

2010

# Proteome-Wide Prediction of Acetylation Substrates

Amrita Basu

Follow this and additional works at: [http://digitalcommons.rockefeller.edu/student\\_theses\\_and\\_dissertations](http://digitalcommons.rockefeller.edu/student_theses_and_dissertations)

 Part of the [Life Sciences Commons](#)

---

## Recommended Citation

Basu, Amrita, "Proteome-Wide Prediction of Acetylation Substrates" (2010). *Student Theses and Dissertations*. 414.  
[http://digitalcommons.rockefeller.edu/student\\_theses\\_and\\_dissertations/414](http://digitalcommons.rockefeller.edu/student_theses_and_dissertations/414)

This Thesis is brought to you for free and open access by Digital Commons @ RU. It has been accepted for inclusion in Student Theses and Dissertations by an authorized administrator of Digital Commons @ RU. For more information, please contact [mcsweej@mail.rockefeller.edu](mailto:mcsweej@mail.rockefeller.edu).



## **PROTEOME-WIDE PREDICTION OF ACETYLTATION SUBSTRATES**

A Thesis Presented to the Faculty of  
The Rockefeller University  
in Partial Fulfillment of the Requirements for  
the degree of Doctor of Philosophy

by

Amrita Basu

June 2010



## **Proteome-wide Prediction of Acetylation Substrates**

Amrita Basu, Ph.D.  
The Rockefeller University 2010

Eukaryotic DNA is found packaged with proteins and RNA, which forms a substance called chromatin. This packaging is dynamic and regulates access to DNA for essential cellular processes such as transcription, replication, and repair. In recent years, studies have shown that regulated changes in the chemical and physical properties of chromatin often lead to dynamic changes in multiple cellular processes by affecting the accessibility of the DNA. These changes can be brought about in part through post-translational modifications of histone proteins, which are involved in disrupting chromatin contacts or by recruiting effector proteins to chromatin.

Acetylation is one of the well-studied post-translational modifications that has been associated with chromatin-associated processes, notably gene regulation. Many studies have contributed to our knowledge of the enzymology underlying acetylation, including efforts to understand the molecular mechanism of substrate recognition by several acetyltransferases, but traditional experiments to determine intrinsic features of substrate and site specificity have proven challenging. In my thesis work, I hypothesize that the primary amino acid sequence surrounding an acetylated lysine plays a critical role in acetylation site selection, and whether there are sequence preferences that enable a lysine acetyltransferase to recognize target

lysines. A computational method was devised to examine this hypothesis, and an experimental approach was taken to test my computationally-derived predictions. In Chapter 2, I describe my basic computational methods, using a clustering analysis of protein sequences to predict lysine acetylation based on the sequence characteristics of acetylated lysines within histones. I define a local amino acid sequence composition that represents potential acetylation sites by implementing a clustering analysis of histone and non-histone sequences. I demonstrate that this sequence composition has predictive power on two independent experimental datasets of acetylation marks. In Chapter 3, I describe the experimental validation approach used to detect acetylation in histone and nonhistone proteins using mass spectrometry. I also report several novel non-histone acetylated substrates in *S. cerevisiae*. My approach, combined with more traditional experimental methods, may be useful for identifying additional proteins in the acetylome. Finally, in Chapter 4, I describe two bioinformatics approaches; one to predict additional chromatin associated effector proteins, and another to further understand the evolutionary history and complexity of the Polycomb Group (PcG) proteins in multicellular organisms in order to infer gene expansion, co-evolution, and deletion events.

## ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor and mentor, Dr. David Allis, for his constant support and guidance. I joined the doctoral program with a desire to apply computational methods to key biological problems, which included experimental validation of my computational approach. While I was not trained as a wet-bench scientist, Dave graciously allowed me join the lab and fulfill my scientific vision. I thank him for being enthusiastic about my project, his scientific wisdom, his continuous optimism, and his generosity. When I felt discouraged on any matter in and outside of science, Dave could lift my spirits magically, and I always felt better after speaking to him.

Next, I want to thank Dr. Eran Segal for serving as a computational mentor on my main project. Eran helped guide the computational component of the project, and mentored me on how to apply bioinformatic approaches to histone biology.

I thank my faculty advisory committee, Dr. Mike Rout, and Dr. Brian Chait for their time, their interest in my work, and all their helpful suggestions along the way. I enjoyed all the scientific discussions that we engaged in regarding my project. I would also like to thank my outside committee member, Dr. Christina Leslie from MSKCC, for providing valuable input on the computational side of my project.

I am grateful to Dr. Sandra Hake. Her suggestions, input, and feedback have been right on the mark throughout this process, and I thank her for her dedication and hard work towards my project.

I thank our collaborators Dr. Donald Hunt, Dr. Kristie Rose, Dr. Beatrix Ueberheide, Dr. Yingming Zhao, Dr. Ronald Beavis, and Dr. Ben Garcia for their excellent mass spectrometry work as well as the pleasant interaction during our collaboration. I also thank Dr. Brian Chait for putting me in touch with Dr. Yingming Zhao and Dr. Ronald Beavis, both of whom were excellent collaborators.

I want to thank Dr. Monika Lachner (a former postdoc in the Allis lab) who was very generous with her time, and who often served as my lab “go-to” person during her tenure at Rockefeller. I enjoyed the numerous scientific conversations with her and friendship that we developed along the way.

Dr. Alex Ruthenburg also helped me at various stages by rendering aesthetically appealing structural images, discussing my projects and ideas with me, and critically reading parts of my thesis and manuscript.

I thank Dr. Laura Banaszynski, Sarah Whitcomb, and Lindsey Baker for critically reading parts of my thesis and for their helpful comments on my thesis. I thank Dr. Jung-Ae Kim for guiding me on the yeast mutagenesis experiments and critically reading parts of this thesis. I also thank Aaron Goldberg for his encouraging words during my tenure in the Allis Lab.

I am very grateful to all past and present members of the Allis lab for their willingness to always share reagents and advice. In particular, Dr. Jason Tanny was very helpful on the yeast experimental aspects of the project.

I thank the lab manager of the Allis lab, Jamie Winshell, for helping me out at countless times, and for running the lab flawlessly. I also want to thank

the laboratory administrator, Chandra Egger for handling all my paper and email requests especially during the time of manuscript publication.

I am indebted to the Tri-Institutional Computational Biology program for providing me with computational resources, funding, and support. I thank David Rockefeller Graduate Program also for funding, and the Dean's Office during thesis preparation times. In particular, I thank Cris Rosario and Marta Delgado for always answering questions and helping with all the thesis scheduling and paperwork.

I want to thank my friends who have supported me throughout this endeavor. In particular, I want to thank my fiancé, Sanjeev Somani, for his moral support the last few months and his incredible calming effect on me.

I want to thank my sister, Sambrita for always listening to everything I have to say, for making me laugh out loud, and for being my best friend.

Finally, I want to thank my parents, Dr. Jayasree Basu and Dr. Sankar Basu, for providing me with unbelievable love, support, and wisdom throughout my life. They have been there with me in each and every step of the way. I could not have pursued my passion without all their hard work, sacrifice, and devotion.



## TABLE OF CONTENTS

<b>Acknowledgments .....</b>	<b>iii</b>
<b>Table of Contents .....</b>	<b>vi</b>
<b>List of Figures .....</b>	<b>vii</b>
<b>List of Tables.....</b>	<b>x</b>
<b>List of Abbreviations .....</b>	<b>xi</b>
<b>Chapter 1: General Introduction.....</b>	<b>1</b>
<b>Chapter 2: Computational Prediction of Acetylation Substrates ..</b>	<b>46</b>
<b>Chapter 3: Experimental Validation of Acetylation Substrates ....</b>	<b>76</b>
<b>Chapter 4: Domain Prediction of Chromatin-associated Proteins .....</b>	<b>101</b>
<b>Chapter 5: General Discussion.....</b>	<b>137</b>
<b>Chapter 6: Materials/Methods.....</b>	<b>155</b>
<b>Appendix.....</b>	<b>171</b>
<b>References .....</b>	<b>192</b>

## LIST OF FIGURES

<b>Figure 1.1:</b> Chromatin organization and Post-translational Modifications .....	3
<b>Figure 1.2:</b> Histone Acetylation Function and Primary Sequence Context .....	8
<b>Figure 1.3:</b> Gcn5 (a histone KAT) coupled to histone H3 peptide and coenzyme A (CoA) .....	12
<b>Figure 1.4:</b> Examples of nonhistone protein acetylation.....	16
<b>Figure 1.5:</b> H3 acetylation in yeast and conservation between Organisms .....	23
<b>Figure 1.6:</b> Crosstalk between lysine methylation and acetylation on H3 and H4 histone tails.....	28
<b>Figure 1.7:</b> Effector proteins and cryptic domains.....	36
<b>Figure 1.8:</b> Polycomb schematic representations .....	44
<b>Figure 2.1:</b> Schematic of computational approach .....	49
<b>Figure 2.2:</b> Computational prediction of human histone acetylation sites.....	54
<b>Figure 2.3:</b> Predictive hierarchical tree of all 56 lysines in human core histones.....	55

<b>Figure 2.4:</b> Performance on test set of human acetylated substrates...	59
<b>Figure 2.5:</b> Frequency distribution of amino acids surrounding lysines in human histone and nonhistone proteins .....	63
<b>Figure 2.6:</b> Performance on Yeast Proteome-wide Dataset.....	67
<b>Figure 2.7:</b> Frequency distribution of amino acids surrounding yeast lysines .....	68
<b>Figure 2.8:</b> PredMod, an acetylation prediction tool .....	70
<b>Figure 3.1:</b> Validation on histone predicted lysines .....	79
<b>Figure 3.2:</b> Novel predictions and in-vivo validations in <i>S. cerevisiae</i> .....	83
<b>Figure 3.3:</b> Mutagenized yeast histones .....	88
<b>Figure 3.4:</b> PTMs as result of flanking residue mutations surrounding H3K14 .....	92
<b>Figure 3.5:</b> Crosstalk Model.....	97
<b>Figure 3.6:</b> Sequence alignment of Gcn5 mediated substrates.....	99
<b>Figure 4.1:</b> Domain and motif structure of selected PRC1 proteins .....	106
<b>Figure 4.2:</b> Domain and motif structure of selected PRC2 proteins .....	108
<b>Figure 4.3:</b> Phylogenetic representation of selected organisms and their PcG homologs.....	112
<b>Figure 4.4:</b> CEC-1 sequence alignment and phylogeny of mouse Cbx proteins .....	115

<b>Figure 4.5:</b> Schematic of domain prediction method .....	121
<b>Figure 4.6:</b> Effector Protein Predictions.....	126
<b>Figure 4.7:</b> Peptide binding assay using Arp8 whole cell extract .....	129
<b>Figure 4.8:</b> H3K4Me0 Known and Putative Effectors.....	134
<b>Figure 5.1:</b> Proposed discovery of KAT substrates .....	140
<b>Figure 5.2:</b> Proposed mutagenesis of H3K36.....	146
<b>Figure 5.3:</b> Schematic of mammalian histone variant proteins.....	150
<b>Figure 5.4:</b> H3K56 and Cse4 .....	151
<b>Figure 5.5:</b> H1 Schematic and Predictions .....	153

## LIST OF TABLES

<b>Table 3.1:</b> Mutations, Rationale, and Acetylation Prediction .....	87
--	----

## ABBREVIATIONS

aa	amino acid
Ac	Acetylation
acetyl-CoA	acetyl-Coenzyme A
AUC	Area Under Curve
bp	base pairs
BD	bromodomain
BPTF	bromodomain PHD finger transcription factor
BSA	bovine serum albumin
C	Celsius
CAD	Collisionally Activated Dissociation
ChIP	Chromatin Immunoprecipitation
CD	chromodomain
3D	three dimensional
DMEM	Dulbecco's Modified Eagle's Medium
DNA	deoxyribonucleic acid
DTT	DL-1,4-dithiothreitol
E-value	expectation value
FOA	fluorouracil-6-carboxylic acid monohydrate
Fp	false positive
GO	gene ontology
H3	histone H3
H31-20	H3 residues 1-20
HAT	histone acetyltransferase
HDAC	histone deacetylase
HeLa	Henrietta Lacks
HMM	Hidden Markov Model
HP1	Heterochromatin Protein 1

HPLC	high performance liquid chromatography
IP	immunoprecipitation
KAT	lysine acetyltransferase
kb	kilo bases
kDa	kilo Dalton
KMT	lysine methyltransferase
LC	liquid chromatography
LOV	Leave One out cross Validation
LTQ-FT	linear quadrupole ion trap-Fourier transform mass spectrometer
Lys	Lysine
l	liter
M	molar
m/z	mass over charge
Me	Methylation
Me0	Unmodified methyl
Me1	mono-methylated
Me2	di-methylated
Me3	tri-methylated
MS	mass spectrometry
MSA	multiple sequence alignment
MW	molecular weight
nt	nucleotide
NURF	Nucleosome Remodeling Factor
OD600	optical density at 600 nm
P	phosphorylation
PHD	Plant Homeodomain
P-value	probability with a value
Pc	Polycomb

PcG	Polycomb group
Ph	Polyhomeotic
PHD	Plant Homeo Domain
PMSF	phenylmethanesulphonyl fluoride
Pr	Probability
PrA	Protein A
PRC	Polycomb Repressive Complex
PRE	Polycomb Responsive element
Psc	Posterior sex combs
PTM	post-translational modification
ROC	Receiving Operating Curve
RP-HPLC	reversed-phase high-performance liquid chromatography
PCR	polymerase chain reaction
rpm	rotations per minute
$S_n$	Sensitivity
$S_p$	Specificity
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel Electrophoresis
SH2	Src-homology 2
TRE	trithorax responsive element
WT	wild-type, original sequence without mutation



## **Chapter 1: General Introduction**

The annotated human genome contains approximately 40,000 genes by current estimates (Karolchik, Baertsch et al. 2003), a number that seems surprisingly low. However, the estimated number of proteins encoded by these genes is two to three orders of magnitude higher (Hsu, Pringle et al. 2005). Because each protein is generally encoded by one gene in the genome, one might expect a one to one correspondence of genes to proteins. However, this is not the case: the estimated number of proteins encoded by the 40,000 human genes is two to three orders of magnitude higher. What is the purpose of this diversity? Proteome complexity can be built by diversification at both the mRNA level (through alternative splicing) and the protein level (through post-translational modification (PTM) of the protein side chains). Greater than 5% of the genes in the human genome encode enzymes that catalyze such modifications, including hundreds of protein kinases, phosphatases, ubiquitinyl ligases, acetylases and deacetylases, methyl transferases and glycosyl transferases. By adding chemical moieties onto one or more amino acids, PTMs can determine a protein's localization, interactions with other proteins, and gene expression. For example, phosphorylation of a protein substrate can propagate downstream signaling events (Burnett and Kennedy 1954; Olsen, Blagoev et al. 2006), ubiquitination marks cyclins and other proteins for degradation at cell-cycle specific time points (Xu, Duong et al. 2009), and methylation can

epigenetically activate (Beisel, Imhof et al. 2002) or repress gene expression (Nakayama, Rice et al. 2001; Grewal and Rice 2004).

PTMs are particularly abundant and well studied in histone proteins, the major protein components of chromatin. Chromatin is the coupling of DNA, RNA, and protein that make up chromosomes. The nucleosome, the fundamental repeating unit of chromatin, consists of 146 base pairs of DNA wrapped around the four core histones (H2A, H2B, H3, H4)(Luger, Mader et al. 1997) (**Figures 1.1A, B**). Multiple nucleosomes joined by stretches of linker DNA form a structure known as “beads on a string” (**Figure 1.1B**). We are only beginning to understand that the changes in chromatin structure that underlie many DNA-templated processes are affected by post-translational modification of histone proteins, including but not limited to methylation, phosphorylation, and acetylation (**Figure 1.1C**). These modifications may contribute to “epigenetic signatures” that are important for diverse processes such as gene regulation, apoptosis, mitosis, and responses to DNA damage. These modifications create a dynamic readout, referred to as the “histone code” (Strahl and Allis 2000; Turner 2008), where PTMs can function as binding sites for specific protein domains while other PTMs alter the net charge on the nucleosomes in a way that alters chromatin structure.

