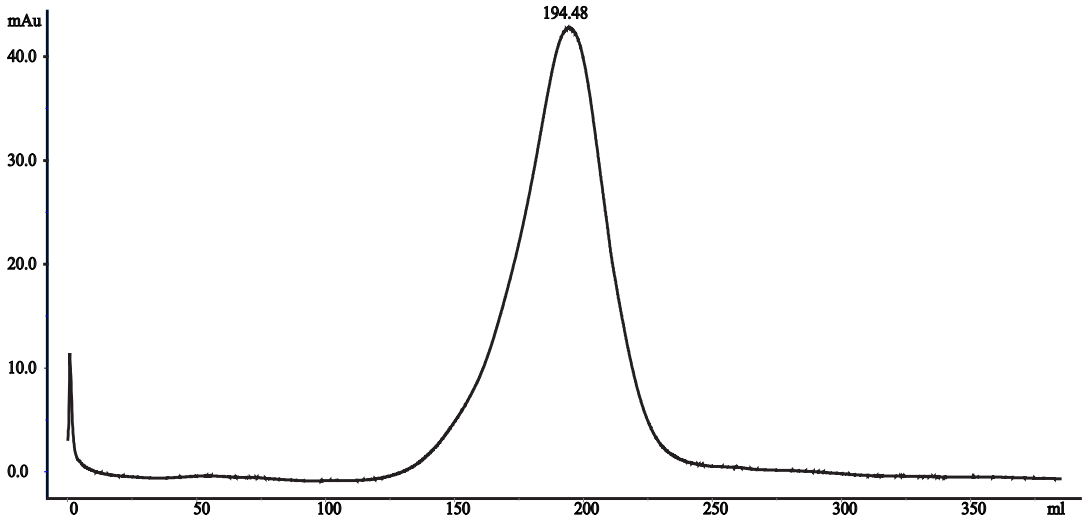


Figure 2.10 (L-N). Chromatograms showing the FPLC profile peaks of purified D1686A, D1717A, and N1675A respectively.

Synthetic Peptides

98% pure Histone H3 synthetic peptide with residues 1-11 (H3-1-11W) was ordered from Genscript for the binding studies of the wild type and the mutants of hBAZ2A-PHD proteins. The W in the synthetic peptide sequence represents C-terminal Tryptophan residue. H3-1-11W. The C-terminal Tryptophan residue in the peptide was used for estimating peptide concentrations by absorbance using the computed molar extinction coefficient[66], which is $5.5M^{-1}cm^{-1}$. The purity of the Histone H3-1-11W peptide was confirmed with mass-spectroscopy.

Synthetic Peptide and Recombinant Protein Concentration Measurements

The wild type and mutant proteins of hBAZ2A-PHD from the FPLC HiLoad 26/600 superdex column were concentrated using the Amicon Ultra-15 Centrifugal Filter Units (Millipore) with a 10kDa molecular weight filter since the size of the recombinant proteins was 18.4kDa. The concentration was determined by using the UV-Vis Nanodrop Spectrophotometer to measure the absorbance of the proteins at 280nm (which is the absorption wavelength for tryptophan) and then the concentration calculated by using the molar extinction coefficient obtained from ExPaSy ProtParam tool (a bioinformatics software for computing the chemical and physical properties of proteins). The molar extinction coefficients of the hBAZ2A-PHD proteins and Histone H3 1-11W peptide are 8.48 $M^{-1}cm^{-1}$ and 5.5 $M^{-1}cm^{-1}$ respectively.

The concentrations were confirmed by using the Bradford Protein Assay standard curve (Figure 2.7) that was generated by the using Quick Start Bradford Protein Assay protocol.

Isothermal Titration Calorimetry (ITC)

The peptides and recombinant proteins (wild type hBAZ2A and mutants) for ITC were both prepared in the same 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 (0.1-0.15 mM) and peptide (10 fold higher) respectively loaded in the cell and syringe at 25°C. Twenty 2µl-injections with a 3-minute injection-interval, with a syringe stirring speed of 1000 rpm were used for the titrations. $\Delta G°$ of peptide binding at 25°C was computed as $-RT\ln(Ka)$ where R, T and Ka are respectively the gas constant, temperature and association

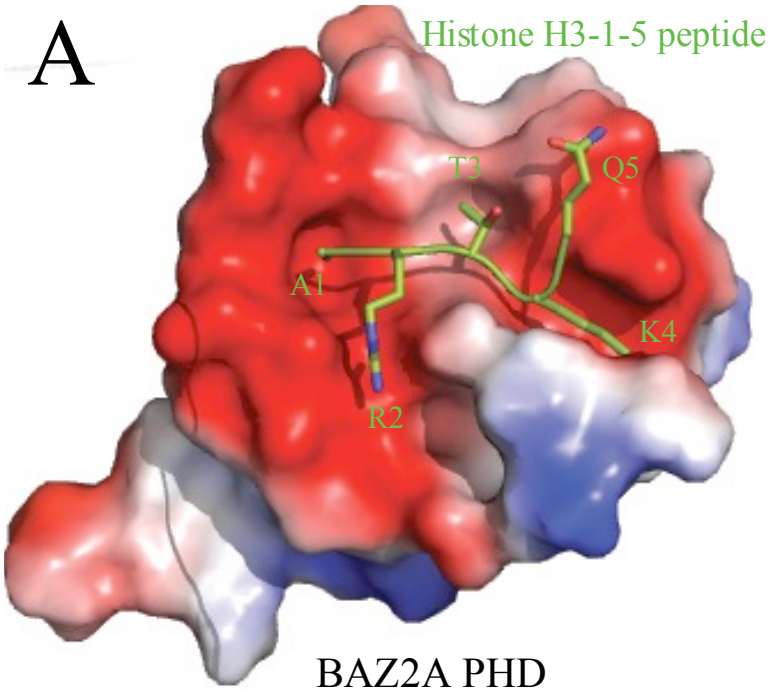
constant. $\Delta\Delta G°$ for a residue's energetic contribution is estimated as $\Delta\Delta G° = \Delta G°_{mutant} - \Delta G°_{wildtype.}$ All titrations were carried out in identical conditions of buffer and temperature. In the case of many of the mutants, no binding was observed (i.e. negligible amount of observed heats) and data fitting was unreliable to the observed enthalpy. In such cases it was assumed that $\Delta\Delta G°$ (>> 1.5 kcal.mol^{-1}) was large. Tryptophan residue of the PHD finger was used for estimating protein concentrations by absorbance using the computed molar extinction coefficient[66]. Titrations were also performed at 250 mM sodium chloride buffer (25 mM phosphate at pH=7.6) to check the influences of electrostatics on the protein-peptide interactions. Binding measurements (titrations) for each protein was run in triplicate to confirm the data.

RESULTS AND DISCUSSIONS

Contribution of electrostatics on peptide binding

Histone H3 peptide is positively charged, and there is a considerable distribution of negatively charged residues around the binding surface of BAZ2A-PHD finger (Figure 2.11A). It is likely that electrostatics would influence peptide-binding interactions. For a qualitative estimate of the influence of electrostatics, we carried out experiments at both 150mM NaCl salt and 250mM NaCl salt for the wild type protein. It was observed that, by increasing the salt concentration by 100mM, the free energy (ΔG) of peptide binding at 25°C becomes unfavorable by ~0.7kcal/mol (or 3.35 fold affinity drop) confirming the role of electrostatics in the overall binding interactions (Figure 2.11A, 2.12). This suggests that the binding is driven by electrostatics hence higher salt concentration

shields the electrostatics. Thus, in selecting the control residues, among several other possible residues, distal negatively charged residues (E1711 and E1715) were preferred. The peptide in complex with BAZ2A-PHD possesses three positive charged centers, which are normally involved in the electrostatic interaction. These centers are: (i) the N-terminal NH_3^+ -group, (ii) the CZ- atom of Arginine-2 and (iii) the NZ- atom of Lysine-4.



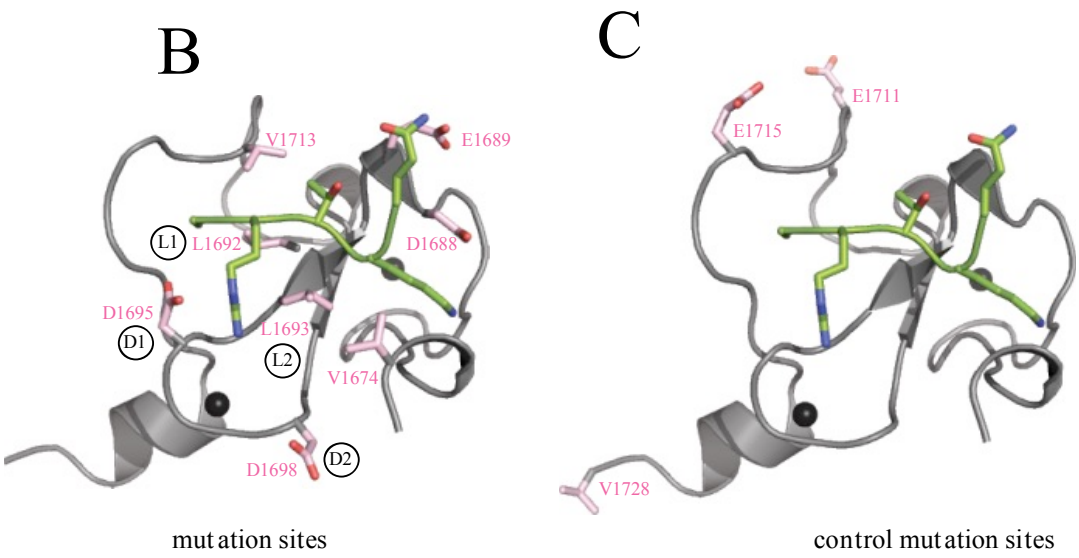
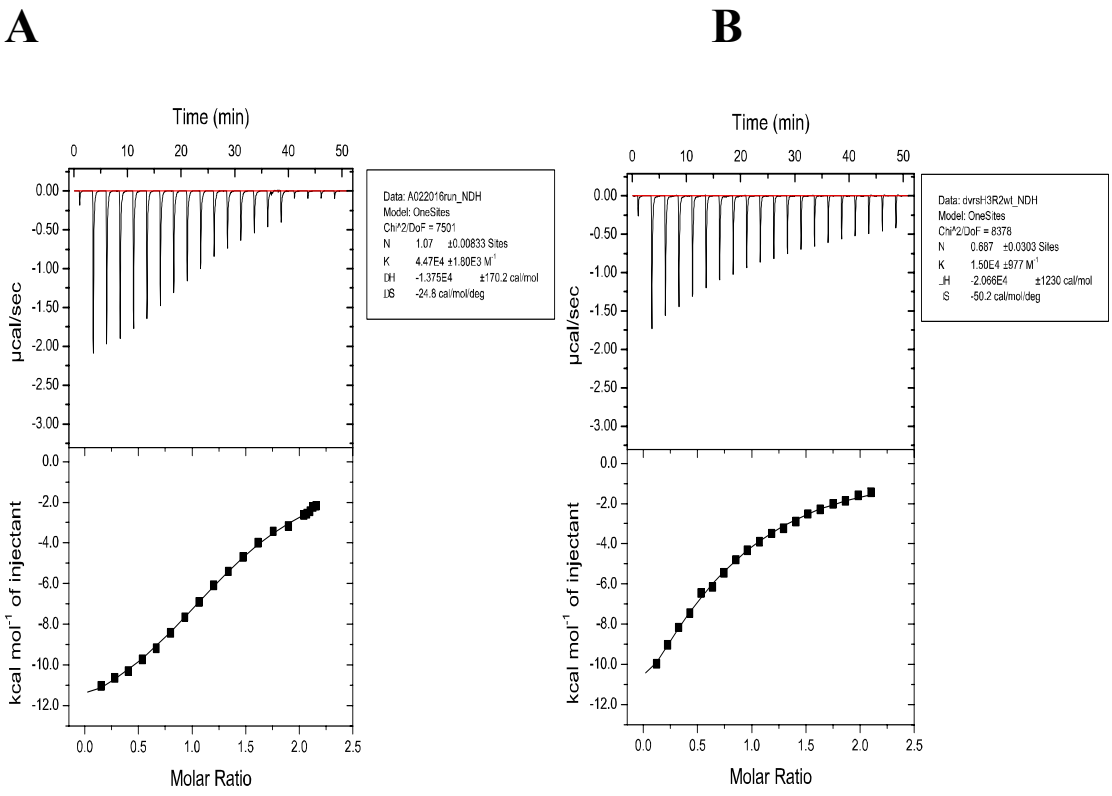


Figure 2.11. (A) Electrostatic potential (isocontour value of ± 76.3 kT/e) surface representation of the BAZ2A PHD bound to the H3-1-5 residue peptide (splitpea), (B) Side-chain atoms of BAZ2A-PHD residues that are mutated to Alanine in this study are represented in sticks (light-pink). Positions of interest are labeled as L1, L2, D1, and D2. (C) Side-chains of positions of control mutations are also represented in sticks.



C

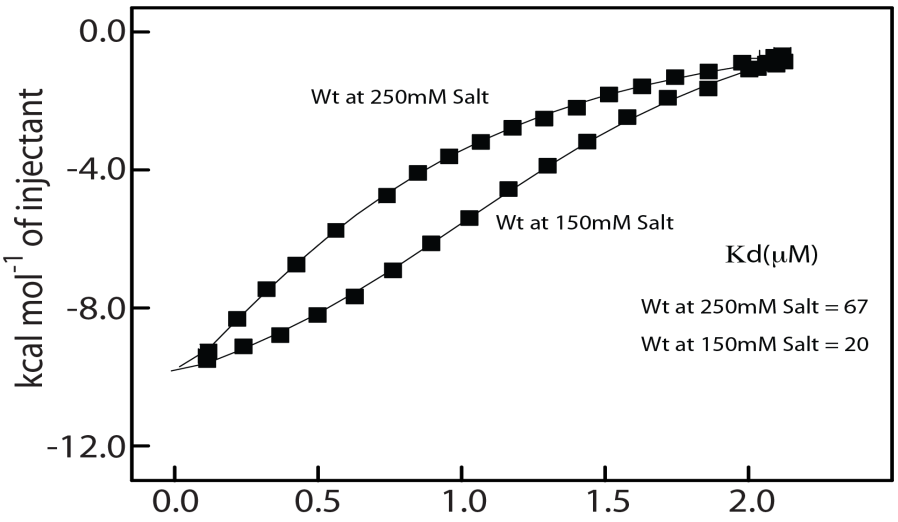


Figure 2.12. ITC binding studies between histone H3 unmodified N-terminal peptide and wild type of BAZ2A-PHD at (A) 150mM NaCl Salt and (B) 250mM NaCl salt. (C) Overlayed titration profiles of the wild types at different salt concentrations. The Kd is reduced by 3.35 fold by changing the salt from 250mM to 150mM.

Dissecting the energetic contributions of BAZ2A-PHD residues

The recently reported 1.9-Å X-ray structure of BAZ2A-PHD[47](*pdb code* 4q6f) in complex with the unmodified histone H3 N-terminal peptide (Figure 2.11A) is used here as the starting point for inferring the energetic contributions of BAZ2A-PHD residues towards peptide binding. BAZ2A-PHD residues that undergo a change in the accessible surface area, ΔASA, ≥ 10Å^2 of the side chain atoms are considered for probing the energetic contributions (Figure 2.11B, *see* Methods) by mutagenesis. Seven BAZ2A-PHD residues lose ~15Å^2 or more of their side-chain surface area upon complex formation, and we consider these residues as the primary contacts for histone H3 peptide (Figure 2.11B). However, D1698 (ΔASA = 0) is also included in the list of residues for mutagenesis (Figure 2.11B). This is because, in an earlier study[7], it had been noted that

D1698 of BAZ2A is the second of the pair of Asp residues in the PHD treble-clef knuckle xCDxCDx sequence pattern, a feature present in a distinct PHD subgroup that includes the *d*ouble *P*HD *f*inger (DPF or also referred as the tandem-PHD-PHD) type of modules (Figure 3.1). To probe the binding energetics by mutagenesis, we also included three *control* residues (E1711, E1715 and V1728) that are either distal from the peptide-binding site ($\Delta ASA \cong 0$) (Figure 2.11A, C). The selected eleven positions were substituted to Alanine, and mutants were then purified for probing histone H3 peptide binding energetics by isothermal titration calorimetry (ITC) (Figure 2.13-2.14 and Table 2.8). For comparison of the wild type protein's peptide-binding behavior with that of the rest of the mutants, ΔG and K_d at 25°C are summarized in table 2.8. We then probed the eleven mutants to learn if factors in addition to electrostatics would contribute to the overall binding energetics.

Mutations at three positions (L1692A, L1693A and D1695A) completely aborted peptide binding, i.e. titrations with these mutants did not show any detectable enthalpy change (Figure 2.13, 2.14(D-F) and Table 2.8). Based on the binding enthalpy, for the rest of the mutants, we observed a range of binding behaviors (Figure 2.13-2.14, Table 2.8). However, none of them completely disrupted peptide binding. Thus, we anticipate that L1692A, L1693A and D1695A make very large energetic contributions towards peptide binding, and these residues likely dominate the peptide binding energetics. It is interesting to note two nonpolar residues (L1692 and L1693) are among the residues dominating the binding energetics between the two oppositely charged interacting partners. We discuss the details of the binding energetic observations of BAZ2A-PHD finger in the context of the PHD superfamily in the next chapter.

Table 2.8. (A) A listing of the structural (ΔASA) and thermodynamic (K_a, K_d, ΔG^o) properties of the mutated BAZ2A-PHD positions listed in the study

	Mutants	ΔASA_{SC}(Å^2)	K_a (M^{-1})	K_d (μM)	ΔG° (kcal/mol)†
	BAZ2A WT	---	$4.47x10^4$	22.37	-6.31
	V1674A	21.71	$4.94x10^4$	20.24	-6.37
	D1688A	42.66	$5.26x10^4$	19.01	-6.41
	E1689A	43.71	$7.44x10^4$	13.44	-6.62
L1	**L1692A**	14.52	X	X	X
L2	**L1693A**	56.83	X	X	X
D1	**D1695A**	35.41	X	X	X
D2	**D1698A**	0.00	$1.41x10^4$	70.92	-5.64
	V1713A	31.68	$0.69x10^4$	144.92	-5.21
	BAZ2A WT (250 mM NaCl)	---	$1.50x10^4$	67.0	-5.70
Control Mutants	**E1711A**	0.00	$3.79x10^4$	26.39	-6.22
Control Mutants	**E1715A**	0.68	$0.90x10^4$	111.11	-5.37
Control Mutants	**V1728A**	0.00	$4.78x10^4$	20.92	-6.36

† ITC at 25°C and 150 mM Nacl

Table 2.8. (B) A listing of the structural (ΔASA) and thermodynamic (K_a, K_d, ΔG^o) properties of the mutated BAZ2A-PHD positions that are remote from the peptide binding site (distal site) but disrupts peptide binding.

Mutant	**ΔASA_{SC} (Å^2)**	**K_a (M^{-1})**	**K_d (μM)**	**ΔG°(Kcal/mol)**
D1686A	0.00	X	X	X
D1717A	0.95	X	X	X
N1675A	20.23	X	X	X

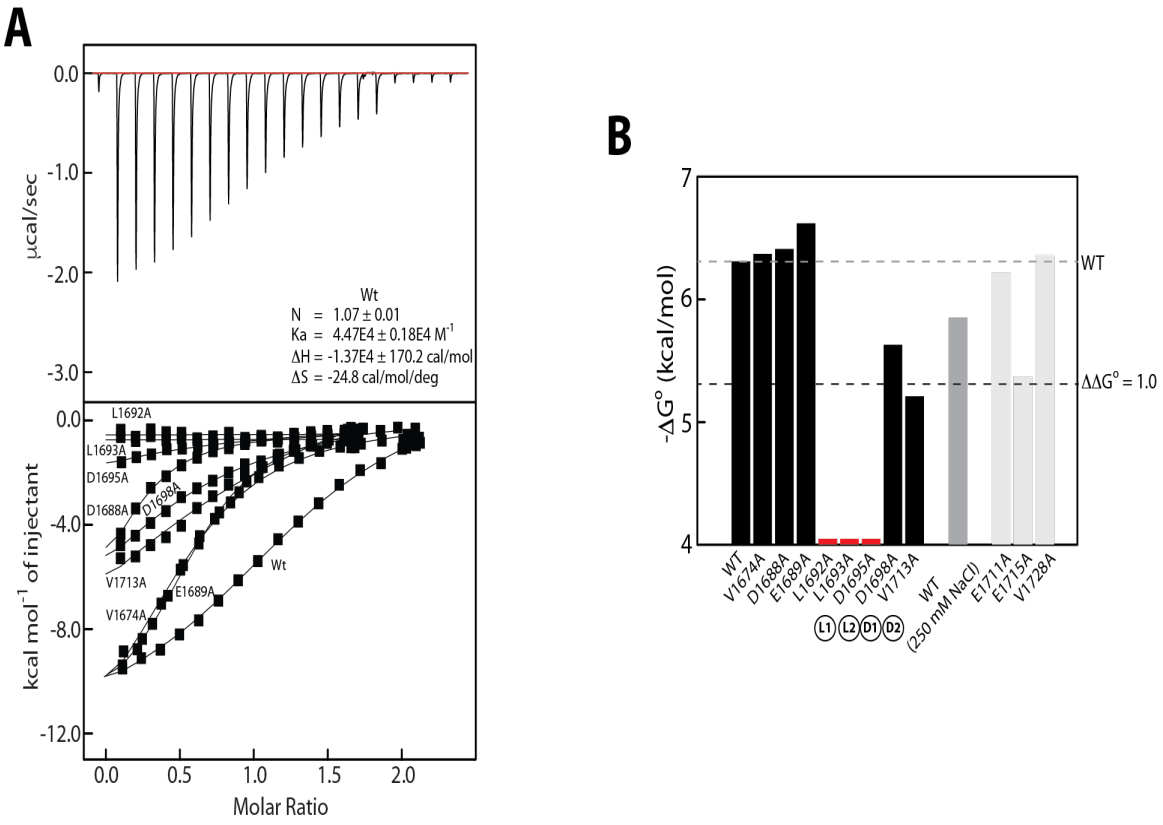
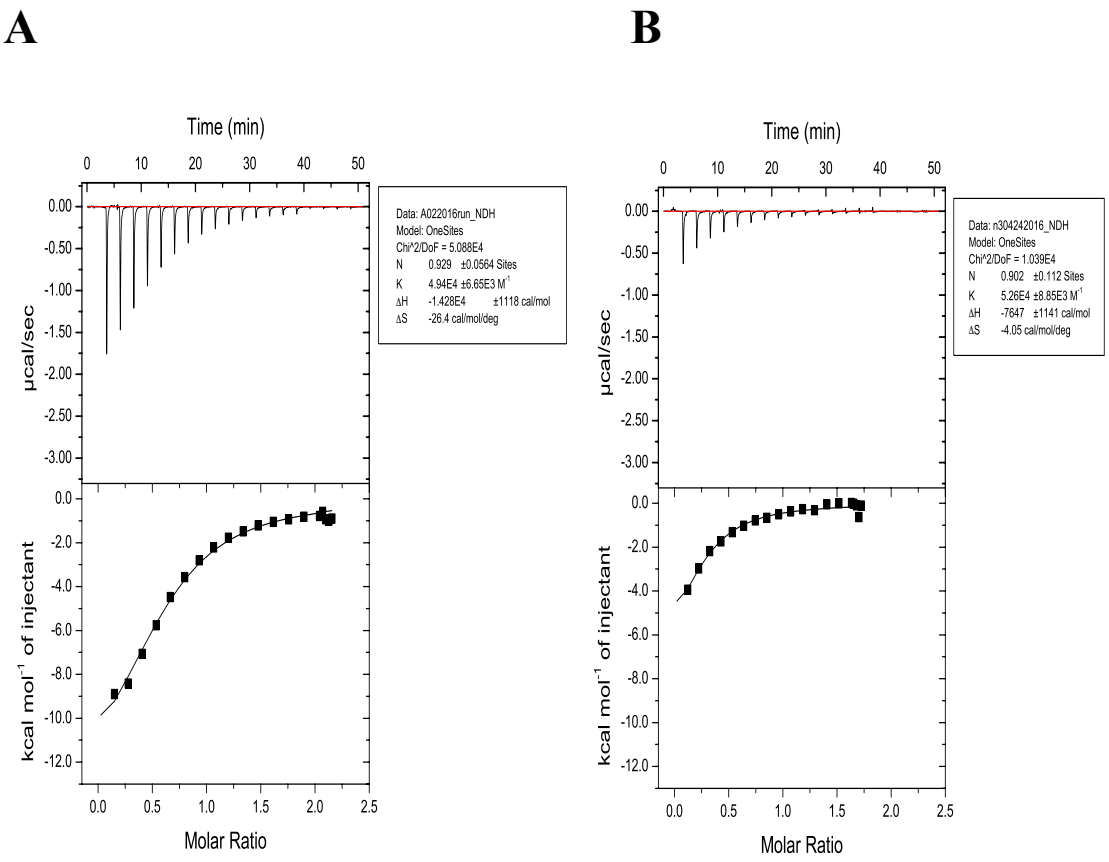
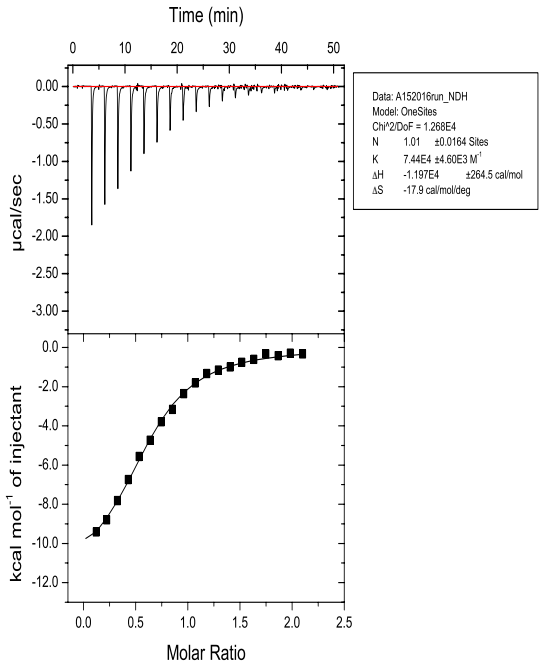


Figure 2.13 (A) ITC binding studies between histone H3 unmodified N-terminal peptide, and wild type, mutants of BAZ2A-PHD. A representative isothermal calorimetric titration (top) of H3-1-9 peptide (syringe) into protein (cell) is shown for the wild type protein while exothermic heats (bottom) exchanged

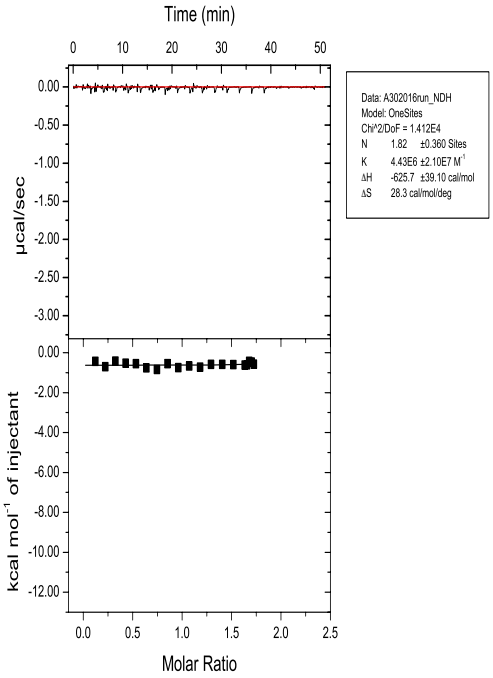
per mol of injectant as a function of the molar ratio of peptide to protein is overlayed for all titrations for convenience. (B) ΔG^o of the forward binding reaction for the wild type and mutants are shown as bar. For convenience $-\Delta G^o$ is shown. Residues that contribute more than 1 kcal/mole of free energy are below the black dotted line. Red bars for positions where binding was not detected in ITC experiments.



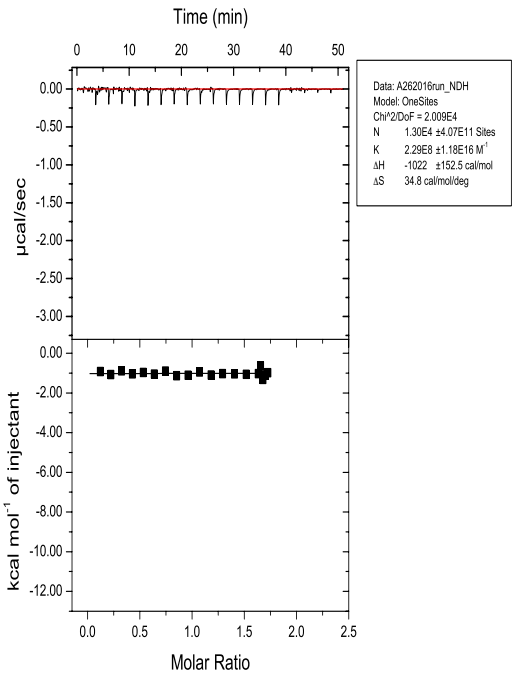
C


Time (min)
µcal/sec
kcal mol^{-1} of injectant
Molar Ratio
Data: A152016run_NDH
Model: OneSites
Chi^2/DoF = 1.268E4
N 1.01 ±0.0164 Sites
K 7.44E4 ±4.60E3 M^{-1}
ΔH -1.197E4 ±264.5 cal/mol
ΔS -17.9 cal/mol/deg

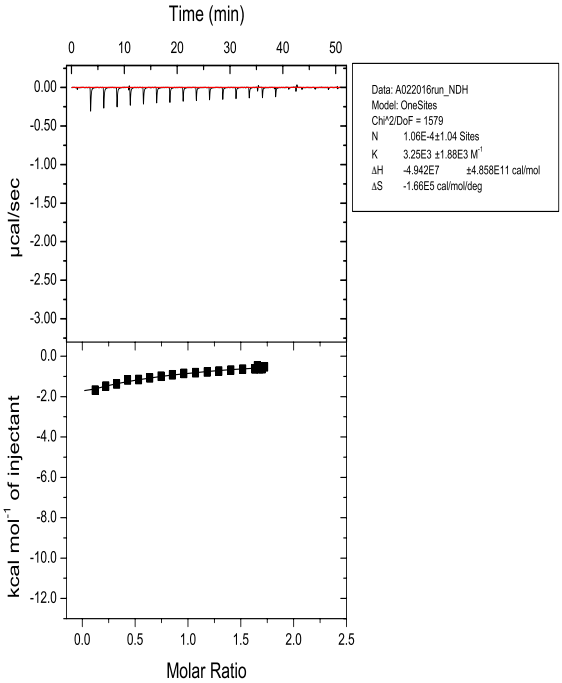
D


Time (min)
µcal/sec
kcal mol^{-1} of injectant
Molar Ratio
Data: A302016run_NDH
Model: OneSites
Chi^2/DoF = 1.412E4
N 1.82 ±0.360 Sites
K 4.43E6 ±2.10E7 M^{-1}
ΔH -625.7 ±39.10 cal/mol
ΔS 28.3 cal/mol/deg

E


Time (min)
µcal/sec
kcal mol^{-1} of injectant
Molar Ratio
Data: A262016run_NDH
Model: OneSites
Chi^2/DoF = 2.009E4
N 1.30E4 ±4.07E11 Sites
K 2.29E8 ±1.18E16 M^{-1}
ΔH -1022 ±152.5 cal/mol
ΔS 34.8 cal/mol/deg

F


Time (min)
µcal/sec
kcal mol^{-1} of injectant
Molar Ratio
Data: A022016run_NDH
Model: OneSites
Chi^2/DoF = 1579
N 1.06E-4±1.04 Sites
K 3.25E3 ±1.88E3 M^{-1}
ΔH -4.942E7 ±4.858E11 cal/mol
ΔS -1.66E5 cal/mol/deg

G **H**

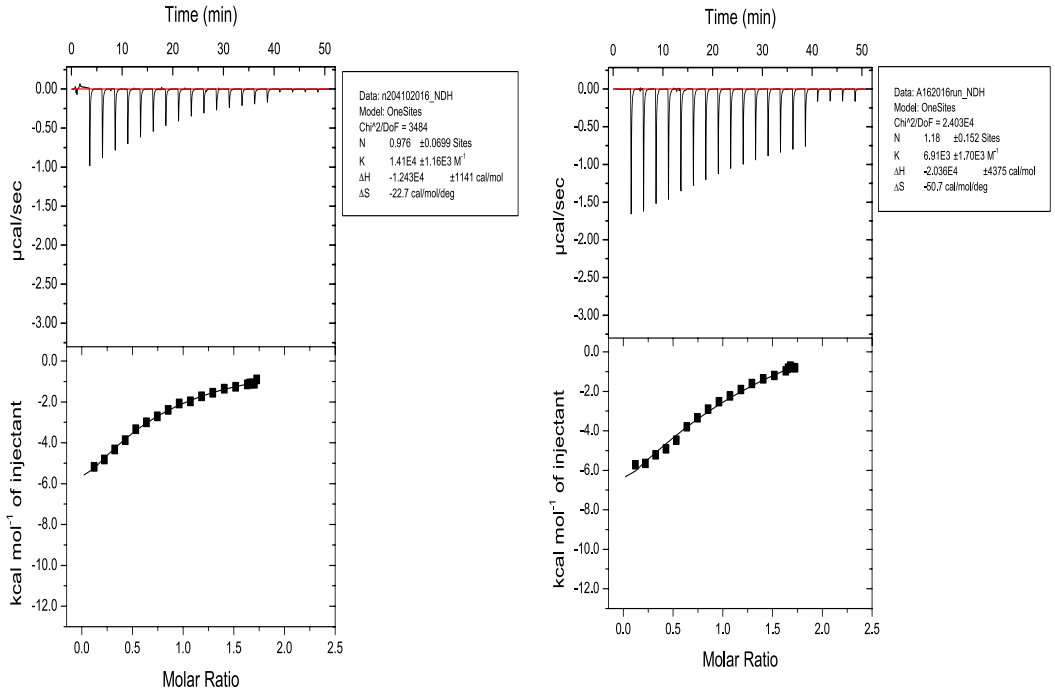


Figure 2.14.(A-H) ITC binding studies between histone H3 unmodified N-terminal peptide and mutants of BAZ2A-PHD. A representative isothermal calorimetric titration (top) of H3-1-9 peptide (syringe) into protein (cell) is shown for the mutant proteins while exothermic heats (bottom) exchanged per mol of injectant as a function of the molar ratio of peptide to mutant proteins. From A to H are V1674A, D1688A, E1689A, L1692A, L1693A, D1695A, D1698A and V1713A respectively.