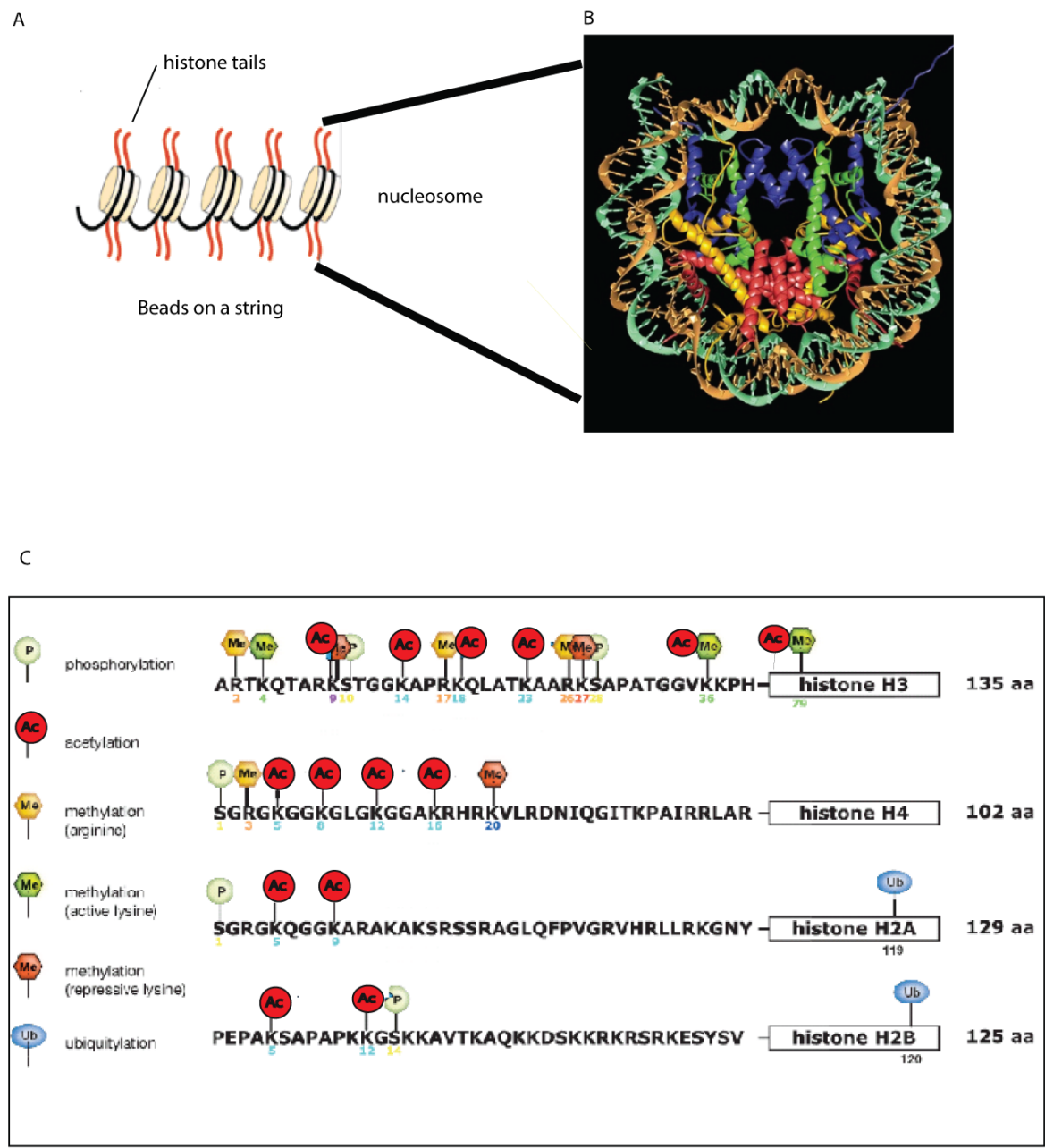
**Figure 1.1: Chromatin organization and post-translational modifications**

**A.** A nucleosome particle is composed of 146 bp of DNA wrapped around an octamer of 2 copies each of histones H2A, H2B, H3, and H4, with the tails of the histone proteins protruding from the core structure. These particles are then linked together by the DNA to form a structure known as “beads on a string”. Image modified from (Marmorstein 2001).

**B.** Crystal structure of nucleosome solved at 2.8Å. Nucleosome core particle: ribbon traces for the 146-bp DNA (Watson strand and Crick strands in brown and turquoise, respectively) and eight histone protein main chains (blue: H3; green: H4; yellow: H2A; red: H2B). Image adapted from (Luger, Mader et al. 1997).

**C.** Shown are the four core histones: H3, H4, H2A, and H2B, and their post-translational modification sites. Acetylation marks are denoted by red filled circles, arginine methylation denoted by yellow hexagons, and lysine methylation denoted by green or red hexagons. For lysine methylation sites, green hexagons denote active marks (such as H3K4me) while red hexagons denote repressive marks, (such as H3K27). Green circles are phosphorylation and blue ovals are ubiquitylation. Boxes surrounding H3, H4, H2A, and H2B represent histone globular domains. Image modified from Epigenetics, 2007.

Figure 1.1



## Histone Acetylation

More than forty years ago, Allfrey and colleagues reported a strong correlation between increased levels of histone acetylation and elevated levels of gene expression (Allfrey, Faulkner et al. 1964). Since then, the field of chromatin biology has advanced considerably with remarkable progress made into mechanistic insights of histone modifications and their biological functions. Histone acetylation has the capacity to destabilize the chromatin polymer through neutralization of the basic charge of the lysine residue, potentially with consequences for chromatin nucleosomal stability and higher-order structure ("cis" effects) (Tse, Sera et al. 1998; Verreault, Kaufman et al. 1998; Taverna, Li et al. 2007). Furthermore, acetylation can potentially affect chromatin dynamics by recruiting specialized "effector" proteins ("trans"-effects) (Jenuwein and Allis 2001).

Lysine acetylation in histones was the first PTM for which the enzymes that both catalyze HATs (histone acetyltransferases) and remove HDACs (histone deacetylases) was identified. These enzymes are responsible for governing a steady-state balance of acetylation (Brownell, Zhou et al. 1996; Taunton, Hassig et al. 1996; Pflum, Tong et al. 2001). In fact, the first transcription-related nuclear histone acetyltransferase (HAT, or KAT renamed in (Allis, Berger et al. 2007)), Gcn5, was isolated from the *Tetrahymena* macronucleus through an in-gel assay (a mixture of proteins

from cell extracts separated on a SDS gel) (Brownell, Zhou et al. 1996). Since then, the discovery of KAT proteins acetylating all four core histones has been a hallmark for the chromatin field as researchers have studied the individual modifications and their functional relevance in the cell (Kurdistani, Tavazoie et al. 2004; Shogren-Knaak, Ishii et al. 2006; Kaplan, Liu et al. 2008; Li, Zhou et al. 2008; Tjeertes, Miller et al. 2009)(**Figure 1.2A**; in red). In addition to the discovery of the first transcription-related histone KAT, the discovery of the first transcription related histone HDAC, HDAC1, (Taunton, Hassig et al. 1996) also led to the formation of a model for gene-specific histone PTMs: activators that are bound by DNA are involved in recruiting KATs to acetylate nucleosomal histones, while repressors recruit HDACs to deacetylate histones. Three years later, the first bromodomain was discovered as a protein that interacts specifically with acetylated lysines in histones (Dhalluin, Carlson et al. 1999), making it the first known protein module to do so. Bromodomains were functionally linked to the HAT activity of co-activators in the regulation of gene transcription. Together, modifying enzymes that “write” the PTM, enzymes that “erase” the PTM, and molecules that interact specifically with the PTM or “readers”, are linked to enzyme activity in gene expression regulation, but how do these molecules recognize their epitopes, and what dictates their recognition machinery? Primary sequence context surrounding a target lysine could be one major factor dictating these types of recognition (**Figure 1.2B**), however this factor is

still relatively understudied. Limited biochemical and structural studies of KATs and bromodomain coupled to histones display some sequence preferences (Mujtaba, Zeng et al. 2007; Liu, Wang et al. 2008), however a rigorous and thorough analysis is still needed to pinpoint specific sequence motifs or patterns on the substrate that help establish appropriate acetyllysine-dependent interactions with chromatin.

**Figure 1.2: Histone Acetylation Function and Primary Sequence Context**

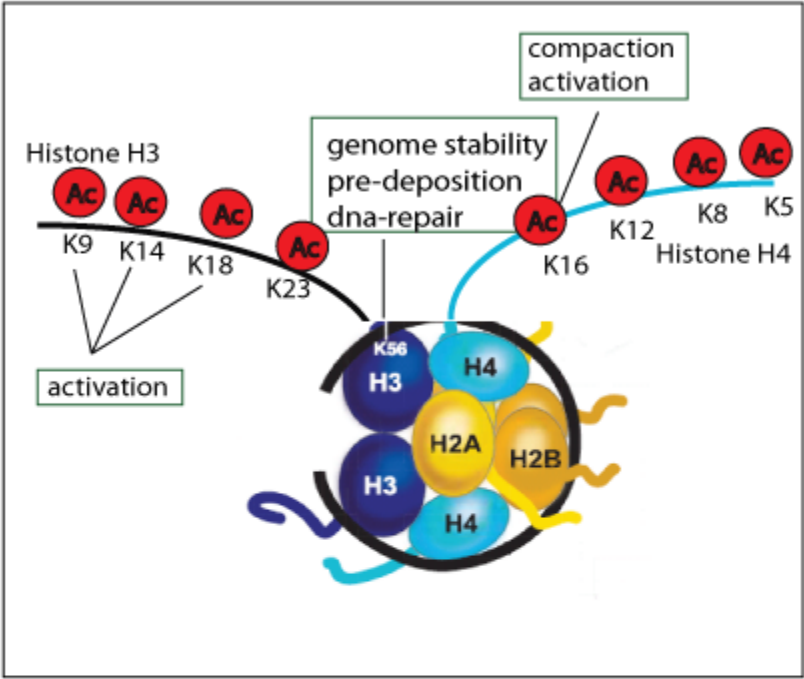
**A.** Acetylation marks are denoted by red circles. Functions associated with the acetyl marks are shown by a rectangular box. Image modified from Epigenetics,2007.

**B.** Lysine in center (red) and letters (in black) surrounding lysine represent the primary sequence context of a PTM. Effector (in green) also can bind the modification based on contacts it makes with residues surrounding the PTM.

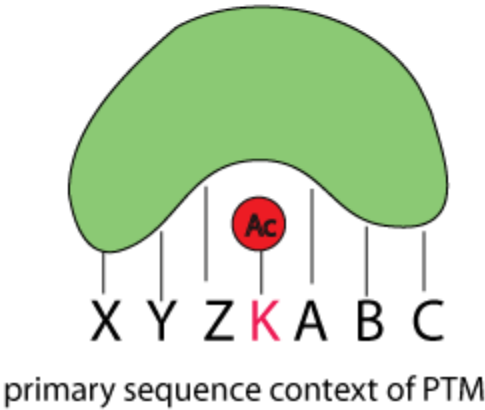


Figure 1.2

A



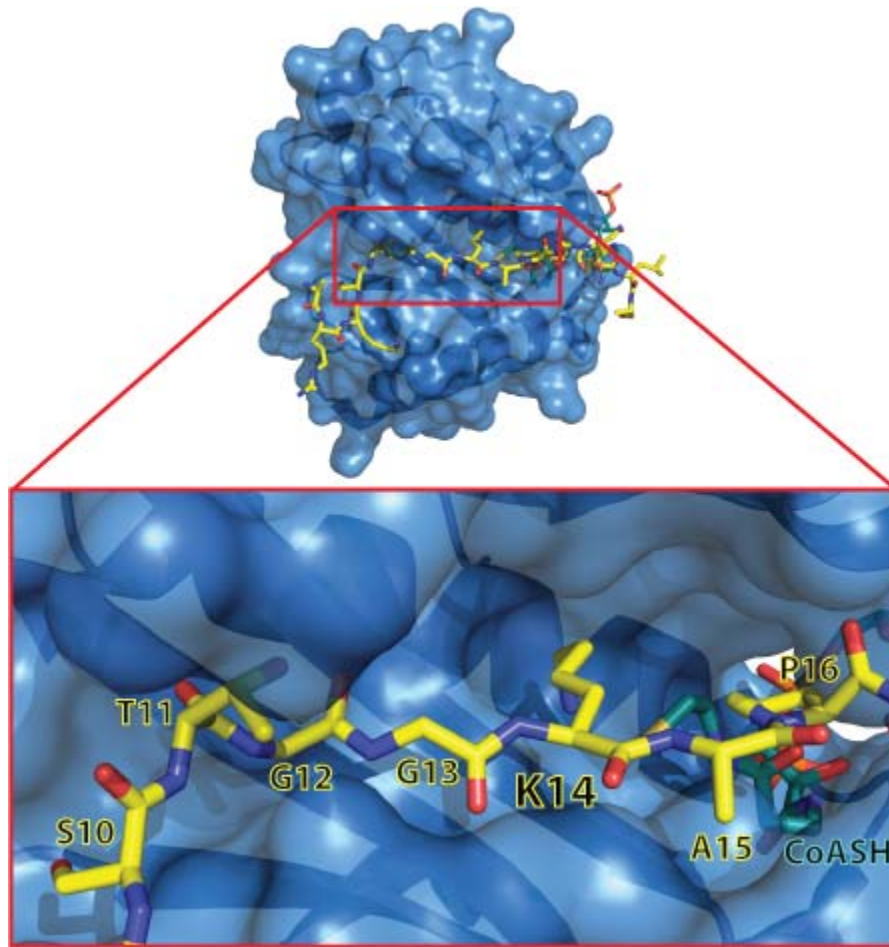
B



Structural analyses of KATs coupled to histone tail peptides have been the subject of intense study. KAT enzymes catalyze the transfer of an acetyl group from the co-factor acetyl-Coenzyme A (acetyl-CoA) to the  $\epsilon$ -amine of a substrate lysine side chain. A large number of KATs have now been identified and characterized. Studies on the divergent histone KAT enzymes Gcn5/PCAF, Esa1 and Hat1 have provided insights into the underlying mechanism of acetylation by KAT proteins (Marmorstein 2001; Marmorstein and Roth 2001). These three histone KAT enzymes contain a conserved core domain that plays a role in binding the co-factor acetyl-CoA in catalysis. More recently, biochemical and structural studies of the metazoan-specific p300/CBP (human homolog of Gcn5) and fungal-specific Rtt109 histone KATs have provided a new understanding into the evolutionary and ancestral relationship between histone KATs and their divergent catalytic and substrate-binding properties (Wang, Tang et al. 2008).

Co-crystallized histone KATs coupled to peptides also provide insight regarding substrate recognition. Positively charged residues are typically present within three to four amino acid residues (about 10 Å) upstream or downstream of the acetylated lysine residues of known p300 (a human homolog of Gcn5)/CBP substrates (Wang, Tang et al. 2008). Moreover, the X-ray crystal structure of p300, in complex with a bi-substrate inhibitor, Lys-CoA reveals the preference for nearby basic residues, such as a lysine in the +2 and -2 positions (Wang, Tang et al. 2008). Specificity for a random-coil

structure containing a G-K-X-P recognition sequence on the histone substrate is revealed when the Gcn5 crystal structure is coupled to the H31-20 peptide (Rojas, Trievel et al. 1999; Liu, Wang et al. 2008) (**Figure 1.3**). Critical contacts between Gcn5 and H3K14 are displayed; the glycines (G13), and proline (P15) preceding and following H3 lysine 14 make the tightest contacts with Gcn5 and suggest residues that are specific to Gcn5 recognition. Other nonhistone Gcn5-mediated substrates also possess a similar G-K-X-P pattern and will be described further in Chapter 3. These studies demonstrate that residues in the core domain of the KAT together with proximal residues surrounding acetyl-lysines on the histone substrate can help achieve substrate specificity (Liu, Wang et al. 2008).



**Figure 1.3: Gcn5 (a histone KAT) coupled to histone H3 peptide and coenzyme A (CoA)**

Solvent accessible surface representation of *Tetrahymena* GCN5 HAT domain coupled to Histone H31-20 peptide (in yellow) and CoA (shown clearer in enlarged box; bottom). Red box (bottom) represents an enlarged view of the Gcn5 HAT domain in contact with the H3S10-P16 residues of the peptide represented by a stick model. CoA is shown on bottom right. Note that G12 and P16 make direct contacts with the catalytic domain. Image provided by A. Ruthenburg.

## **Nonhistone protein acetylation and methods in detecting acetylated substrates**

To a lesser extent, nonhistone protein acetylation has also been implicated in a wide variety of biological processes, such as DNA binding or the stabilization of multi-subunit complexes (Glozak, Sengupta et al. 2005) (Sterner and Berger 2000; Yang and Seto 2008). One of the most famous nonhistone proteins that was discovered as acetylated was p53, where pioneering studies by Roeder and colleagues demonstrated that p53 acetylation was critical for the regulation of its binding to the DNA (**Figure 1.4A,B**) (Gu and Roeder 1997). Since then, KATs, such as p300, have been shown to acetylate multiple other nonhistone transcription-related proteins. Additional studies have identified additional acetylated proteins, such as HIV1 integrase (Cereseto, Manganaro et al. 2005), an HIV protein whose acetylation is required for its viral integration. Acetylation of alpha tubulin was also recently shown to be critical for the formation of cortical neurons in the developing mouse brain (Wynshaw-Boris 2009). Intriguingly, several transcription factors, such as E2F1, and MYC implicated in cancer pathways, have also been shown to be acetylated (Sterner and Berger 2000; Ozaki, Okoshi et al. 2009).

Conventional experiments, such as mutagenesis of potential acetylated lysines, acetylation-specific antibodies, metabolic labeling, mass

spectrometry (MS), and in-vitro histone acetyltransferase assays (**Figure 1.4C**) have typically been used in order to identify acetylated lysines in substrate proteins. More recently, large scale proteomic studies have emerged as a result of high- throughput technology. In one study, the KAT NuA4 (the essential nucleosome acetyltransferase of H4) was incubated with yeast proteome microarrays in the presence of radioactive acetyl-CoA. Many non-chromatin substrates of complex were identified and validated, including acetylation (Lys19 and Lys514) of phosphoenolpyruvate carboxykinase (Pck1p). Acetylation at these sites was then shown to be important for yeast glucogenesis (Lin, Lu et al. 2009). Additionally, advanced proteomic tools have enabled identification of several hundred acetylation sites in approximately two hundred proteins, using samples derived from HeLa cells, mouse liver and bacteria (Kim, Sprung et al. 2006). In addition to regulators of chromatin-based cellular processes, non-nuclear proteins with diverse functions were also identified. Most strikingly, acetylated lysines were found in more than 20% of mitochondrial proteins, including many metabolic enzymes (Kim, Sprung et al. 2006). Another high resolution mass spectrometry study published recently revealed that lysine acetylation preferentially targets large macromolecular complexes involved in diverse cellular processes, such as chromatin remodeling, cell cycle, splicing, nuclear transport, and actin nucleation. The study also reveals that acetylation substrates had enriched residues flanking the target lysine depending on

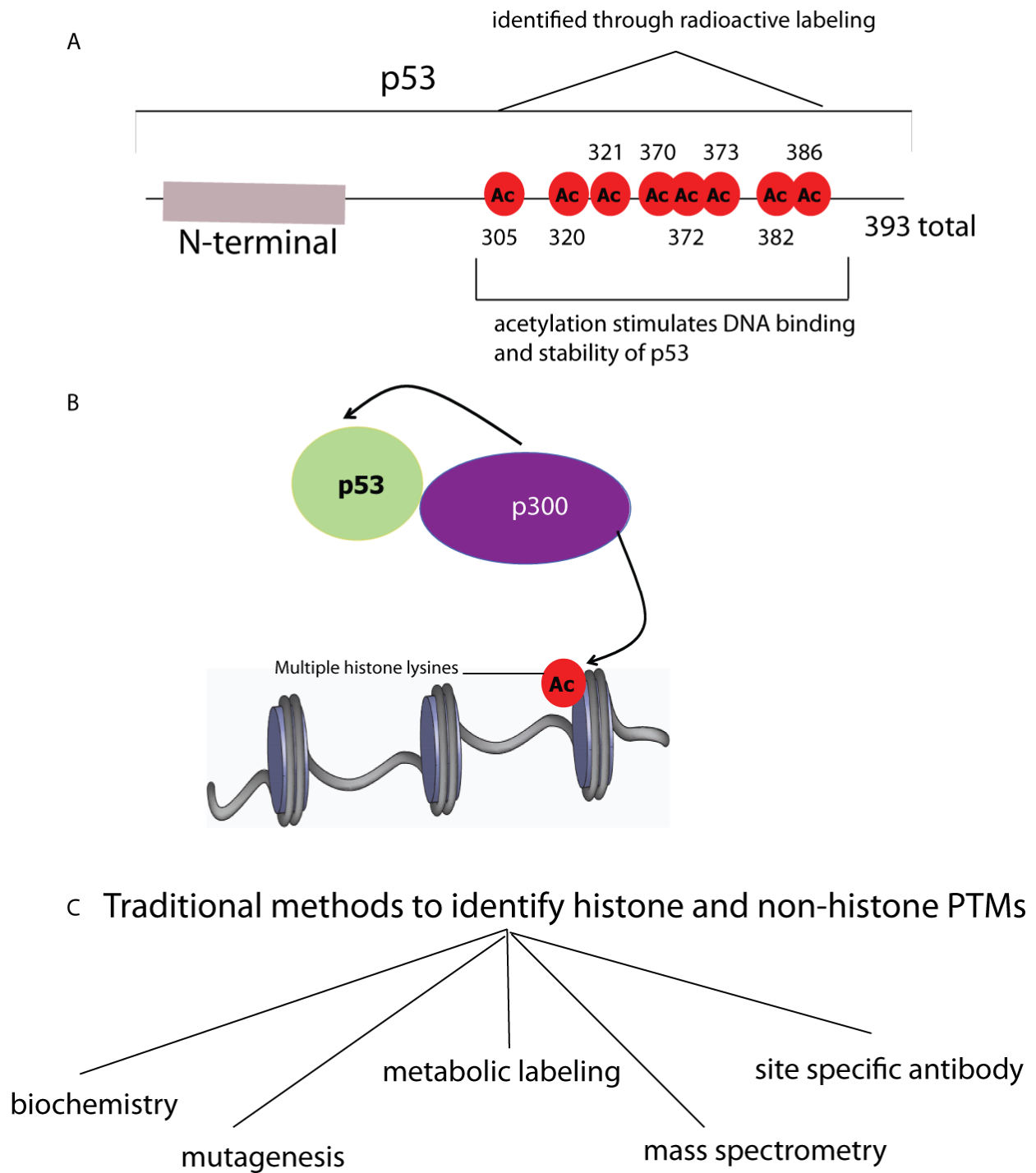
whether the protein resided in the nuclear, cytoplasm, or mitochondria compartment in the cell (Choudhary, Kumar et al. 2009).

**Figure 1.4: Examples of nonhistone protein acetylation**

- A.** p53 was one of the first nonhistone acetylated proteins discovered via radioactive labeling (Gu and Roeder 1997). The protein (total of 393 amino acids) is heavily acetylated protein in the C-terminal tail. Eight lysines have have been identified as acetylated via mass spectrometry, immunoblotting, and mutagenesis experiments. p53 acetylation stimulates sequence specific DNA binding activity *in vivo* and in *vitro* and increases stability of protein.
- B.** A histone KAT can also acetylate lysines in nonhistone proteins. p53 is a classic example, where p300 ( the histone KAT) acetylates multiple residues of p53 as shown in B, but also lysines in the core histones.
- C.** Some of the traditional methods used to identify PTMs on histone and nonhistone proteins. Shown are biochemistry methods, metabolic labeling, site-directed mutagenesis of potential PTM sites, mass spectrometry, and site specific acetyl antibodies.



Figure 1.4



Thus, the discovery of acetylation on histone and nonhistone proteins has been a key advancement in the chromatin field, and has allowed for a better understanding of the roles acetylation plays in human biology and disease. Promising advances have been made in the treatment of certain cancers by developing drug therapies that target HDACs (Marks 2007). While knowing which amino acids in the proteome are acetylated has clear biological and disease applications, currently, only a subset of the total acetylated residues in the proteome have been identified. Methods that allow for rapid identification of all acetylated amino acids in the proteome would therefore be of great importance to the biological community. A computational tool that is predictive of acetylation events in histone and nonhistones could contribute to a more complete understanding of what substrates are physiologically relevant, as more insights are gained into acetylation-mediated pathways.

### **Computational studies**

A limited number of studies suggest that there may be sequence recognition target(s) for certain KATs (Kimura and Horikoshi 1998; Kimura and Horikoshi 1998). A putative “rule” for lysine selection in a primary sequence by a KAT has been proposed previously by examining the N-terminal tail of histones (Kimura and Horikoshi 1998). For example, TIP60 (a mammalian KAT), recognizes specific glycine-lysine G-K patterns in human proteins *in*

*vitro* and *in vivo* (Kimura and Horikoshi 1998). Additionally, the acetylation of Rch1 (a nuclear importin factor) is severely inhibited when a glycine adjacent to the modified lysine is mutated, supporting the view that G-K is part of a recognition motif for acetylation (Bannister, Miska et al. 2000). Moreover, the KAT CLOCK1, acetylates H3 lysine 14, which bears a similar sequence environment to the acetylated lysine of the CLOCK-mediated substrate BMAL1 (Hirayama, Sahar et al. 2007). The term “histone mimic” was recently put forward to describe short stretches in nonhistone proteins that closely resemble histone sequences containing PTMs (Sampath, Marazzi et al. 2007). Observations such as these provide early indications of specific “rules” that can potentially define enzyme recognition of target substrates. More recently, a prediction program was published illustrating that acetylation could be predicted via an analysis of lysine acetylation motifs in a large human proteome-wide dataset (Schwartz, Chou et al. 2009). These motifs were scanned against proteomic sequence data using a newly developed tool called scan-x to globally predict other potential modification sites within multiple organisms. Ten-fold cross-validation was used to determine the sensitivity and minimum specificity for each set of predictions, where a specificity (proportion of of actual positives which are correctly identified) 94%, and a sensitivity (proportion of negatives which are correctly identified) of 17% was achieved. In Chapter 2, I describe a computational model which I have developed to predict human additional

histone and nonhistone substrates in the acetylome. In addition to a computational approach, I experimentally validated my target predictions in budding yeast *in vivo* as described in Chapter 3.

### **Yeast PTMs and conservation of histones**

Histones are among the most highly conserved proteins across species, allowing for their study in multiple organisms. In particular, acetylation in budding yeast histones is a well-studied phenomenon. Extensive studies mapping acetylation and the responsible enzymatic pathways using the yeast core histones have been performed (Grunstein 1990; Clarke, O'Neill et al. 1993; Brownell, Zhou et al. 1996; Grant, Duggan et al. 1997; Smith, Eisen et al. 1998; Bird, Yu et al. 2002). These studies were possible through traditional biochemical methods (**Figure 1.4C**) and highly specific antibodies against histone acetyl-lysines that have been widely used in yeast chromatin studies (Suka, Suka et al. 2001). Genome wide mapping studies of yeast acetylation (via ChIP-Chip) illustrate that H3K9ac peaks at the predicted transcriptional start sites of active genes and that this modification correlates with transcription rates genome-wide (**Figure 1.5A**) (Pokholok, Harbison et al. 2005). Similarly, acetylation of histone H3 at lysine 14 peaks over the start sites of active genes and correlates with transcription rates genome-wide. Additionally, H3K36 was discovered as acetylated (**Figure 1.5A**) and the pattern of H3K36ac localization is similar to that of other sites

of H3 acetylation, including H3K9ac and H3K14ac, although the peak of H3K36me<sub>2</sub>, a well described PTM in yeast, is within the active gene coding region (**Figure 1.5A**) (Morris, Rao et al. 2007). Set2 (KMT3)-dependent methylation of histone H3 at lysine 36 (H3K36) was also shown to promote deacetylation and repression (Youdell, Kizer et al. 2008).

Perhaps one of the better described and major acetylation marks in budding yeast is H3K56, an acetylation mark in the globular domain of H3, which is critical for the recruitment of the nucleosome remodeling factor Snf5 and subsequent transcription (Xu, Zhang et al. 2005). These findings indicated to the chromatin field that histone H3 K56 acetylation enables recruitment of the SWI/SNF nucleosome remodeling complex to regulate gene activity. Moreover, H3K56ac has been shown as a DNA-damage-responsive mark, important for genomic stability in mammals (Yuan, Pu et al. 2009). In addition, acetylation of H3K56 is increased in multiple types of cancer, correlating with increased levels of ASF1A, a key protein in nucleosome assembly (Tjeertes, Miller et al. 2009; Yuan, Pu et al. 2009). A former postdoc in the Allis Lab, Sandra Hake, along with the Don Hunt laboratory at UVA (Charlottesville, VA), performed an in-depth analysis of methylation and acetylation profiles of multiple organisms ranging from *Tetrahymena* to mammals (**Figure 1.5B**). Their work illustrates that higher eukaryotes (multicellular) contain a higher number of methylation or silencing marks, whereas lower eukaryotes possess more acetylation or

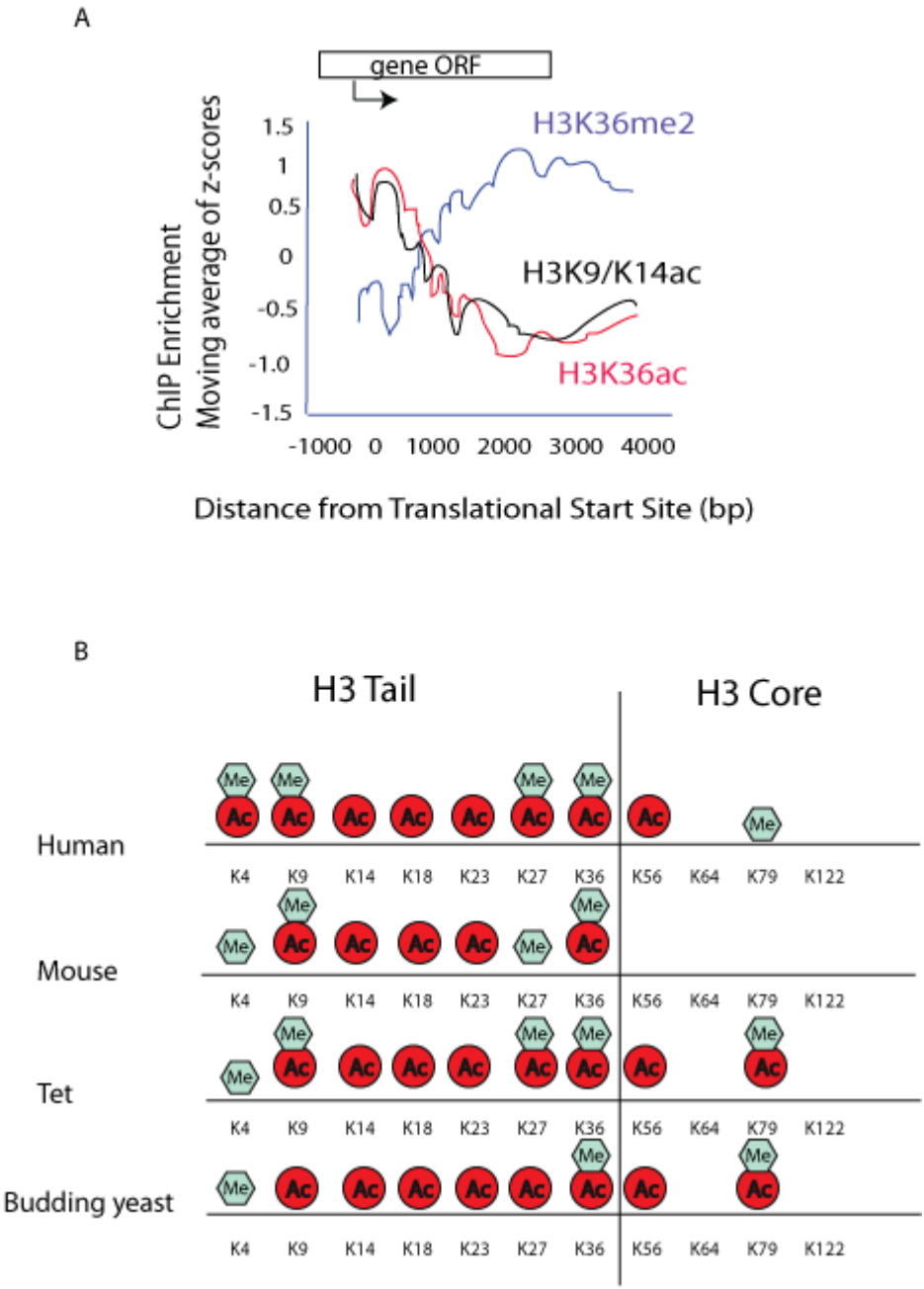
activation marks (Garcia, Hake et al. 2007). Moreover, within the histone globular domain, PTMs were far less conserved from unicellular to multicellular organisms (ie. K56ac), whereas the H3 tail acetylation marks were more conserved between the two groups (**Figure 1.5B**) (Garcia, Hake et al. 2007).