Contributions of negatively charged residues at BAZ2A interface

Since BAZ2A belongs to a subtype that is characterized by the presence of a pair of negatively charged residues at the treble clef knuckle, we first wanted to understand the contributions of the BAZ2A D1695 (**D1** position) and D1698 (**D2** position) residues as a first step so we can better understand in the context of subtype structure in the next chapter. The substitution of D1695 severely compromised binding. Side-chain carboxylate at **D1** is engaged in a salt bridge mediated interaction with the H3Arg2

residue. The complete loss of binding upon D1695 substitution was expected due to the loss of the electrostatic interaction between D1695 and H3Arg2, and thus D1695 has a very large energetic contribution. ΔΔG for BAZ2A D1698 (**D2** position) was however, 0.67 kcal/mole (Figure 2.13, 2.14F-G and Table 2.8). The contribution of D1698 (**D2** position, 0.67 kcal/mole) could be less than that of the **D2** position of other members because of a structural deviation of the treble clef xCDxCDx knuckle in BAZ2A PHD noted in Figure 2.15. In comparison with what is typically observed in other PHD fingers, the structural deviation of the knuckle places BAZ2A D1698 farther away from the H3Arg2 residue. This structural deviation is also reflected in the loss of the surface area upon complex formation.

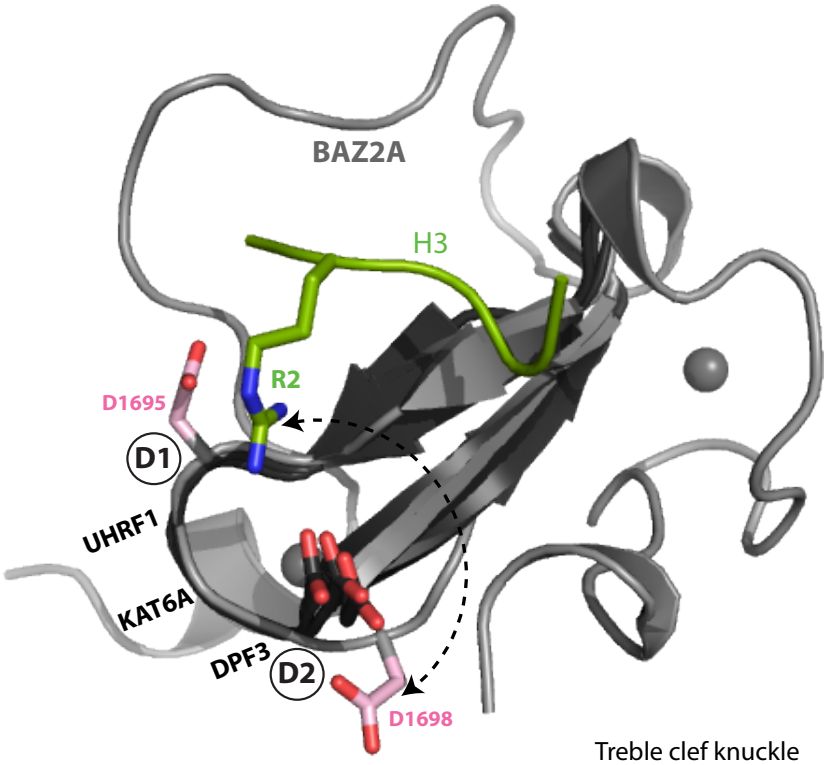
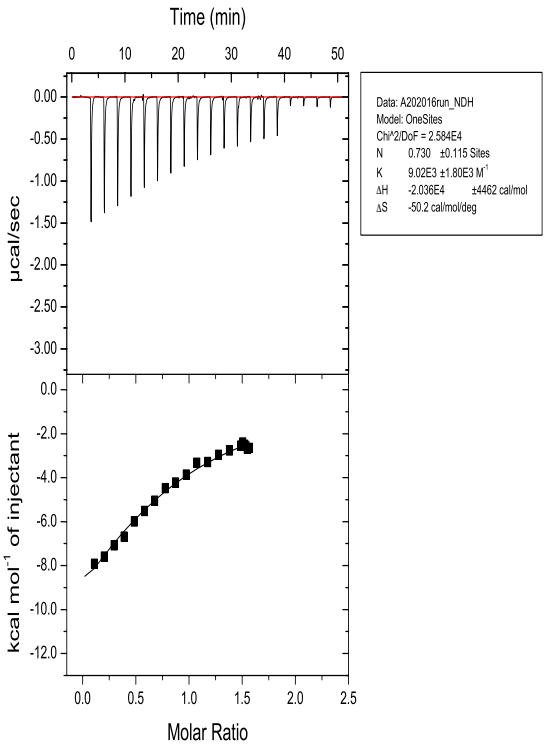


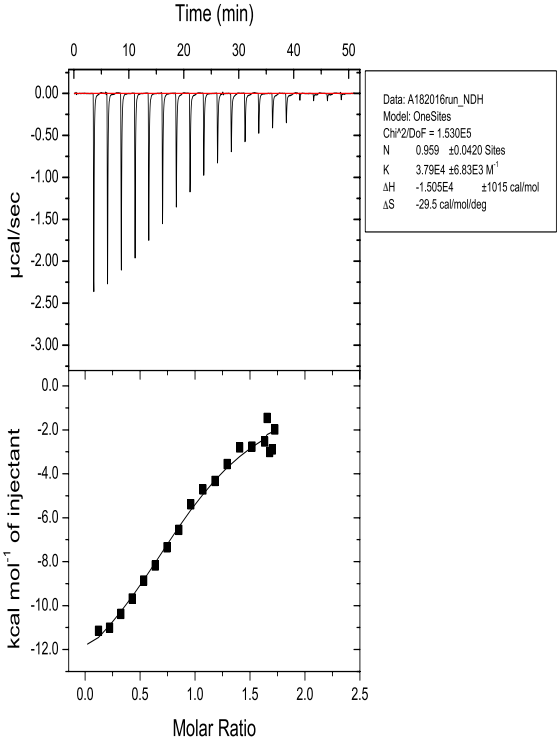
Figure 2.15. BAZ2A-PHD treble clef knuckle Asp pair: Structural deviation and distances between oppositely charged atoms among the PHD_nW_DD subtype structures anchoring histone H3-Arginine 2 residue.

BAZ2A D1698, with 0.0Å^2 ΔASAsc, having a non-zero energetic contribution towards peptide binding also suggested that other negatively charged residue control mutations (Figure 2.11C) might have non-zero energetic contributions possibly through long-range electrostatics. The ΔΔG for BAZ2A E1711A and E1715A control mutations are respectively 0.09 and 0.94 kcal/mol (Figure 2.17-2.18, Table 2.8). The peptide in complex with BAZ2A-PHD possesses three positive charge centers: (i) the N–terminal NH_3^+–group, (ii) the CZ–atom of Arg2 and (iii) the NZ–atom of Lys4. For the E1715 residue, closest of these three peptide positive charge centers is the N–terminal NH_3^+–group at 6.4Å away (Figure 2.17A). For comparison, the closest positive charge center for the D1 residue carboxylate oxygen atoms is the H3R2-CZ atom at 3.46 ± 0.04Å in BAZ2A where we note large energetic contribution from the D1 position. Noting the distances of these distal site positions (E1711 and E1715) in BAZ2A, it is unlikely that the energetic contributions of these sites can be explained merely by electrostatics. Recent reports on distal site perturbations modulating binding in small sized domains suggest that[67, 68], in addition to electrostatics, other structural effects such as the influence on protein dynamics and/or change in the structure could be associated with contributions of these distal sites. Influence of distal site perturbations on binding is not uncommon for PHD fingers, e.g. binding of BCL9 peptide at a distal site on Pygo-PHD (Figure 2.17B) enhances the PHD canonical site's affinity for histone H3 by 2-3 fold[69]. Probing distal site perturbations of mutations will however require further study. The next chapter will try to determine if these positions show a subtype specific enrichment for the distal negatively charged residue.

A


Time (min)
µcal/sec
kcal mol^{-1} of injectant
Molar Ratio
Data: A202016run_NDH
Model: OneSites
Chi^2/DoF = 2.584E4
N 0.730 ±0.115 Sites
K 9.02E3 ±1.80E3 M^{-1}
ΔH -2.036E4 ±4462 cal/mol
ΔS -50.2 cal/mol/deg

B


Time (min)
µcal/sec
kcal mol^{-1} of injectant
Molar Ratio
Data: A182016run_NDH
Model: OneSites
Chi^2/DoF = 1.530E5
N 0.959 ±0.0420 Sites
K 3.79E4 ±6.83E3 M^{-1}
ΔH -1.505E4 ±1015 cal/mol
ΔS -29.5 cal/mol/deg

C

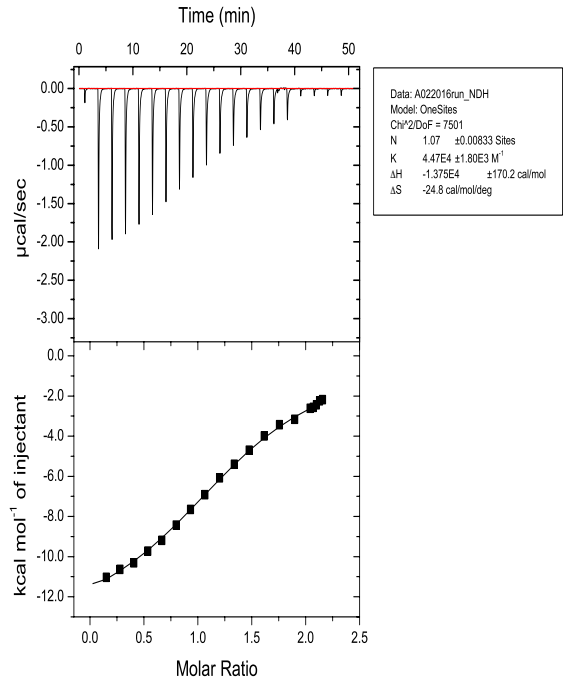

Time (min)
µcal/sec
kcal mol^{-1} of injectant
Molar Ratio
Data: A022016run_NDH
Model: OneSites
Chi^2/DoF = 7501
N 1.07 ±0.00833 Sites
K 4.47E4 ±1.80E3 M^{-1}
ΔH -1.375E4 ±170.2 cal/mol
ΔS -24.8 cal/mol/deg

Figure 2.16A. Negatively charged residue control mutants. ITC binding studies between histone H3 unmodified N-terminal peptide and mutants of BAZ2A-PHD. A representative isothermal calorimetric titration (top) of H3-1-9 peptide (syringe) into protein (cell) is shown for the mutant proteins while exothermic heats (bottom) exchanged per mol of injectant as a function of the molar ratio of peptide to mutant proteins. (A) on left is E1711A, and (B) on right is E1715A, (C) below V1728A.

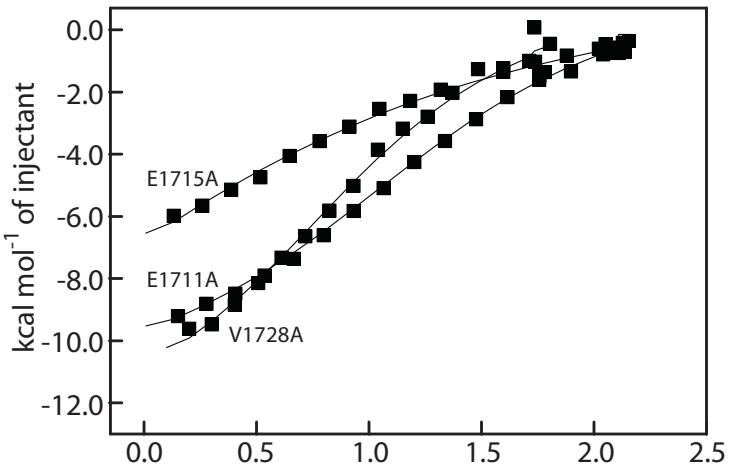


Figure 2.16B. Overlayed titration profiles of the control mutants: ITC binding studies between histone H3 unmodified N-terminal peptide and mutants of BAZ2A-PHD

A

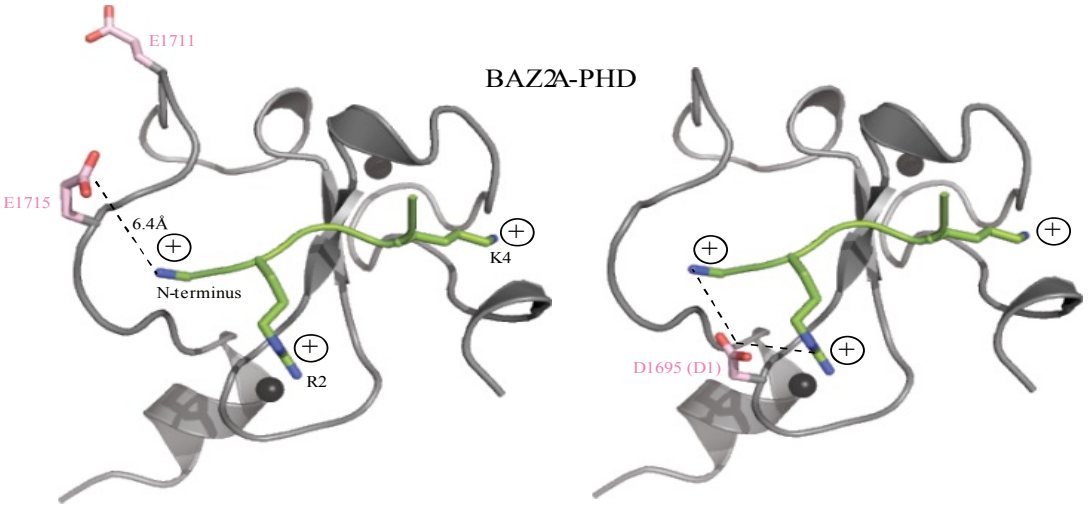


Figure 2.17A. Distances of positive charge centers from the site of mutation (left), and for comparison, distances for D1 position (right).

B

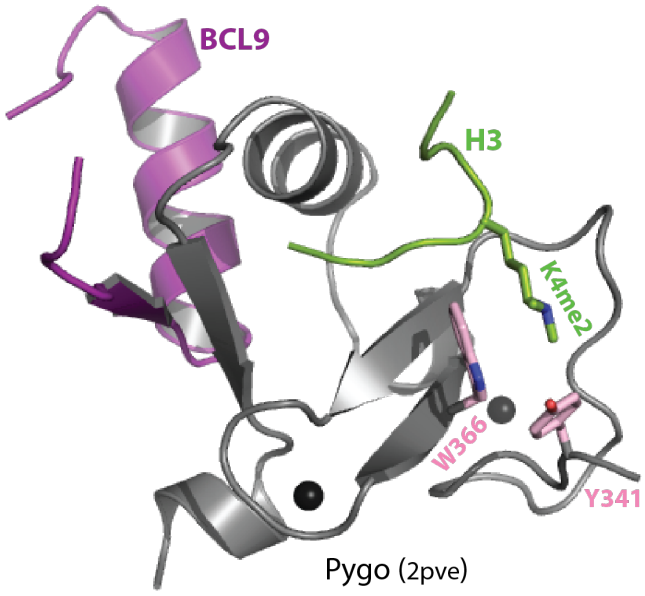


Figure 2.17B. Structural example of the distal site perturbation of Pygo-PHD by BCL9 peptide.

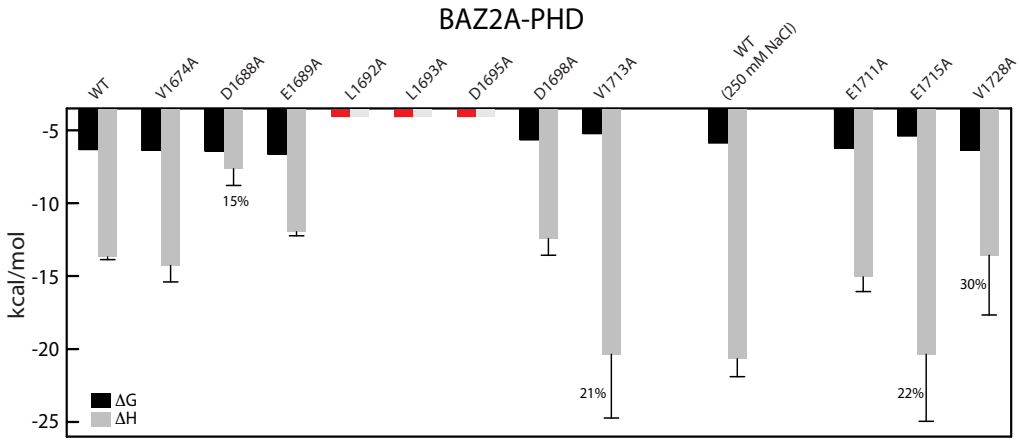
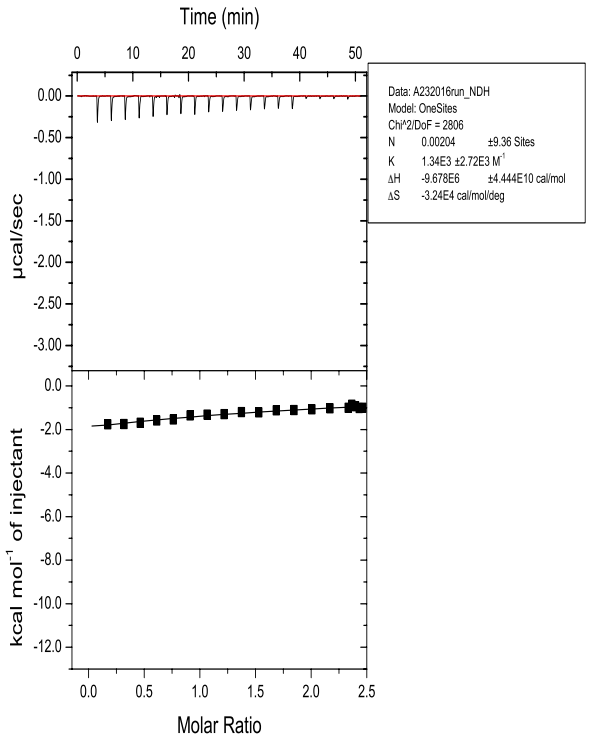


Figure 2.18. The fitted values of ΔH and ΔG obtained using ITC experiments are graphically represented for each BAZ2A mutation. The number next to the error bar represents the error (in percentage) associated with the fitted enthalpy.

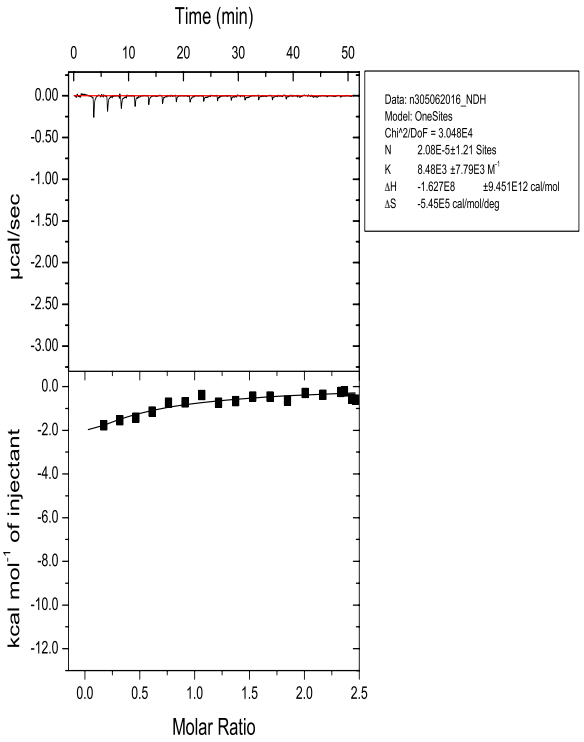
Distal Mutations that Disrupts Histone H3 Peptide binding

We looked at 3 other positions D1686, D1717 and N1675 (Figure 2.19) on the hBAZ2A-PHD surface, which were at a distal site from the histone H3 binding site. Our observation from the energetics was however quite interesting. All these 3 positions had significant contribution towards the protein-peptide complex formation since their mutations caused the interaction to be aborted. It is however not clear whether the influence is due to a long range electrostatic[70, 71] hence a mutation leads to the significant drop in the energetics. The observation could also be as a result of a dynamic motion of these positions hence their mutations causes the interruption of their coupling effect on the protein-peptide complex formation[68, 72-79]. All the possible scenarios could be confirmed by doing a structural study analysis, which will be conducted as the next phase of the work.

A


Time (min)
µcal/sec
kcal mol^{-1} of injectant
Molar Ratio
Data: A232016run_NDH
Model: OneSites
Chi^2/DoF = 2806
N 0.00204 ±9.36 Sites
K 1.34E3 ±2.72E3 M^{-1}
ΔH -9.678E6 ±4.444E10 cal/mol
ΔS -3.24E4 cal/mol/deg

B


Time (min)
µcal/sec
kcal mol^{-1} of injectant
Molar Ratio
Data: n305062016_NDH
Model: OneSites
Chi^2/DoF = 3.048E4
N 2.08E-5±1.21 Sites
K 8.48E3 ±7.79E3 M^{-1}
ΔH -1.627E8 ±9.451E12 cal/mol
ΔS -5.45E5 cal/mol/deg

C

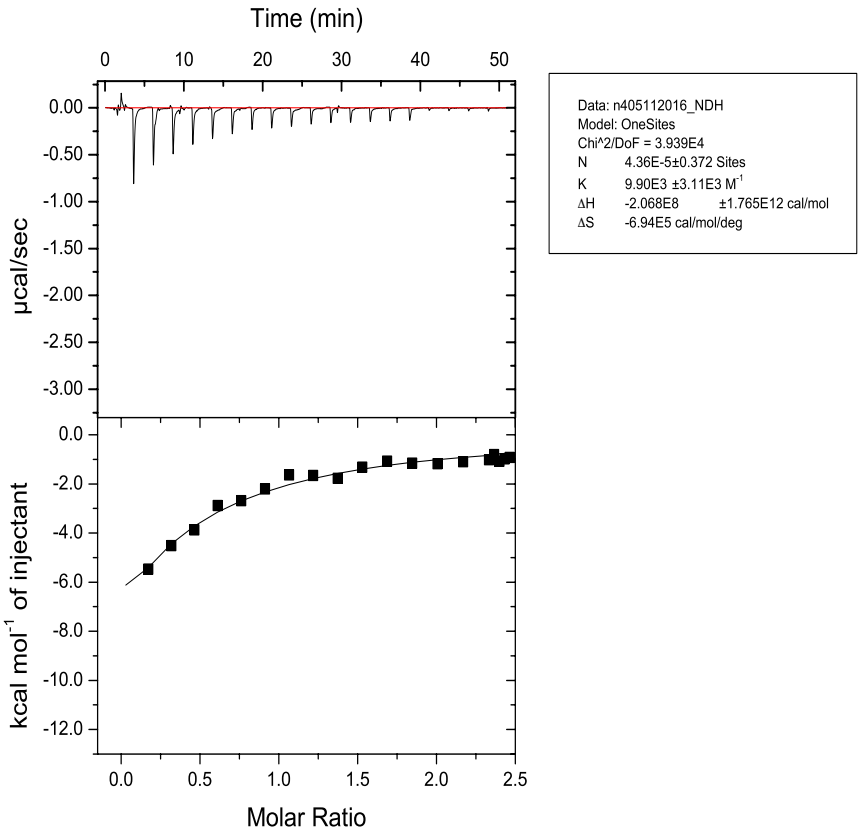


Figure 2.19. Distal site residue mutants: ITC binding studies between histone H3 unmodified N-terminal peptide and mutants of BAZ2A-PHD. A representative isothermal calorimetric titration (top) of H3-1-9 peptide (syringe) into protein (cell) is shown for the mutant proteins while exothermic heats (bottom) exchanged per mol of injectant as a function of the molar ratio of peptide to mutant proteins. (A) D1686A (B) D1717A (C) N1675A- The Stoichiometry (N) could not be fitted for this mutant hence the data fitting is unreliable to the observed enthalpy. As a result, $\Delta\Delta G°$ (>> 1.5 kcal.mol^{-1}) is large or the observed amount of heat is considered to be negligible.

CHAPTER III

BAZ2A AND PHD ZINC FINGER SUBTYPE

ABSTRACT

Although the *plant homeodomain* (PHD), a binuclear interleaved Zn-chelating domain finger superfamily is well known for its site-specific readouts of histone tails, mechanistic details of the *readouts* remain poorly understood. Starting with BAZ2A-PHD (one member of the PHD_nW_DD subtype), key histone peptide anchoring residue positions were first identified by site-specific mutagenesis. Loss of interfacial packing, due residue substitution, contributes to the observed binding energetics. The peptide anchoring residue positions ($\Delta\Delta G \geq 1.0$kcal/mol) of BAZ2A-PHD, interestingly, are enriched in specific type of residues in a subtype specific manner. The energetic contribution of the identified positions were further confirmed by mutagenesis in other three members of the subtype (UHRF1-PHD, KDM5B-PHD, KAT6A-PHD) that included pairs sharing even less than 40% sequence identity with each other. Despite low sequence similarity, mutations cause similar consequences in histone H3 binding suggesting a strong similarity in the binding mechanism.

The experimental data suggests that the binding mechanism of the PHD_nW_DD likely originated early to anchor residue 1-3 of histone H3, featuring a characteristic treble clef knuckle Aspartate residues for anchoring the peptide Arginine hotspot while the acidic patch residues contribute to the helical conformation of N-terminal histone H3. A set of non-polar amino acids is among the residues enriched in the PHD_nW_DD subtype, and these non-polar amino acids tightly pack against the peptide residues with small side-chains, present on either side of the peptide hotspot Arginine residue.

INTRODUCTION

From chapter two, we experimentally discovered two categories of mutations that disrupt peptide binding: (1) Type-A: positions that are in contact with the peptide and (2) Type-B: positions that are remote from the peptide-binding site (distal site). For this chapter of the dissertation, we focused on understanding the biochemical basis of the effects of Type-A mutations using recombinant protein chemistry and biophysical chemistry. Moreover, having observed the energetics of the histone binding mechanism of histone H3 peptide and human BAZ2A-PHD, we wanted to ascertain if the features we observed were features that could be said to be true of the PHD_nW_DD subtype or they were features that were unique to BAZ2A-PHD.

To confirm this, three members of the PHD_nW_DD subtype were selected. One member, KDM5B-PHD[7, 18, 80] had sequence identity of 60% but the other two, UHRF1-PHD[19] and KAT6A-PHD[81] had sequence identity of less than 40% (Figure 3.1B and 3.2). Wild types of these proteins were cloned and purified for the biophysical measurements using ITC. Site-specific mutations were also performed on these selected members (Type-A mutations). For simplicity, the D1695, D1698, L1692, and L1693 positions in BAZ2A-PHD will be referred to as D1, D2, L1, and L2 respectively in corresponding positions in selected members of the subtype.

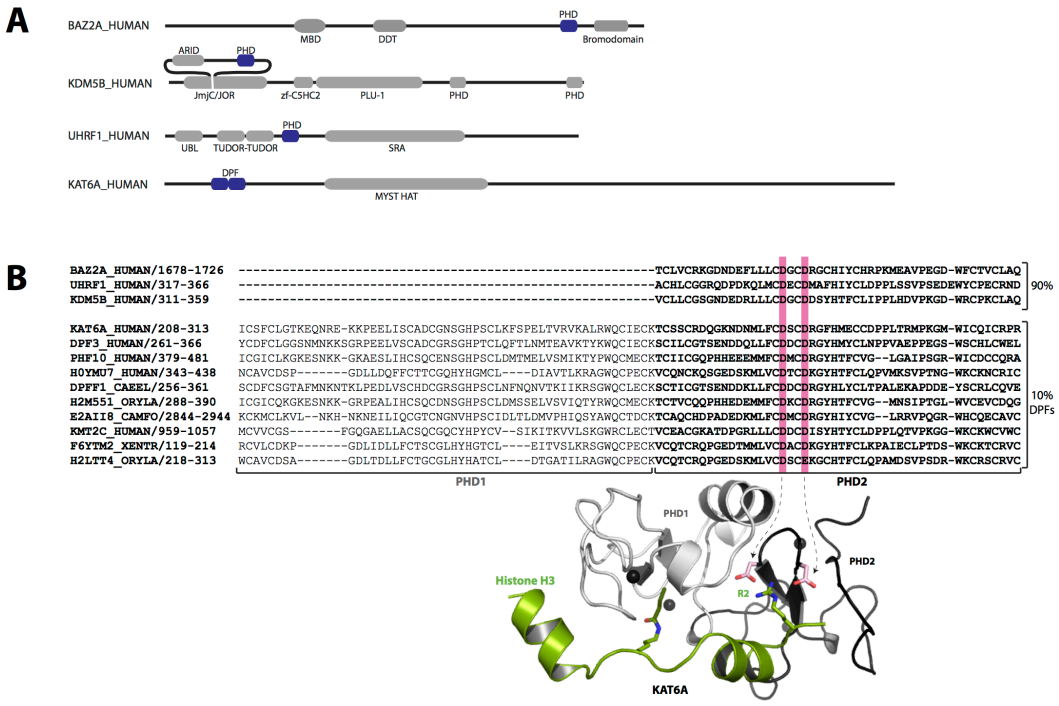


Figure 3.1. The chosen PHD fingers: (A) Cartoon representation of the domain organization of proteins bearing the distinct subfamily of PHD (blue). This PHD subtype occurs in combination with variety of other domains. (B) Double PHD Finger (DPF) are aligned with single PHD finger sequences to highlight the common feature of treble clef Aspartate residues. UniProt accession ids and residue boundaries are indicated left of the alignment. Non-redundant sequences (85% sequence identity) are used to generate the alignment using muscle. PHD1 and PHD2 of DPFs respectively are in gray and black, with KAT6A DPF (bottom, right) highlighting the location of treble clef Asp residues in DPF-PHD2. Among subtype, 10% are DPFs and 90% are isolated PHD.