**Figure 1.5: H3 acetylation in yeast and conservation between organisms**

**A.** ChIP study showing that H3K9/K14ac and H3K36ac (black and red curves, respectively) both peak at transcriptional start sites, whereas K36me2 (blue curve) peaks within active coding regions. Image modified and chart lines were superimposed due to low figure quality from (Morris, Rao et al. 2007).

**B.** Histone H3 modifications detected in MS/MS experiments from human, mouse, *Tetrahymena*, and yeast. Acetyl marks are denoted by red circles, and methyl marks are denoted by green hexagons. Note that histone PTMS in tail region are more conserved in multicellular species and PTMs in globular region are more conserved in unicellular species. Image adapted (Garcia, Hake et al. 2007).

Figure 1.5





## Mutagenesis and crosstalk experiments in yeast

Since yeast is easy to genetically manipulate, many labs including ours have used it for mutagenesis and functional studies of post-translational modifications of histones. Groups such as the Michael Grunstein lab have performed tail deletions of H2A and H2B and swapping experiments of the H3 and H4 tails in yeast, which resulted in the disrupted regulation of specific genes as well as silencing of yeast mating type cassettes (Schuster, Han et al. 1986; Ling, Harkness et al. 1996). More recently, the Shilatifard group systematically generated a complete library of the alanine substitutions of all of the residues of the four core histones in *S. cerevisiae* (Nakanishi, Sanderson et al. 2008). Several cases where one mark was required for another on the histone H3 N-terminal tail were identified, including histone H3K14ac (H3K14ac), a mark which is required for normal levels of H3K4 trimethylation (Nakanishi, Sanderson et al. 2008). Additional data on crosstalk, or one mark is required for the presence or absence of another mark of covalent modifications of histones in yeast reveal that phosphorylation of serine 10 in histone H3 is functionally linked in vitro and in vivo to Gcn5-mediated acetylation at H3K14 (Lo, Trievel et al. 2000). These types of observations suggest that transcriptional regulation can occur through multiple linked covalent modifications in histones (Lo, Trievel et al. 2000; Strahl and Allis 2000; Latham and Dent 2007) (**Figure 1.6A**).

The Boeke laboratory recently created a versatile library of 486 systematic histone H3 and H4 substitution and deletion mutants that probes the contribution of each residue to nucleosome function. Their findings suggest that there are surprisingly a few residues essential for cell viability considering the very well conserved sequences of H3 and H4 among different organisms. This observation implies a possibility that multiple residues on histones can function redundantly (Dai, Hyland et al. 2008). Typically, in order to detect phenotypes and analyze how PTMs can cross-regulate one another (as described in the previous section), lysines which are positively charged, acetylatable residues, are mutated to small (uncharged residues such as alanine or glycine), arginines (positively charged, and unacetylatable), or glutamines (negatively charged, acetyl mimics) (**Figure 1.6B**). Single lysine mutations have demonstrated phenotypic defects in yeast, such as H3K36R producing a transcriptional elongation defect (Carrozza, Li et al. 2005), H3K56R producing genome stability defects, chromosomal breaks, and cell lifespan reduction (Recht, Tsubota et al. 2006; Driscoll, Hudson et al. 2007; Dang, Steffen et al. 2009), and H4K16Q significantly reducing cell lifespan (Dang, Steffen et al. 2009). Multiple mutations on identical or different histones have also demonstrated phenotypes, such as H3K56R combined with H4K5,8,12R for replication (Li, Zhou et al. 2008), and H4K5,8,12,16R for cell viability (Megee, Morgan et al. 1990). While larger mutagenesis studies and screens have focused on

designing experiments to abrogate levels of specific PTM marks, only a few studies to my knowledge have been aimed at mutating the surrounding residues of the target lysine to detect a different PTM than was previously known, or to observe an increase or decrease of the PTM level of the target lysine (Bannister, Miska et al. 2000; Nelson, Santos-Rosa et al. 2006). In my own work described in Chapter 3, I have mutagenized the flanking residues of H3K14 to in order to determine whether the mutation of its flanking residues has any effect on H3K14ac levels, or whether mutagenesis can induce an additional PTM such as methylation.

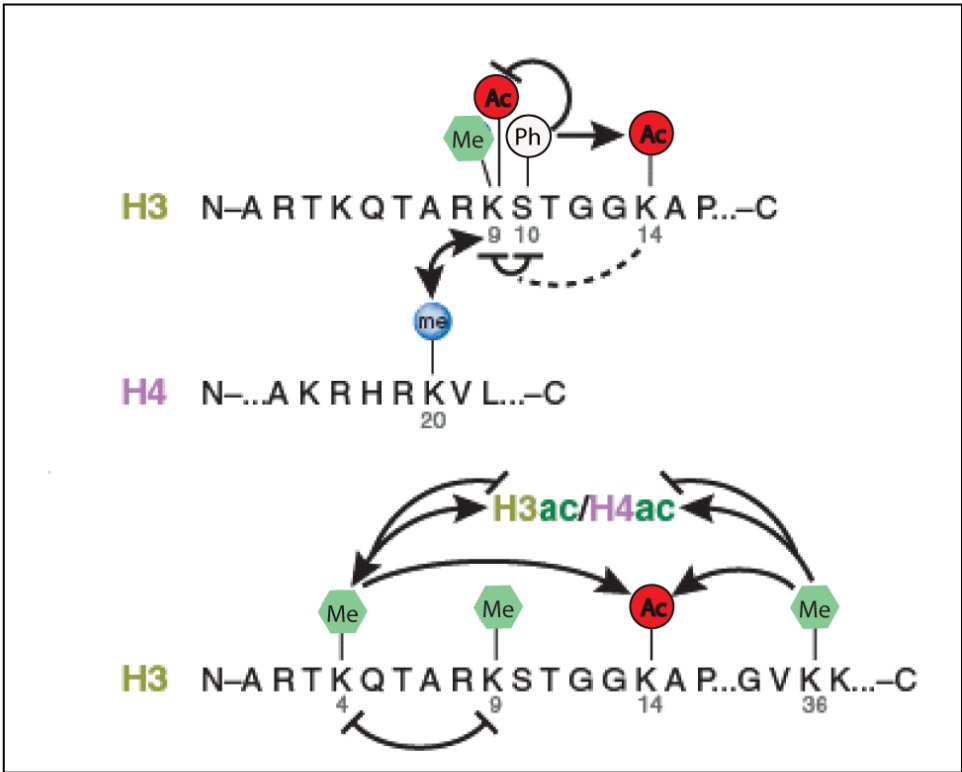
**Figure 1.6: Crosstalk between lysine methylation and acetylation on H3 and H4 histone tails.**

**A.** Crosstalk between H3K9, H3S10, and H3K14 based on previous literature. The H3 modifications also have an effect on H4K20 methylation (top). H3ac/H4ac crosstalk with H3K4me (bottom). H3ac/H4ac refer to multiple acetylation sites on H3 and H4. Red circles represent acetylation marks, green hexagons represent methylation marks, and white circles represent phosphorylation marks. Image adapted from (Latham and Dent 2007)

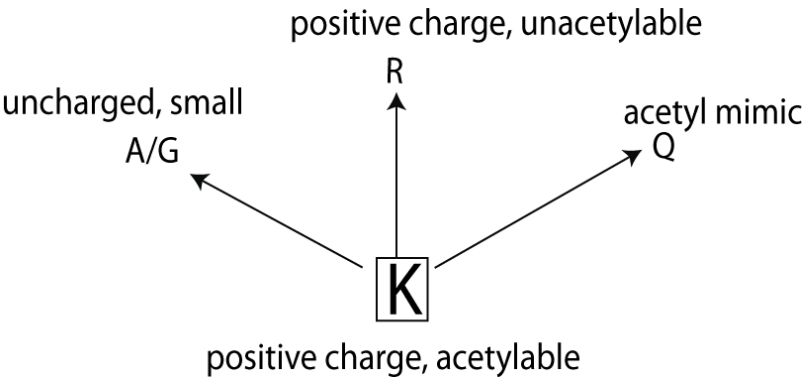
**B.** Lysines (positively charged, acetylatable residues) typically mutated into small uncharged residues (glycine or alanine), charged, unacetylatable residues (arginine), and uncharged, acetyl mimics (glutamines) for phenotypic analysis, PTM detection, or to detect effects on other modifications.

Figure 1.6

A



B



In the previous sections, I discussed protein acetylation, focused largely on histone acetylation “writers”. In the next section, I will next discuss “readers,” proteins that interact specifically with a PTM and the historic challenges and methods used to uncover additional domains within these readers. I will also briefly introduce the domain structure and functional importance of an array of chromatin-associated proteins.

### **Domain Prediction Challenges**

Domain prediction from primary amino acid sequence has historically been a challenging task. However, improved domain prediction programs are enabling high accuracy structure predictions, more directed mutagenesis and binding studies, and a deeper scope to perform evolutionary analyses. Many protein motifs characteristically associated with chromatin have been shown to have affinity for modified histone tails, acting as “effectors” or readers for histone PTMs, notably lysine acetylation and methylation. These motifs have been discovered through experimental cloning and in-vitro peptide binding assays, however since there are multiple modifications on the protein with a multitude of interaction partners, a number of these effector molecules have yet to be discovered. Using domain prediction tools and computational techniques can thus help us achieve a higher number of chromatin binding domains that are still undiscovered. In Chapter 4, I aim to describe the bioinformatic method I used to predict additional effectors in chromatin

associated proteins, and present data on the evolutionary domain analysis of PcG proteins.

## **Domain definitions and statistical methods used to characterize domains**

To understand the roots of the domain prediction challenge, one needs to delve deeply into the definition of a domain. Domains are considered to be the building units of protein structure. A protein can contain a single domain or multiple domains, each one typically associated with a specific function. The architecture of a protein's domain determines the function of the protein, its subcellular localization and the intermolecular interactions it is involved in. However, over the years, domains have been defined in several different ways, each definition focusing on a different aspect of the domain hypothesis (Ingolfsson and Yona 2008):

- A domain is a protein unit that can fold independently
- It forms a specific fold in 3D space
- It performs a specific task/function
- It is a movable unit that was formed early on in the course of evolution

Most of these definitions are widely accepted, but some are more subjective than others. For example, the definition of a cluster in 3D space is dependent on the algorithm and experimental data used to define the

clusters and the parameters of that algorithm. Furthermore, while the first two definitions focus on structural aspects, the other definitions do not necessarily contain a structural constraint. Domain prediction programs such as SMART and PFAM (Schultz, Milpetz et al. 1998; Finn, Tate et al. 2008) have aimed to overcome these challenges by using statistical prediction methods. However because the precise definition of domain is subjective, predicted domains can be missed in these programs largely for a number of reasons: (1) The protein of interest might contain new domains that have not been characterized or studied yet, and therefore the protein might be poorly represented in existing domain databases with limited information about its domain. For most proteins (and especially newly sequenced ones) the structure is unknown, thus structure-focused methods are excluded. (2) Protein should have homologs in order for the prediction method to be effective. Methods may only work using databases of proteins containing homology, but not on individual proteins. (3) Existing domain prediction algorithms can be inconsistent in their domain annotations. Without accurate structural information, it is difficult to validate the performance of prediction algorithms. When the structure is known, determining the constituent domains might not be straightforward (Ingolfsson and Yona 2008).

Statistical prediction methods have tried to solve the multifaceted problem of domain prediction by implementing well-constructed multiple



sequence alignments (MSAs), which are a positive source of information for domain predictions. These MSAs feed into a statistical model, typically a Hidden Markov Model (HMM), which computes the probability of a given sequence by using the consensus MSA to predict additional domains of interest. MSA alignments are much easier and faster to generate than three-dimensional models and can thus be used in large-scale domain predictions. Moreover, MSAs can also be used to detect remote homologies. On the other hand, sequences that are highly similar usually contain little information on domain boundaries and bias predictions as they mask other equally crucial but less represented sequences.

Several atypical or cryptic domains, domains which are dissimilar to the canonical domain by sequence, yet contain some or all the key residues are typically needed for the binding surface are often missed by domain prediction programs. These types of domains are challenging for prediction algorithms since domains often function through intermolecular interactions that depend on chemical and structural properties of the interacting surface that need to be compensated for, and primary amino acid sequence is insufficient. RAG2 (an essential component of the RAG1/2 V(D)J recombinase) is an example of a protein that contains a plant homeodomain (PHD) finger that specifically recognizes histone H3 trimethylated at lysine 4 (H3K4me3) (Matthews, Kuo et al. 2007). Though RAG2 has a conserved tryptophan that constitutes a key structural component of the K4me3-

binding surface and is essential for Rag2's recognition of H3K4me3, the overall similarity of RAG2's PHD finger with the canonical domain is only 25% similar by amino acid sequence and therefore is missed by domain prediction programs. These examples display that with a better training set, and manipulation of key sequence parameters, we can improve reliability and the number of predictions generated on single and multidomain proteins.

While many of the proteins known today contain a single domain, it has been noted that the majority of proteins contain multiple domains. Consistent with the evolutionary aspect of the domain hypothesis, more complex organisms have a higher number of multi-domain proteins (Tordai, Nagy et al. 2005). Multidomain proteins represent a substantial fraction of the proteome: about 27% of proteins in bacteria and 39% of proteins in metazoa are multi-domain proteins (Tordai, Nagy et al. 2005). Multidomain proteins are structurally (and often functionally) more complex than single-domain proteins, which lends itself to additional complexity at the protein signaling level. Multidomain proteins can evolve through gene duplication and domain shuffling (insertion, deletion, rearrangement, and internal duplication of domains), which can be caused by duplication and insertion of a domain into a novel genomic context, tandem duplication of domains, domain deletion, gene fusion and fission (Song, Sedgewick et al. 2007).

## Chromatin associated domains and importance

A great deal of evidence suggests that PTMs function through the recruitment of downstream effector proteins, which in turn perform a task on chromatin (Shi, Hong et al. 2006) (Wysocka, Swigut et al. 2006) (**Figure 1.7A**). A specific role for post-translational modifications on histones (histone marks) is to stabilize the binding of effector proteins, which can influence gene expression by modifying chromatin structure, recruiting the components of the transcription machinery, or establish additional modifications (**Figure 1.7A**).

**Figure 1.7: Effector proteins and cryptic domains**

**A.** Well characterized bromodomain (BD), chromodomain (CD), and PHD finger domains in histones H3 and H4. Shown is the *Drosophila* NURF 301 PHD finger bound to H3K4me3, HP1 chromodomain bound to H3K9me, Polycomb (Pc) bound to H3K27me, and the bromodomain of GCN5 bound to H4K16 in yeast.

**B.** Multiple sequence alignment of ING2 PHD fingers. Aromatic cage residues, zinc coordination residues, and histone H3 arginine 2-interacting residues are colored in purple, red, and orange, respectively. Mouse RAG2 is highlighted (in red box) and displays a cryptic PHD finger due to the sequence dissimilarity with coordinated Zn and cage residues. Image adapted from (Ruthenburg, Allis et al. 2007).