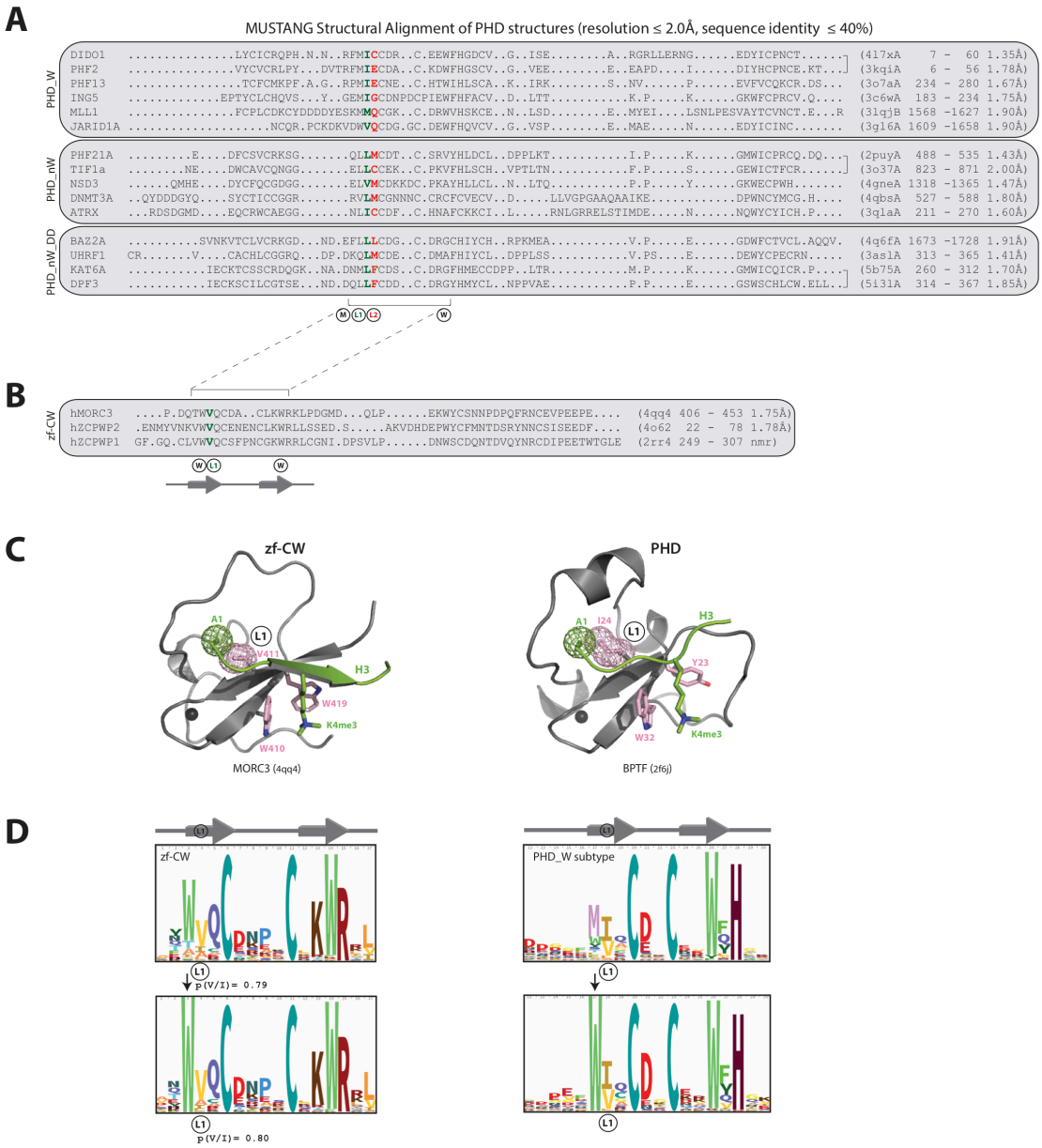


Figure 3.2. PHD, zf-CW and the L1 position: (A) MUSTANG structural alignment of PHD finger. The protein names are on the left of the alignment, and pdb ids, residue boundary and resolution are on the right of the alignment. PHD fingers with < 40% sequence identity are chosen, and the '] ' symbol to the right of the alignment indicates some pairs with sequence slightly above 40% sequence identity. (B) MUSTANG structural alignment of zf-CW domains that have < 40% sequence identity with one another. The dotted lines between the zf-CW and PHD structural alignments indicate the segment containing the strand-knuckle-strand. L1 and L2 positions in the alignment are respectively in green and red color. (C) The

structure of MORC3 zf-CW domain in complex with histone H3K4me3 peptide (left). The surface mesh represents contact between H3A1-CH_3 and L1. The structure BPTF PHD finger (right) with a similar pattern of aromatic cage residues as that of zf-CW. (D) Similarity in the sequence logo between zf-CW and PHD_W subtype for the pair of strands that augments the peptide.

MATERIALS AND METHODS

DNA constructs and Mutagenesis of Subtype proteins

The wild type proteins selected for this section of the work had already being cloned and sequences verified from GenScript by Dr. Chakravarty's lab[7]. Briefly, 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 BamHI and XhoI restriction sites. The pET28A modification is that the Thrombin protease cleavage site DNA sequence was modified to incorporate BamHI restriction site without altering the protein sequence. PCR (polymerase chain reaction) amplification of synthetic DNA representing the sequences of hUHRF1 PHD and hKAT6A DPF were first carried out for creating the desired constructs. Synthetic DNA molecules were obtained from GeneScript. Mutants of hUHRF1 PHD (L344A, M345A, and D350A) and hKAT6A DPF (L279A, F280A, and D285A) were created by site directed mutagenesis (refer to chapter 2 for procedure and conditions). The recombinant GST fusion hKDM5B PHD (residues 305-366 of UniProt ID Q9UGL1/KDM5B_Human) and hAIRE PHD (residues of 294-343 UniProt ID O43918/AIRE_HUMAN) used here was taken from an earlier study done by the lab[7]. hKDM5B PHD and hAIRE PHD were originally cloned into pGEX-4T3 vector. hKDM5B PHD mutants (D331A, L325A, L326A) and hAIRE PHD mutant

(C310L) were created using site directed mutagenesis (refer to chapter 2). In total, 14 proteins (4 wild types and 10 mutants) created in this section of the study are used to probe the peptide binding mechanisms of PHD finger subtypes.

Table 3.1. Primers used for the Site Directed Mutagenesis of the subtype Proteins. The letter colored red is the site of mutation. Primers were purchased from Integrated DNA Technologies (IDT).

Protein	Forward Primer	Reverse Primer	Protein sequence
UHRF1_Wildtype	CGCGGATCCG GTCCGAGCTG CAAGCACTG	CCGCTCGAGTT ACGCGTCGTTA CGGCATTCCG	GPSCKHCKDDVNRLCR VCACHLCGGRQDPDK QLMCDECDMAFHIYC LDPPLSSVPSEDEWYCP ECRNDA
UHRF1_D350A	ATGTGCGACG AGTGCGCTAT GGCGTTCCAC ATC	GATGTGGAAC GCCATAGCGC ACTCGTCGCAC AT	GPSCKHCKDDVNRLCR VCACHLCGGRQDPDK QLMCDECAMAFHIYC LDPPLSSVPSEDEWYC PECRNDA
UHRF1_L344A	CAGGACCCGG ATAAGCAAGC GATGTGCGAC GAGTGCGAT	ATCGCACTCG TCGCACATCG CTTGCTTATC CGGGTCCTG	GPSCKHCKDDVNRLCR VCACHLCGGRQDPDK QAMCDECDMAFHIYC LDPPLSSVPSEDEWYCP ECRNDA
UHRF1_M345A	GACCCGGATA AGCAACTGGC GTGCGACGAG TGCGATATG	CATATCGCAC TCGTCGCACG CCAGTTGCTT ATCCGGGTC	GPSCKHCKDDVNRLCR VCACHLCGGRQDPDK QLACDECDMAFHIYC LDPPLSSVPSEDEWYC PECRNDA
KAT6A_Wildtype	CGCGGATCCC ACATGCTGGA GCTGCCGCAC	CCGCTCGAGT TACGGACGGC AAATTTGGCA GATC	HMLELPHEKDKPVA EPIPICSFCLGTKEQN REKKPEELISCADCG NSGHPSCLKFSPELT VRVKALRWQCIECK TCSSCRDQGKNADN MLFCDSCDRGFHME CCDPPLTRMPKGMW ICQICRP
KAT6A_D285A	TTCTGCGAC AGCTGCGCT CGTGGCTTT CACATG	CATGTGAAAG CCACGAGCGC AGCTGTCGCA GAA	HMLELPHEKDKPVA EPIPICSFCLGTKEQN REKKPEELISCADCG NSGHPSCLKFSPELT

			VRVKALRWQCIECK TCSSCRDQGKNADN MLFCDSCARGFHME CCDPPLTRMPKGMW ICQICRP
KAT6A_L279A	GGTAAAAACG CGGATAACAT GGCGTTCTGC GACAGCTGC	GCAGCTGTCG CAGAACGCCA TGTTATCCGCG TTTTTACC	HMLELPHEKDKPVA EPIPICSFCLGTKEQN REKKPEELISCADCG NSGHPSCLKFSPELT VRVKALRWQCIECK TCSSCRDQGKNADN MAFCDSCDRGFHME CCDPPLTRMPKGMW ICQICRP
KAT6A_F280A	AACGCGGATA ACATGCTGGC CTGCGACAGC TGCGATCGT	ACGATCGCAG CTGTCGCAGG CCAGCATGTT ATCCGCGTT	HMLELPHEKDKPVA EPIPICSFCLGTKEQN REKKPEELISCADCG NSGHPSCLKFSPELT VRVKALRWQCIECK TCSSCRDQGKNADN MLACDSCDRGFHME CCDPPLTRMPKGMW ICQICRP
KDM5B_Wildtype	CGCGGATCC AATGCTGTG GACCTGTATG	GAATGTAGT AAGCCACAA GAATAA CTCGAGCGG	NAVDLYVCLLCGS GNDEDRLLLCDGC DDSYHTFCLIPPLH DVPKGDWRCPKCL AQECSKPQE
KDM5B_D331A	TGTGATGGCT GTGCGGACAG TTACCAT	ATGGTAACT GTCCGCACA GCCATCACA	NAVDLYVCLLCGS GNDEDRLLLCDGC ADSYHTFCLIPPLH DVPKGDWRCPKCL AQECSKPQE
KDM5B_L325A	AATGATGAAG ACCGGCTAGC GTTGTGTGAT GGCTGTGAT	ATCACAGCCA TCACACAACG CTAGCCGGTC TTCATCATT	NAVDLYVCLLCGS GNDEDRLALCDGC DDSYHTFCLIPPLH DVPKGDWRCPKCL AQECSKPQE
KDM5B_L326A	GATGAAGACC GGCTACTGGC GTGTGATGGC TGTGATGAC	GTCATCACAGC CATCACACGCC AGTAGCCGGTC TTCATC	NAVDLYVCLLCGS GNDEDRLLACDGC DDSYHTFCLIPPLH DVPKGDWRCPKCL AQECSKPQE
AIRE-	CGCGGATCC AAGAATGAG GACGAGTGTG	CCGCTCGA GTTACTCC TGGACTGT	KNEDECAVCRDG GELICCDGCPRAF HLACLSPPLREIPS

PHD1_Wildtype		TGCCTG	GTWRCSSCLQATVQE
AIRE- PHD1_C310L	GACGGCGGGG AGCTCATCCT CTGTGACGGC TGCCCTCGG	CCGAGGGCAG CCGTCACAGA GGATGAGC TCCCCGCCGTC	KNEDECAVCRDG GELILCDGCPRAF HLACLSPPLREIPS GTWRCSSCLQATVQE

Table 3.2. The cDNA UniProt ID and nucleotide sequence of the PHD finger subtype proteins

cDNA UniProt ID	**Nucleotide Sequence**
hUHRF1 PHD (residues 299-367 UniProt ID Q96T88/UHRF1_HUMAN)	GGTCCGAGCTGCAAGCACTGCAA AGACGATGTGAACCGTCTGTGCC GTGTTTGCGCGTGCCACCTGTGC GGTGGCCGTCAGGACCCGGATAA GCAACTGATGTGCGACGAGTGCG ATATGGCGTTCCACATCTATTGCC TGGACCCGCCGCTGAGCAGCGTG CCGAGCGAGGATGAATGGTATTG CCCGGAATGCCGTAACGACGCG
hKAT6A DPF (residues 196-319 UniProt ID Q92794/KAT6A_HUMAN)	CACATGCTGGAGCTGCCGCACGAA AAGGACAAACCGGTGGCGGAGCCG ATCCCGATTTGCAGCTTCTGCCTGG GTACCAAGGAGCAGAACCGTGAAA AGAAACCGGAGGAACTGATCAGCTG CGCGGATTGCGGTAACAGCGGCCAC CCGAGCTGCCTGAAGTTTAGCCCGG AGCTGACCGTGCGTGTTAAAGCGCT GCGTTGGCAGTGCATTGAATGCAAG ACCTGCAGCAGCTGCCGTGACCAAG GTAAAAACGCGGATAACATGCTGTT CTGCGACAGCTGCGATCGTGGCTTTC ACATGGAATGCTGCGACCCGCCGCTG ACCCGTATGCCGAAAGGCATGTGGAT CTGCCAAATTTGCCGTCCG
hKDM5B PHD (residues 305-366 of UniProt ID Q9UGL1/KDM5B_Human)	AATGCTGTGGACCTGTATGTCTGTCTTT TATGTGGCAGTGGCAATGATGAAGACC GGCTACTGTTGTGTGATGGCTGTGATGA CAGTTACCATACCTTTTGCTTGATCCCAC CTCTCCATGATGTTCCCAAGGGAGACTGG AGGTGTCCTAAGTGTTTGGCTCAGGAATG TAGTAAGCCACAAGAA
hAIRE PHD (residues of 294-343 UniProt ID O43918/AIRE_HUMAN)	AAGAATGAGGACGAGTGTGCCGTGT GTCGGGACGGCGGGGAGCTCATCTG

	CTGTGACGGCTGCCCTCGGGCCTTCC ACCTGGCCTGCCTGTCCCCTCCGCTC CGGGAGATCCCCAGTGGGACCTGGAG GTGCTCCAGCTGCCTGCAGGCAACAG TCCAGGAG

Expression and purification of proteins and their Mutants

The recombinant subtype PHD finger proteins and their mutants were purified as described in chapter 2. Briefly, E. coli BL21 (DE3) cells expressing the protein from pET28a vector were induced at OD600 = 0.6-1.0 with 1 mM IPTG grown overnight at 16°C. The fusion proteins were isolated using His-tag affinity chromatography. The proteins were then purified to homogeneity by gel filtration using HiLoad 26/600 superdex column using FPLC. Protein purity was confirmed by SDS-PAGE. Recombinant hUHRF1 PHD and its mutants (L344A, M345A, and D350A), hKAT6A DPF and its mutants (L279A, F280A, and D285A) were all purified similarly as described for hBAZ2A PHD following the His-tag affinity chromatography step. Recombinant hKDM5B PHD and its mutants (D331A, L325A and L326A) were purified as described in Chakravarty et al. [7] using the pGEX-4T3 vector constructs. Similarly, recombinant hAIRE PHD and its mutant (C310L) were purified 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 FPLC. Remaining purification steps were same as that of the hBAZ PHD in chapter 2.

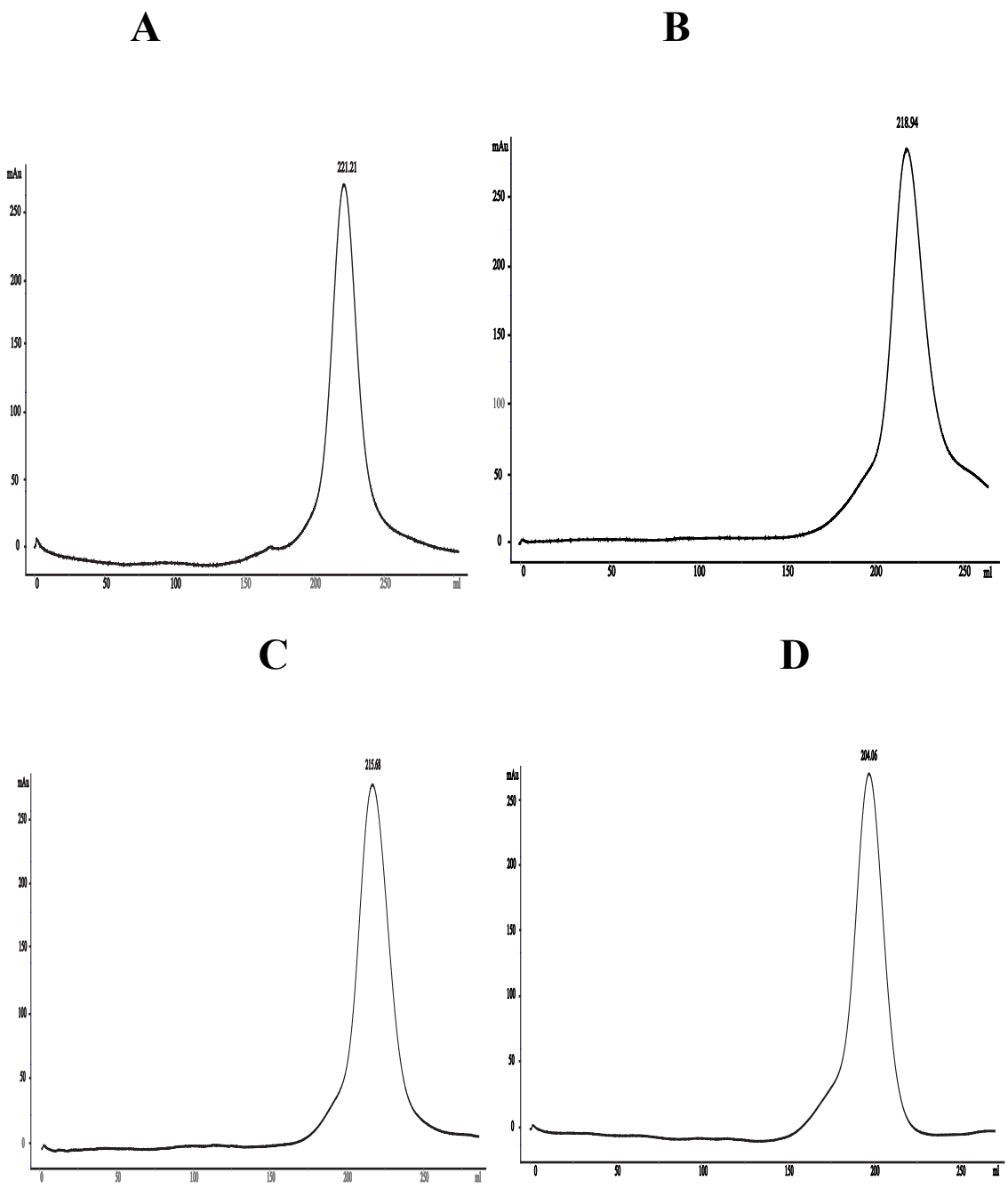


Figure 3.3. (A-D) Chromatograms showing the FPLC profile peaks of purified hUHRF1-PHD, (A) Wildtype, (B) D350A, (C) L344A, and (C) M345A respectively. All the proteins were eluted with 100ml 150mM imidazole in 1x phosphate buffer and concentrated to 4ml before loading into the AKTA prime FPLC at 4°C. The tubes corresponding to the peaks were concentrated to 0.1mM for the ITC measurement.

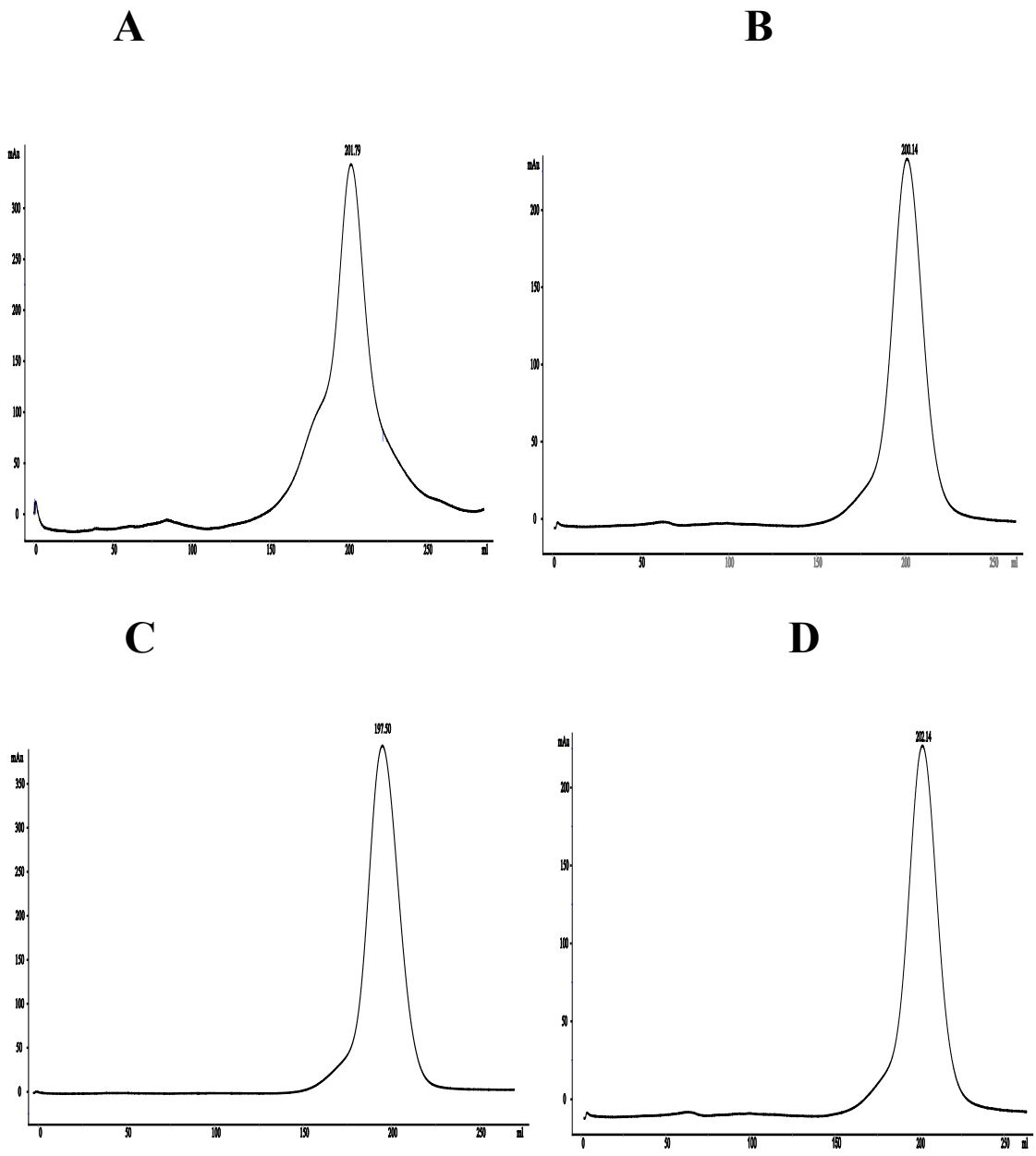


Figure 3.4. (A-D) Chromatograms showing the FPLC profile peaks of purified hKAT6A-PHD, (A) Wildtype, (B) D285A, (C) L279A, and (D) F80A respectively. All the proteins were eluted with 100ml 150mM imidazole in 1x phosphate buffer and concentrated to 4ml before loading into the AKTA prime FPLC at 4°C. The tubes corresponding to the peaks were concentrated to 0.1mM for the ITC measurement.

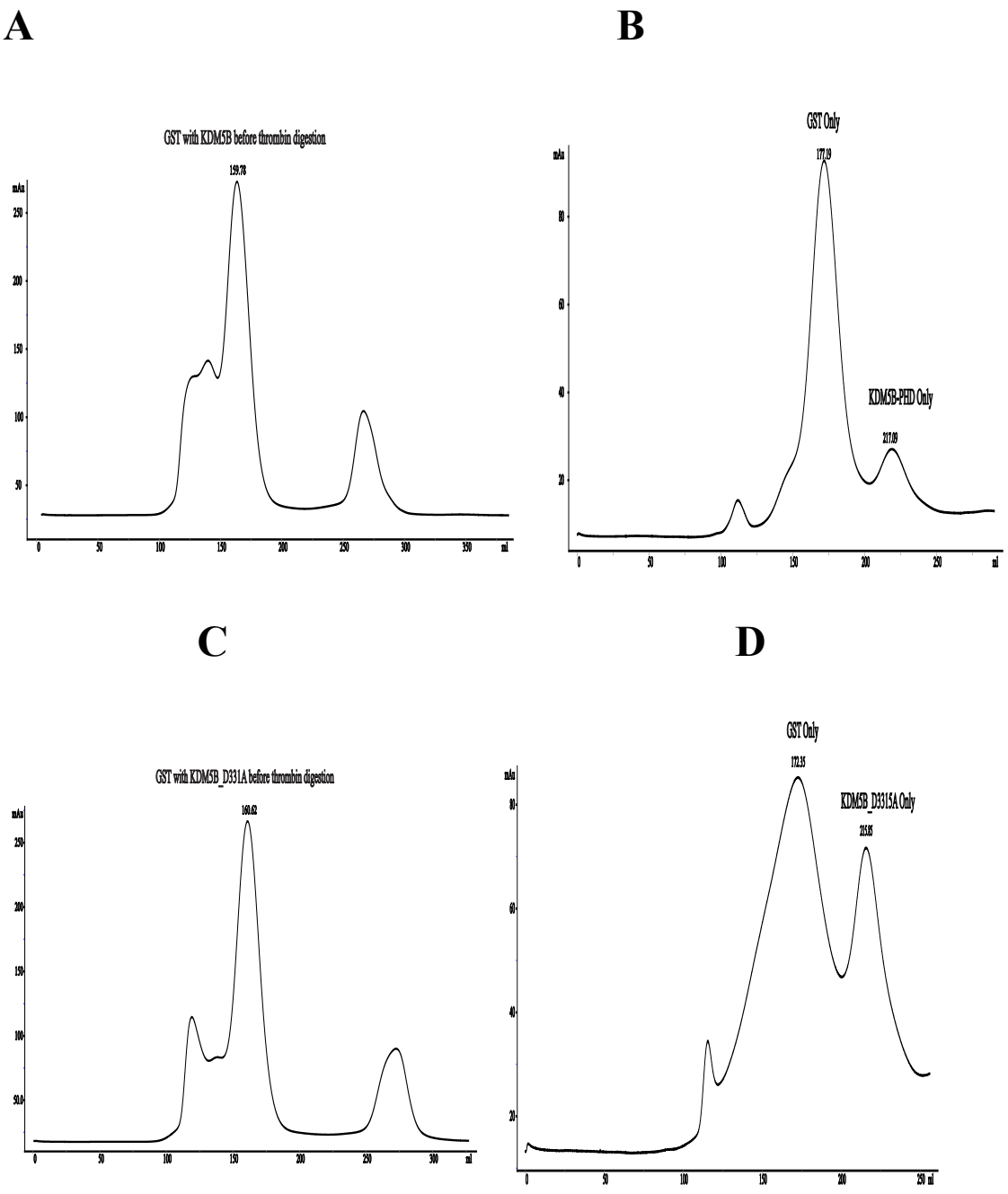


Figure 3.5A. (A-D). Chromatograms showing the FPLC profile peaks of purified hKDM5B-PHD, (A) Wildtype before GST cleavage, (B) Wildtype after GST cleavage, (C) D331A mutant before GST cleavage and (D) D331A mutant after GST cleavage.

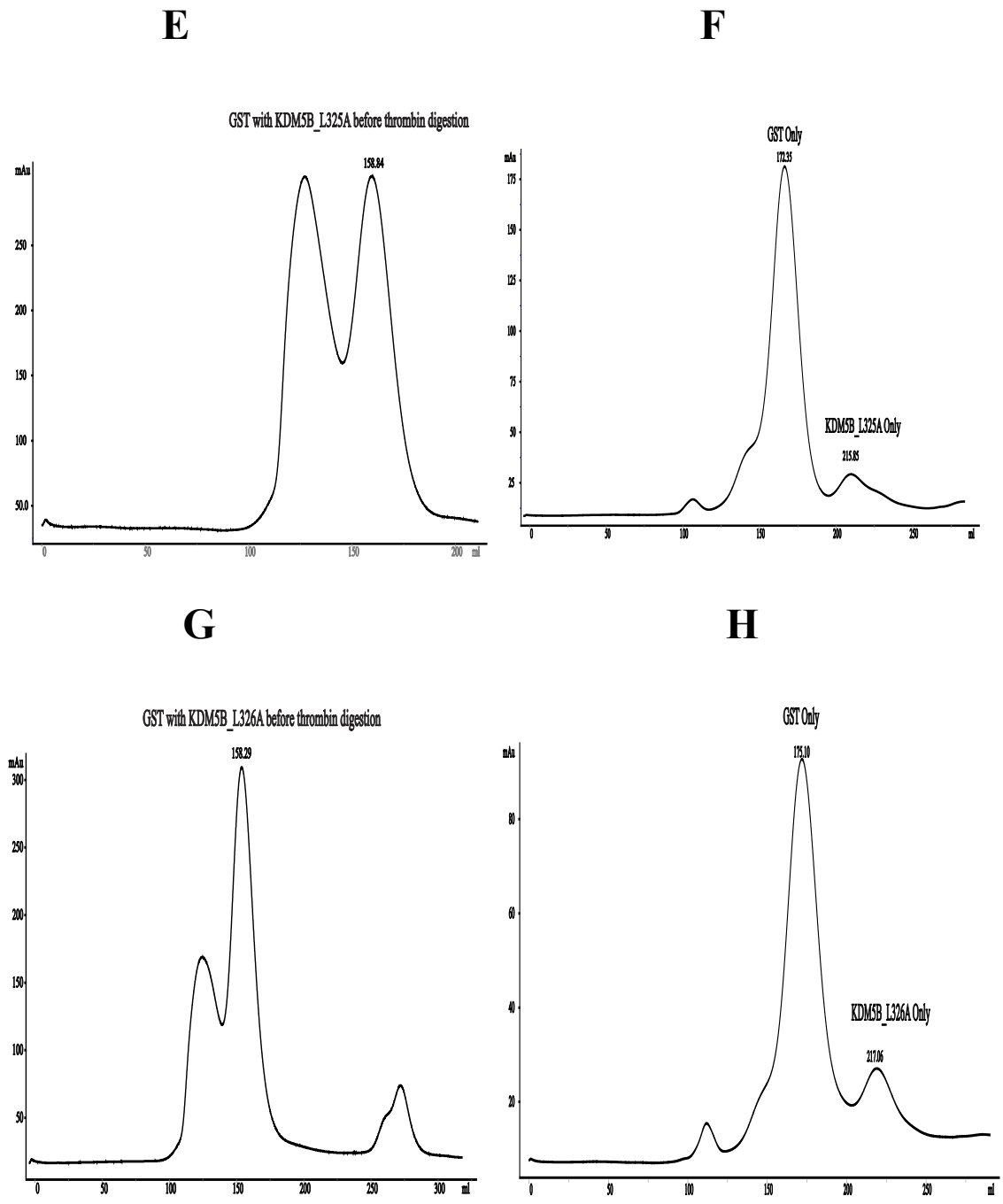


Figure 3.5B. (E-F). Chromatograms showing the FPLC profile peaks of purified hKDM5B-PHD mutants, (A) L325A mutant before GST cleavage, (B) L325A mutant after GST cleavage, (C) L326A mutant before GST cleavage and (D) L326A mutant after GST cleavage.