A

37

Wysocka et al. show that a plant homeodomain (PHD) finger of nucleosome remodeling factor (NURF), an ISWI-containing ATP-dependent chromatin-remodeling complex, interacts with H3K4me3. When the PHD finger is deleted in *Xenopus*, a loss-of-function phenotype is observed, and compromises spatial control of *Hox* gene expression (Wysocka, Swigut et al. 2006). Additional PHD finger proteins have more recently been shown to be implicated in disease, such as MLL (Wang, Song et al. 2009), RAG2 (Matthews, Kuo et al. 2007) (**Figure 1.7B**; red box), and Kap1 (Zeng, Yap et al. 2008). Crystal studies of the numerous PHD finger domains coupled to H3 peptides have revealed high conservation within the domain and an aromatic cage that is important for histone binding selectivity for H3K4me3. Moreover, crystal structures of the BPTF PHD finger coupled to the H3 peptide display that the residues in the N-2 and N+2 positions respective to the H3K4 mark, are often critical determinants of binding specificity (Taverna, Li et al. 2007).

Another domain of interest is the bromodomain which recognizes acetylated lysine residues such as those on the N-terminal tails of histones. Bromodomain binding perpetrates a pivotal mechanism for regulating protein-protein interactions in histone-directed chromatin remodeling and gene transcription (**Figure 1.7A**) (Mujtaba, Zeng et al. 2007). Several reports indicate that the overall fold of bromodomains is highly conserved, but the subtle structure of the non-conserved loop region is crucial to their

function (Mujtaba, Zeng et al. 2007). The chromodomain which recognizes methyl-lysines (the most famous being HP1 which binds to H3K9me (**Figure 1.7A**) is part of the Royal Superfamily of protein folds, which also includes the Tudor, Chromo barrel, and MBT domains (Taverna, Li et al. 2007). The HP1 and Polycomb chromodomains, which are similar to the chromodomains of CDY family proteins, distinguish two methylated lysine residues within ARKS motifs in the H3 tail (H3K9me and H3K27me) (**Figure 1.8A**; boxed in red).

While the Polycomb “readers” and “writers” are fairly well understood in *Drosophila* (**Figure 1.8A**), in higher organisms, the specificity of effector interaction, such as chromodomain recognition becomes murky and less understood. In particular, the Chromobox (Cbx) chromodomain-containing proteins underwent a massive expansion in mammals, with a total of five Cbx paralogs compared to one protein in *Drosophila*, known as Pc. It has been shown that not all chromodomains (CDs) within the Cbx family of highly conserved chromodomains within the Polycomb family display affinity towards both histone H3 trimethylated at K9 and H3K27 (Bernstein, Duncan et al. 2006). Some display preferential affinity towards histone H3K9me<sub>3</sub> and some towards H3K27me<sub>3</sub>. While H3K9 and H3K27 are identical in surrounding sequence profile, the distinct functions of these marks is intriguing. It has been suggested that residues immediately preceding the ARK(S/T) motif impact on the specificity of chromodomain interactions

(Fischle, Wang et al. 2003). Amino acids proximal to the substrates, or motifs on the effectors outside of key conserved domains could be contributing to the specificity of these interactions (Fischle, Wang et al. 2003). Thus, in order to detect additional unannotated domains, and motifs in these complex proteins, a bioinformatics sequence based method and phylogenetics approach is useful to gain insight about these set of highly conserved, yet functionally divergent proteins. In Chapter 4, I discuss the domain finding method I performed to detect additional unannotated domains within the PcG family of proteins in multiple organisms in order to infer gene expansion and diversification. In the remainder of this chapter, I briefly introduce PcG complexes and their core functions.

### **PcG complexes and their evolutionary history**

The PcG proteins are structurally and functionally diverse and form large multimeric complexes of two general types: Polycomb repressive complex 1 (PRC1) and PRC2 (Ringrose and Paro 2004) (**Figure 1.8B**). These complexes post-translationally modify histone tails and are believed to cooperate in transcriptional repression of target genes by altering local, higher-order chromatin structure (Francis, Kingston et al. 2004; Sparmann and van Lohuizen 2006).

Polycomb group complexes, which are known to regulate homeotic genes, have now been found to control hundreds of other genes in mammals



and insects. Polycomb complexes function as global enforcers of epigenetically repressed states, balanced by an antagonistic state that is mediated by Trithorax. These epigenetic states must be reprogrammed when cells become committed to differentiation. PcG proteins were originally identified in *Drosophila melanogaster* as factors necessary to maintain cell-fate decisions throughout embryogenesis by repressing *Hox* genes in a body-segment-specific manner (Kennison 1995). Now recognized as a large family of chromatin-associated proteins conserved from plants to humans, the PcG is involved in many cellular memory processes including body patterning X inactivation in female mammals (Heard 2004) and vernalization in plants (Sung and Amasino 2004).

Core PRC1 is composed of Polycomb (Pc), dRing, Posterior sex combs (Psc) and Polyhomeotic (Ph) (**Figure 1.8B**) (Ringrose and Paro 2004). Already annotated, Pc has an N-terminal chromodomain (CD) and a C-terminal Pc box CDs, which are found in many chromatin-associated proteins and are well-characterized methyllysine-binding modules (Eissenberg 2001). Specifically, the CD of *Drosophila* Pc binds most strongly to H3K27me<sub>3</sub>, the modification generated by PRC2 (Fischle, Wang et al. 2003). The Pc box is a 15-amino acid motif necessary for transcriptional repression of target genes and for interaction with dRing, the catalytically active subunit of PRC1 (Muller 1995). dRing, named for its RING-type zinc finger (**Figure 1.8B**), is an E3 ubiquitin ligase that monoubiquitylates histone H2A at lysine 119

(H2AK119ub) (Wang, Wang et al. 2004). This modification, along with H3K27me3, is important for PcG-mediated gene repression (de Napoles, Mermoud et al. 2004; Wang, Wang et al. 2004). The precise function of Ph in PRC1 complexes remains to be characterized, but it has been speculated that Ph might influence the spreading of PcG complexes (Kim, Gingery et al. 2002). Additional new plant studies display significant evidence of sequence similarity between the C-terminal region of the PRC1 Ring finger proteins and the ubiquitin (Ubq)-like family of proteins, thus defining a new Ubq-like domain, the RAWUL domain. Analysis of the conserved domain architecture among PRC1 Ring finger proteins revealed the existence of long sought PRC1 protein orthologs in these organisms, suggesting the functional conservation of PRC1 throughout higher eukaryotes (Sanchez-Pulido, Devos et al. 2008). Additionally, these proteins have multiple domains and not all have been annotated in the protein databases. Thus, being able to detect these domains through bioinformatic methods can potentially help us identify the distinct roles of PcG proteins and their cooperative mechanisms. The Cavalli lab bioinformatically performed a phylogenetics analysis of PcG proteins on a broad spectrum of eukaryotes, and Hox gene clusters were mapped onto the species. Phenotypes of PcG mutants and the strong binding of PRC1 to Hox gene clusters in flies and vertebrates suggested that these clusters are important PRC1 targets. Thus, one hypothesis might be that PRC1 genes can be lost as a consequence of the disintegration of the Hox gene cluster,

which occurred repeatedly during evolution (Schuettengruber, Chourrout et al. 2007). In Chapter 4, I will describe my aim to address the evolutionary expansion and diversification of the PcG family of proteins.

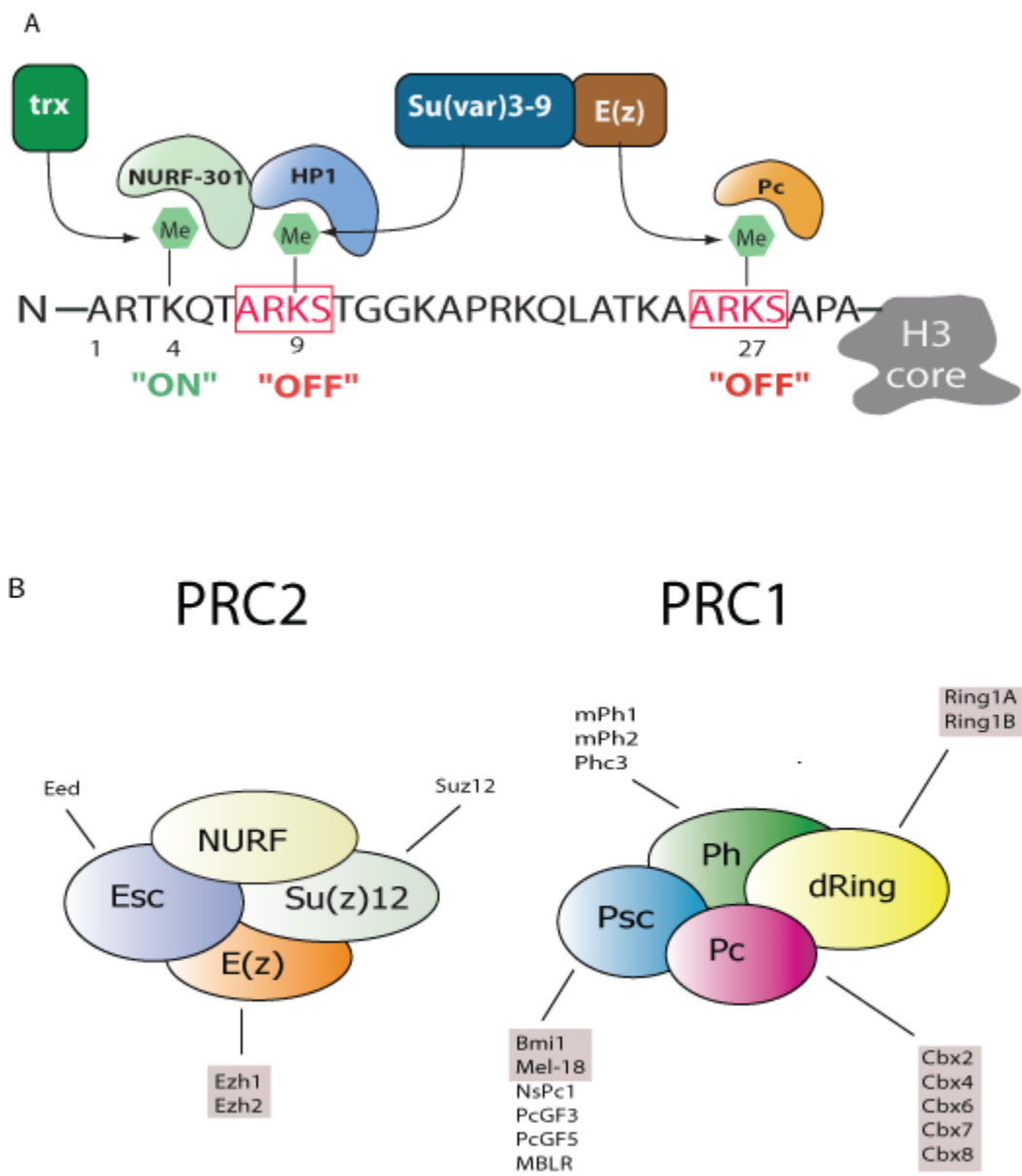
**Figure 1.8: Polycomb schematic representations.**

**A.** Schematic representation of the H3 N-terminal tail. Shown are *Drosophila* proteins with rectangles representing the “writer” enzymes, and half moons representing effector “reader” molecules. The “ARKS” motif is boxed in red. Note although the ARKS sequence motif is identical in the two H3 positions (9 and 27), a different set of readers and writers exist for these two lysines. Note the gene “ON” and “OFF” mechanisms associated with the specific methylation marks.

**B.** Schematic representation of known core members of PRC1 and PRC2 complexes.

*Drosophila* proteins are shown as colored ovals; mouse homologs of these proteins are listed adjacently. Gray boxes denote the mammalian homologs discussed in detail throughout the text.

Figure 1.8



## Chapter 2: Computational Prediction of Acetylation Substrates

### Summary

Acetylation is a well-studied post-translational modification that has been associated with a broad spectrum of biological processes, notably gene regulation. Many studies have contributed to knowledge of the enzymology underlying acetylation, including efforts to understand the molecular mechanism of substrate recognition by several acetyltransferases, but traditional experiments to determine intrinsic features of substrate site specificity have proven challenging. In this project, in collaboration with Dr. Eran Segal at The Weizmann Institute of Science (Rehovot, Israel), I performed a clustering analysis of protein sequences to predict protein acetylation based on the sequence characteristics of acetylated lysines within histones. I utilized the local amino acid sequence composition that represents potential acetylation sites by implementing a clustering analysis of histone and nonhistone sequences. I demonstrated that this sequence composition has predictive power on two independent experimental datasets of acetylation marks. Finally, I detected acetylation for selected putative substrates using mass spectrometry as described Chapter 3, and report several novel nonhistone acetylated substrates in budding yeast. My approach, combined with more traditional experimental methods, may be useful for identifying acetylated substrates proteome-wide. In this chapter,

I adapted text from the publication (Basu, Rose et al. 2009).

## **Results**

### **Training, key assumptions, and method**

I used histones as a training set because of the wealth of information known about their PTM patterns and well-developed purification and analytical detection methods and focused on the major human core histones bearing a total of 56 lysines (H2A: 13, H2B: 19, H3: 13, H4: 11) (**Figure 2.1A**). To date, MS and antibody data suggest that there are 23 “validated” acetylated lysines and 33 lysines that have “not yet been observed as acetylated” in human histones based on literature (see **Appendix** for table). I sought to uncover additional acetylation sites within the “not observed” class of lysines in a systematic, rigorous manner via my computational method. I selected parameters that could influence my ability to predict acetylation sites on histones by making a series of assumptions. First, I focused my attention on short stretches of amino acids N- and C-terminal of all 56 lysines. Since structural studies of published KAT domains coupled with peptide substrates typically do not exceed 14-20 amino acids in length (Marmorstein 2001), a sliding window of a maximum number of 12 residues flanking each lysine was chosen (**Figure 2.1B**). Residues most proximal to the lysine were given the highest weight (**Figure 2.1B**), assuming that these residues are most important for enzyme recognition, as several studies have shown

(Marmorstein 2001; Marmorstein and Roth 2001). The weight function and additional details on how I weighted residues is described in the **Appendix**. Second, I varied standard Blast sequence alignment parameters including gap penalty, extension, insertion, and deletion scores (**Figure 2.1C**). For lysines in the extreme N- and C-terminal region, such as H3K4 or H2AK129, I normalized the raw alignment score based on the length of the sequence. Additionally, both orientations of the protein sequence (N-terminal to C-terminal or vice versa) were weighted equally. For sequences with lysines located in close proximity to each other, such as H3K36 and H3K37, I restricted the alignment matrix so that these sequences did not receive an alignment score. This restriction prevented my training set to be overrepresented with sequences from overlapping fragments of the same protein. Finally, I compensated for structural accessibility by penalizing buried lysines, while improving the score of accessible lysines (Luger, Mader et al. 1997). This, however, did not influence my ability to predict acetylation sites on histones, and therefore was not included in my further computations.



**Figure 2.1: Schematic of computational approach**

**A.** Human core histone proteins (H2A: orange, H2B: red, H3: blue, H4: green) containing 56 lysines (black) were taken as input data for computational training.

**B.** A sliding window of amino acids (black bars) flanking the input lysine (at position 0) is used to train the model. Not all window lengths are shown. Weights (calculated as inversely proportional to distance [d]) are applied to amino acids based on the distance from the input lysine to the amino acid in positions -12 to +12.