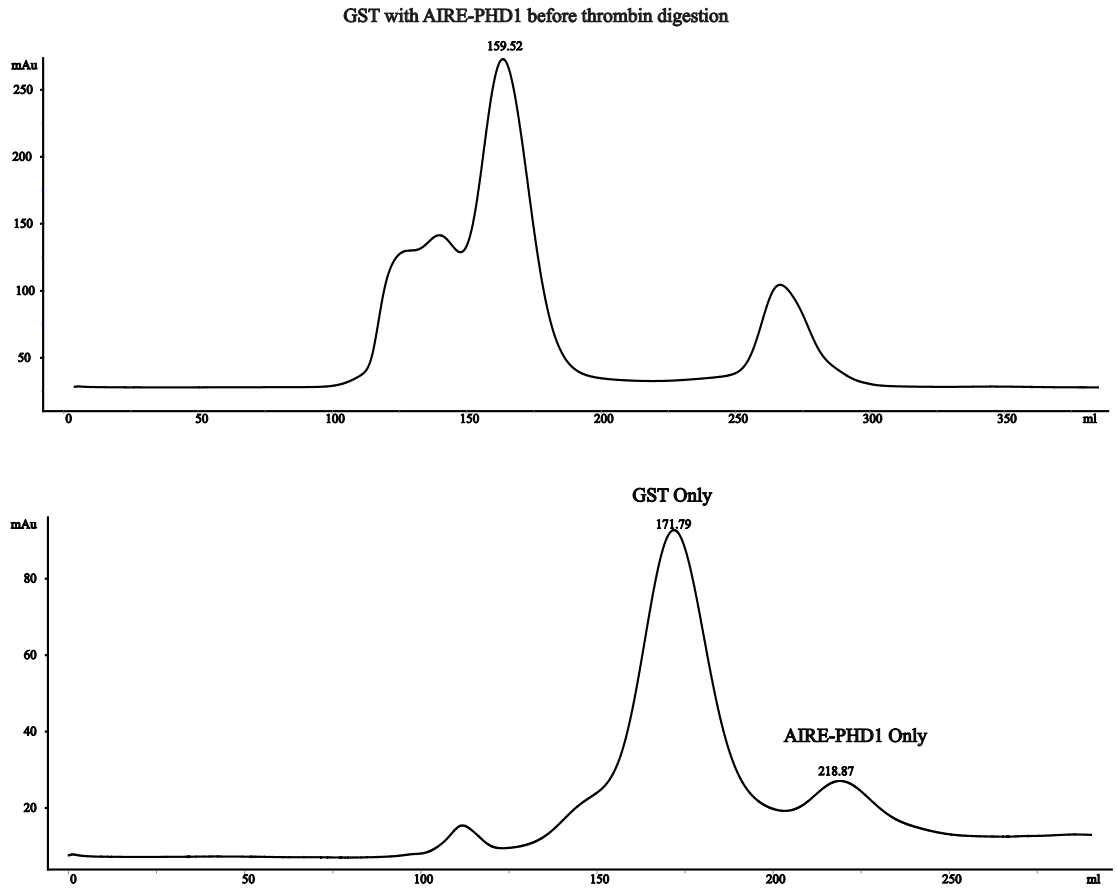
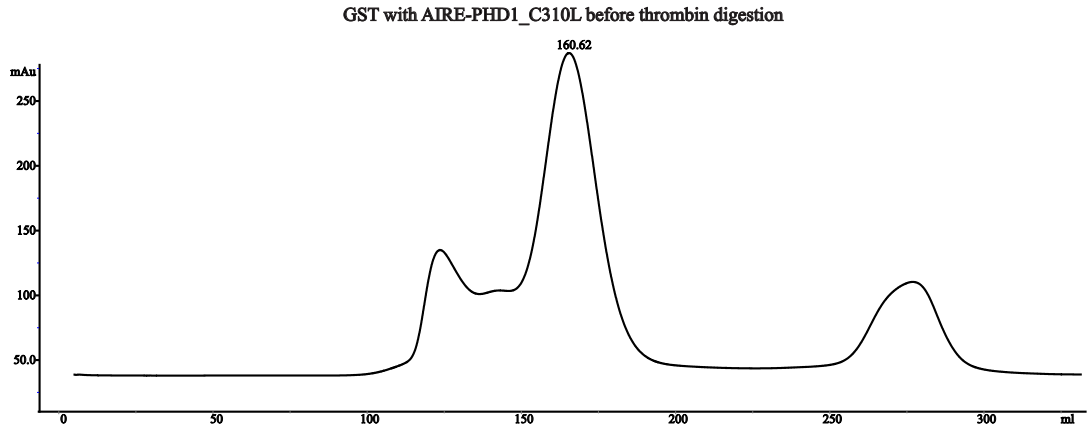


Figure 3.6A. Chromatograms showing the FPLC profile peaks of purified hAIRE-PHD1 wild type. (Top) Before GST cleavage (Bottom) After GST cleavage



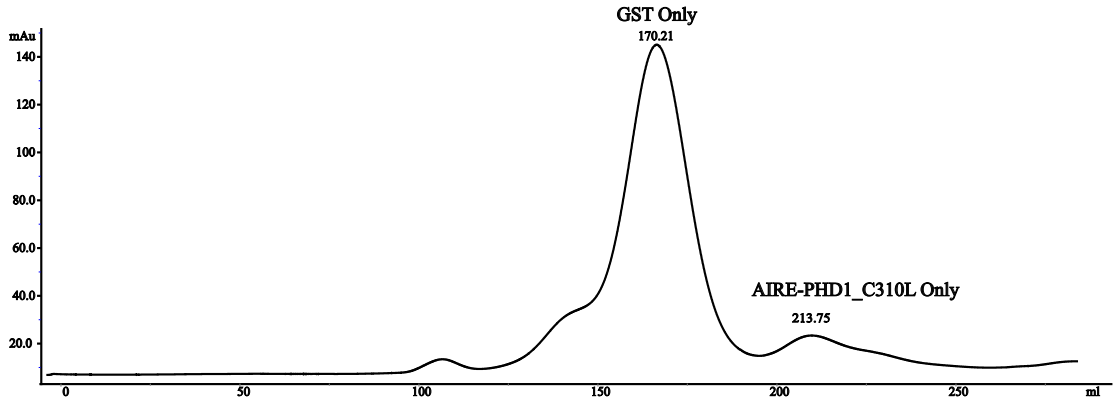


Figure 3.6B. Chromatograms showing the FPLC profile peaks of purified hAIRE-PHD1 C310L Mutant. (Top) Before GST cleavage (Bottom) After GST cleavage

Synthetic Peptides

We used two synthetic peptides for this section of the study. They are: (i) Histone H3 residues 1-11 (H3-1-11W), and (ii) Histone H3 residue 1-16 with acetylation at K14 (H3K14ac-1-16W). The W in the synthetic peptide sequence represents C-terminal Tryptophan residue. H3-1-11W was used for the binding studies of the wild type and the mutants of hUHRF1 PHD and hKDM5B PHD proteins. H3K14ac-1-16W peptide was used for the binding studies of the wild type and the mutants of hKAT6A DPF. 98% pure (confirmed with mass-spec) synthetic peptides were obtained from GeneScript. The C-terminal Tryptophan residue in the peptide was used for estimating peptide concentrations by absorbance using the computed molar extinction coefficient[66]. These peptides were used for ITC (Isothermal Titration Calorimetry) experiments. While this work was in progress we had learned that hKAT6A DPF binds crotonylated H3K14cr-1-25 peptide with the highest affinity (Kd = 5.8μM)[24] that is ~4.0 fold higher than that for the H3K14ac peptide. Though we realize that it would have been ideal to have the

peptide mutational analysis for hKAT6A DPF in the crotonylated H3K14cr-1-25 peptide background, we carried out our experiments based on the earlier structural report of hKAT6A DPF in complex with H3K14ac-1-16[81]. The ITC experiments, reported in the earlier study of the interaction of hKAT6A DPF with H3K14ac-1-16[81], were carried out in 50 mM NaCl, while our ITC experiments were all performed at 150 mM NaCl. Under these conditions, we do not observe the difference in the binding enthalpy between H3-1-16 and H3K14ac-1-16 peptides reported earlier[81].

Alignments, PHD subtypes and Occurrence Frequency

10,858 PFAM-aligned sequences of PHD finger superfamily (PFAM[82] 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-Hit[83], 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 3.7. 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 3.7). A breakdown of these 3,469 sequences into subtypes: (i) PHD_W subtype (1,425 sequences), and (ii) PHD_nW subtype (2,044 sequences). The PHD_nW subtype

sequences were further segregated into: (a) PHD_nW_DD (692 sequences), (b) the remaining 1362 PHD_nW sequences. PHD_nW_DD subtype refers to PHD_nW sequences having a pair of Asp residues in the treble-clef knuckle of the PHD scaffold. As described in Chakravarty et al.[7], double PHD finger (DPF) scaffolds belong to this subtype, and therefore, are categorized as a distinct subtype. As a control for the PHD_nW_DD subtype, PHD_W subtype sequences were also segregated into: (a) PHD_W_DD (234 sequences) and (b) the remaining 1191 PHD_W sequences. PHD_W_DD 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 3.7), 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 3.7). 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 W/Y/F for the 26th position in the PFAM alignment (Figure 3.1) resulted in PHD_W subtype and the remaining sequences are considered PHD_nW. Similarly, selecting PHD_nW sequences with residue D at the 24th position in the PFAM alignment (Figure 3.1) resulted in PHD_nW_DD subtype. The frequency of occurrence (pji) of the jth amino acid (or residue type) in the ith alignment position (column) is the ratio between the number of times the jth residue type appears in the ith alignment position and the total number of sequences in the alignment (e.g. subtype alignment). Entropy for

the ith position in this alignment was computed as, Entropy i = 4.39 - Σ (pji)×(log2 (pji)), where pji is the probability of the jth amino acid at the ith alignment position (column) and 4.39 is the value of maximum entropy. Entropy and pji for each PHD-subtype and the PHD superfamily are calculated separately using the subtype specific alignments shown in Figure 3.7.

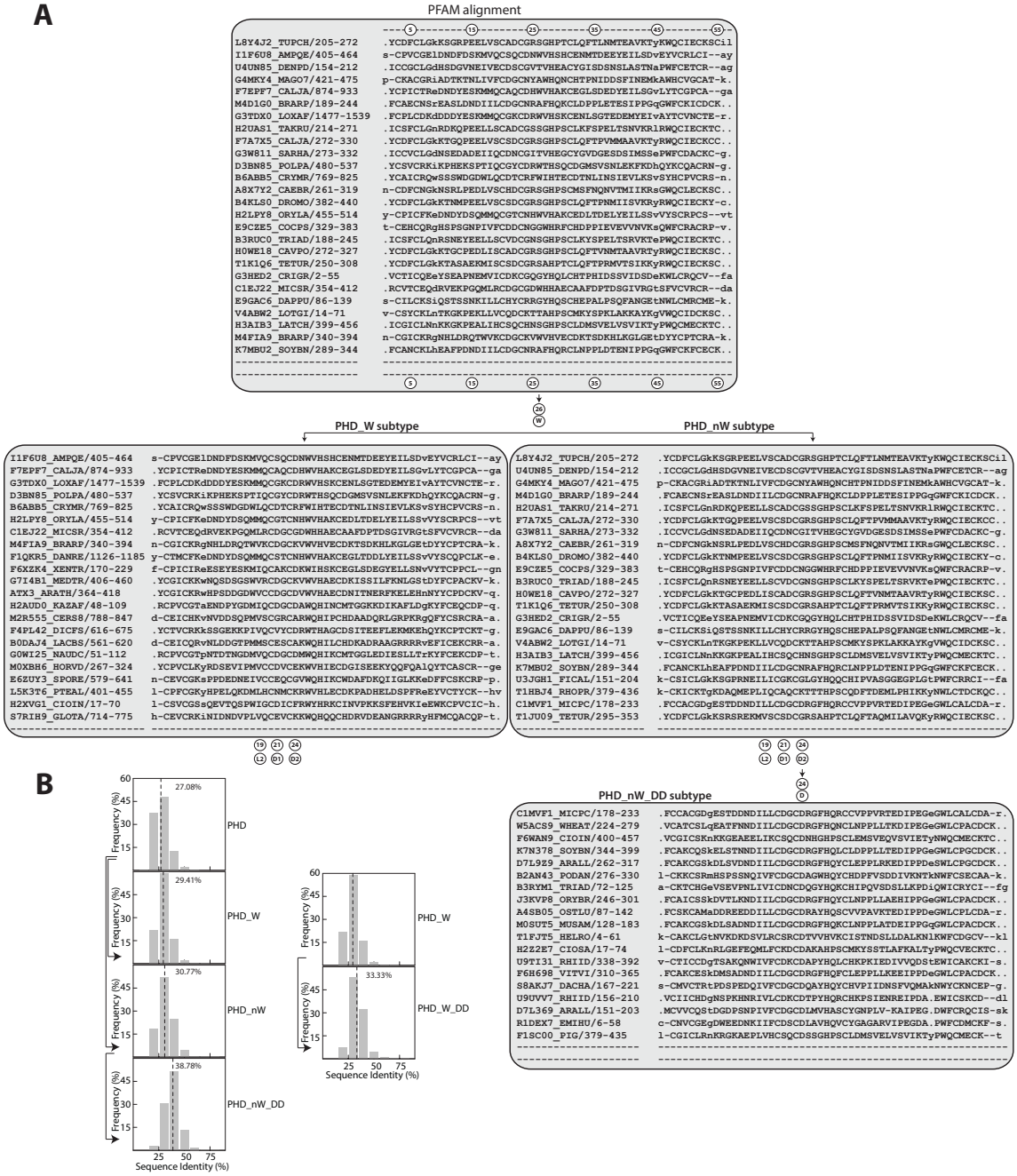


Figure 3.7. Master pfam alignment of PHD subtypes: (A) PFAM alignments of non-redundant (85% sequence identity) PHD finger sequences (top) are further segregated into subtype specific sequences (middle row and bottom). Sequences in the middle row subtypes are grouped based on the position number '26' as having W/aromatic-residue (left, PHD_W subtype) and not having W/aromatic-residue (right, PHD_nW). The bottom row sequences (PHD_nW_DD subtype) are extracted from PHD_nW with the position number '24' as having D. (B) The distributions of sequence identity in each subtype with the dashed lines representing the median of sequence identity for each subtype.

The sequence logo (e.g., Figure 3.2, 3.15 and 3.17) was created using the online version of the Skylign program[84] by uploading PHD subtype specific alignments. For sequence identity distribution (Figure 3.7B), sequence identity between sequences, i and j, was calculated by comparing residue identity between the aligned positions of master alignment. The observed median of the sequence identity distributions (Figure 3.7B, dashed line) of a subtype (e.g. PHD_nW_DD) was used to set the sequence identity cutoff (≤ 40%) for structural analysis.

All structural alignments and structural superposition operations (e.g., Figure 3.2, 3.8, 3.21), discussed here, were generated using MUSTANG[85]. MUSCLE[86] sequence alignment was also used for aligning: (a) double PHD finger (DPF) sequences (Figure 3.1A) with the remaining PHD_nW_DD sequences, and (b) sequences of a PFAM domain with the GREMLIN[87, 88] consensus sequence of the PFAM domain (see below, Figure 3.8C). The sequences of the DPFs were first compiled as described earlier[7] and the sequences with 85% identical residues were eliminated after MUSCLE alignment.

Mutual Information and PFAM domains

The mutual information (MI) of a residue pair of PFAM domain alignments was calculated using the MISTIC[89] web server by manually uploading the master PFAM MSA (see above) in fasta format. The server calculates MI between pairs of columns in the MSA. The calculation of frequency for each amino acid pair uses sequence weighting and low count corrections in addition to comparing the computed frequency to the expected pair-frequency with the assumption that the amino acids are non-correlated[90]. The MI is represented as a weighted sum of the log-ratios between the observed and expected amino acids pair frequencies[90].

The server returns an MI value for each pair of residue positions in an MSA[89]. The server also returns a cumulative mutual information score (CMI) for each MSA position as the sum of MI values above a certain threshold (measured in terms of the MI Z-score[91]) for every amino acid pair where the particular residue appears. The CMI value represents the degree of participation of a given MSA position in the mutual information network[91]. The scores returned by the server correspond to positions of the first sequence in the alignment. For convenience, in all our alignments for MI calculations by the MISTIC server, the first sequence was reserved for the protein of interest having a known structure, e.g. BAZ2A PHD in the PHD alignments. For the PDZ domain (PFAM accession number PF00595.71), Drosophila inactivation-no-after-potential (InaD) protein (INAD_DROME, pdbid: 1ihj) was used as the first sequence for MI analysis. An earlier large-scale mutational analysis of PDZ peptide binding site had noted that 16-18 positions around the binding site confer binding specificity[40]. The binding site in that study was defined by PDZ pdbids (1ihj, 1n7t, 2h2b, 1i92, etc.), and therefore, we used

1ihj (INAD_DROME) as reference sequence as its structure is in complex with the peptide. We used these 16-18 residue positions of INAD_DROME around its binding site as a representative of the PDZ binding site (Figure 3.22A). To avoid confusion, the residue numbering is that of INAD_DROME (Figure 3.22A, B, C), and for convenience, their correspondence to PDZ family alignment is indicated in Figure 3.22B. The MI score text file was reformatted as a matrix for creating MI density plot (e.g., Figure 3.21B, 3.22A) using the Matrix2png[92] program. For display of the CMI scores mapped onto the structure, the CMI score of a residue replaced the b-factor column of the corresponding residue in the coordinate file. The 85% non-redundant alignments of PHD, PDZ and Bromodomains (PFAM accession number PF00439.22) respectively had 3469, 4950 and 3477 sequences. For MI calculation of artificial sequences (see below), a set of 1000 sequences was used, and for PHD, BAZ2A sequence was once again placed at the first sequence among the set of 1000 artificial sequences.

GREMLIN restraint score analysis

The pre-computed GREMLIN results (residue pair covariance and the consensus sequence) of each PFAM domain were manually downloaded from the GREMLIN[87, 88] webserver. The GREMLIN result text file, consisting residue pair covariance information, contains three scores: (i) raw score (r_sco), (ii) scaled score (s_sco), and (iii) probability score (prob). The scaled score (s_sco) represents covariance normalized among the residue pair scores of a PFAM domain family[87, 88]. These paired scores correspond to residues of GREMLIN consensus sequence. We use MUSCLE alignment (see above) to get the correspondence between residues of the GREMLIN consensus

sequence and the sequence of protein of interest (Figure 3.8C). GREMLIN also imposes a sequence separation cutoff of 3 residues for a pair of residues to have covariance score. Therefore, we do not have s_sco score for the D2−W pair, as the sequence separation is 2 in the PHD alignment (Figure 3.21B). A residue can have more than one s_sco score as it can be in contact with several residues, and the highest s_sco score of a residue was mapped on to the structure by manipulating the b-factor column of the coordinate file (Figure 3.8A). The root mean squared deviation (RMSD) of Cα-atoms between the corresponding positions in MUSTANG structural alignment is mapped to the structure by replacing the b-factor column with the RMSD. PFAM domain structures with resolution better than 2.2Å and having a sequence identity ≤ 40% with each other were considered for the GREMLIN−RMSD analysis (Figure 3.8B). Even though there are several new structures available for the PFAM domains analyzed here, for the GREMLIN−RMSD analysis, pdbids listed as "DR PDB" record within the header of pfam alignment file in STOCKHOLM 1.0 format were used for the convenience of running automated scripts, and the PFAM alignment files were downloaded in 2015 November.

GREMLIN restraints for other PFAM peptide-binding domains (e.g., the PDZ domain and bromodomain) were also examined for a comparative study. Here we utilized the normalized GREMLIN restraint score (s_sco, see above) to map on to the structure to visually infer the relative contributions of residues to folding (Figure 3.23A). Compared with other PHD positions, the s_sco scores of the binding-site positions (including W, L2, and D2) rank high (Figure 3.23A). In comparison, the s_sco scores of regions remote from the binding site in the PDZ domain and bromodomain (Figure 3.23A) are higher than that of the respective binding sites. This could also be attributed to the larger size of

the PDZ domain (~80 residues), the bromodomain (~85 residues). Compared with the ~50-residue PHD scaffold, PDZ and bromodomain have many more inter-residue contacts, which are necessary to fold the scaffold. The PHD peptide-binding site is also positioned between two Zn-chelation centers, the key folding nucleus of the scaffold. The higher s_sco score for binding-site residues of the PHD scaffold is therefore likely justified. Consistent with the GREMLIN score, a comparison of the RMSD of crystal structures of PFAM domain members sharing less than 40% sequence identity shows that the structural deviation of the PHD peptide-binding site is smaller (Figure 3.23B).

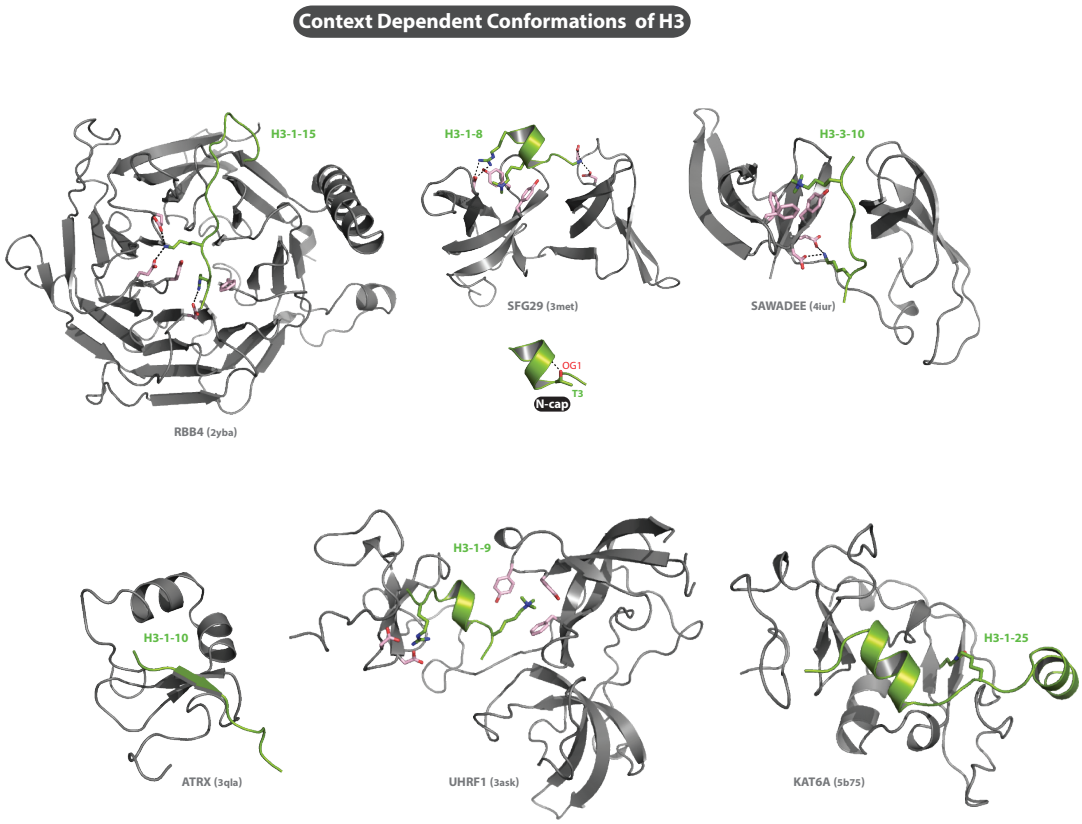


Figure 3.8. Context dependent conformations of the H3 peptide: Structural representations of the conformations of histone H3 involving regular (mediated by backbone hydrogen bonds) as well as non-regular secondary structures. This structural malleability of the disordered histone H3 segment thus likely contributes to the *one-to-many* interactions[93] of chromatin network hubs. Protein names and pdb ids are

indicated. The role of H3T3 OG1 atom as N-cap is also observed in other H3 complexes. Additional factors (other than those listed in Figure 3.18D bottom) might play important role in the context dependent conformations of the H3 peptide. For example, with UHRF1, only the first helical turn of H3 is observed where the *t*andem *t*udor *d*omain (TTD) anchors the H3K9me3 residue, while 3 helical turns are observed with DPFs where PHD1 of DPFs anchor the alkylated H3K14 residue.

Perturbed and Artificial Sequences

The multiple sequence alignment (MSA) of a PFAM domain consisting of 85% non-redundant sequences is considered as the master MSA of the family. 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 3.7, left) 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, Δp = (conditional probability) - (probability) (see Figure 3.9A). For Δp, a positive value is considered as positive enrichment ($\uparrow$) and a negative value is considered as negative enrichment ($\downarrow$) (Figure 3.9A, 3.7C). For convenience, Δp is calculated for the most frequent amino acid in a column. The probability of the most frequent residue at each master PHD alignment position is shown in Figure 3.10A (right, top) and the

corresponding conditional probability when perturbed by the D2 position in Figure 3.10A (right, middle) as an example. We note negative enrichment for the W position when perturbed by positions characteristic of the PHD_nW_DD subtype (Figure 3.9A, 3.10A bottom). 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 the vice versa. As PHD_nW_DD do not have W in the W-position, but a G, we use both symbols in the discussion. We were cautious in interpreting enrichment for regions with a poor quality of alignment. For example, the quality of alignment in the region housing BAZ2A V1713, particularly, in the PHD_W group is less reliable (Figure 3.19B, 3.19D). This is due to the insertions and deletion, as can be noted with PHD_W structures having helices of varying lengths in this region (Figure 3.17D). Therefore, the estimate of occurrence frequency of specific nonpolar residue ($p = 0.23$) for this region in the PHD_W subtype is less reliable (Figure 3.19B).

We generated artificial sequences in the following manner. Amino acids, in the row positions of each column of the master alignment, are assigned randomly such that the probability of occurrence of an amino acid in that column remained the same as that of the master alignment. In one iteration, 1,000 artificial sequences were generated, and 100 iterations were carried out. The following two conditions were checked: (i) E-value of an artificial sequence was lower than a preset value (Figure 3.11A), and (ii) ΔEntropy (the numerical difference between the entropy of the corresponding columns in the alignments of artificial and natural sequences) was lower than a preset value (Figure 3.10C, right). The E-value of sequences was calculated by hmm2search of HMMER2[94] package using PFAM PHD hmm. The set of 1,000 artificial sequences were perturbed similarly, and the

probability and conditional probability of the most frequent amino acids in the corresponding columns of artificial and perturbed artificial sequences calculated (Figure 3.10B, C). In Figure 3.10B (right), the vertical arrow represents the conditional probability of the indicated position when perturbed by the D2 position while the horizontal arrows indicate the direction of the deviation of the conditional probability obtained from artificial sequences from that obtained from real sequences. The direction of the horizontal arrow also indicates, positive enrichment (right-hand direction) and negative enrichment (left-hand direction).

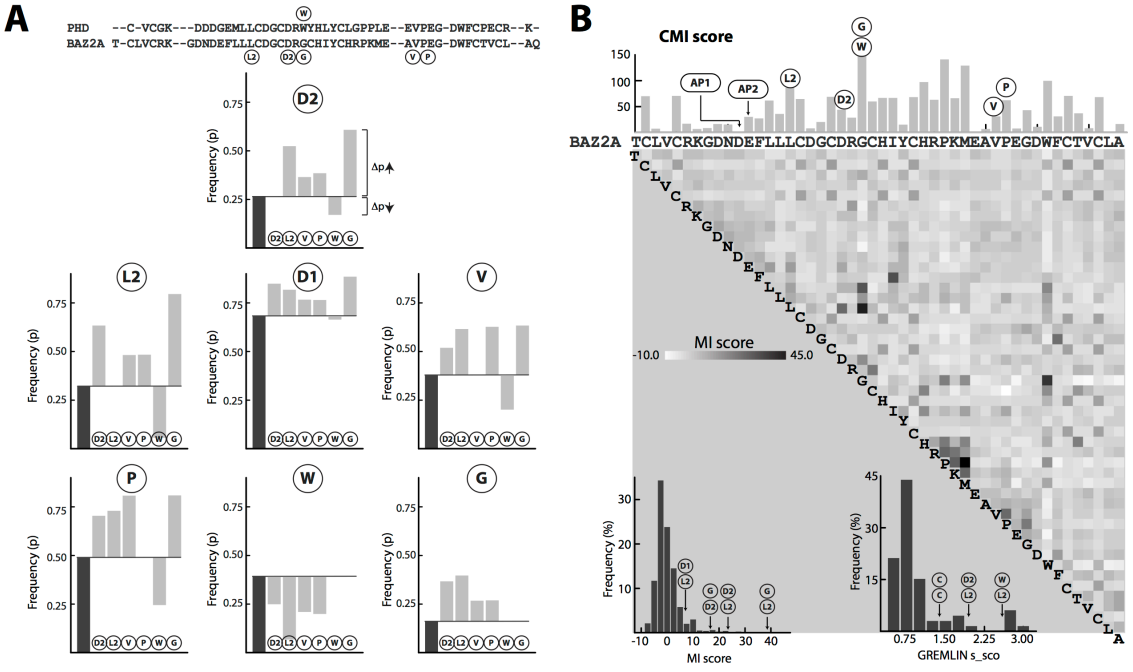


Figure 3.9. Reciprocal enrichment and mutual information (MI) of PHD: (A) The occurrence frequency (black bar) at an indicated position is altered (gray bar) as a result of perturbation. The gray bars represent change in frequency (Δp) with respect to the horizontal line (black bar height). The perturbed positions are indicated with encircled smaller fonts while encircled larger fonts represent positions undergoing frequency change. For convenience, the most frequent amino acid at each position in the 85% non-redundant PFAM and corresponding BAZ2A PHD sequences are at the top. (B) The mutual information (MI) between

positions in the PHD finger family. The CMI score for each position is at the top and the matrix represents the MI scores between positions. The distribution of the MI values (left) and the GREMLIN s_sco values (right) between positions are at the bottom.