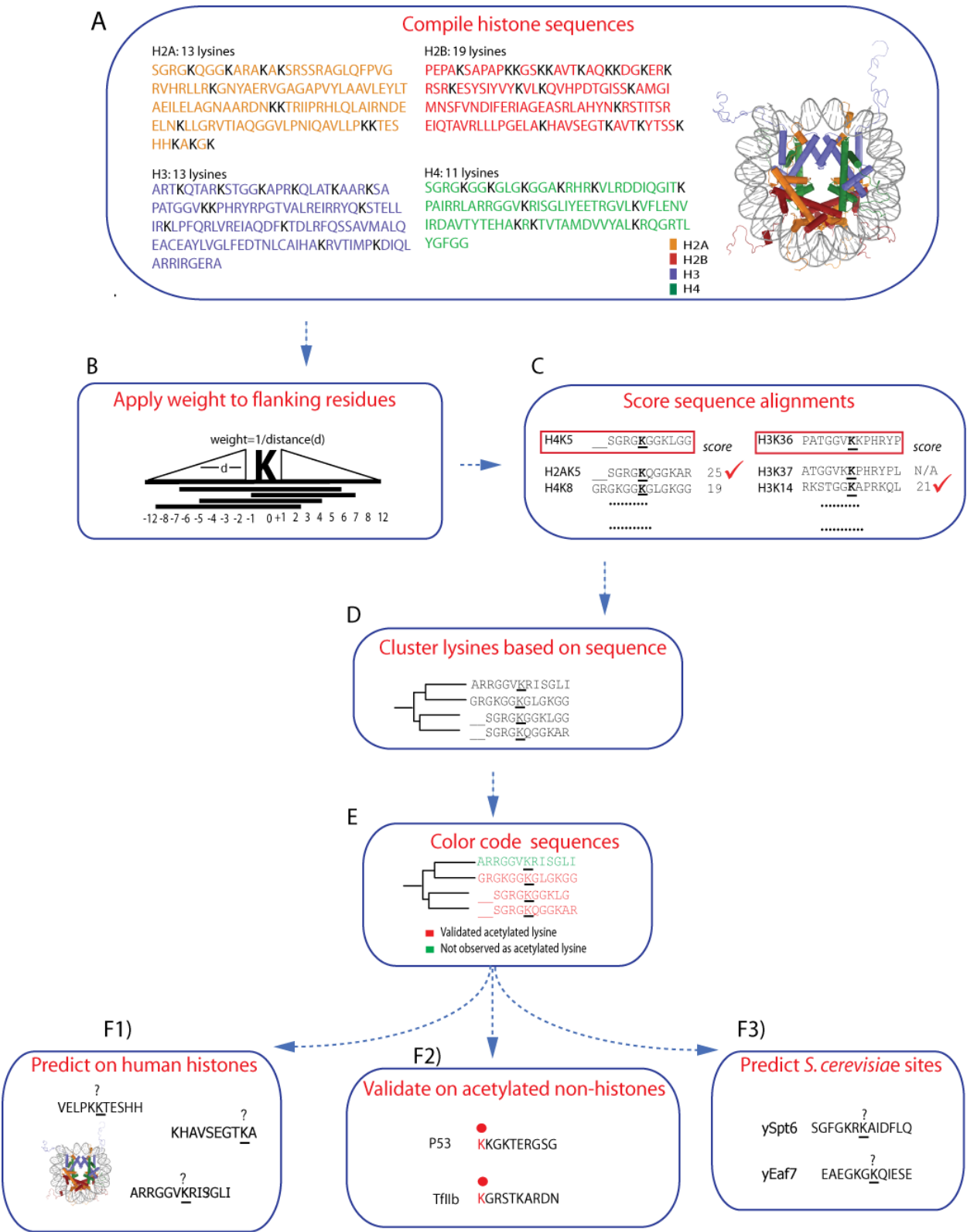
**C.** BLAST sequence alignments are performed between all 56 lysines and surrounding sequences and the highest scoring alignment is selected to begin the clustering analysis. Shown are sequences H4K5 and H3K36 (boxed in red) spanning positions -6 to +6 and their highest scoring match (denoted by a checkmark). Note that H4K5 and H2AK5 do not have six residues flanking the lysine N-terminally; scores are normalized based on length in these cases.

**D.** Lysines clustered together based on sequence alignment scores creating a fully predictive hierarchical tree (four sequences are shown here; all 56 sequences are shown in Fig. 2.2).

**E.** Sequences are color coded according to published data on their modification state. Red: “validated” evidence of the lysine being acetylated, green: this lysine was “not observed” as being acetylated in literature.

**F.** After establishing PredMod, predictions were made on lysines in human core histones. The algorithm was then validated using a set of human acetylated proteins reported in literature, substrates detected using a pan-acetyl IP approach, and a yeast proteome-wide dataset. Finally, predictions were made on yeast nonhistone sites and validated *in vivo*.

Figure 2.1



I performed a hierarchical clustering of core histone lysines based on the sequences surrounding each of these given lysines. All 56 histone core sequences were aligned to each other creating a matrix of pairwise alignment scores; generating a hierarchical tree of histone sequences (**Figure 2.1D**). I next classified each lysine into one of two categories based on its acetylation status reported in literature: “validated” (23 lysines), or “not observed” (33 lysines) (**see Appendix**). Finally, I visually categorized each of the 56 lysines by color-coding my tree based on the acetylation status of each lysine (**Figure 2.1E**).

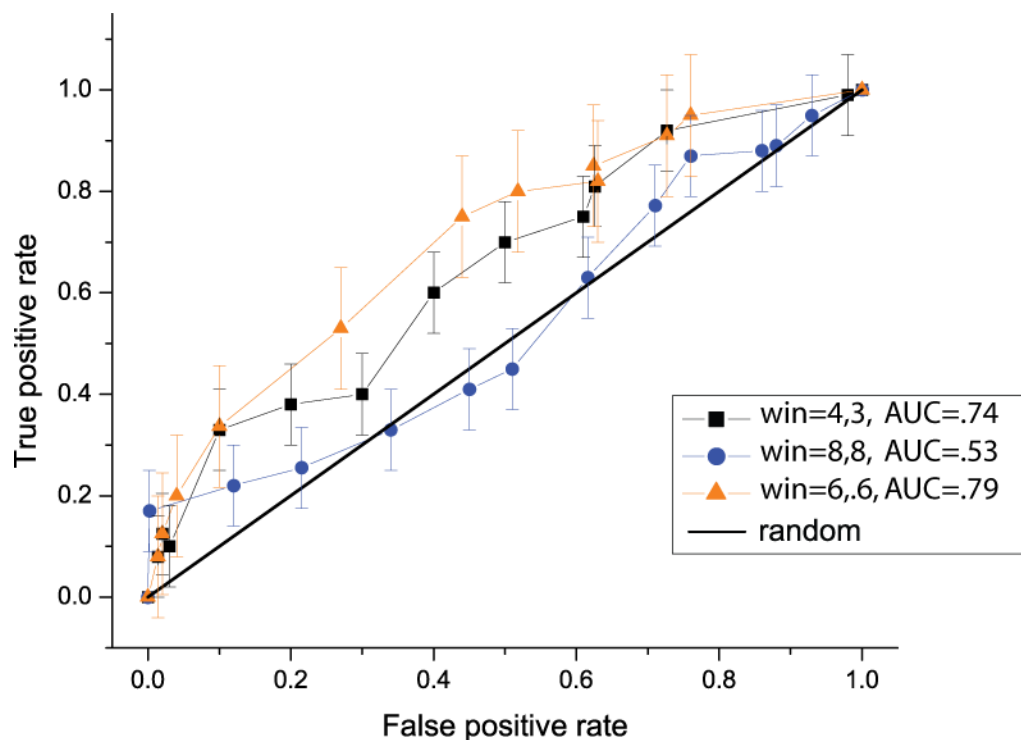
To assess how robust the clustering was and how well it could actually predict lysine acetylation (**Figure 2.1F1**), I took all 56 lysines and performed a Leave One out cross-Validation (LOV) (Cooper, Aliferis et al. 1997) by iteratively excluding one lysine from my training set. Next, I reconstructed the hierarchical tree with the remaining 55 lysines, and incorporated the excluded single lysine observation as test data. For each set and combination of predefined parameters (stated above) and in a single run, I performed a LOV analysis to examine the predictive power on all 56 lysines to discover which set of parameters best optimized classification power. If two lysines were in overlapping fragments of the same protein, I excluded both of these lysines from my training set when either lysine was a test case. I took each test lysine (56) and traversed through my training

tree to find which subgroup of sequences my target sequence formed the tightest cluster with.

A Receiving Operating Curve (ROC) analysis was performed on my test dataset (**Figure 2.2**), where the statistics measure used was the Area Under Curve (AUC). An AUC of 1 represents a perfect prediction and an AUC of 0.5 random predictions. Each point on a single curve of the ROC plot was calculated by measuring the false positive versus true positive rate of the performance on all 56 lysines for a given parameter(s) under a cutoff alignment score. If the “test lysine” clustered within a group of “validated” acetylated lysines (**Figure 2.3**; red color) above the cutoff score, the lysine was predicted to be acetylated. Conversely, if the test lysine clustered within a group of “not observed” lysines (**Figure 2.3**; green color) above the alignment score, the lysine was predicted as not acetylated. The default status of the lysine when it did not fall into the above criteria was “not acetylated.” The best ROC plot achieved an AUC of 0.80, and the parameters in this case included six weighted residues to both the left and right of the tested lysine (**Figure 2.2**). A threshold for prediction was also determined based on this plot. To test the significance of this score, I applied the above procedure to 1000 random permutations of the labels of the observed and not observed lysines. The median AUC in these permutations was 0.64 and the maximum score was 0.79, and thus, my AUC was statistically significant ( $p < 0.001$ ).

## Computational prediction of novel human histone acetylation marks and *in vivo* validation by mass spectrometry

After hierarchical clustering of all the lysine-embedded histone sequences, I next sought to predict novel acetylation sites in the human core histones. As the tree illustrates in **Fig. 2.3**, “not observed” lysines that clustered tightly with “validated” acetylated lysines (black arrows) were potential acetylation targets because of their similar sequence constitution. Based on the threshold, as determined by the ROC plot, I selected these as candidate sites. The method described above predicted seven novel acetylation sites in the human core histones; four in H2A (K9, K13, K125, K127), one in H2B (K116), one in H4 (K44), and one in H3 (K37) (**Figure 2.3**; black arrows, for enlarged view of tree, please see **Appendix**). This large number of predictive sites was unexpected, since histones have been intensely investigated for PTMs in recent years. I describe all details of my *in vivo* validation in Chapter 3. In summary, I correctly predicted four of these seven acetyl-lysine sites suggesting that the algorithm is capable to identify acetylation sites in human histone proteins. I also selected a small number of lysines that I predicted with high confidence as “not acetylated”. These lysines were H3K64 and H3K115, and preliminary results display that these lysines to my knowledge have not been observed as acetylated (personal communication, K. Rose).



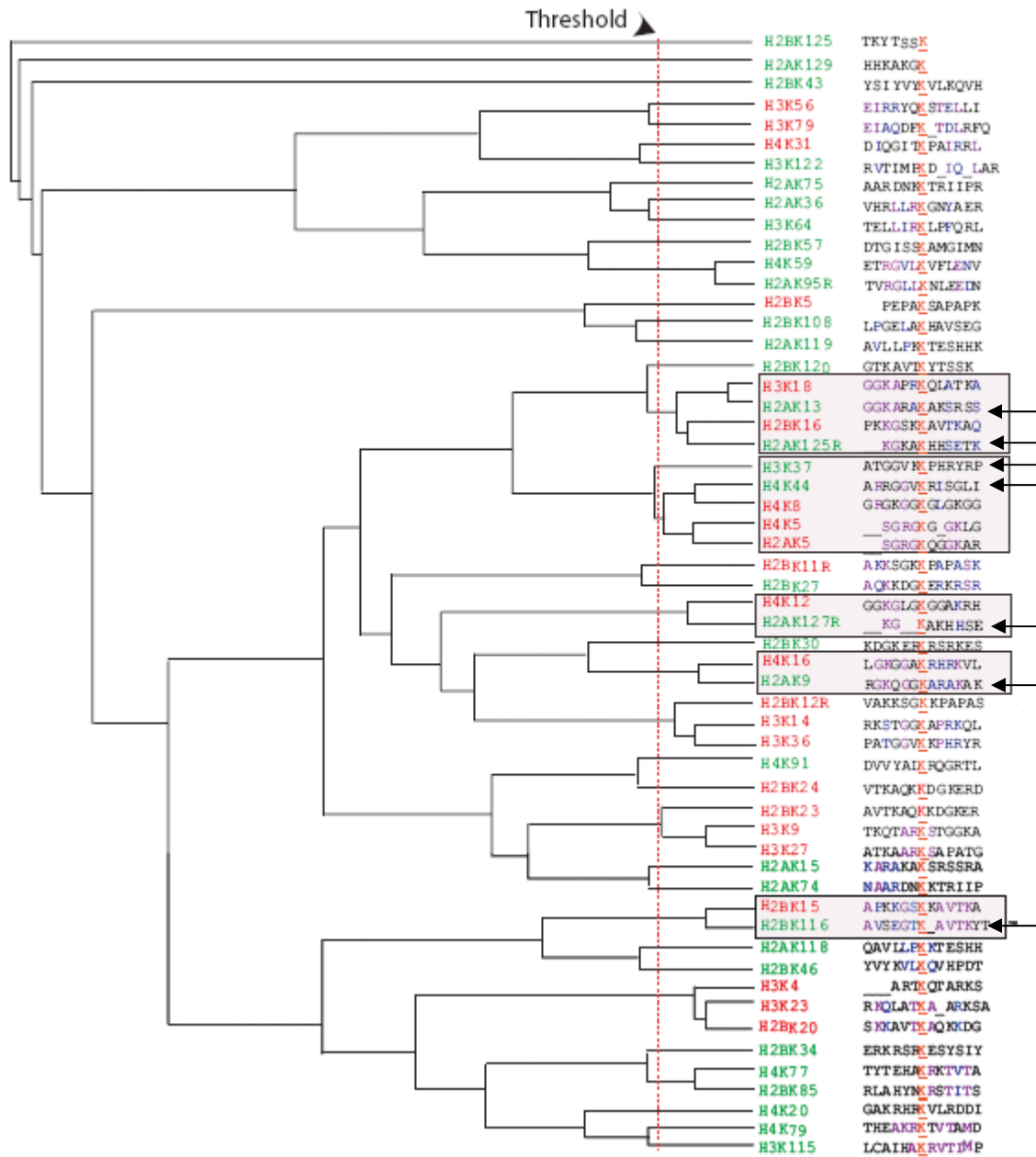
**Figure 2.2: Computational prediction of human histone acetylation sites**

ROC curve using LOV on the 56 human histone lysines for selected parameters. True positive rate (y axis) versus false positive rate (x axis). Win=(x,y) denotes the length of residues spanning the lysine; x represents the number of residues N-terminal to the lysine and y represents the number of residues C-terminal to the lysine. Win=4,3, AUC=0.74; Win=6,6, AUC=0.79; Win=8,8, AUC=0.53. Diagonal line represents a random prediction.

**Figure 2.3: Predictive hierarchical tree of all 56 lysines in human core histones**

Predictive tree of all 56 lysines from human core histone sequences using hierarchical clustering. Histone lysines (in red or green) are color coded according to published data on their modification state as described in **Fig. 2.1E**. For each pair of sequences under a single node, amino acids are colored in light purple (identical residues) or dark blue (in accordance with the BLOSUM matrix). Underlined red lysines represent the residue that was used for training the algorithm. Dashed red vertical line represents the selected threshold used to make predictions. Grey boxes represent lysines that cluster together. Black arrows represent those lysines predicted as acetylated. For an enlarged view of tree, please see **Appendix**. An "R" next to the lysine indicates that a C- to N-terminal arrangement was used in the alignment.

Figure 2.3





## Nonhistone sequence based dataset prediction and validation

Since my computational analysis revealed a high level of sequence homogeneity among acetylated lysines within histone proteins, leading to the successful prediction of novel modified residues, I next wondered if my approach might also enable us to predict nonhistone acetylation sites as well (**Figure 2.1F2**). In my first approach, I included a dataset that contained both nuclear and cytosolic proteins from HeLa cells, which were immunoprecipitated with a pan-acetyl antibody (**Figure 2.4A**) and identified by MS (Kim, Sprung et al. 2006) (see **Appendix** for full list). The precipitate contained peptides with a total of 1413 lysines, and 51 previously validated acetylation sites. With PredMod, I was able to predict 34 (67%) of these sites correctly (**Figure 2.4C**) when they were surrounded by six residues to the left and right (AUC= 0.75, sensitivity ( $S_n$ )=0.60, specificity ( $S_p$ )=0.91) (**Figure 2.4C**; orange curve). In total, 6% (85) of the total number of lysines were predicted that were not validated as acetylated ( $F_p < 6\%$ ).  $F_p$  is a maximum false positive rate since a true negative count cannot be accurately determined because many of these lysines could potentially be acetylated, but not detected under the experimental procedures used.