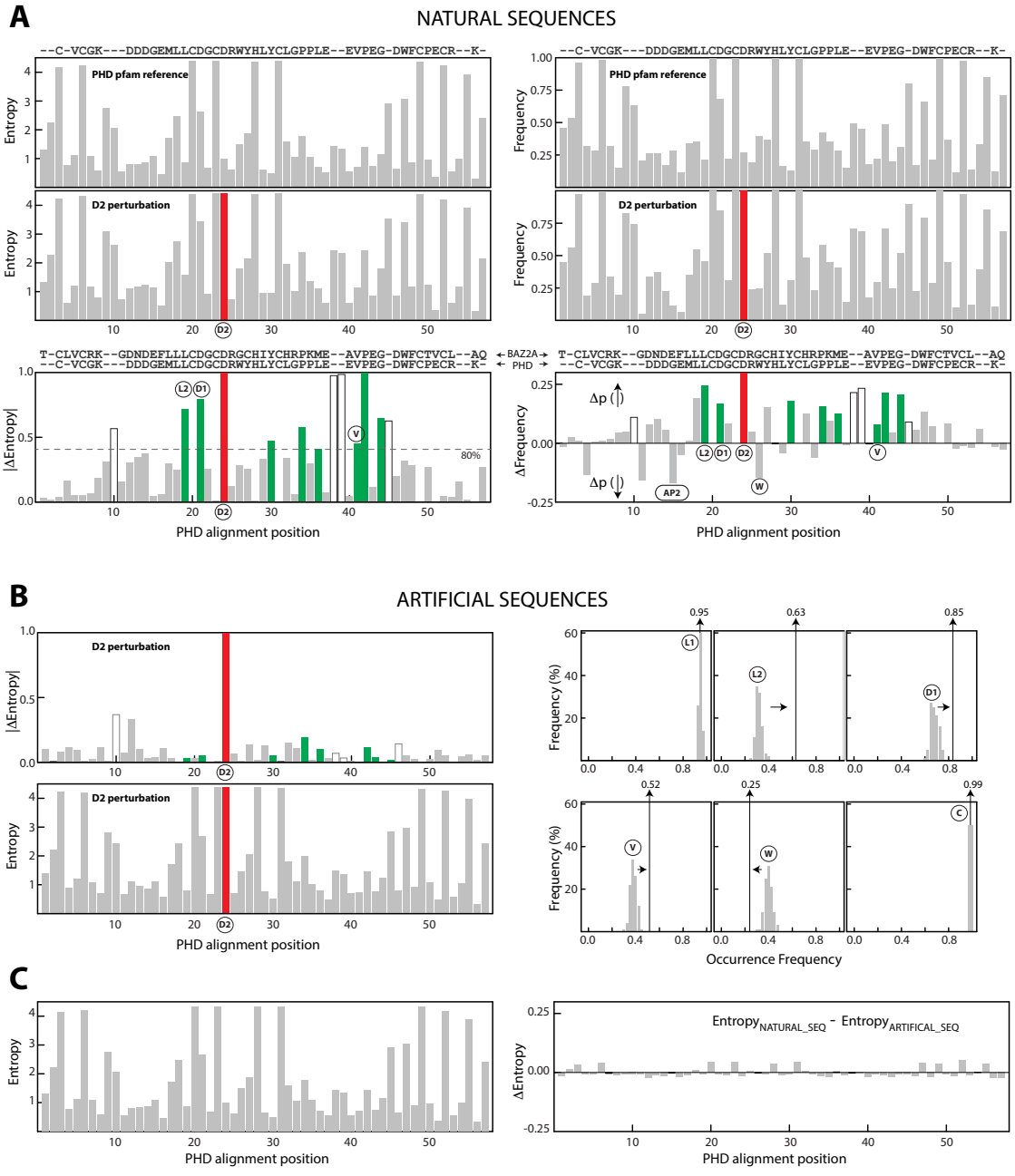


Figure 3.10. Characteristics of PHD Natural and artificial sequences: (A) The characteristics of each position of the aligned PHD natural sequences are categorized by entropy (left panel) and occurrence

frequency of the most frequent residue (right panel). The top and the middle panels respectively are for natural and perturbed natural sequences. As an example, the D2 position (red bar) is perturbed here and the perturbation is achieved by retaining only Asp at the D2 position. The bottom panel is respectively for changes in entropy (ΔEntropy, left) and frequency (Δp, right) upon perturbation at D2 position. For convenience, modulus of ΔEntropy (|ΔEntropy|), and positive enrichment ($\Delta p > 0$, symbol '↑') and negative enrichment ($\Delta p < 0$, symbol '↓') are indicated. Green bars in the bottom panel are for positions showing ΔEntropy higher than that of the remaining 80% of the positions (dashed line). The aligned sequence of PHD (top) is represented by the most frequent amino acid at each position in the 85% non-redundant PFAM sequences. The aligned BAZ2A PHD sequence representing the correspondence between positions is also shown for convenience. Empty bars are for positions where a gap is the most frequent symbol for an alignment column. (B) Characteristics of the artificial sequences categorized by entropy (left) and occurrence frequency (right). For the left panel, only one iteration of the set of 1000 sequences is shown. The right panel includes 100 iterations of the set of 1000 sequences. The distribution of the conditional probability in the 100 iterations of the set of 1000 artificial sequences is compared with the conditional probability (vertical arrow) of obtained from natural sequences. The vertical arrow represents the conditional probability of the indicated position when perturbed (e.g., the D2 position) while the horizontal arrows indicate the direction of the deviation of the conditional probability obtained from artificial sequences from that obtained from natural sequences. The direction of the horizontal arrow also indicates, *positive enrichment* (right-hand direction) and *negative enrichment* (left-hand direction). For reference, the behavior of Zn-chelating-Cys residue is included to show that highly conserved positions do not undergo change between natural and artificial sequences. (C) Entropy (left) and ΔEntropy (right) from the average of 100 iterations show that entropy of a column of aligned artificial sequences is very similar to that of natural sequences.

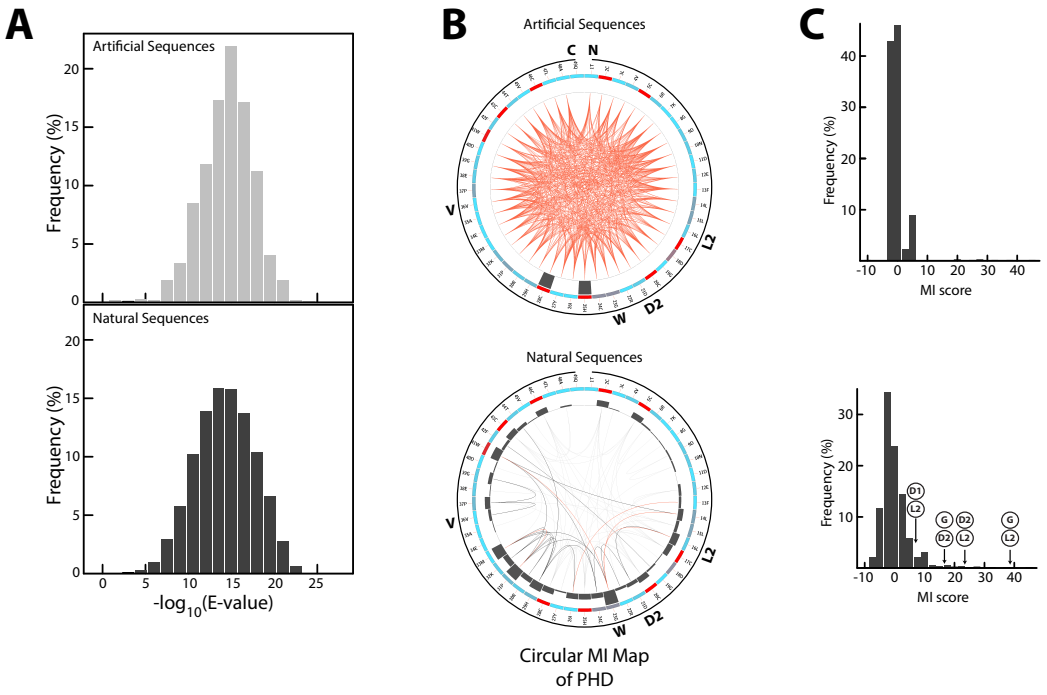


Figure 3.11. E-values and MI of PHD artificial sequences: (A) The $-\log_{10}$(E-value) of the artificial (top) and natural (bottom) of PHD sequences. (B) The inter-residue MI values represented in a circular form along the protein sequence (created by the MISTIC web server) are shown for the artificial (top) and natural (bottom) sequences. The positions of interest (L2, D2, W etc.) are indicated on the BAZ2A sequence, while N and C respectively represent the N- and C-terminus of the protein. (C) Comparison of the inter-residue MI scores between artificial (top) and natural (bottom) sequences.

Peptide Binding-site Hotspot Propensity

Using the pepx[95] high-resolution peptide-protein complex structural dataset (described in Chakravarty et al.)[7], we carried out peptide hotspot propensity calculations following the procedure of London et al[46]. Briefly, Rosetta in silico alanine scanning mutagenesis[96] was used to estimate the energetic contribution of peptide residue in the selected peptide-protein complexes. Peptide residues with ΔΔG of 1 kcal/mol or more were considered hotspots. 270 peptide-protein complexes, where the peptide had at least

one Arg (in 140 complexes) or one Lys residue (in 130 complexes), were selected for calculations. The frequency of occurrence (Fr_e) of peptide amino acids when making large energetic contributions towards overall binding ($\Delta\Delta G \geq 1$ kcal/mol) and the frequency of occurrence peptide amino acids (Fr_b) were used to compute the reference hotspot propensity ratio (Fr_e/Fr_b) (Figure 3.12B, top). Peptide residue with the propensity ratio (Fr_e/Fr_b) >1.0 was considered to be a hotspot. Propensity of peptide amino acids to appear adjacent to an Arg hotspot residue (ARG-Nearness) was calculated similarly by diving the occurrence frequency of peptide amino acids appearing adjacent to hotspot Arg residue (i.e. ± 2 residue positions) by Fr_b (Figure 3.12B, bottom). Frequency of occurrence of Arg hotspot residues was computed considering a 5-residue stretch spanning ± 2 residue positions on either side of the Arg hotspot residue (Figure 3.12A). Eighty five percent of peptide Arg hotspot residues were observed to be at the center of sequence segments spanning five residues (Figure 3.12A), while for the remaining 15%, the Arg residue is either close to the peptide terminus (e.g., H3R2 in PHD recognition) or is within a motif consisting of less than five residues (e.g., the WDR5-interacting [Win] –AR– peptide motif)[97].

Propensity of protein binding site hotspot residues making non-polar contact with peptide residues present at ± 2 residues of the peptide hotspot residue (see Figure 3.12C, right cartoon) was computed in the following way. LIGPLOT[98] was used to determine protein binding-site residues in contact with the peptide ligand. Non-polar carbon atom contacts were taken from LIGPLOT's ligand-protein contact list. The frequency of occurrence (Fr_np) of protein amino acids having non-polar contacts with peptide residues present at ± 2 residues of peptide hotspot residues (Figure 3.12C, right cartoon)

that make large energetic contributions ($\Delta\Delta G \geq 1$ kcal/mol) was computed. The frequency of occurrence of protein binding-site amino acids (Fr_pb) was used to compute the propensity ratio (Fr_np/Fr_pb) (Figure 3.12C). A binding-site residue is one that has $\Delta ASA_{SC} > 0.0$ $Å^2$ (see below). A protein Arg residue side chain atom can also have non-polar contacts using its $-(CH_2)_3-$ group, and therefore, polar residues can also be present in Figure 3.12C.

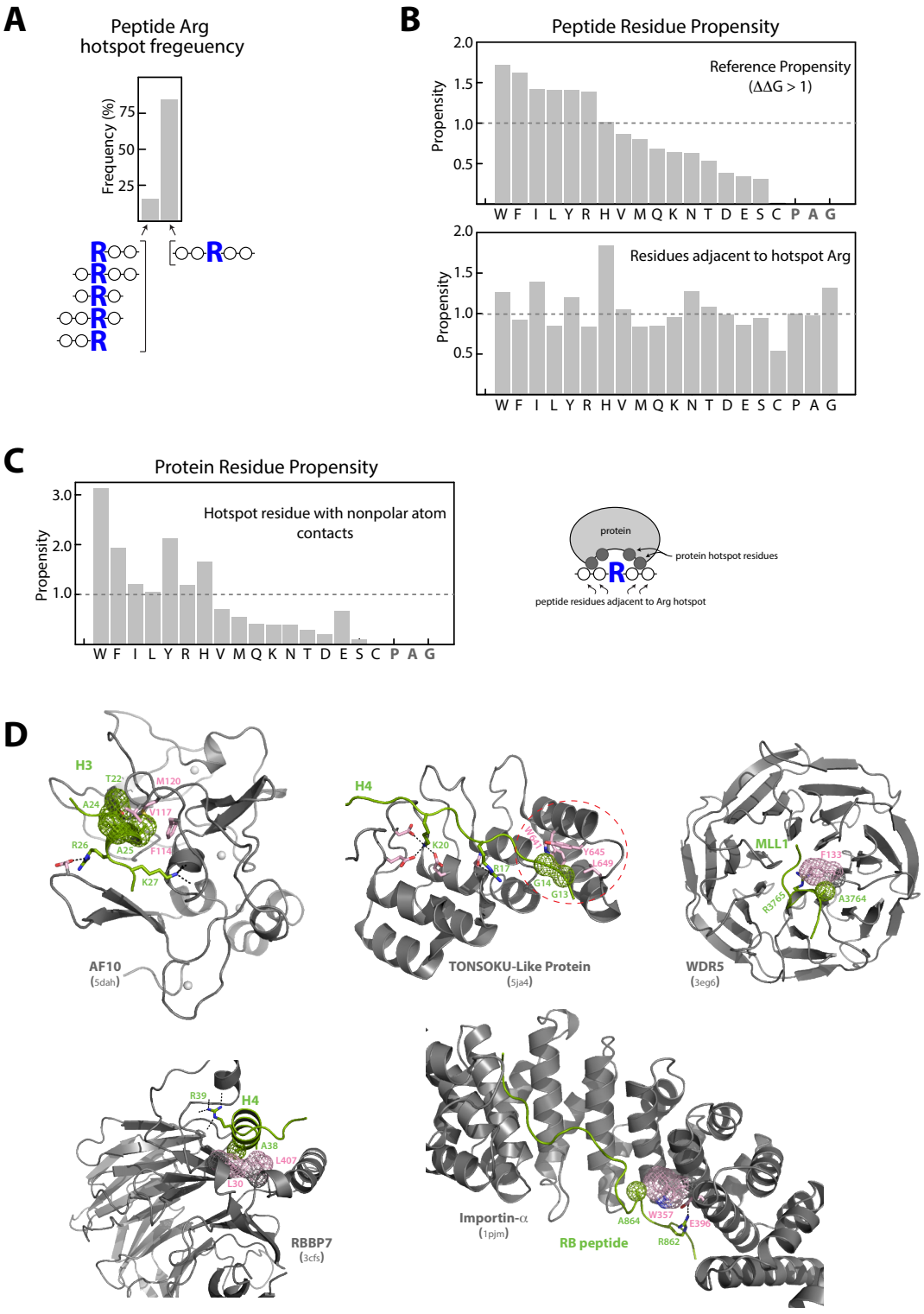


Figure 3.12. Recognition characteristics of Arg-rich peptides: (A) The frequency of occurrence of peptide Arg (R) hotspot (rossetta ΔΔG > 1.0) residues in peptides. Fifteen percent of times, the Arg-hotspot residue

appears close to the peptide terminal or within motifs smaller than five residues, while the remaining 85% of times it appears at the center of the sequences spanning five residues. (B) The propensity of peptide hotspot residues (top) and the propensity of residues to appear adjacent to Arg hotspot residue (± 2 residues) (bottom). (C) The propensity of protein hotspot residues (left) making non-polar contacts with peptide residues adjacent to the peptide Arg hotspot residue (right cartoon). (D) Examples of proteins (gray) recognizing histone/histone-like Arg-rich peptides (green) where peptide residues with small side chains (Gly, Ala and Thr) are in contact with bulky nonpolar (light pink) protein residues. Mesh surface represents the peptide residues with small side chain. Protein name, residue name and the pdb ids are indicated.

Packing Density, Surface Area and Secondary Structure

We used the Voronoia program[99] to compute packing density of atoms. Voronoia program reports packing density (PD) of an atom as the ratio, PD = VvdW/(VvdW+ Vse), where VvdW is the volume inside the van der Waals sphere of an atom, and Vse is the remaining solvent excluded volume assigned to each atom[99]. For PD of a set of atoms, we average the PDs of the set of atoms of interest. Only high-resolution structures (resolution ≤ 2.0Å) were considered for packing calculations. The packing density of atoms varies depending on their distance from the surface[100]. Therefore, atom depth[101] was computed for packing analysis. Atom depth was computed here as the distance from the nearest surface atom (atoms with non-zero surface area). Reference distribution of packing density as a function of atom depth (see Figure 3.13) was calculated using the PISCES[102] non-redundant set (≤ 26% sequence identity and resolution ≤ 1.6 Å) of 3700 pdb single chains with size ≥ 50 residues. For discussion of the energetic contribution residues in the context of PHD scaffold structures, a set of 15 high-resolution (≤ 2.0 Å) PHD–peptide complexes belonging to different subtypes (Figure 3.2A) 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. As peptide–protein interfaces have been observed to be very tightly packed, even more tightly packed than protein–protein interfaces[46], we rationalized the contributions of nonpolar residues at L1 and L2 with respect to interfacial packing. Having observed the disruptive consequences of substitution of the tightly packed H3A1 residue, the $-CH_3$ group in several PHD–peptide complexes in an earlier study[7] encouraged us to look in a similar manner at interfacial nonpolar group packing in BAZ2A proteins.

NACCESS[103] program was used to compute the accessible surface area (ASA). The amount of surface area lost by the side chain atoms of a protein residue upon formation of a peptide–protein complex is measured by ΔASA_{SC}. ΔASA_{SC} of the ith residue is, ΔASA_{SC}_res_i = (residue_i_ASA_{SC}_protein) – (residue_i_ASA_{SC}_complex) where residue_i_ASA_{SC}_protein and residue_i_ASA_{SC}_complex are respectively the side chain atom surface area of the ith residue in the free and bound form of the protein. Protein residues with $\Delta ASA_{SC} \geq 10 Å^2$ are considered to make contact with the peptide (Figure 2.14A). ΔASA_{SC} of protein hotspot residues (Figure 3.14) is also computed this way. Assignment of the secondary structural elements (SSE) of coordinate files was carried out using the SEC-STR program[104], and the SSE symbol 'G' was taken as the 3_{10}-helix. Amino acid frequencies of segments with 3_{10}-helix conformation were used to check the preference of Asp residues within the 3_{10}-helical segments of the PISCES[102] non-redundant set.

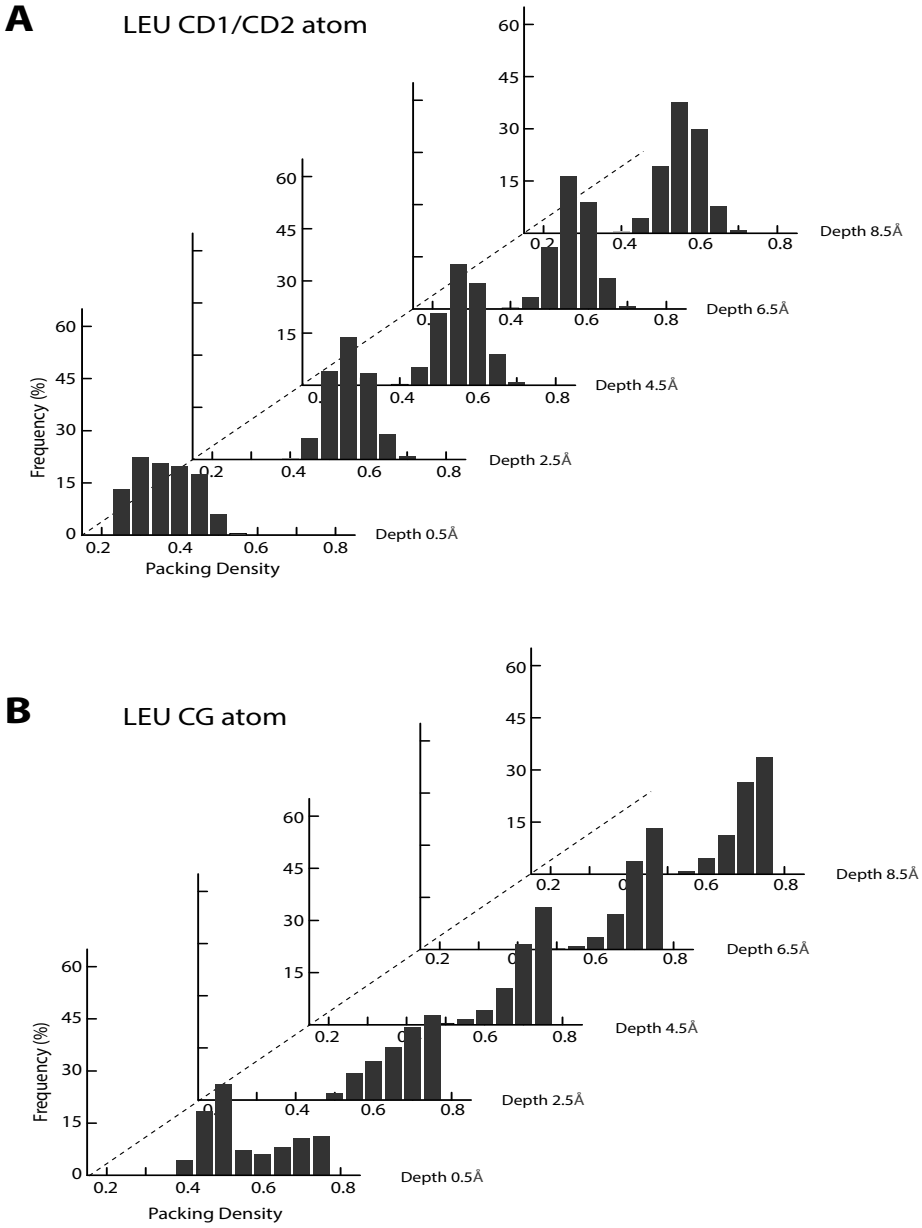


Figure 3.13. Reference packing density of nonpolar atoms: The reference packing density distribution of specific type of atoms as a function of their distance from the surface. These distributions were used here (e.g., Figure 3.17 and 3.19).

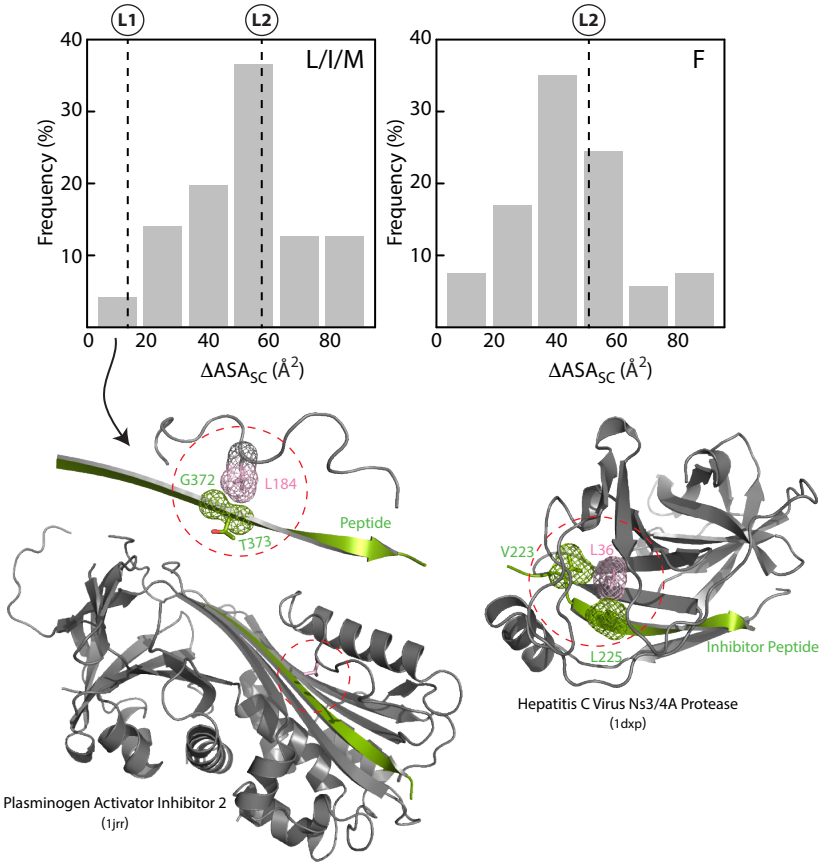


Figure 3.14. Characteristics of ΔASA_{SC} of protein hotspot residues: The distribution of the ΔASA_{SC} of Leu/Ile/Met (left, top) and Phe (top, right) of protein hotspot residues upon peptide binding. The dashed lines indicate the ΔASA_{SC} of L1 and L2 residues in PHD_nW_DD subtype. Above are examples where L1 like ΔASA_{SC} of peptide complexes are noted. Mesh surface represents the peptide and protein hotspot residues in contact with each other. Protein name, residue name and the pdb ids are indicated.

Peptide Swapping and Peptide Backbone Clash

Using the MUSTANG[85] structural alignment (Figure 3.2A), the coordinates of one PHD–peptide complexes were superposed and then transformed onto another. For convenience, we use one PHD–peptide complex as a reference, and the remaining complexes are superposed onto the reference molecule. After coordinate transformations,

the peptide coordinates of one complex are then placed onto another. For example, the peptide coordinates of the transformed NSD3-PHD complex are placed onto transformed coordinates of KAT6A-PHD2. We refer to this as peptide swapping. This is a simple but a crude approach, as it does not optimize backbone hydrogen bonding. The extended conformations of peptides in the PHD_nW subtype (2puy, 4gne, 4qbs and 3qla) were placed onto the PHD_nW_DD subtype (4q6f, 3asl, 5b75 and 5i3l). PHD–peptide inter-molecular backbone carbon–carbon contact distances for H3 residues 4-7 were determined using LIGPLOT[98] (Figure 3.21D). 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%) contributed by the crude swapping approach (Figure 3.21C). 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.

RESULTS AND DISCUSSIONS

Binding energetics of the BAZ2A PHD finger in the context of the PHD superfamily

Though diverse sequences populate this family, PHD finger sequences can be grouped into two major functional subtypes[7]: (i) the PHD_W subtype and (ii) the PHD_nW subtype (Figure 3.7). The justification for this grouping is that functionally related positions show enrichment in specific residue types among the subtype sequences; e.g., aromatic cage-forming positions are enriched with aromatic residues in the PHD_W subtype[7]. Positions in the PHD finger superfamily corresponding to the BAZ2A PHD were similarly checked for residue-type enrichment. The overall result for the occurrence

frequency and residue enrichment at the L1 and L2 positions of the PHD superfamily is represented as sequence logos in Figure 3.15. It is clear that bulky nonpolar residues (Ile, Val, and Leu) are always present at the L1 position irrespective of the subtype; i.e., p = 0.95 is observed for the entire superfamily (Figure 3.15A, B). However, this is not the case for the L2 position, as the occurrence frequency of bulky nonpolar residues is progressively enriched (0.32 → 0.51 → 0.82) as one moves from the superfamily level to specific subtypes (Figure 3.15B). Interestingly, the value of p for the bulky nonpolar residues at the L2 position in PHD_W subtype sequences is negligibly small (p = 0.05). The PHD treble clef knuckle xCDxCDx sequence pattern mentioned above refers to the PHD_nW_DD subtype, which includes the BAZ2A PHD, KDM5B PHD, UHRF1 PHD, KAT6A DPF, and DPF3 DPF domains. A bulky nonpolar residue at L2 seems to be a key feature of the PHD_nW_DD subtype. As a control, we also checked the xCDxCDx motif-containing sequences (e.g., TAF3 PHD, DIDO PHD) in the PHD_W subtype, and we observed that they never include bulky nonpolar residues (p = 0.03) at L2 (Figure 3.15B), suggesting that PHD_nW_DD is likely to be a distinct subtype. In other words, in addition to having the pair of Asp residues in a treble clef knuckle, the PHD_nW_DD subtype likely has additional features, such as a bulky nonpolar residue at L2. To confirm the energetic contribution of residues at L2 for the PHD_nW_DD subtype, we also mutated bulky nonpolar residues to Ala at the L2 position in hUHRF1 PHD (M345A), hKDM5B PHD (L326A), and hKAT6A DPF (F280A) domains (Figure 3.16A and see Figure 3.2 alignment). The L2 position mutants of the hUHRF1 PHD, hKDM5B PHD, and hKAT6A DPF did not show any detectable peptide binding according to ITC (Figure 3.16A), 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 sequence identity distribution within subtypes (Figure 3.7B) suggests that, although the median value increases from the superfamily level to the PHD_nW_DD subtype, there are diverse sequences in the PHD_nW_DD subtype. For example, the hBAZ2A PHD, hUHRF1 PHD, and hKAT6A PHD2 share less than 40% sequence identity with each other. Yet these members retain similar peptide-anchoring residues, indicating a similar binding mechanism and therefore providing a reason to select them for the subtype study.

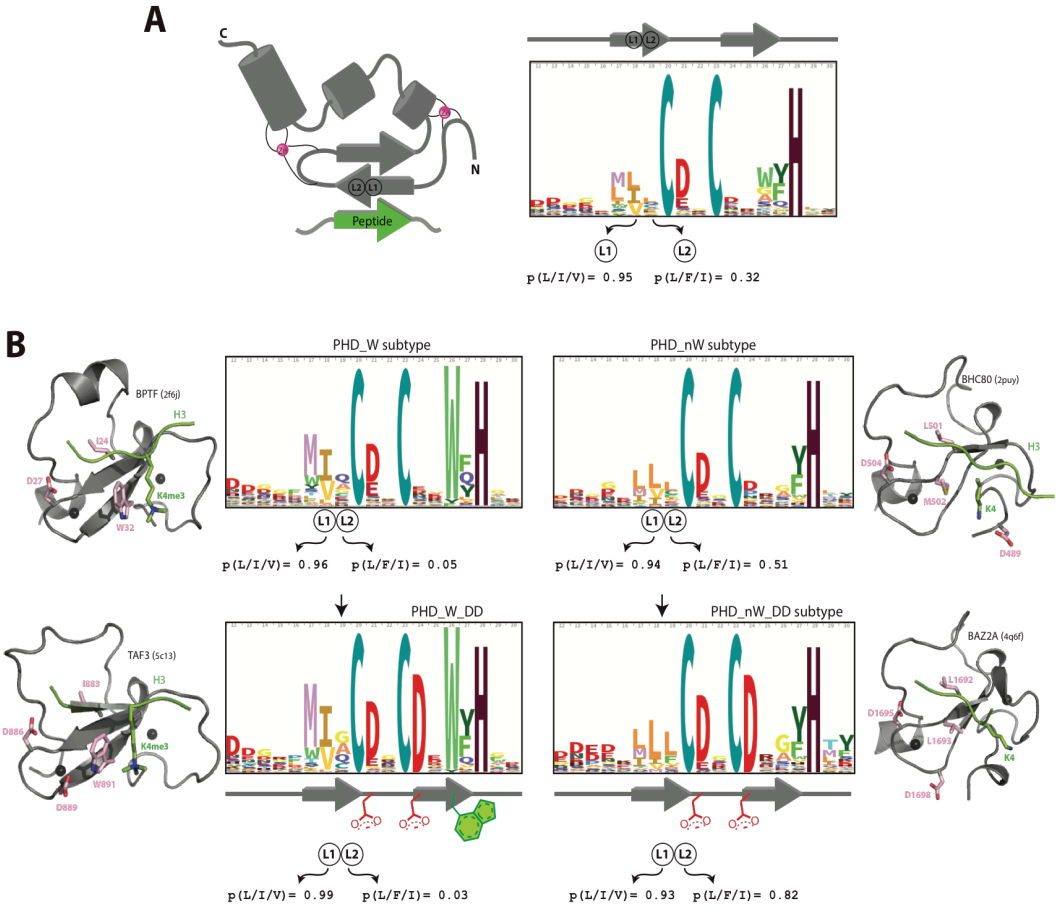


Figure 3.15. PHD subtypes, and L1 and L2 enrichment: (A) PHD topology cartoon (left) showing the strand that augments the peptide (green) harboring the L1 and L2 positions. Sequence logo (right) showing strand-knuckle-strand segment of the PHD fold with the secondary structural elements on top. The occurrence frequency of bulky nonpolar residues at L1 and L2 positions is below the logo. (B) Sequence logo of the two major PHD subtypes, PHD_W (left) and PHD_nW (right). Structures of representative members of these subtypes are adjacent to the logo showing the characteristic key residues in stick (light pink). The lower panel (shown with arrow) indicates logo of sequences extracted from the two respective subtypes with the knuckle having xCDxCDx sequence pattern. The occurrence frequency of bulky nonpolar residues at L1 and L2 positions is similarly below the respective logos. The pdb ids are in brackets, and convenience H3C4 atom in TAF3 PHD is represented as N (blue).