In my second dataset, I compiled a list of 32 proteins containing 1378 lysines with 73 of these reported in literature to be acetylated *in vivo* and/or

*in vitro* (**Figure 2.4B**, see **Appendix**). With PredMod, I predicted 39 out of 73 (53%) lysine marks accurately with  $F_p < 6.5\%$  ( $AUC = 0.74$ ,  $S_n = 0.58$ ,  $S_p = 0.91$ ) when these were surrounded by six residues to the left and right (**Figure 2.4D**; orange).

Both test datasets exhibited a decrease in performance when larger numbers of residues N- and C-terminal to the target lysine were used (blue line), suggesting that KATs may recognize a smaller and defined set of residues. Overall, my results from both approaches revealed that my selected parameters for histones were also valid for the prediction of acetylated nonhistone substrates using a ROC analysis approach.

**Figure 2.4: Prediction performance on test set of human acetylated substrates.**

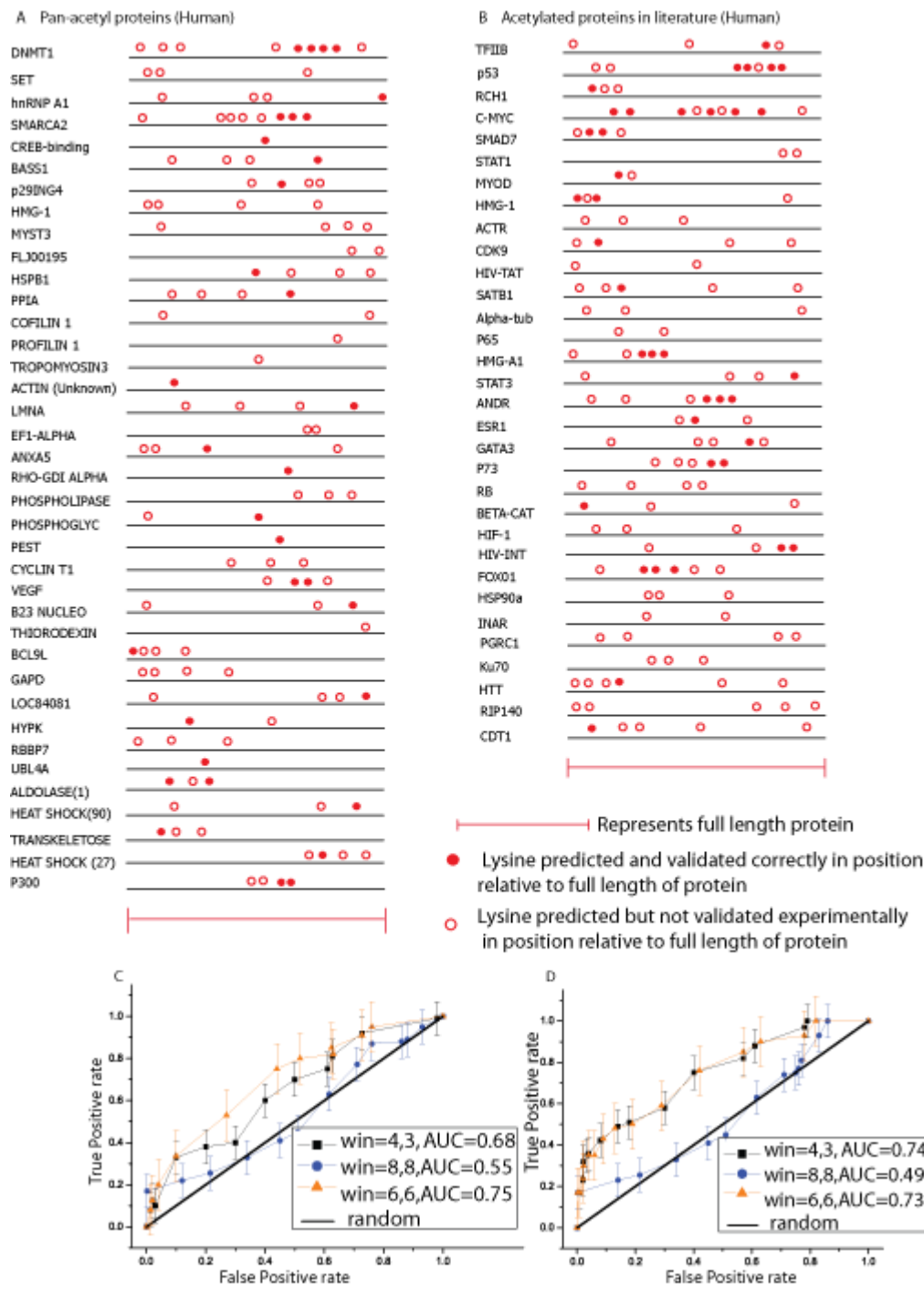
**A.** Pan-acetyl immunoprecipitated substrates. Black line represents the full-length protein. Red filled circles indicate lysines that are predicted and correctly validated in their positions relative to the full-length of the protein. Red empty circles indicate lysines predicted, but not validated under the experimental conditions tested in their positions relative to the full-length of the protein.

**B.** Literature validated human acetylated proteins. Symbols are the same as in A.

**C.** ROC curve for human pan-acetyl IP substrate test set. Y axis represents the true positive rate and X axis the false positive rate. Win=(x,y) denotes the length of residues spanning the lysine; x: number of residues N-terminal to the lysine; y: number of residues C-terminal to the lysine. Diagonal line represents a random prediction.

**D.** ROC curve for human literature-validated test set. Symbols are the same as in C.

Figure 2.4



## Analysis of acetylation motifs

I next sought to understand which amino acids play a critical role in acetylation site selection and asked whether there were preferences for certain amino acids near the target acetylated lysines in my datasets. Notably, when I examined the surrounding residues (six residues to the left and right) of a “validated” acetylated lysine versus a “not observed” one in human histone and nonhistone proteins I discovered an enrichment for small residues (G/A in pink), lysines (K in green), and phosphorylatable residues (S/T in blue) (**Figure 2.5**). To test whether the observed enrichment of G, K, and S was statistically significant, I determined the frequency of these residues flanking a lysine in the entire human proteome. I noticed that on average, these residues were of significantly higher frequency in my datasets than in the human proteome. I employed the hypergeometric test to measure the statistical relevance of this observation. For a table with all the p-values of these statistical observations (see **Appendix** for full list of values).

My results display that the most significant p-values were found in the category of small residues ( $p < 0.01$  in multiple flanking positions, **Figure 2.5**, tick marks), suggesting that small amino acids, perhaps due to their sterically undemanding side chains, could accommodate the flexibility of the substrate thus allowing protein docking and catalysis. This observation was

in agreement with a previous study (Schwartz, Chou et al. 2009), which revealed that glycine preceding lysine was common among acetylated lysines. In conclusion, I was able to identify a significant enrichment of mainly small amino acids and lysines surrounding “validated” acetylated lysines in comparison to “not observed” ones, suggesting that KAT enzymes have a general need for specific residues for recognition and/or activity. These observations are in agreement with studies of several KATs with test substrates (Marmorstein 2001).

**Figure 2.5: Frequency distribution of amino acids surrounding lysines in human histone and nonhistone proteins.**

**A.** Frequency of amino acids (Y-axis) spanning positions -6 to +6 (X-axis) in validated acetylated lysines in histone proteins (23 lysines). Residues in green: basic, red: hydrophobic, pink: small, blue: S/T, black: all other residues. Underlined red K: lysine that has been validated experimentally as acetylated, and an underlined green K: lysine that has not been experimentally observed as acetylated. "X" denotes that no amino was present in that position.

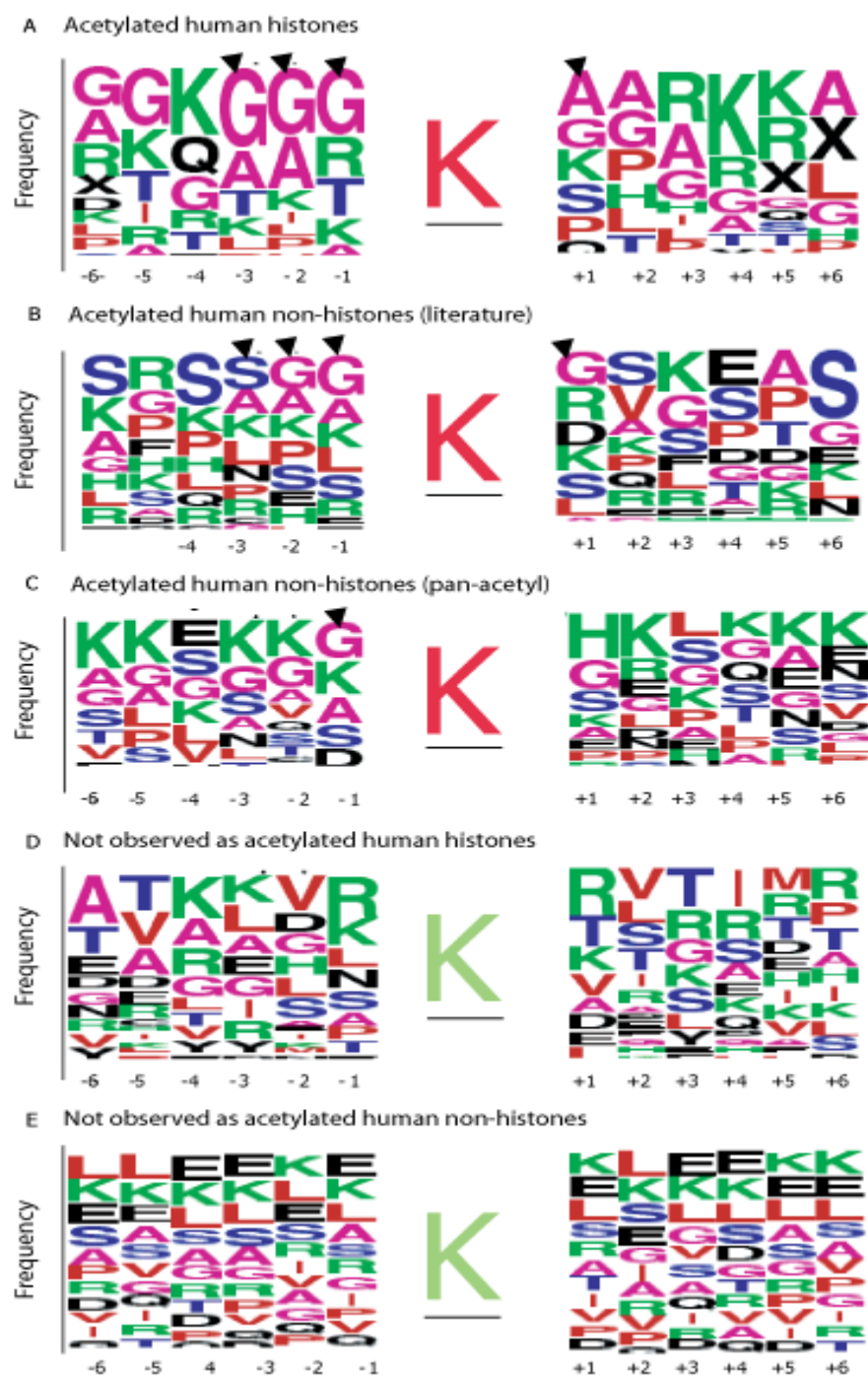
**B.** Frequency of amino acids (Y-axis) spanning positions -6 to +6 (X-axis) in validated acetylated lysines within proteins in literature (73 lysines). Colors represented same as in A.

**C.** Frequency of amino acids (Y-axis) spanning positions -6 to +6 (X-axis) in validated lysines in the pan-acetyl IP substrates (51 lysines). Colors represented same as in A.

**D.** Frequency of amino acids (Y-axis) spanning positions -6 to +6 (X-axis) in "not observed as acetylated" lysines in histones (33 lysines). Colors represented same as in A.

**E.** Frequency of amino acids (Y-axis) spanning positions -6 to +6 (X-axis) not observed as acetylated lysines in proteins as reported in literature and not observed as acetylated lysines in pan-acetyl IP substrates (3493 lysines). Colors represented same as in A.

Figure 2.5





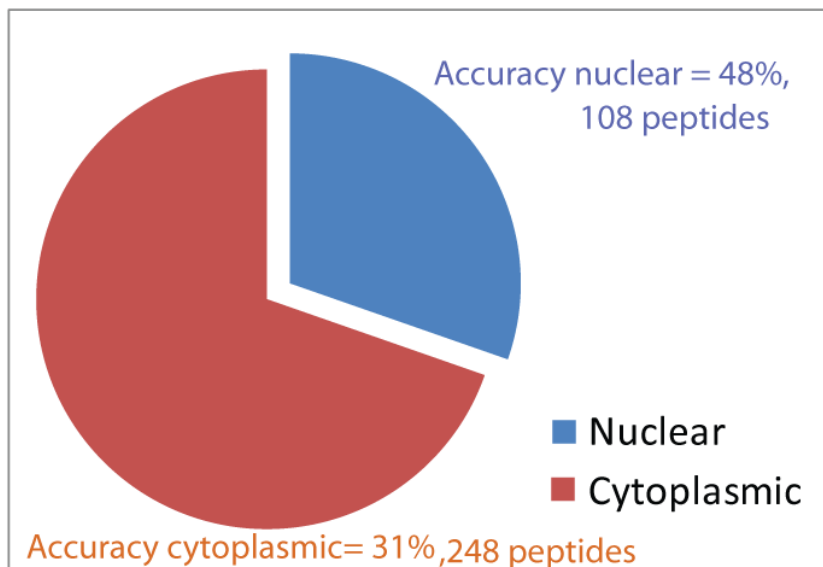
## ***S.cerevisiae* proteome-wide prediction**

The previous predictions were performed with human proteins, and I therefore wondered whether my algorithm would also be able to predict acetylation sites in proteins from other organisms (**Figure 2.1F3**). Since histone acetylation has been studied extensively in budding yeast, I assessed the performance of my model on a proteome-wide dataset that included acetylated peptides in *S. cerevisiae* (Craig, Cortens et al. 2006) (see **Appendix** for details). In addition, I experimentally validated my predicted acetylation sites in candidate yeast nonhistone proteins *in vivo*.

In my first approach, I examined an *in vitro* proteome-wide dataset of acetylated peptides of *S. cerevisiae* that contains 356 peptides including also acetylated histone peptides. This dataset allowed me to approximate the number of yeast acetylation events on a global level (0.6%; 356 acetylated peptides out of approx. 50,000 total non-redundant peptides), and the substrates themselves allowed me to further validate my prediction algorithm. I filtered these protein-derived peptides according to their cellular compartment (nuclear vs. cytoplasmic) (Huh, Falvo et al. 2003), and correctly predicted 48% of acetylation events on nuclear proteins (79 lysines total, AUC= 0.71,  $S_n=0.63$ ,  $S_p=0.92$ , Fp < 4%), and 31% on the cytoplasmic proteins (248 lysines total, AUC= 0.70,  $S_n=0.61$ ,  $S_p=0.90$ , Fp < 5%) (**Figure 2.6A, B**). I also noted that nuclear yeast proteins showed a

similar enrichment for small residues surrounding the target lysine, as found in the human substrates (**Figure 2.7**; tick marks). Given the mass resolution of many of the spectra used to create the yeast library, it was not possible to distinguish a priori between acetyl-lysine and trimethyl-lysine on the basis of the tandem spectra alone. Thus my yeast proteome-wide dataset could potentially contain tri-methylated peptides. Since there are few reported trimethyl marks across the yeast proteome, and my dataset does not contain the previously validated yeast histone trimethylated sites H3K36, H3K79, and H3K4 (Garcia, Hake et al. 2007), I believe that the number of trimethyl sites in the dataset could be small. In accordance, my dataset contains the following peptides: H3K9 (Suka, Suka et al. 2001), H3K14 (Suka, Suka et al. 2001), H3K18 (Suka, Suka et al. 2001), H3K27 (Suka, Suka et al. 2001), H3K56 (Xu, Zhang et al. 2005), H2BK11 (Suka, Suka et al. 2001), H2BK16 (Suka, Suka et al. 2001), H4K5 (Suka, Suka et al. 2001), H4K8 (Suka, Suka et al. 2001), H4K12 (Suka, Suka et al. 2001), and H4K16 (Suka, Suka et al. 2001); all yeast acetyl marks validated in literature. Of note, my current prediction accuracy could be limited by an underrepresentation of acetyl-lysines, and it would be interesting to see how my accuracy improves with datasets obtained using more sensitive MS techniques.