Justifying the contribution of the L1 position

For the BAZ2A complex, the L1692 Leu CD1 atom is in van der Waals contact with the H3A1 residue $–CH_3$. The packing density of the BAZ2A L1692 Leu CD1 atom in the BAZ2A complex is observed to be higher than what is typically observed for the Leu CD1 atom at similar distances from the protein surface (Figure 3.16B), justifying the role of the L1692 side chain in tightly anchoring the H3A1 residue $–CH_3$. The packing density of the BAZ2A L1692 Leu side chain atoms in the complex is 0.62. Irrespective of the PHD subtype, L1 residues ($p = 0.95$) in the 15 PHD–peptide complexes are also in van der Waals contact with the H3A1 $–CH_3$ functional group. The average side chain packing density (0.623 ± 0.032) of nonpolar L1 residues in these complexes suggests that the L1 residues are packed very similarly as in the BAZ2A complex (Figure 3.16C). The similarity between H3A1–CH_3 anchorage by BAZ2A L1692 and by the L1 residue in the rest of the complexes can further be appreciated by noting that ΔASA_{SC} for the L1 residue in 15 complexes (17.15 ± 4.4 Å^2) is close to that of BAZ2A L1692 (~15 Å^2). In

other words, despite low sequence identity, diverse members of the PHD superfamily utilize the L1 nonpolar residue in nearly identical fashion in anchoring H3A1–CH_3, and its substitution thus likely leads to a poorly packed interface, resulting in the disruption of the complex. The replacement of L1 nonpolar residues to Ala in the BAZ2A, UHRF1, KAT6A, and KDM5B (Figure 3.16A) PHD fingers therefore severely compromises peptide binding. In the recently acquired structures of the zf-CW domains[105, 106] (e.g., MORC3, ZCPWP2, and ZCPWP1) in complex with the histone H3K4me3 peptide, one also observes that H3A1–CH_3 (Figure 3.2C) is in van der Waals contact with a highly conserved nonpolar residue ($p = 0.80$) in the zf-CW scaffold, located at a position corresponding to that of the PHD L1 position (Figure 3.2B–D). As both PHD and zf-CW domains require the peptide NH_2-terminal for backbone-mediated hydrogen bonding, it is interesting to note that both have adopted very similar mechanisms for anchoring the terminal Ala –CH_3 group by utilizing a nonpolar L1 residue (Figure 3.2B–D). These observations are also important in the context of histone reader reengineering. Reengineering the PHD and zf-CW domains for effectively reading marks other than at the H3-terminal would most likely require manipulation of the L1 position.

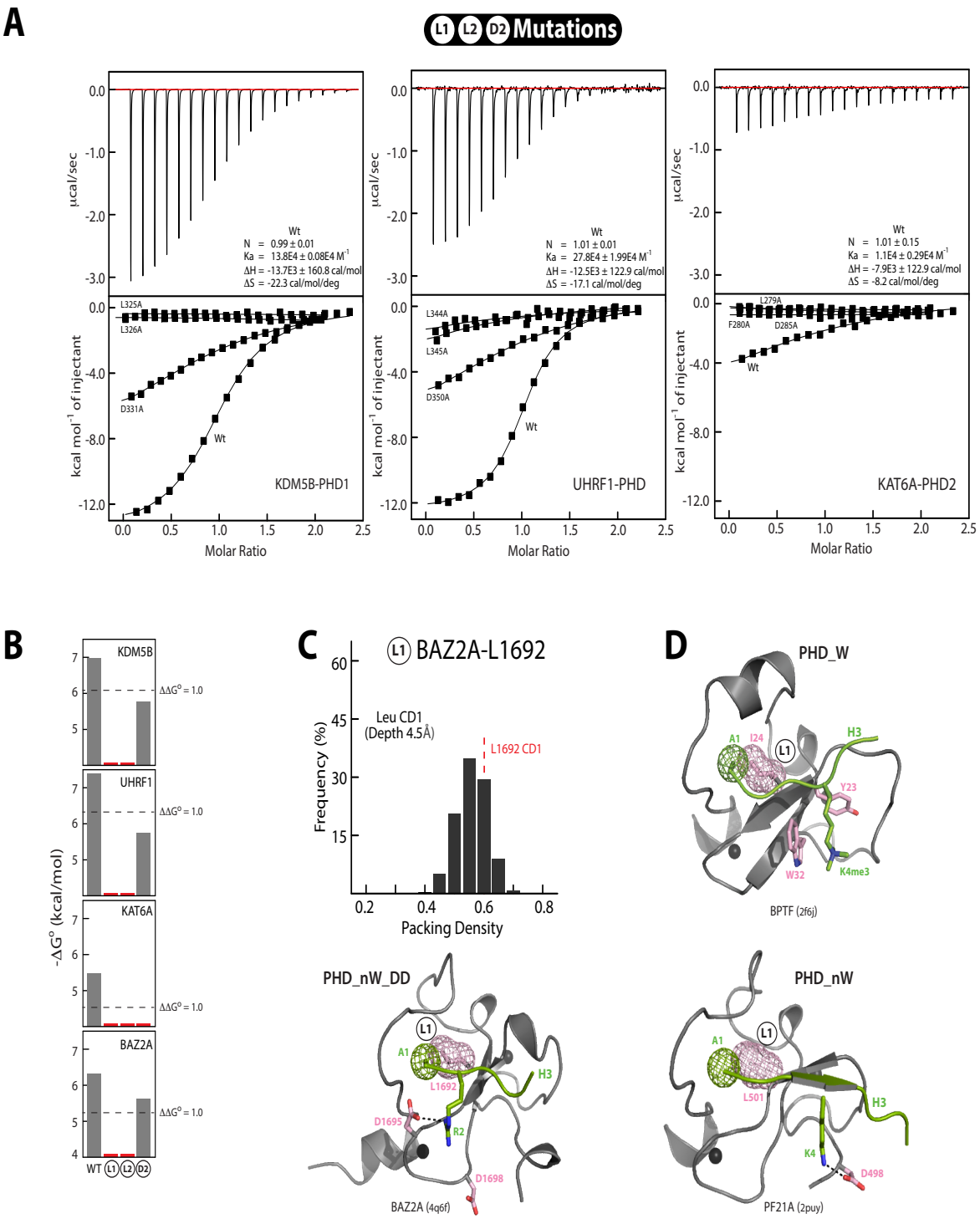


Figure 3.16. PHD_nW_DD subtype mutations: (A) ITC titration profiles of wild type and mutants of hKDM5B PHD (left), hUHRF1 PHD (middle) and hKAT6A DPF (right) with histone H3 peptide. (B) ΔG^{o}

of the forward binding reaction is shown (as $-\Delta G^{o}$) for the three *readers*, hKDM5B PHD (top), hUHRF1 PHD (middle) and hKAT6A DPF (bottom). For convenience of comparison, BAZ2A mutants are shown at the bottom. Residues that contribute more than 1 kcal/mole of free energy are below the dotted line. Red bars are for positions where binding was not detected in ITC experiments. (C) Distribution of packing density of Leu-CD1 atom (top) in proteins within the shell of 3.5 – 5.5Å from the protein surface, and the packing density of BAZ2A L1692 CD1 (red dashed line) within similar distances from the surface in the BAZ2A complex. Surface mesh (bottom) representing the contact between H3A1-CH_3 and L1 position in BAZ2A. (D) Surface mesh representing the contact between H3A1-CH_3 and L1 position in different subtypes. In the mesh representations, protein carbon atoms are in light pink and peptide carbon atoms are in split pea color.

Justifying the contribution of the L2 position

In a similar manner, we then looked at the BAZ2A PHD L1693 residue. The CD1 and CG atoms of this residue (L2) are in van der Waals contact with the histone H3 Thr3 backbone C atom (Figure 3.18B). Moreover, the packing density of the CD1 and CG atoms of this residue in the BAZ2A complex is similarly found to be higher than typically observed for other Leu CD1 and CG atoms at their respective distances from the surface (Figure 3.18A), and their contacts with the peptide backbone C atom maintain a well-packed environment for the PHD–peptide complex. The structural changes for the BAZ2A L2 position upon peptide binding are, however, different from those at the L1 position. The ΔASA_{SC} values for the L1 and L2 positions are ~20 $Å^2$ and ~55 $Å^2$, respectively, for BAZ2A (Figure 3.14). A value of 20 $Å^2$ for ΔASA_{SC} of a hotspot residue is, however, far less common (ASA ≅ 20 $Å^2$, Figure 3.14). A more common scenario for protein hotspot residues is the one observed for the L2 position, where the residue is nearly fully exposed and then undergoes a large change in surface area while

contributing to tight interfacial packing. The change in packing density (ΔPD) of the BAZ2A L2 side chain atoms is 0.16, while that of those at the L1 position is 0.08. Similar to the BAZ2A L1693 residue, bulky nonpolar residues at the L2 position in UHRF1 (M345), KAT6A (F280), and DPF3 (F333) complexes pack against peptide backbone carbon atoms (Figures 3.17A, 3.18B). In general, in the PHD_nW_DD subtype, L2 residue side chain atoms contribute toward tightly packing the peptide backbone atoms. This is in contrast to the contacts made by L2 residues in other subtypes. In the PHD_W subtype, L2 residues, with smaller side chains (e.g., Gly, see Figure 3.2A), are typically not in contact with H3 peptide backbone atoms (Figure 3.18C). Based on the side chain size, the PHD_nW subtype members may or may not retain contacts with the peptide backbone (Figure 3.17B). The side chain packing density of L2 residues in the three subtypes, 0.520 ± 0.045 (PHD_W), 0.542 ± 0.029 (PHD_nW), and 0.572 ± 0.038 (PHD_nW_DD), also shows a trend of enhancing packing for residues at L2 in the PHD_nW_DD subtype. In short, backbone contacts by the L2 residue are likely to contribute in a subtype-specific manner.

Why should there be a need for a contribution from the L2 residue in a subtype-specific manner? We will try to rationalize the answer for this in the following. Anchorage of the peptide side chain H3Lys4 quaternary ammonium ion ($–CH_2N^+(CH_3)_3$) by the PHD aromatic cage and the terminal H3A1–CH_3 by the PHD L1 likely dominate the peptide-binding energetics in the PHD_W subtype (Figure 3.17C). Salt bridge-mediated H3Lys4 and H3Arg2 anchorage and H3A1–CH_3 anchorages by L1 likely dominate the binding energetics in the PHD_nW subtype (Figure 3.17C). For the PHD_nW_DD subtype, we had earlier discovered that peptide binding might not require

an energetic contribution from the H3Lys4 residue[7]. The lack of energetic contribution from the H3Lys4 residue must be compensated by additional contacts, such as those made with bulky L2 residues. Thus, the L2 contribution is likely to be an essential feature for peptide binding in the PHD_nW_DD subtype, as we showed that substitutions of bulky L2 residues with Ala in the BAZ2A, KDM5B, UHRF1, and KAT6A proteins (Figure 3.16A) completely disrupt peptide binding for all of them. Other subtypes may not require L2 contacts ($p \cong 0.05$), 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 and structures of the PHD_nW subtype, one observes that the DNMT3A ATRX–DNMT3–DNMT3L (ADD) domain retains the bulky nonpolar L2 residue, but the ATRX and AIRE proteins do not (Figures 3.2A, 3.17D). It is interesting to note that the treble clef knuckle 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 3.2A, 3.17B). For AIRE, substitution at the L2 position (C310L) by introduction of a Leu residue enhances affinity for the unmodified H3 peptide by ~15 fold (Figure 3.18D). A gain in free energy ($\Delta\Delta G$) of ~1.5 kcal/mol could compensate for the loss of another interaction of similar strength. It is thus temping to speculate that the loss of one type of contact is likely to be compensated by another, such as the L2 contact. However, these speculations will need to be confirmed by detailed experimentation as part of a further investigation.

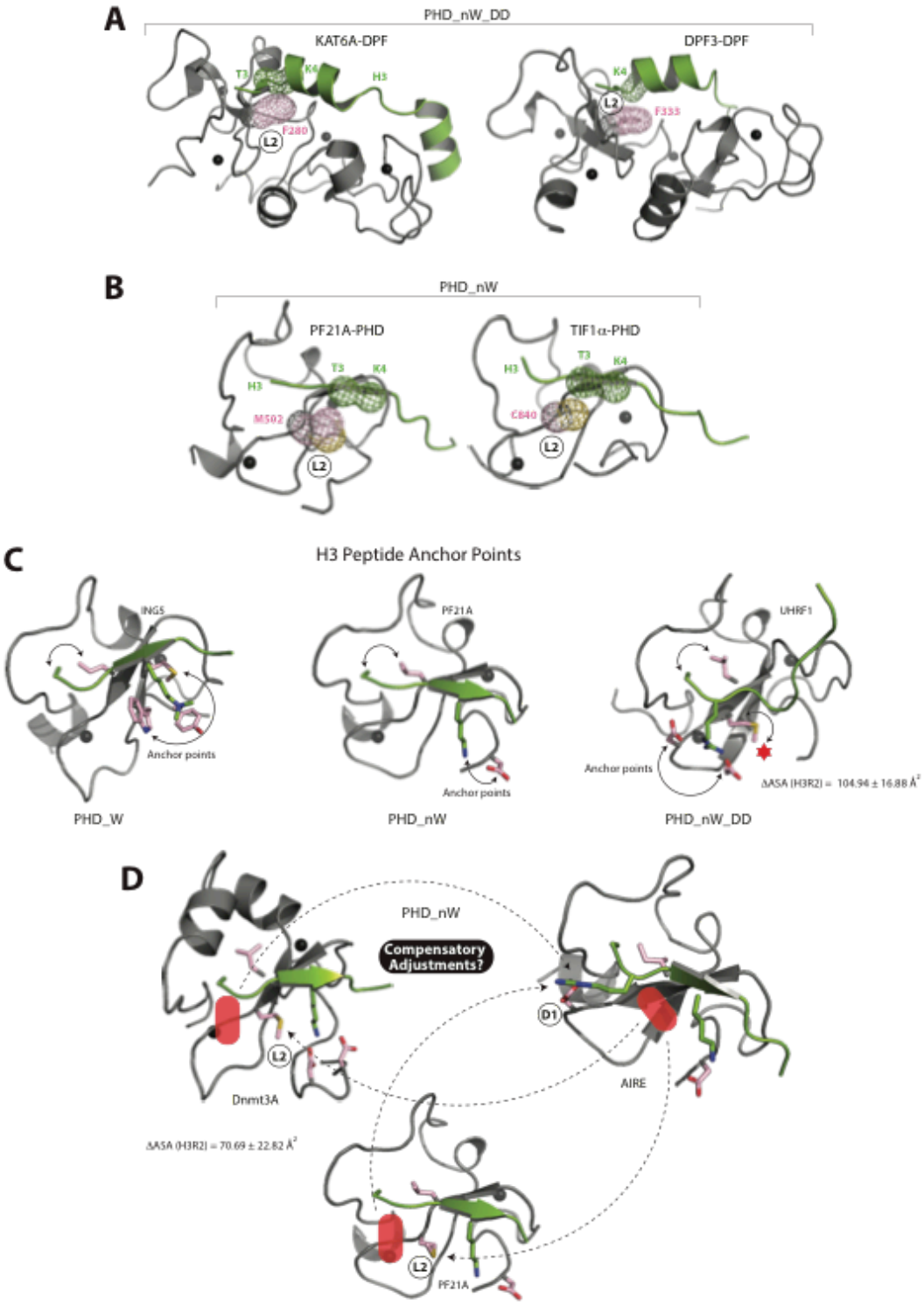


Figure 3.17. L2 contacts in other subtypes: L2 contacts in DPFs (A) and PHD-nW (B). (C) Anchor points in in different subtypes are represented by double headed arrow. (D) A symbolic representation of

adjustments due to position specific residue substitution. A red ‘capsule’ represents absence while the arrow (dashed line) indicates presence of a specific residue at the same site.

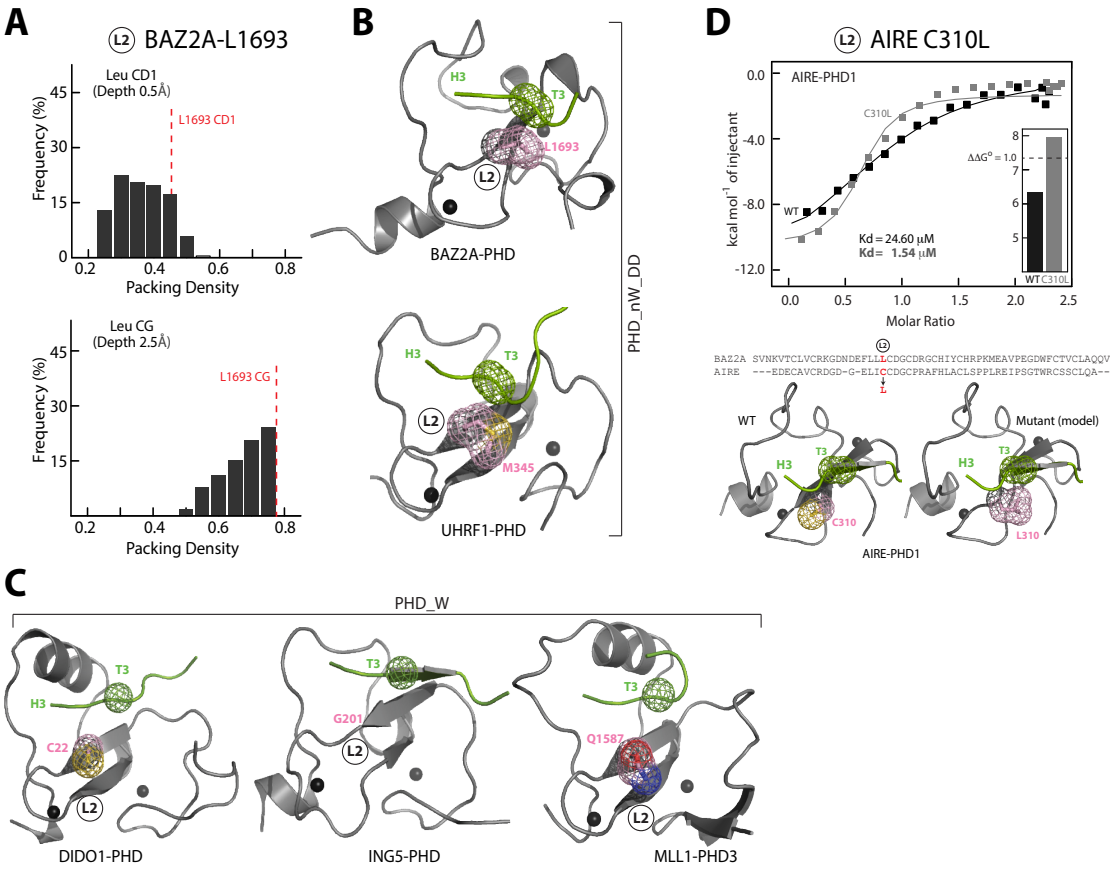


Figure 3.18. L2 mutations in PHD_nW_DD subtype: (A) Packing density of BAZ2A L1693 CD1 (top) and CG (bottom) are shown as red dashed lines among the distributions of Leu CD1 and CG atoms respectively within the indicated shells in protein structures. (B) Surface mesh representing the contact between BAZ2A L1693 side-chain and the peptide backbone (top) and similar L2 contacts for UHRF1 (bottom). (C) Mesh surface representations of the L2 contacts in PHD_W subtype for comparison. (D) ITC titration profiles of the wild type and L2 (C310L) mutant of AIRE PHD with histone H3-1-11 peptide. Mesh coloring is same as that in Figure 3.16.

Other nonpolar residue contributions at the interface of BAZ2A and the Subtype

Substitution of V1713 to Ala in BAZ2A (Figure 2.14, Table 2.8) also weakens peptide binding ($\Delta\Delta G$ = 1.1 kcal/mole). BAZ2A V1713 is in contact with histone H3Thr3–CH3 (the CG1 atom) and thus contributes to packing with the Thr3–CH3 group (Figure 3.19B, C). We similarly observed the enrichment for nonpolar residues at a position corresponding to BAZ2A V1713 (Figure 3.19B); e.g., the occurrence frequency is enhanced (0.38 → 0.64) in the PHD_nW_DD subtype (Figure 3.19B). However, the packing density of the V1713 side chain atom (Figure 3.19C) in contact with the peptide is lower than that observed for the L1 and L2 atoms. This is also consistent with the observation that the energetic contributions of L1 and L2 towards peptide binding are greater than that of V1713, suggesting that the tighter the packing, the larger the energetic contribution to be expected. Interestingly, the enrichment at this position is smaller than at the L2 position (0.32→0.81) and is also consistent with its respective energetic contribution. For the PHD_nW_DD subtype, however, it is interesting to note that, despite low sequence identity, BAZ2A, UHRF1, and KAT6A PHD2 have near-identical structures in this region (Figure 3.19B). Thus, the observed occurrence frequency of nonpolar residues in a position corresponding to that of BAZ2A V1713 in this subtype is more reliable than that of the PHD_W subtype, which includes insertions/deletions in this region (Figure 3.19B). We therefore believe that it is a subtype-specific adaptation to strengthen the anchorage of H3T3–CH3, as UHRF1 V365 and KAT6A M300 are also in van der Waals contact with H3T3–CH3, like BAZ2A V1713 (Figure 3.19D). This is also consistent with an earlier observation that substitution of H3Thr3 to Ala severely compromises peptide binding in BAZ2A and KDM5B[7]. With respect to PHD reader

contact with the peptide, the contribution of H3Thr3‑CH3 (CG1 atom) is anticipated to be different from that of the ‑OH group (OG1 atom) for the observed effects of the H3 (Thr→Ala) 3 substitution. This is further substantiated by the recently obtained DPF structures of KAT6A[24, 81] and DPF3[106], in which H3Thr3–OH is engaged as the N-cap[107] for the peptide helix, while H3Thr3–CH3 is oriented towards the enriched nonpolar residue side chain (Figure 3.19D). Thus, enrichment of multiple nonpolar residues (including Pro, L2) at specific positions is observed in the PHD_nW_DD subtype, which shows a strong preference for contacts with the H3Thr3 backbone (L2) and side chain atoms. As discussed above, a gain in specific contacts is likely to be a compensatory adjustment for the loss of another contact, and the PHD_nW_DD subtype is enriched with residues for making contact with the first three H3 residues (H3A1-T3) at the expense of other contacts. Tracing the orthologs of human PHD fingers, it had earlier been noted that KDM5B PHD and UHRF1 (of the PHD_nW_DD subtype) are likely among the early "inventions" of the PHD[7]. Among the orthologs of the KDM5B PHD, tracking the positions that show a gain in specific contacts with the peptide suggest that anchoring histone H3 by the first three residues (H3A1-T3) could have been an early adaptation by utilizing position-specific nonpolar residues (Figure 3.20A). The DPFs, in addition to these features, also extended their contacts with H3 well beyond the H3-1–3 residues by utilizing the non-canonical surface of the PHD1 of the DPFs[24, 81, 106, 108, 109]. Enrichment of multiple nonpolar residues among histone readers was also observed earlier; e.g., the histone H3K9me3 and H3K27me3 readout by the chromodomains of human CBX paralogues[110]. CBX2, -4, -6, -7, and -8 retain a pair of position-specific clasping nonpolar residues at the peptide binding site (Figure 3.20C) in place of a pair of

clasping negatively charged residues in the corresponding positions in CBX1, -3, and -5[110]. 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 H3K27me3[110]. Thus, the nonpolar residues in the CBX example contribute to lowering affinity and specificity. The packing density of the CBX2 binding-site nonpolar residue (V11 and L50) side chain atoms in complex with the H3K27me3 groups (Figure 3.20C) are only 0.39 and 0.43, respectively, which are much lower than at the L1, L2, and V positions of PHD_nW_DD. The enriched nonpolar residues observed here certainly contribute to strengthening specific interactions due to tight packing and thus are distinct in behavior from those of the CBX example.

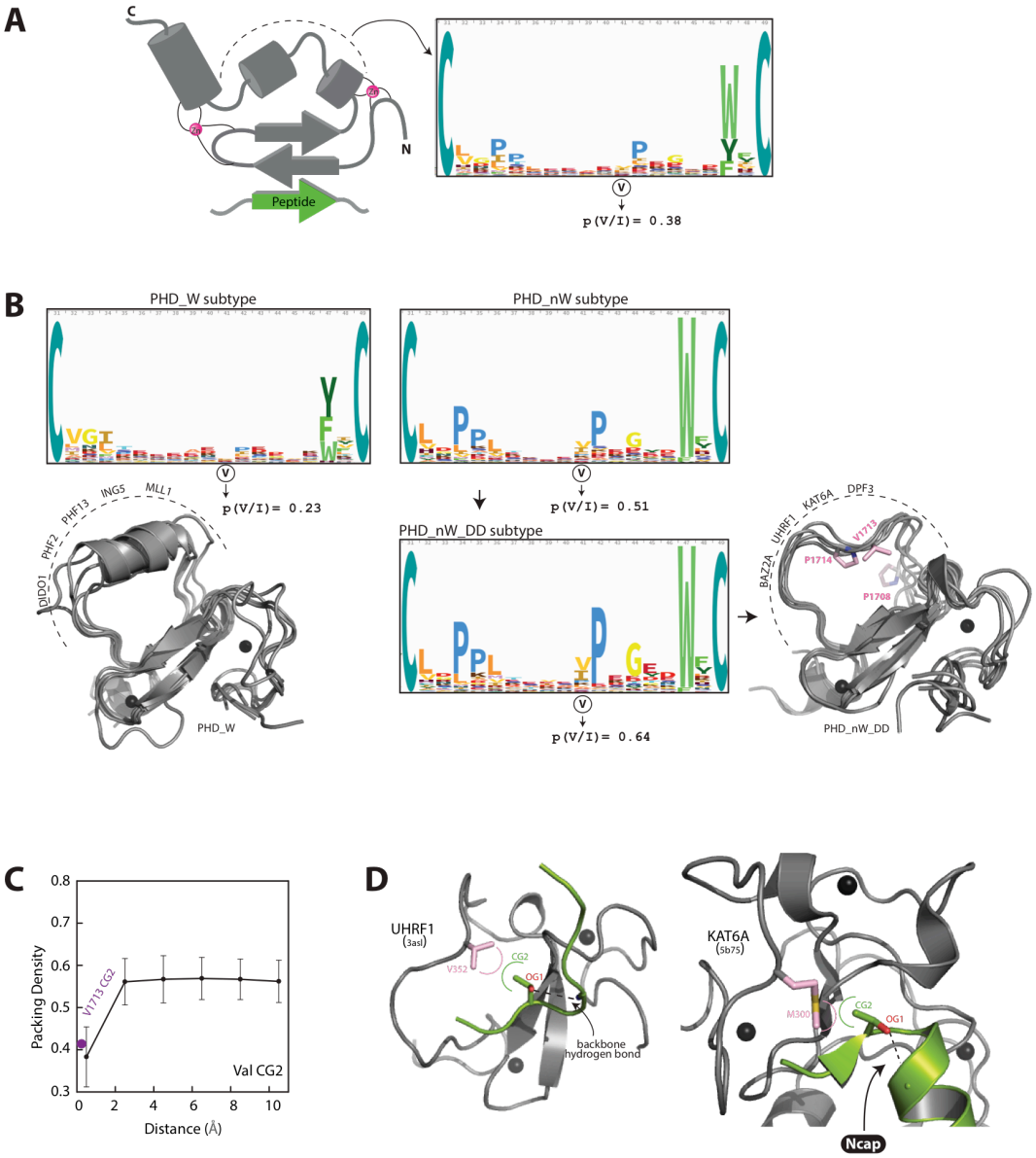


Figure 3.19. Enrichment at other positions: (A) The sequence logo (right) of the segment of interest (dashed line in the topology cartoon, left). (B) Sequence logo of the same segment respectively of PHD_W (top, left), PHD_nW (top, right) and PHD_nW_DD (bottom, right) subtypes. Structures of the segment of interest in PHD_nW_DD subtype (bottom, right). Occurrence frequency at the positions of interest is below the logo, and the pdb ids of the proteins are in Figure 3.2. Side-chains of position in BAZ2A that are enriched in this segment are highlighted with sticks (light-pink) in the structures (bottom, right). Structures of the segment of interest in PHD_W subtype structures (bottom, left). The segment of interest indicated in

(A) is similarly indicated with a circular dashed line. (C) Packing density of BAZ2A Val-CG2 atom is compared with that of Val-CG2 atom (reference) within similar distances from protein surface. (D) Orientations of H3Thr3-CG2 and H3Thr3-OG1 atoms in UHRF1 (left) and KAT6A (right). Light pink and green curved lines represent non-polar contacts. 'N-cap' hydrogen bonded capping of helix N-terminus.

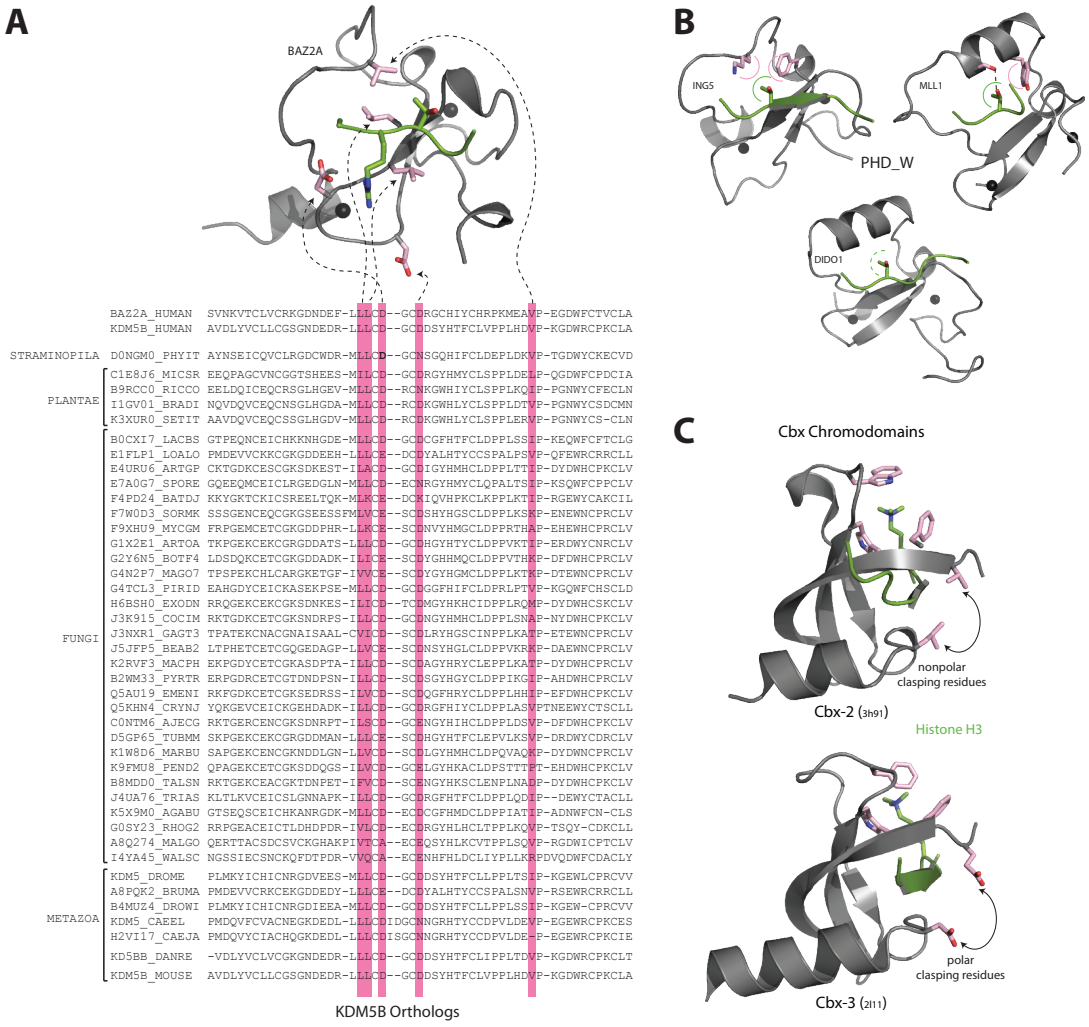


Figure 3.20. KDM5B orthologous sequences and other nonpolar contacts: (A) Tracing BAZ2A residues (light pink) in contact with H3 peptide among the KDM5B orthologs. (B) Residue contacts of histone H3Thr3-CH_3 in the PHD_W subtypes. The PHD_W subtype may or may not retain contacts with H3Thr3-

CH_3. (C) The pair of clasping residues of human CBX chromodomain proteins at the histone peptide-binding site.

Contributions of negatively charged residues at the interface of the Subtype

As the PHD_nW_DD subtype is characterized by the presence of a pair of Asp residues (in the xCDxCDx motif), we wanted to understand the contributions of the BAZ2A D1695 (D1 position) and D1698 (D2 position) residues in the context of subtype structure. The substitution of D1695 severely compromises binding, consistent with the observed effects of D1 substitution in the UHRF1[19], KAT6A[81], and KDM5B[7] proteins. The side chain carboxylate at D1 is engaged in a salt bridge-mediated interaction with the H3Arg2 residue. The complete loss of binding upon D1695 substitution is expected, due to the loss of the electrostatic interactions between D1695 and H3Arg2, and thus D1695 has a very large energetic contribution, similar to that of the D1 positions in UHRF1, KAT6A, and KDM5B. ΔΔG for BAZ2A D1698 (D2 position) is 0.67 kcal/mole (Figure 2.14B). ΔΔG for substitution of Asp residues at the D2 position in UHRF1 and KDM5B are 1.61 and 1.23 kcal/mole, respectively (Figure 3.16A, B). For KAT6A, we expect the energetic contribution of D2 to be as large as we observed for complete disruption of binding upon D2 substitution. In short, the D2 position also has a substantial contribution (ΔΔG >> 1.0 kcal/mole) in anchoring the H3R2 residue for the PHD_nW_DD subtype. This 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 hot spots. For the BAZ2A PHD, however, the contribution of D1698 (D2 position, 0.67 kcal/mole) is less than that of the D2 position in other members as mentioned in chapter 2. A structural deviation of the treble clef knuckle xCDxCDx

pattern in the BAZ2A PHD is observed (Figure 2.16). In comparison with what is typically observed in other PHD fingers, the structural deviation of the knuckle places BAZ2A D1698 farther away from the H3Arg2 residue. This structural deviation is also reflected in the loss of surface area upon complex formation. The ΔASA_{SC} of BAZ2A D1698 is 0.0 Å2 (no H3R2 contact), while that for the D2 positions in the UHRF1, KAT6A, and DPF3 complexes is 17.14 ± 2.55 Å2. The smaller contribution of D1698 is likely due to the smaller contact (enlarged inter-residue distance) in the BAZ2A complex. The D1 ΔASA_{SC} (in BAZ2A, UHRF1, KAT6A, and DPF3) is 34.90 ± 3.93 Å2, while that of D2 is 17.14 ± 2.55 Å2. The amount of contact between the oppositely charged groups is also consistent with the larger energetic contribution of the D1 position to peptide recognition in the PHD_nW_DD subtype.

We then focused on the contributions of the acidic patch[38] residues of BAZ2A. However, substitution of acidic patch residues to Ala (D1688A, E1689A) did not indicate whether they might contribute to the energetics of peptide binding. We actually observed a slightly more negative ΔG of binding for these mutants (Table 2.8A). We therefore focused on the structural contributions of the acidic patch, consistent with the recent structural report on the BAZ2A PHD (pdb code 5t8r)[38]. For the available PHD_nW_DD subtype structures, the conformation of the acidic patch segment is observed to be a 3_{10}-helix (or consecutive turns, Figure 3.21B). Early conformational analysis[111] showed that the occurrence frequency of Asp residues within the first three positions of the 3_{10}-helix or turns are higher than for other amino acids (Figure 3.21B, bottom). The enrichment of negatively charged residues at the acidic patch for the PHD_nW_DD subtype thus suggests that the sequences of this subtype likely retain a 3_{10}-helical conformation. The

sequences of other subtypes prefer Gly for the corresponding segment, and the conformation of this segment in other subtypes is distinctly different from a 3_{10}-helix (Figure 3.21C). These observations therefore suggest a structural, rather than an energetic, role for the acidic patch residues. The adoption of the 3_{10}-helical conformation by the acidic patch has a consequent impact on the cognate peptide conformation (Figure 3.21C). The peptide, in 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 "clash" with the 3_{10}-helical conformation of the acidic patch (Figure 3.21D). Therefore, it is tempting to believe that the H3 peptide adopts a helical conformation in order to evade clashes with the acidic patch. Histone H3 thus adopts a context-dependent conformation consistent with earlier structural observations (Figure 3.9), and a few sequence features of H3 likely contribute to adopting a helical conformation. They are: (a) the N-cap H3T3-OH[107], (b) H3G11, a helix-terminating Gly[112], and (c) the K9 positive charge for accommodating the helix dipole[113] (Figure 3.21D). As we observed the N-cap role of H3T3-OH with other H3 readers (Figure 3.9), the H3T3ph mark in these readers would impede H3 helix initiation. These readers would thus not be suitable candidates for redesigning as H3T3ph readers. Overall, we observed distinct roles of the pairs of acidic residues for the PHD_nW_DD subtype as: (a) treble clef knuckle acidic residues anchoring H3R2 and (b) acidic patch residues influencing H3 conformation.

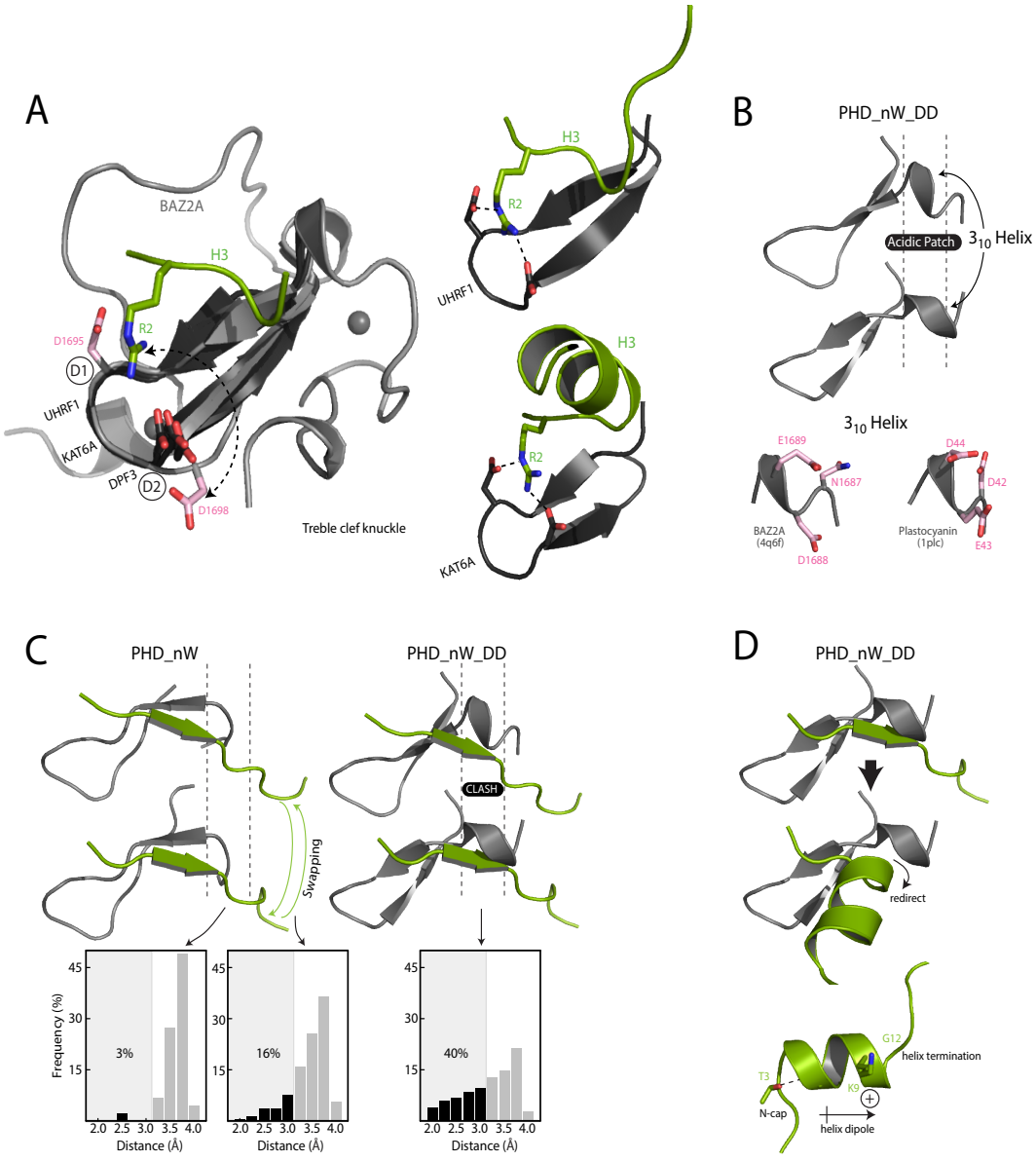


Figure 3.21. BAZ2A PHD treble clef knuckle Asp pair**:** (A) Structural deviation and distances between oppositely charged atoms among the PHD_nW_DD subtype structures (left), and the treble clef knuckle Asp residues pair in UHRF1 (top, right) and KAT6A (bottom, right) anchoring histone H3-Arg2 residue. (B) The 3_{10}-helical conformation of PHD_nW_DD subtype (top), and the similarity between 3_{10}-helical segments of unrelated proteins having negatively charged residues (bottom). (C) Conformation of the corresponding segment in PHD_nW subtype (left) with negligible intermolecular backbone atom clashes (black bar, below) with the strand conformation of the H3 peptide. A much larger proportion of clashes (black bar, below) expected for PHD_nW_DD (right) when interacting with the strand conformation of H3.

(D) Symbolic representation of the context dependent folding of H3 peptide upon interaction with PHD_nW_DD (top), and the helix punctuating features of N-terminal H3 (bottom).

Mutual information between residues at the BAZ2A interface and Subtype

The observed enrichment of nonpolar residues around the peptide-binding site encouraged us to further probe the relationship between the treble clef knuckle Asp pair and the enriched residue positions in the PHD_nW_DD subtype. We examined the following two relationships first: (a) the enrichment in artificial sequences and (b) reciprocal enrichment. For the enrichment analysis, artificial PHD finger sequences were generated with E-values consistent with that of natural PHD sequences (Figure 3.11A) and with characteristics similar to that of the PFAM master alignment (Figure 3.10C). Enrichment analyses carried out with the artificial sequences clearly show that results differ from those observed with natural sequences (Figure 3.10B, right). For example, positive enrichment of the L2 position ($p = 0.31 \rightarrow 0.63$) when the D2 position is perturbed in natural sequences is very different from the distribution of L2 position enrichment probabilities observed with artificial sequences for D2 perturbation (Figure 3.10B, right). This observation suggests the nonrandom nature of the enrichment observed here. For reciprocal enrichment, we wanted to check whether the D2 position would similarly show positive enrichment ($\Delta p > 0$) if other nonpolar residue positions (e.g., the L2 position) were perturbed (Figure 3.9A). Asp was enriched at the D2 position when other peptide-anchoring positions were perturbed in natural sequences (Figure 3.9A). Like the D2 positions, other peptide-anchoring positions were also checked to determine whether there is a mutual relationship among a set of positions in the PHD superfamily fold (Figures 3.9A, 3.7D).

Although enrichment analysis is simple with respect to computation, it may lack mathematical rigor and be unable to distinguish stochastic relationships from real ones. Therefore, we applied mutual information (MI) theory to probe the relationship between the peptide-binding positions of PHD fingers. This was motivated by the earlier observation that a set of PDZ peptide binding-site positions (5–6 positions) was observed to have higher MI, and these residues were concluded to be coevolving[45]. Using the MISTIC[89] server, we estimated the MI for every residue pair in the PHD PFAM alignment positions (Figure 3.9B). Many of the residue pairs showing mutual enrichment (Figure 3.9A) have MI scores that are much higher than the rest of the PHD residue pairs (Figure 3.9B). The MI behavior of PDZ-binding residues was also estimated (Figure 3.22A) for comparison. Despite differences in the construction of the dataset and approach from the earlier study[45], it was interesting to observe that residue pairs consisting of only 5–6 positions (Figure 3.22A) show MI values well above the rest, as reported earlier. Our approach is thus acceptable, and the MI behavior of enriched PHD binding-site residues is comparable to that of the PDZ domain (Figure 3.22A). It is therefore tempting to believe that PHD binding-site residues (at the L2, D2, G/W, AP2 positions) are also coevolving like the PDZ binding site[45]. With regard to coevolving binding-site residues, the similarity between the PHD and PDZ domain can further be appreciated by considering the following. Like ADD redesign (for switching peptide specificity)[28, 29], desirable substitutions at the coevolving PDZ binding-site positions can also switch peptide specificity: e.g., in redesigning Tiam1 to behave like Tiam2[114]. Thus, identification of coevolving positions can also be useful for subtype switching.

Cumulative MI (CMI), reported by the MISTIC[89] server, represents the degree of participation of a position in the mutual information network[91]. For the PHD finger, the W/G position has the highest CMI value (Figure 3.9B), as the position shows a strong negative enrichment compared with many other nonpolar residue peptide-anchoring positions (Figure 3.9A), and the W/G position also shows a positive enrichment for aromatic cage residues[7]. Therefore, using the W/G position to partition the PHD finger sequences as a first step in PHD subtype analysis (see Figure 3.15) is justified. The L2 position also has a large CMI value, suggesting its contribution in PHD subtype classification. However, artificial PHD sequences (see above) do not show a positional correlation (Figure 3.11B, C).

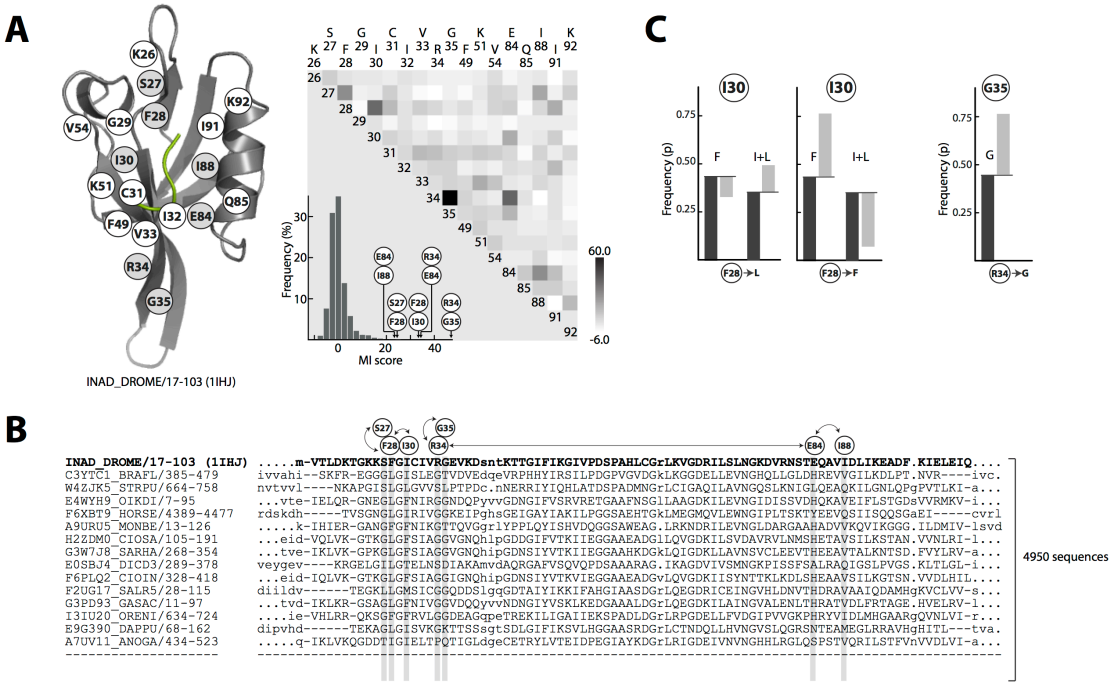


Figure 3.22. Mutual Information of the PDZ binding site: (A) MI score of PDZ binding-site positions (right) with the residues mapped on the INAD_DROME PDZ structure (left). Only 18 binding-site

positions (left) mapped on the structure are analyzed for MI (right). Seven positions with higher inter-residue MI value are indicated in gray on the structure (left, e.g. S27, F28, I30 etc.) and their MI values are indicated with '↓' arrow on the MI histogram. Protein name and pdb id are indicated. (B) A part of the 85% non-redundant PFAM alignment of PDZ domains used for MI calculation shows the positions of the seven binding site positions for convenience. The double-headed arrows indicate interacting positions. (C) Reciprocal enrichment for some of the selected binding-site positions.

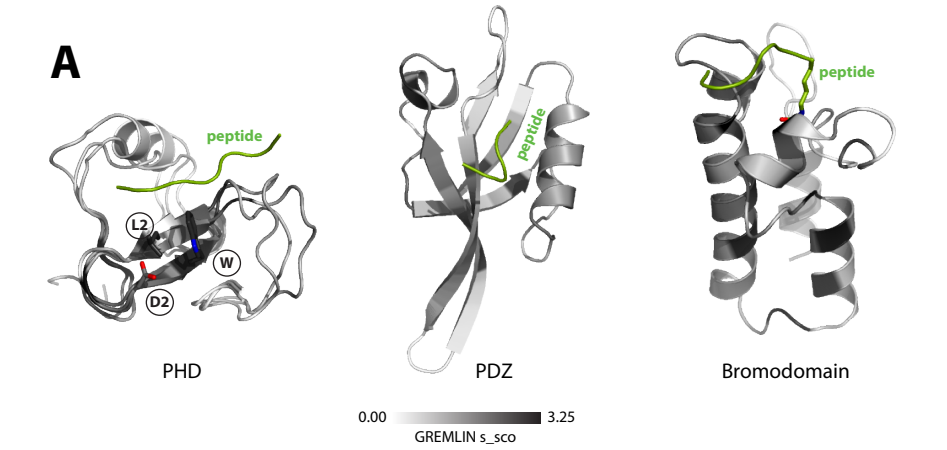
GREMLIN distance restraints and RMSD

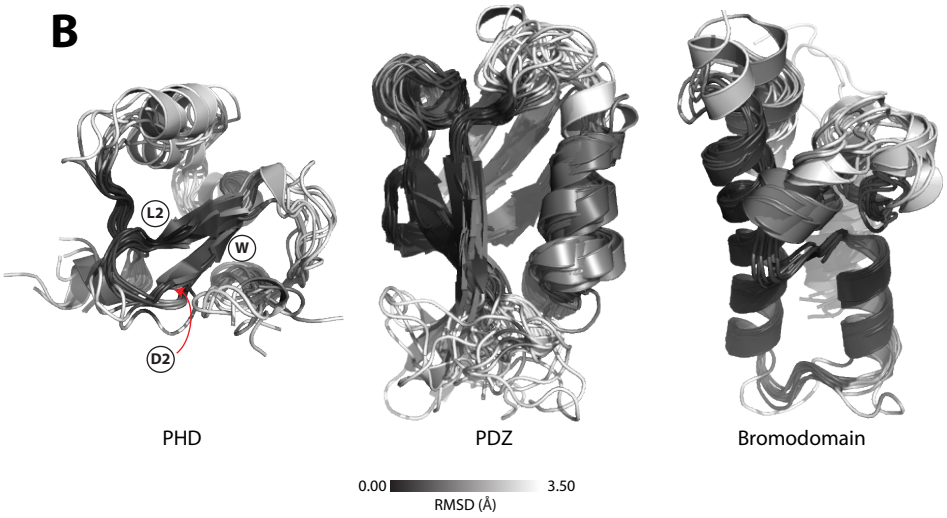
Mutual information extracted from multiple alignments to infer coevolution by correlated mutation has recently been successful in protein structure prediction[87, 88, 115, 116]. In other words, the correlation between alignment positions can provide distance restraints for folding the polypeptide chain representing a protein family[87, 88, 115, 116]. The PHD finger binding-site positions that we found to be correlated are likely to be correlated for biochemical function (e.g., peptide binding) rather than for the folding of the PHD scaffold. With GREMLIN[87, 88] pre-computed distance restraint scores for PFAM domain residue pairs, we wanted to check the scores of PHD finger positions that are related by function (e.g., L2–D2 or L2–W pair positions, Figure 3.9B). GREMLIN distance restraint scores for these functionally related positions are actually observed to be much higher than the rest of the pairs (Figure 3.9B), even though these residues are not in van der Waals contact with each other. Physically interacting positions that contribute to folding are expected to coevolve as a mechanism for compensating the replacement of interacting partner residues. In the PHD example, the restraint score does not seem to distinguish between pairs physically interacting with one another from those

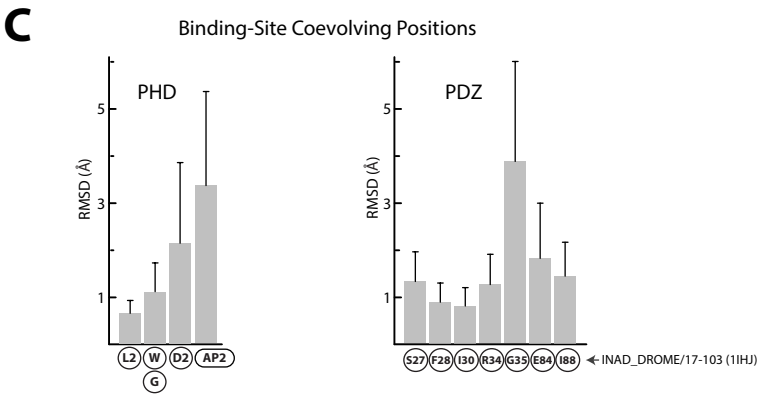
related by function. The pre-computed GREMLIN score therefore, encouraged us to experimentally probe the roles of the correlated binding-site positions by mutagenesis. A comparison of the RMSD of crystal structures (≤ 2.2 Å) of PFAM domain members sharing less than 40% sequence identity shows that the structural deviation of the PHD peptide binding-site is smaller (Figure 3.23B). The implication is that this smaller structural deviation may be useful in design applications. The introduction of desired substitutions at coevolving binding-site positions would likely confer small structural perturbations and thus enable achieving the expected outcome without having to carry out extensive optimization of the rest of the scaffold, e.g., repurposing of the ADD module[28, 29].

Here we were also able to enhance the affinity of the AIRE PHD (~15-fold lower Kd) by merely incorporating a desired substitution at a coevolving position showing a very small structural deviation (e.g., L2 in Figure 3.23C). In addition to altering affinity and specificity, sequence-based subtype classification can also benefit from the observed small structural deviation. As the rest of the PHD sequence positions show larger structural deviations (Figure 3.23B), inclusion of non-binding site sequence positions in subtype classification may not provide reliable groupings. For example, an earlier study[117] on PHD finger classification, using the full-length PHD sequence had grouped BAZ2A (and BAZ1A, BAZ1B, BAZ2B) with AIRE (and TIF1A, CHD3_1, CHD3_2, PF21A), while MYST DPFs were placed in a separate group. As observed in an earlier work[7], the binding behavior of BAZ2A, a PHD_nW_DD subtype, is distinct from that of AIRE, and we found that DPF binding behavior is similar to PHD_nW_DD. Thus, our

experiments suggest that grouping based on a set of coevolving binding-site residues, especially showing small structural deviations, would likely be more reliable.







D

```
YCPVC-GQSDDDGD-----MIQCD-SCNRWYHQECVGPPEK----AEEIPSGD-WYCPNCRPK GREMLIN CONSENSUS
ACHLC-GGRQDPDK-----QLMCD-ECDMAFHIYCLDPP------LSSVPSEDEWYCPECRND 3ASLA 317-366
FCPLC-DKCYDDDDYES-KMMQCG-KCDRWVHSKCENLSDEMYEILSNLPESVAYTCVNCTER 3LQHA 1568-1627
TC-FC-MKPFAGRP-----MIECN-ECHTWIHLSCAKIR------KSNVPEV--FVCQKCRDS 3O7AA 234-280
ICSFCLGTKEQNREKKPEELISCA-DCGNSGHPSCLKFSPE----LTVRVKALRWQCIECKTC 3V43A 208-265
TCSSCRDQGKNADN-----MLFCD-SCDRGFHMECCDPP------LTRMPKGM-WICQICRPR 3V43A 264-313
RC-IC-GFTHDDGY-----MICCD-KCSVWQHIDCMGID------RQHIPDT--YLCERCQPR 4L58A 4-50
YC-IC-RQPHNNRF-----MICCD-RCEEWFHGDCVGISEARGRLLERNGED--YICPNCT-- 4L7XA 8-58
PCGACRSEVNDDQD-----AILCEASCQKWFHRECTGMTESAYGLLTTEASAV-WACDLCVYV 4UP0A 329-1237
```

Figure 3.23. GREMLIN score and RMSD of peptide binding sites: (A) GREMLIN s-sco score mapped on to the structure of PHD (left), PDZ (middle) and Bromodomain (right). (B) Structural deviation (RMSD) between corresponding positions among X-ray structures (resolution ≤ 2.2Å) of a superfamily sharing less than 40% sequence identity. The color ranges between black (RMSD = 0.0 Å) to white (RMSD = 3.5Å). PHD positions are indicated on the structure. (C) Mean and SD of the corresponding position's Cα-Cα atom distances for the binding- site coevolving positions. (D) Alignment between GREMLIN consensus sequence and the protein sequence is presented for the convenience of mapping the s_sco score on the structure.

Characteristics of Arg-rich peptide recognition

Histone peptides are rich in positively charged residues, and therefore the observation of enriched nonpolar residues making significant energetic contributions to histone peptide recognition encouraged us to further determine the recognition characteristics of peptides rich in Arg and Lys residues. Using Rosetta alanine-scanning mutagenesis[96] on the dataset of peptide−protein complexes used in an earlier work[7], we computationally probed peptide-recognition characteristics. Peptide hotspot propensity scores computed using this dataset also show strong preferences for W, F, I, L, Y, and R (propensity score > 1.0), and some of the remaining (disfavored) amino acids have scores <0.75 (Figure 3.12B). Therefore, for the analysis we focused on peptide Arg hotspot residues (Figure 3.12B). ARG-Nearness, the preference of an amino acid to be adjacent (± 2 positions) to the peptide Arg hotspot residue was also computed (Figure 3.12B). The ARG-Nearness propensity score is quite different from that of the reference peptide hotspot scores (Figure 3.12B). For example, irrespective of side chain size, the ARG-Nearness propensity scores of several amino acids are close to 1 (Figure 3.12B); i.e., amino acids with even small side chains (e.g., G, A, or T) are thus not disfavored to be

near an Arg hotspot residue in peptide−protein complexes. This is important, as the presence of amino acids with small side chains adjacent to Arg residues is common among histone-tail sequences. For comparison, Arg in the kinase-docking D motifs is close to one or more bulky nonpolar residues[118, 119]. Therefore, to better understand nature's design principles, we wanted to know whether there is a preference for amino acids (at the peptide binding site of a protein, e.g., L1 or L2) for making contact with residues with small side chains adjacent to peptide Arg hotspot residues (Figure 3.12C).

Hotspot propensity scores of protein residues (Figure 3.12C) also showed that bulky nonpolar amino acids dominate the residue preferences. This is also consistent with our observations for several histone/histone-like peptide complexes (Figure 3.12D) that bulky nonpolar residues at the binding site establish contacts with the small-volume residues adjacent to the peptide Arg hotspot residue. Therefore, the mechanism of having bulky residues at the L1 and L2 positions for contacting Ala and Thr of the H3 peptide adjacent to the hotspot R2 residue is likely to be a well-tested natural mechanism, and therefore the same strategy could be implemented in designing diagnostic proteins to target site-specific Arg residues of histone tails. Grafting a complementary peptide segment onto an antibody scaffold has shown promise for targeting specific peptide segments within intrinsically disordered segments[120]. Our observations on nonpolar residues could thus be useful for such a strategy for targeting intrinsically disordered unmodified histone peptide segments.

CHAPTER IV

GENERAL DISCUSSION AND CONCLUSION

Next-generation sequencing efforts on diagnosed patient genomes or cancer tissues now show that aberrations (mutations or overexpression) in proteins harboring the PHD_nW_DD subtype are associated with different pathological consequences (Figure 4.1). Therefore, as a first step toward eventual comprehension of the possible mechanisms underlying pathological consequences, detailed characterization of the binding mechanism of the PHD_nW_DD subtype was undertaken here. Starting with one member (BAZ2A PHD) of the subtype, key histone peptide-anchoring residue positions were first identified by mutagenesis. Loss of interfacial packing due to residue substitution contributes to the observed binding energetics. Interestingly, the peptide-anchoring residue positions ($\Delta\Delta G \geq 1.0$ kcal/mol) of the BAZ2A PHD are enriched in specific types of residues in a subtype-specific manner. The energetic contributions of the identified positions were further confirmed by mutagenesis in three other members of the subtype (UHRF1, KDM5B, KAT6A), which included pairs sharing <40% sequence identity with each other. Despite low sequence similarity, mutations cause similar consequences in histone H3 binding, suggesting a strong similarity in the binding mechanism and thus supporting the classification of the subtype.

The positions of peptide-anchoring residues of the PHD_nW_DD subtype are also mutually correlated, while corresponding positions in artificial PHD sequences (E-value $\leq 10^{-5}$) are not. Based on our experimental observations, it is thus temping to believe that the correlation is for biochemical function (peptide binding) rather than for folding. The three-dimensional distances between two correlated positions in a protein sequence

predicted by powerful computational approaches[87, 88, 115, 116] can thus be complemented by a systematic mutagenesis approach to dissect the true origin of the correlation, particularly for positions at the functional site. Like the coevolving binding-site residues of the versatile PDZ domain, the correlated binding-site residues of the versatile PHD finger are also coevolving. The positions of some of the coevolving residues show very small structural deviations among diverse sequence homologs, and this observation will be useful for redesigning the PHD binding site for enhancing affinity and switching specificity. The binding mechanism of the PHD_nW_DD subtype likely originated early to anchor residues 1–3 of histone H3. This mechanism features the characteristic treble clef knuckle Asp residues for anchoring the peptide Arg hotspot, while the acidic patch residues contribute to the helical conformation of N-terminal histone H3. A set of nonpolar amino acids is among the residues enriched in the PHD_nW_DD subtype, and these nonpolar amino acids tightly pack against the peptide residues with small side chains present on either side of the peptide hotspot Arg residue. This strategy of employing tightly packed, bulky nonpolar residues to anchor residues with small side chains adjacent to the peptide hotspot can be useful for histone diagnostic design. In general, the lessons learned from the current study will be useful in the functional characterization of large protein families.

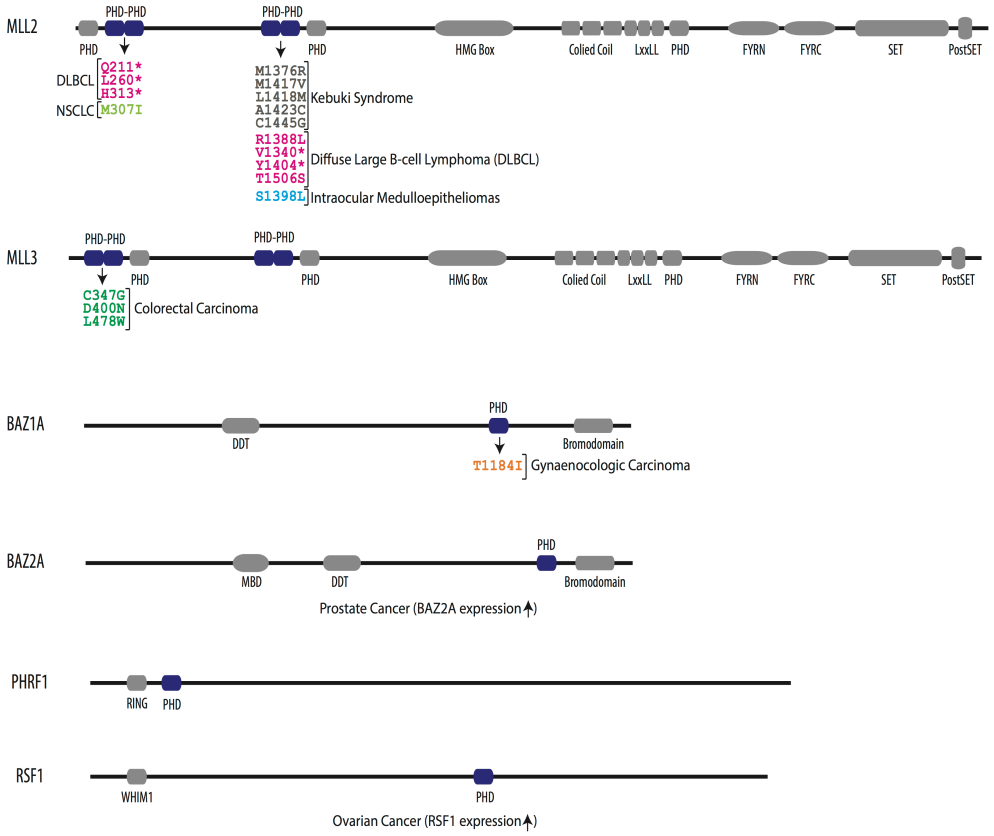


Figure 4.1. Human disorders associated with PHD_nW_DD: Kebuki syndrome[105-108], diffuse large B-cell lymphoma (DLBCL)[108], colorectal[109], gynecological carcinoma[110], intraocular medulloepitheliomas[19], non–small cell lung cancer (NSCLC)[111] mutations respectively are colored in gray, pink, green, yellow, cyan, split pea, and the upward arrow '↑' indicates increase in protein expression in prostate[112] and ovarian[113] cancer. Protein names are on the left and star '*' denotes nonsense mutations. PHRF1 is included here for recent reports on the association of the tumor suppressor role of PHRF1[89, 114]. All these mutations are taken from the references cited in above, and for convenience only selected proteins and their PHD_nW_DD mutations (among other mutations) are listed here.

REFERENCES

[1] Onuchic, J. N., and Wolynes, P. G. (2004) Theory of protein folding, *Curr Opin Struct Biol 14*, 70-75.

[2] Baker, D., and Eaton, W. A. (2004) Folding and binding, *Curr Opin Struc Biol 14*, 67-69.

[3] Phillips, D. C. (1966) The three-dimensional structure of an enzyme molecule, *Sci Am 215*, 78-90.

[4] Chakravarty, S., Zeng, L., and Zhou, M. M. (2009) Structure and site-specific recognition of histone H3 by the PHD finger of human autoimmune regulator, *Structure 17*, 670-679.

[5] Capili, A. D., Schultz, D. C., Rauscher, I. F., and Borden, K. L. (2001) Solution structure of the PHD domain from the KAP-1 corepressor: structural determinants for PHD, RING and LIM zinc-binding domains, *EMBO J 20*, 165-177.

[6] Aasland, R., Gibson, T. J., and Stewart, A. F. (1995) The Phd Finger - Implications for Chromatin-Mediated Transcriptional Regulation, *Trends in Biochemical Sciences 20*, 56-59.

[7] Chakravarty, S., Essel, F., Lin, T., and Zeigler, S. (2015) Histone Peptide Recognition by KDM5B-PHD1: A Case Study, *Biochemistry 54*, 5766-5780.

[8] Sanchez, R., and Zhou, M. M. (2011) The PHD finger: a versatile epigenome reader, *Trends Biochem Sci 36*, 364-372.

[9] Essel, F. (2015) Discovery of Novel Readers for Interpretation of the Epigenome, Chemistry and Biochemistry Department, South Dakota State University.

[10] Schindler, U., Beckmann, H., and Cashmore, A. R. (1993) HAT3.1, a novel Arabidopsis homeodomain protein containing a conserved cysteine-rich region, *Plant J 4*, 137-150.

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

[12] Kwan, A. H., Gell, D. A., Verger, A., Crossley, M., Matthews, J. M., and Mackay, J. P. (2003) Engineering a protein scaffold from a PHD finger, *Structure 11*, 803-813.

[13] Pascual, J., Martinez-Yamout, M., Dyson, H. J., and Wright, P. E. (2000) Structure of the PHD zinc finger from human Williams-Beuren syndrome transcription factor, *J Mol Biol 304*, 723-729.

[14] Wysocka, J., Swigut, T., Xiao, H., Milne, T. A., Kwon, S. Y., Landry, J., Kauer, M., Tackett, A. J., Chait, B. T., Badenhorst, P., Wu, C., and Allis, C. D. (2006) A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling, *Nature 442*, 86-90.

[15] Matthews, A. G., Kuo, A. J., Ramon-Maiques, S., Han, S., Champagne, K. S., Ivanov, D., Gallardo, M., Carney, D., Cheung, P., Ciccone, D. N., Walter, K. L., Utz, P. J., Shi, Y., Kutateladze, T. G., Yang, W., Gozani, O., and Oettinger, M. A. (2007) RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination, *Nature 450*, 1106-1110.

[16] Taverna, S. D., Ilin, S., Rogers, R. S., Tanny, J. C., Lavender, H., Li, H., Baker, L., Boyle, J., Blair, L. P., Chait, B. T., Patel, D. J., Aitchison, J. D., Tackett, A. J., and Allis, C. D. (2006) Yng1 PHD finger binding to H3 trimethylated at K4 promotes NuA3 HAT activity at K14 of H3 and transcription at a subset of targeted ORFs, *Mol Cell 24*, 785-796.

[17] Shi, X., Hong, T., Walter, K. L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Pena, P., Lan, F., Kaadige, M. R., Lacoste, N., Cayrou, C., Davrazou, F., Saha, A., Cairns, B. R., Ayer, D. E., Kutateladze, T. G., Shi, Y., Cote, J., Chua, K. F., and Gozani, O. (2006) ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression, *Nature 442*, 96-99.

[18] Zhang, Y., Yang, H., Guo, X., Rong, N., Song, Y., Xu, Y., Lan, W., Zhang, X., Liu, M., Xu, Y., and Cao, C. (2014) The PHD1 finger of KDM5B recognizes unmodified H3K4 during the demethylation of histone H3K4me2/3 by KDM5B, *Protein & Cell 5*, 837-850.

[19] Rajakumara, E., Wang, Z., Ma, H., Hu, L., Chen, H., Lin, Y., Guo, R., Wu, F., Li, H., Lan, F., Shi, Y. G., Xu, Y., Patel, D. J., and Shi, Y. (2011) PHD finger recognition of unmodified histone H3R2 links UHRF1 to regulation of euchromatic gene expression, *Mol Cell 43*, 275-284.

[20] Grishin, N. V. (2001) Treble clef finger--a functionally diverse zinc-binding structural motif, *Nucleic Acids Res 29*, 1703-1714.

[21] Allis, C. D., and Jenuwein, T. (2016) The molecular hallmarks of epigenetic control, *Nat Rev Genet 17*, 487-500.

[22] Smith, E., and Shilatifard, A. (2010) The Chromatin Signaling Pathway: Diverse Mechanisms of Recruitment of Histone-Modifying Enzymes and Varied Biological Outcomes, *Molecular Cell 40*, 689-701.

[23] Ruthenburg, A. J., Allis, C. D., and Wysocka, J. (2007) Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark, *Mol Cell 25*, 15-30.

[24] Xiong, X., Panchenko, T., Yang, S., Zhao, S., Yan, P., Zhang, W., Xie, W., Li, Y., Zhao, Y., Allis, C. D., and Li, H. (2016) Selective recognition of histone crotonylation by double PHD fingers of MOZ and DPF2, *Nat Chem Biol 12*, 1111-1118.

[25] Andrews, F. H., Strahl, B. D., and Kutateladze, T. G. (2016) Insights into newly discovered marks and readers of epigenetic information, *Nat Chem Biol 12*, 662-668.

[26] Patel, D. J., and Wang, Z. (2013) Readout of epigenetic modifications, *Annual review of biochemistry 82*, 81-118.

[27] Musselman, C. A., Lalonde, M. E., Cote, J., and Kutateladze, T. G. (2012) Perceiving the epigenetic landscape through histone readers, *Nat Struct Mol Biol 19*, 1218-1227.

[28] Noh, K.-M., Allis, C. D., and Li, H. (2016) Reading between the Lines: "ADD"-ing Histone and DNA Methylation Marks toward a New Epigenetic "Sum", *Acs Chem Biol 11*, 554-563.

[29] Noh, K. M., Wang, H., Kim, H. R., Wenderski, W., Fang, F., Li, C. H., Dewell, S., Hughes, S. H., Melnick, A. M., Patel, D. J., Li, H., and Allis, C. D. (2015) Engineering of a Histone-Recognition Domain in Dnmt3a Alters the Epigenetic Landscape and Phenotypic Features of Mouse ESCs, *Mol Cell 59*, 89-103.

[30] Kungulovski, G., Kycia, I., Tamas, R., Jurkowska, R. Z., Kudithipudi, S., Henry, C., Reinhardt, R., Labhart, P., and Jeltsch, A. (2014) Application of histone modification-specific interaction domains as an alternative to antibodies, *Genome Research 24*, 1842-1853.

[31] Moore, K. E., Carlson, S. M., Camp, N. D., Cheung, P., James, R. G., Chua, K. F., Wolf-Yadlin, A., and Gozani, O. (2013) A general molecular affinity strategy for global detection and proteomic analysis of lysine methylation, *Mol Cell 50*, 444-456.

[32] Grummt, I. (2007) Different epigenetic layers engage in complex crosstalk to define the epigenetic state of mammalian rRNA genes, *Hum Mol Genet 16 Spec No 1*, R21-27.

[33] Zhou, Y., and Grummt, I. (2005) The PHD finger/bromodomain of NoRC interacts with acetylated histone H4K16 and is sufficient for rDNA silencing, *Current biology : CB 15*, 1434-1438.

[34] Santoro, R., and Grummt, I. (2005) Epigenetic mechanism of rRNA gene silencing: temporal order of NoRC-mediated histone modification, chromatin remodeling, and DNA methylation, *Mol Cell Biol 25*, 2539-2546.

[35] Guetg, C., Lienemann, P., Sirri, V., Grummt, I., Hernandez-Verdun, D., Hottiger, M. O., Fussenegger, M., and Santoro, R. (2010) The NoRC complex mediates the heterochromatin formation and stability of silent rRNA genes and centromeric repeats, *Embo j 29*, 2135-2146.

[36] Zhou, Y., Schmitz, K. M., Mayer, C., Yuan, X., Akhtar, A., and Grummt, I. (2009) Reversible acetylation of the chromatin remodelling complex NoRC is required for non-coding RNA-dependent silencing, *Nat Cell Biol 11*, 1010-1016.

[37] Gu, L., Frommel, S. C., Oakes, C. C., Simon, R., Grupp, K., Gerig, C. Y., Bar, D., Robinson, M. D., Baer, C., Weiss, M., Gu, Z. G., Schapira, M., Kuner, R., Sultmann, H., Provenzano, M., Yaspo, M. L., Brors, B., Korbel, J., Schlomm, T., Sauter, G., Eils, R., Plass, C., Santoro, R., and Prostate, I. P. E. O. (2015) BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer and its overexpression predicts disease recurrence, *Nat Genet 47*, 22-+.

[38] Bortoluzzi, A., Amato, A., Lucas, X., Blank, M., and Ciulli, A. (2017) Structural Basis of Molecular Recognition of Helical Histone H3 Tail by PHD Finger Domains, *Biochem J*.

[39] Ivarsson, Y. (2012) Plasticity of PDZ domains in ligand recognition and signaling, *FEBS Lett 586*, 2638-2647.

[40] Tonikian, R., Zhang, Y., Sazinsky, S. L., Currell, B., Yeh, J. H., Reva, B., Held, H. A., Appleton, B. A., Evangelista, M., Wu, Y., Xin, X., Chan, A. C., Seshagiri, S., Lasky, L. A., Sander, C., Boone, C., Bader, G. D., and Sidhu, S. S. (2008) A specificity map for the PDZ domain family, *PLoS Biol 6*, e239.

[41] Remaut, H., and Waksman, G. (2006) Protein-protein interaction through beta-strand addition, *Trends Biochem Sci 31*, 436-444.

[42] Ernst, A., Sazinsky, S. L., Hui, S., Currell, B., Dharsee, M., Seshagiri, S., Bader, G. D., and Sidhu, S. S. (2009) Rapid evolution of functional complexity in a domain family, *Sci Signal 2*, ra50.

[43] Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A., and MacBeath, G. (2008) Predicting PDZ domain-peptide interactions from primary sequences, *Nat Biotechnol 26*, 1041-1045.

[44] Stiffler, M. A., Grantcharova, V. P., Sevecka, M., and MacBeath, G. (2006) Uncovering Quantitative Protein Interaction Networks for Mouse PDZ

Domains Using Protein Microarrays, *Journal of the American Chemical Society 128*, 5913-5922.

[45] Sakarya, O., Conaco, C., Egecioglu, O., Solla, S. A., Oakley, T. H., and Kosik, K. S. (2010) Evolutionary expansion and specialization of the PDZ domains, *Mol Biol Evol 27*, 1058-1069.

[46] London, N., Movshovitz-Attias, D., and Schueler-Furman, O. (2010) The structural basis of peptide-protein binding strategies, *Structure 18*, 188-199.

[47] Tallant, C., Valentini, E., Fedorov, O., Overvoorde, L., Ferguson, F. M., Filippakopoulos, P., Svergun, D. I., Knapp, S., and Ciulli, A. (2015) Molecular basis of histone tail recognition by human TIP5 PHD finger and bromodomain of the chromatin remodeling complex NoRC, *Structure 23*, 80-92.

[48] Vakoc, C. R., Wen, Y. Y., Gibbs, R. A., Johnstone, C. N., Rustgi, A. K., and Blobel, G. A. (2009) Low frequency of MLL3 mutations in colorectal carcinoma, *Cancer genetics and cytogenetics 189*, 140-141.

[49] Pasqualucci, L., Trifonov, V., Fabbri, G., Ma, J., Rossi, D., Chiarenza, A., Wells, V. A., Grunn, A., Messina, M., Elliot, O., Chan, J., Bhagat, G., Chadburn, A., Gaidano, G., Mullighan, C. G., Rabadan, R., and Dalla-Favera, R. (2011) Analysis of the coding genome of diffuse large B-cell lymphoma, *Nat Genet 43*, 830-837.

[50] Ng, S. B., Bigham, A. W., Buckingham, K. J., Hannibal, M. C., McMillin, M. J., Gildersleeve, H. I., Beck, A. E., Tabor, H. K., Cooper, G. M., Mefford, H. C., Lee, C., Turner, E. H., Smith, J. D., Rieder, M. J., Yoshiura, K., Matsumoto, N., Ohta, T., Niikawa, N., Nickerson, D. A., Bamshad, M. J., and Shendure, J. (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome, *Nat Genet 42*, 790-793.

[51] Miyake, N., Koshimizu, E., Okamoto, N., Mizuno, S., Ogata, T., Nagai, T., Kosho, T., Ohashi, H., Kato, M., Sasaki, G., Mabe, H., Watanabe, Y., Yoshino, M., Matsuishi, T., Takanashi, J., Shotelersuk, V., Tekin, M., Ochi, N., Kubota, M., Ito, N., Ihara, K., Hara, T., Tonoki, H., Ohta, T., Saito, K., Matsuo, M., Urano, M., Enokizono, T., Sato, A., Tanaka, H., Ogawa, A., Fujita, T., Hiraki, Y., Kitanaka, S., Matsubara, Y., Makita, T., Taguri, M., Nakashima, M., Tsurusaki, Y., Saitsu, H., Yoshiura, K., Matsumoto, N., and Niikawa, N. (2013) MLL2 and KDM6A mutations in patients with Kabuki syndrome, *American journal of medical genetics. Part A 161a*, 2234-2243.

[52] Micale, L., Augello, B., Fusco, C., Selicorni, A., Loviglio, M. N., Silengo, M. C., Reymond, A., Gumiero, B., Zucchetti, F., D'Addetta, E. V., Belligni, E., Calcagnì, A., Digilio, M. C., Dallapiccola, B., Faravelli, F., Forzano, F., Accadia, M., Bonfante, A., Clementi, M., Daolio, C., Douzgou, S., Ferrari, P., Fischetto, R.,

Garavelli, L., Lapi, E., Mattina, T., Melis, D., Patricelli, M. G., Priolo, M., Prontera, P., Renieri, A., Mencarelli, M. A., Scarano, G., Monica, M. d., Toschi, B., Turolla, L., Vancini, A., Zatterale, A., Gabrielli, O., Zelante, L., and Merla, G. (2011) Mutation spectrum of MLL2 in a cohort of kabuki syndrome patients, *Orphanet Journal of Rare Diseases 6*, 38-38.

[53] Jones, S., Stransky, N., McCord, C. L., Cerami, E., Lagowski, J., Kelly, D., Angiuoli, S. V., Sausen, M., Kann, L., Shukla, M., Makar, R., Wood, L. D., Diaz, L. A., Lengauer, C., and Velculescu, V. E. (2014) Genomic analyses of gynaecologic carcinosarcomas reveal frequent mutations in chromatin remodelling genes, *Nature communications 5*, 5006-5006.

[54] Semple, T. U., Quinn, L. A., Woods, L. K., and Moore, G. E. (1978) Tumor and lymphoid cell lines from a patient with carcinoma of the colon for a cytotoxicity model, *Cancer Res 38*.

[55] Neve, R. M., Chin, K., Fridlyand, J., Yeh, J., Baehner, F. L., Fevr, T., Clark, L., Bayani, N., Coppe, J. P., Tong, F., Speed, T., Spellman, P. T., DeVries, S., Lapuk, A., Wang, N. J., Kuo, W. L., Stilwell, J. L., Pinkel, D., Albertson, D. G., Waldman, F. M., McCormick, F., Dickson, R. B., Johnson, M. D., Lippman, M., Ethier, S., Gazdar, A., and Gray, J. W. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes, *Cancer Cell 10*.

[56] Natarajan, T. G., Kallakury, B. V., Sheehan, C. E., Bartlett, M. B., Ganesan, N., Preet, A., Ross, J. S., and FitzGerald, K. T. (2010) Epigenetic regulator MLL2 shows altered expression in cancer cell lines and tumors from human breast and colon, *Cancer Cell International 10*, 13.

[57] Gong, W., Suzuki, K., Russell, M., and Riabowol, K. (2005) Function of the ING family of PHD proteins in cancer, *Int J Biochem Cell Biol 37*.

[58] Glaser, S., Lubitz, S., Loveland, K. L., Ohbo, K., Robb, L., Schwenk, F., Seibler, J., Roellig, D., Kranz, A., Anastassiadis, K., and Stewart, A. F. (2009) The histone 3 lysine 4 methyltransferase, Mll2, is only required briefly in development and spermatogenesis, *Epigenetics Chromatin 2*.

[59] Elston, C. W., and Ellis, I. O. (1991) Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up, *Histopathology 19*.

[60] Yin, S., Yang, J., Lin, B., Deng, W., Zhang, Y., Yi, X., Shi, Y., Tao, Y., Cai, J., Wu, C.-I., Zhao, G., Hurst, L. D., Zhang, J., Hu, L., and Kong, X. (2014) Exome sequencing identifies frequent mutation of MLL2 in non–small cell lung carcinoma from Chinese patients, *Scientific Reports 4*, 6036.

[61] Jones, S., Stransky, N., McCord, C. L., Cerami, E., Lagowski, J., Kelly, D., Angiuoli, S. V., Sausen, M., Kann, L., Shukla, M., Makar, R., Wood, L. D., Diaz, L. A., Jr., Lengauer, C., and Velculescu, V. E. (2014) Genomic analyses of gynaecologic carcinosarcomas reveal frequent mutations in chromatin remodelling genes, *Nat Commun 5*, 5006.

[62] Egan, S. M., and Schleif, R. F. (1993) A regulatory cascade in the induction of rhaBAD, *J Mol Biol 234*, 87-98.

[63] Wegerer, A., Sun, T., and Altenbuchner, J. (2008) Optimization of an E. coli L-rhamnose-inducible expression vector: test of various genetic module combinations, *BMC biotechnology 8*, 2.

[64] Klock, H. E., Koesema, E. J., Knuth, M. W., and Lesley, S. A. (2008) Combining the polymerase incomplete primer extension method for cloning and mutagenesis with microscreening to accelerate structural genomics efforts, *Proteins-Structure Function and Bioinformatics 71*, 982-994.

[65] Haldimann, A., Daniels, L. L., and Wanner, B. L. (1998) Use of new methods for construction of tightly regulated arabinose and rhamnose promoter fusions in studies of the Escherichia coli phosphate regulon, *Journal of Bacteriology 180*, 1277-1286.

[66] Pace, C. N., Vajdos, F., Fee, L., Grimsley, G., and Gray, T. (1995) How to measure and predict the molar absorption coefficient of a protein, *Protein Sci 4*, 2411-2423.

[67] Smith, C. A., Ban, D., Pratihar, S., Giller, K., Paulat, M., Becker, S., Griesinger, C., Lee, D., and de Groot, B. L. (2016) Allosteric switch regulates protein-protein binding through collective motion, *Proc Natl Acad Sci U S A 113*, 3269-3274.

[68] Hudson, W. H., Kossmann, B. R., de Vera, I. M., Chuo, S. W., Weikum, E. R., Eick, G. N., Thornton, J. W., Ivanov, I. N., Kojetin, D. J., and Ortlund, E. A. (2016) Distal substitutions drive divergent DNA specificity among paralogous transcription factors through subdivision of conformational space, *Proc Natl Acad Sci U S A 113*, 326-331.

[69] Fiedler, M., Sanchez-Barrena, M. J., Nekrasov, M., Mieszczanek, J., Rybin, V., Muller, J., Evans, P., and Bienz, M. (2008) Decoding of methylated histone H3 tail by the Pygo-BCL9 Wnt signaling complex, *Mol Cell 30*, 507-518.

[70] Sinha, N., and Smith-Gill, S. J. (2002) Electrostatics in protein binding and function, *Curr Protein Pept Sci 3*, 601-614.

[71] Fadrna, E., Hladeckova, K., and Koca, J. (2005) Long-range electrostatic interactions in molecular dynamics: an endothelin-1 case study, *J Biomol Struct Dyn 23*, 151-162.

[72] Schreiber, G., and Fersht, A. R. (1995) Energetics of protein-protein interactions: Analysis ofthe Barnase-Barstar interface by single mutations and double mutant cycles, *Journal of Molecular Biology 248*, 478-486.

[73] Yang, L.-Q., Sang, P., Tao, Y., Fu, Y.-X., Zhang, K.-Q., Xie, Y.-H., and Liu, S.-Q. (2014) Protein dynamics and motions in relation to their functions: several case studies and the underlying mechanisms, *Journal of Biomolecular Structure & Dynamics 32*, 372-393.

[74] Pascoe, H. G., Gutowski, S., Chen, H., Brautigam, C. A., Chen, Z., Sternweis, P. C., and Zhang, X. (2015) Secondary PDZ domain-binding site on class B plexins enhances the affinity for PDZ-RhoGEF, *Proc Natl Acad Sci U S A 112*, 14852-14857.

[75] Li, Y., Wei, Z., Yan, Y., Wan, Q., Du, Q., and Zhang, M. (2014) Structure of Crumbs tail in complex with the PALS1 PDZ-SH3-GK tandem reveals a highly specific assembly mechanism for the apical Crumbs complex, *Proc Natl Acad Sci U S A 111*, 17444-17449.

[76] Riising, E. M., Comet, I., Leblanc, B., Wu, X., Johansen, J. V., and Helin, K. (2014) Gene silencing triggers polycomb repressive complex 2 recruitment to CpG islands genome wide, *Mol Cell 55*, 347-360.

[77] Chen, S., Yang, Z., Wilkinson, A. W., Deshpande, A. J., Sidoli, S., Krajewski, K., Strahl, B. D., Garcia, B. A., Armstrong, S. A., Patel, D. J., and Gozani, O. (2015) The PZP Domain of AF10 Senses Unmodified H3K27 to Regulate DOT1L-Mediated Methylation of H3K79, *Mol Cell 60*, 319-327.

[78] Graham, S. E., Tweedy, S. E., and Carlson, H. A. (2016) Dynamic behavior of the post-SET loop region of NSD1: Implications for histone binding and drug development, *Protein Sci 25*, 1021-1029.

[79] de Oliveira, P. S., Ferraz, F. A., Pena, D. A., Pramio, D. T., Morais, F. A., and Schechtman, D. (2016) Revisiting protein kinase-substrate interactions: Toward therapeutic development, *Sci Signal 9*, re3.

[80] Kristensen, L. H., Nielsen, A. L., Helgstrand, C., Lees, M., Cloos, P., Kastrup, J. S., Helin, K., Olsen, L., and Gajhede, M. (2012) Studies of H3K4me3 demethylation by KDM5B/Jarid1B/PLU1 reveals strong substrate recognition in vitro and identifies 2,4-pyridine-dicarboxylic acid as an in vitro and in cell inhibitor, *Febs j 279*, 1905-1914.

[81] 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 alpha-helical structure of the histone H3 tail to facilitate acetylation and methylation sampling and modification, *Nucleic Acids Res 42*, 822-835.

[82] Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A., and Finn, R. D. (2012) The Pfam protein families database, *Nucleic Acids Res 40*, D290-301.

[83] Li, W., Jaroszewski, L., and Godzik, A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases, *Bioinformatics 18*, 77-82.

[84] Wheeler, T. J., Clements, J., and Finn, R. D. (2014) Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models, *Bmc Bioinformatics 15*, 7.

[85] Konagurthu, A. S., Whisstock, J. C., Stuckey, P. J., and Lesk, A. M. (2006) MUSTANG: a multiple structural alignment algorithm, *Proteins 64*, 559-574.

[86] Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Research 32*, 1792-1797.

[87] Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era, *Proc Natl Acad Sci U S A 110*, 15674-15679.

[88] Ovchinnikov, S., Kamisetty, H., and Baker, D. (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information, *Elife 3*, e02030.

[89] Simonetti, F. L., Teppa, E., Chernomoretz, A., Nielsen, M., and Marino Buslje, C. (2013) MISTIC: Mutual information server to infer coevolution, *Nucleic Acids Res 41*, W8-14.

[90] Buslje, C. M., Santos, J., Delfino, J. M., and Nielsen, M. (2009) Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information, *Bioinformatics 25*, 1125-1131.

[91] Marino Buslje, C., Teppa, E., Di Domenico, T., Delfino, J. M., and Nielsen, M. (2010) Networks of high mutual information define the structural proximity of catalytic sites: implications for catalytic residue identification, *PLoS Comput Biol 6*, e1000978.

[92] Pavlidis, P., and Noble, W. S. (2003) Matrix2png: a utility for visualizing matrix data, *Bioinformatics 19*, 295-296.

[93] Hsu, W. L., Oldfield, C. J., Xue, B., Meng, J., Huang, F., Romero, P., Uversky, V. N., and Dunker, A. K. (2013) Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding, *Protein Sci 22*, 258-273.

[94] Eddy, S. R. (1998) Profile hidden Markov models, *Bioinformatics 14*, 755-763.

[95] Vanhee, P., Reumers, J., Stricher, F., Baeten, L., Serrano, L., Schymkowitz, J., and Rousseau, F. (2010) PepX: a structural database of non-redundant protein-peptide complexes, *Nucleic Acids Res 38*, D545-551.

[96] Kortemme, T., Kim, D. E., and Baker, D. (2004) Computational alanine scanning of protein-protein interfaces, *Science's STKE : signal transduction knowledge environment 2004*, pl2.

[97] Dharmarajan, V., Lee, J. H., Patel, A., Skalnik, D. G., and Cosgrove, M. S. (2012) Structural basis for WDR5 interaction (Win) motif recognition in human SET1 family histone methyltransferases, *J Biol Chem 287*, 27275-27289.

[98] Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions, *Protein Eng 8*, 127-134.

[99] Rother, K., Hildebrand, P. W., Goede, A., Gruening, B., and Preissner, R. (2009) Voronoia: analyzing packing in protein structures, *Nucleic Acids Research 37*, D393-D395.

[100] Rother, K., Preissner, R., Goede, A., and Frommel, C. (2003) Inhomogeneous molecular density: reference packing densities and distribution of cavities within proteins, *Bioinformatics 19*, 2112-2121.

[101] Chakravarty, S., and Varadarajan, R. (1999) Residue depth: a novel parameter for the analysis of protein structure and stability, *Structure 7*, 723-732.

[102] Wang, G., and Dunbrack, R. L., Jr. (2003) PISCES: a protein sequence culling server, *Bioinformatics 19*, 1589-1591.

[103] Hubbard, S. J., and Thornton, J. M. (1993) 'NACCESS', computer program.

[104] Zhang, W., Dunker, A. K., and Zhou, Y. (2008) Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks, *Proteins 71*, 61-67.

[105] He, F., Umehara, T., Saito, K., Harada, T., Watanabe, S., Yabuki, T., Kigawa, T., Takahashi, M., Kuwasako, K., Tsuda, K., Matsuda, T., Aoki, M., Seki, E., Kobayashi, N., Guntert, P., Yokoyama, S., and Muto, Y. (2010) Structural insight into the zinc finger CW domain as a histone modification reader, *Structure 18*, 1127-1139.

[106] Li, W., Zhao, A., Tempel, W., Loppnau, P., and Liu, Y. (2016) Crystal structure of DPF3b in complex with an acetylated histone peptide, *J Struct Biol 195*, 365-372.

[107] Aurora, R., and Rose, G. D. (1998) Helix capping, *Protein Sci 7*, 21-38.

[108] Klein, B. J., Simithy, J., Wang, X., Ahn, J., Andrews, F. H., Zhang, Y., Cote, J., Shi, X., Garcia, B. A., and Kutateladze, T. G. (2017) Recognition of Histone H3K14 Acylation by MORF, *Structure 25*, 650-654.e652.

[109] 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.

[110] Kaustov, L., Ouyang, H., Amaya, M., Lemak, A., Nady, N., Duan, S., Wasney, G. A., Li, Z., Vedadi, M., Schapira, M., Min, J., and Arrowsmith, C. H. (2011) Recognition and specificity determinants of the human cbx chromodomains, *J Biol Chem 286*, 521-529.

[111] Karpen, M. E., de Haseth, P. L., and Neet, K. E. (1992) Differences in the amino acid distributions of 3(10)-helices and alpha-helices, *Protein Science : A Publication of the Protein Society 1*, 1333-1342.

[112] Gunasekaran, K., Nagarajaram, H. A., Ramakrishnan, C., and Balaram, P. (1998) Stereochemical punctuation marks in protein structures: glycine and proline containing helix stop signals, *J Mol Biol 275*, 917-932.

[113] Shoemaker, K. R., Kim, P. S., York, E. J., Stewart, J. M., and Baldwin, R. L. (1987) Tests of the helix dipole model for stabilization of alpha-helices, *Nature 326*, 563-567.

[114] Liu, X., Speckhard, D. C., Shepherd, T. R., Sun, Y. J., Hengel, S. R., Yu, L., Fowler, C. A., Gakhar, L., and Fuentes, E. J. (2016) Distinct Roles for Conformational Dynamics in Protein-Ligand Interactions, *Structure 24*, 2053-2066.

[115] Hopf, T. A., Colwell, L. J., Sheridan, R., Rost, B., Sander, C., and Marks, D. S. (2012) Three-dimensional structures of membrane proteins from genomic sequencing, *Cell 149*, 1607-1621.

[116] Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., and Sander, C. (2011) Protein 3D Structure Computed from Evolutionary Sequence Variation, *PLoS ONE 6*, e28766.

[117] Slama, P., and Geman, D. (2011) Identification of family-determining residues in PHD fingers, *Nucleic Acids Research 39*, 1666-1679.

[118] Lowe, E. D., Tews, I., Cheng, K. Y., Brown, N. R., Gul, S., Noble, M. E., Gamblin, S. J., and Johnson, L. N. (2002) Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A, *Biochemistry 41*, 15625-15634.

[119] Sharrocks, A. D., Yang, S. H., and Galanis, A. (2000) Docking domains and substrate-specificity determination for MAP kinases, *Trends Biochem Sci 25*, 448-453.

[120] Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015) Rational design of antibodies targeting specific epitopes within intrinsically disordered proteins, *Proc Natl Acad Sci U S A 112*, 9902-9907.