A



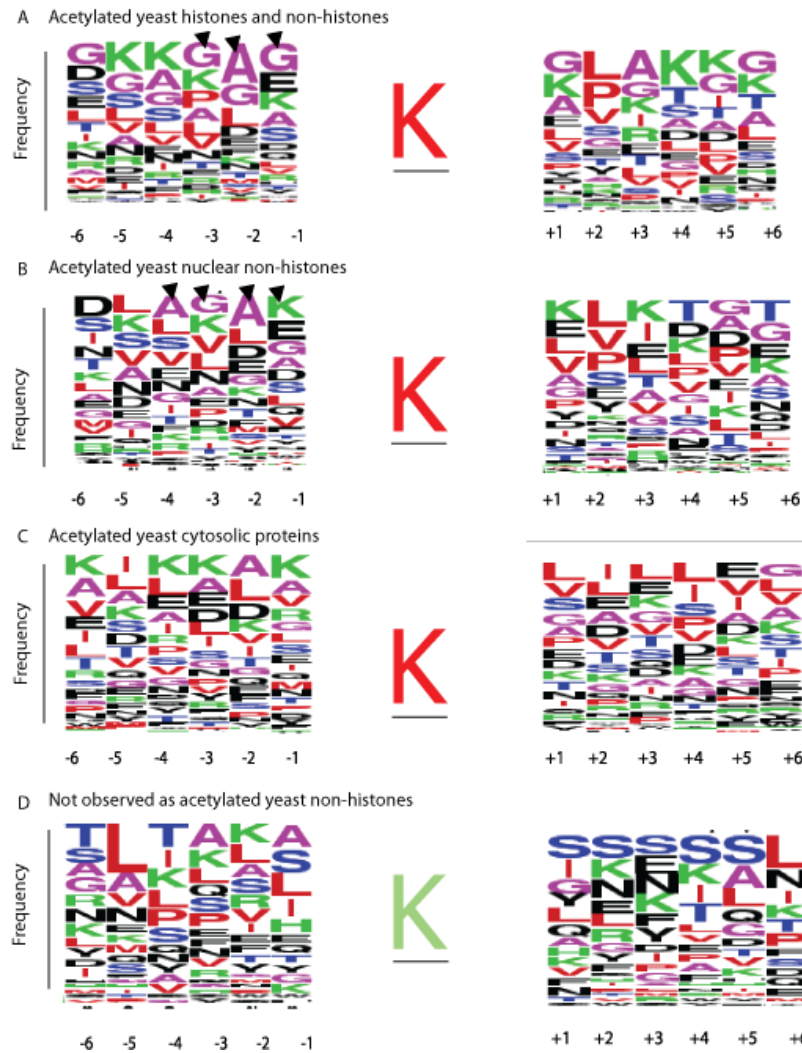
B

	Prediction Nuclear	Prediction Cytosolic
Training set		
Histone	48%   4%	31%   6%
Nuclear	45%   5%	28%   6%
Cytosolic	22%   7%	36%   6%

**Figure 2.6: Performance on Yeast Proteome-wide Dataset**

**A.** Pie chart reflecting correctly predicted lysines on nuclear versus cytosolic peptides in *S. cerevisiae*. Peptides were generated from Global Proteome Machine database. Nuclear peptides (108 peptides) and cytosolic peptides (248 peptides) shown.

**B.** Performance accuracy (#correctly predicted as acetylated/total number of positives) and false positive rate on nuclear peptides versus cytosolic peptides. Individual training sets (histones, nuclear, and cytosolic proteins) used to perform accuracy. The first number in table cell reflects accuracy, and second number after the | reflects the percentage of lysines predicted as acetylated in the test dataset (Fp).



**Figure 2.7: Frequency distribution of amino acids surrounding yeast lysines**

**A.** Frequency of amino acids (Y-axis) spanning positions -6 to +6 (X-axis) in validated acetylated lysines in yeast nuclear histone and nonhistone proteins (108 lysines). Residues in green: basic, red: hydrophobic, pink: small, blue: S/T, black: all other residues.

Underlined red K: lysine that has been validated experimentally as acetylated, and an underlined green K: lysine that has not been experimentally observed as acetylated. Tick marks represent residues described in text.

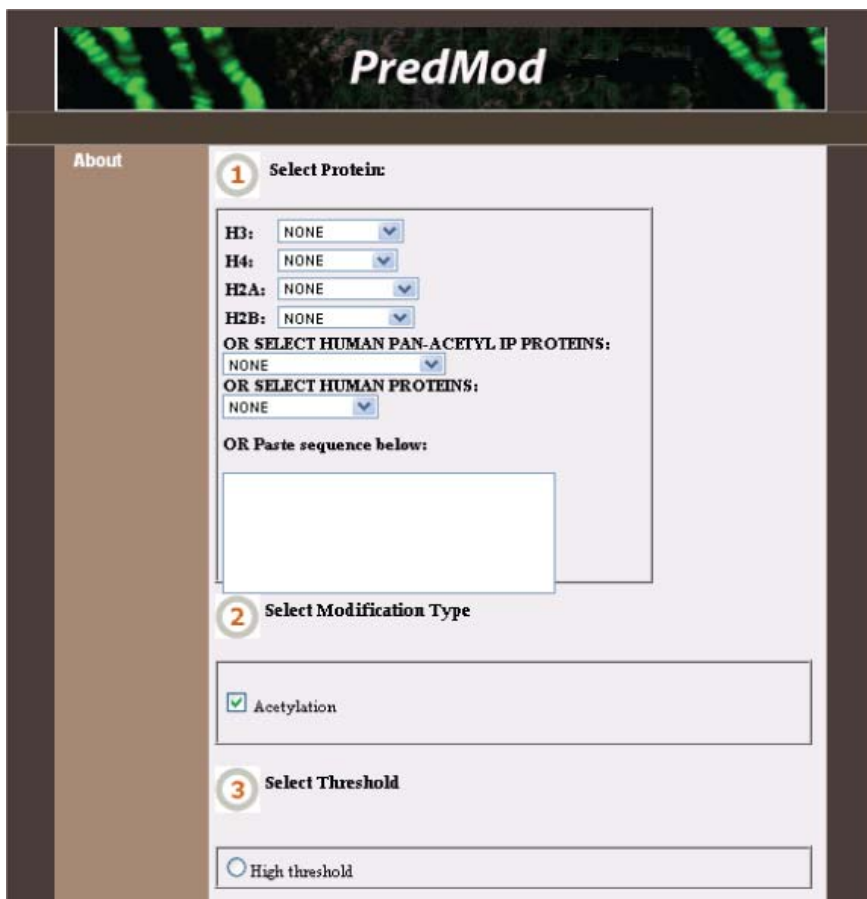
**B.** Frequency of amino acids in validated acetylated lysines in yeast nuclear proteins excluding histones (79 lysines). Colors are as in A.

**C.** Frequency of amino acids in validated acetylated lysines in yeast cytosolic proteins (248 lysines). Colors are as in A.

**D.** Frequency of amino acids in not observed as acetylated lysines in yeast nuclear and cytosolic nonhistones (13814 lysines). Colors are as in A.

In my second approach, I validated my predictions on three yeast candidate proteins that had previously not been published to contain acetylated sites (Spt6 (Clark-Adams and Winston 1987), Sir3 (Gasser and Cockell 2001), and Eaf7 (Krogan, Baetz et al. 2004). With PredMod, I predicted 15 sites to be acetylated out of 416 total lysines in my three candidate proteins combined. Four of these, within my top six ranked predicted sites were validated as acetylated by MS and therefore predicted correctly (more detail in Chapter 3). Since the total number of acetylated lysines in the yeast proteome is approximately 0.6% my *in vivo* hit rate of approximately 25% is of reasonable accuracy.

Finally, in order to allow users to view predictions of their favorite protein, I developed a software tool, PredMod, which allows a user to enter a protein of choice and view all lysines, with their respective confidence scores with which they are predicted (**Figure 2.8**)(described further in **Appendix**). Users can also view lysine predictions of substrates that were contained in both independent validation sets as well as histone proteins including variants in multiple organisms. In the future, PredMod will potentially be powerful for identifying bonafide acetylation sites in nonhistone proteins, and further display the strength of using histone sequences as a useful guide for nonhistone acetylation prediction.



**PredMod**

[About](#)

**1 Select Protein:**

H3:

H4:

H2A:

H2B:

OR SELECT HUMAN PAN-ACETYL IP PROTEINS:

OR SELECT HUMAN PROTEINS:

OR Paste sequence below:

**2 Select Modification Type**

☒ Acetylation

**3 Select Threshold**

☐ High threshold

**Figure 2.8: PredMod, an acetylation prediction tool**

PredMod, an acetylation prediction tool was developed for a user to enter a protein of choice and obtain a list of predicted acetyl lysines with a confidence score for the lysine.

## Chapter 2 Discussion

My results suggest that the “sequence environment” in both histone and nonhistone proteins contributes to the likelihood of acetylation. Consistent across both human and yeast acetyl datasets, I noticed an enrichment of small residues, particularly glycine, and charged amino acids flanking validated acetylated lysines (**Figure 2.5, Figure 2.7**; tick marks). It is possible that I am perhaps achieving a higher accuracy for nuclear proteins, since the KAT substrates (histones) I used for training PredMod, mostly reside in the nucleus. My results also suggest that nuclear versus cytoplasmic KATs in yeast could possess unique substrate recognition profiles as illustrated by the differences in preferred flanking residues C-terminal to the lysine (**Figure 2.7**). In order to address whether histones were the best training set for nonhistone prediction, I retrained my algorithm with nonhistone lysines exclusively. I compiled lysines in both the literature scanned and pan-acetyl datasets together, and observed whether I would perform better using the nonhistone set for training purposes. I used five-fold cross validation on the datasets, thereby training on 4/5 of the dataset, and retesting on 1/5 of the dataset. I obtained an accuracy of 65% when training on histones, and 60% when training on nonhistones. When I trained on nonhistones to predict histones sites also using cross validation, I obtained 54%. Interestingly, I achieved a higher accuracy when training on histone versus nonhistone proteins. This could be due to the fact that in my

nonhistone dataset, I have a heterogeneous distribution of enzymes, due to the fact that this set contains both nuclear and cytosolic proteins. A compartmentalization between nuclear and cytoplasmic proteins could help achieve better results.

In my computational analysis, as shown in **Figure 2.5**, I observed that 15 out of the 51 pan-acetyl lysine substrates contained a histidine adjacent to the lysine (C-terminal) (**Figure 2.5C**). It is possible that this pan-acetyl antibody has a specific preference for histidine immediately adjacent to the acetylated lysine (**Figure 2.5C**). Whether the observed enrichment of histidines is due to a bias of the antibody or whether it could be part of a KAT recognition motif needs to be further explored. In contrast to the enrichment of small residues flanking validated acetylated lysines, I did not observe an enrichment of small amino acids surrounding “not observed” lysines in human histones and nonhistones therefore strengthen my confidence in the pronounced signal among acetylated lysines (**Figure 2.5D,E**).

An area that was unexplored that would be interesting to delve into is whether structural information could add to the predictive power of the algorithm. Thus, if a given lysine is located within a loop, beta-strand or alpha-helix, I can add this information onto the algorithm, where a lysine within the loop region would be more exposed than a lysine within a beta-strand. Current structural prediction programs such as Psi-Pred and other



prediction programs can aid in predicting these regions, and as a result the misclassified data may become true positives.

The prediction program, PredMod that I developed is a promising step in detecting additional novel acetylation sites. A study that has been recently published involves a quantitative proteomics approach to detect acetylation substrates (Choudhary, Kumar et al. 2009). One of the major findings of the study is that lysine characteristics of sequences in mitochondria, nuclear, and the cytoplasm are divergent from each other (Choudhary, Kumar et al. 2009). They observed the following results: all acetylated substrates have high frequency of tyrosines in the +1 position. I suspect that this could be due to the bias of pan-acetyl antibody that was used in their study, which resembles the high frequency of histidine in the +1 position on my pan-acetyl antibody dataset. In agreement with my analysis, they find an enrichment for glycine in the -1 position within their nuclear acetyl sites, unlike the cytosolic sites that do not show this type of preference. They also discover a high frequency of lysines surrounding acetylated lysines, which is also in agreement with my studies. Conclusions such as this from these types of large scale studies makes me curious as to whether incorporating other modifications that are potentially on the same peptide could be included into the algorithm. Since it is known that there is cross-regulation occurring between modifications on histone tails, and even nonhistones, a parameter that I could add to the algorithm would be to

include other modifications that are present within the flanking residues of a target lysine. Information on additional peptide modifications from the Global Proteome Machine Database (GPMDB) would be useful to incorporate on a proteome-wide scale. In the early stages when I was developing the algorithm, I utilized this same approach on histone modifications in order to derive whether multiple modifications within the same region of the histone would influence my histone predictions overall. For example, if the target lysine was H3K9ac, instead of representing the +1 residue relative to H3K9 as H3S10, this residue was represented as phosphorylated H3S10. Using this method, I achieved an accuracy of 70%, which is lower than my current prediction accuracy. One possibility of why I may have been achieving a lower prediction accuracy could be that I was overspecifying the training set, resulting in some of the true positives being missed. More complete quantitative data such as the abundance levels of additional PTMs, could be crucial for my algorithm performance.

Overall, my results suggest that KATs target specific sequence patterns, and that the predictive knowledge about histone acetylation provides a platform for studying both histone and nonhistone lysine acetylation. My model and results represent a step towards gaining a framework for predicting lysine acetylation sites in both human and yeast proteomes. It will be of interest in future studies to see whether my algorithm is capable of predicting lysine acetylation sites in many other

organisms. As more substrates in the acetyl-proteome are discovered (Yang and Seto 2008), it is likely that the predictive power of my approach will be strengthened, leading to more accurate confidence in the predicted site. Though my training dataset is 10-100 times in magnitude lower than other PTM datasets including acetylation (9, 21, 22), my approximate sensitivity measure of 60% is comparable and often higher than other prediction algorithms that achieve as low as 16-18% sensitivity (Blom, Sicheritz-Ponten et al. 2004; Saunders, Brinkworth et al. 2008; Schwartz, Chou et al. 2009). It would be interesting to see whether similar approaches could be applied to the prediction of other widespread histone modifications such as lysine methylation.

## **Chapter 3: Experimental validation of predicted acetylation targets**

### **Summary**

In this chapter, I describe the validation of my computationally predicted targets through wet-bench experiments that I performed in the Allis laboratory. In collaboration with the Don Hunt laboratory at the University of Virginia (Charlottesville, VA), and the Yingming Zhao laboratory at the University of Texas Southwestern Medical Center (Dallas, TX), I describe several novel acetylation marks in human histones and yeast nonhistone proteins that were correctly validated through my computationally oriented algorithm (as described in Chapter 2) via mass spectrometry. These marks further validate my computational algorithm and illustrate that primary sequence context of histone and nonhistone lysines are a driving factor in acetylation target and recognition. In addition, in order to determine acetyltransferase rules and pinpoint residues that could be critical for enzymatic recognition, I present data on yeast mutagenesis experiments that I performed in histone H3 in collaboration with the Ben Garcia laboratory at Princeton University (Princeton, NJ). These results have led to testable models and hypotheses that may be further explored.

## **Results**

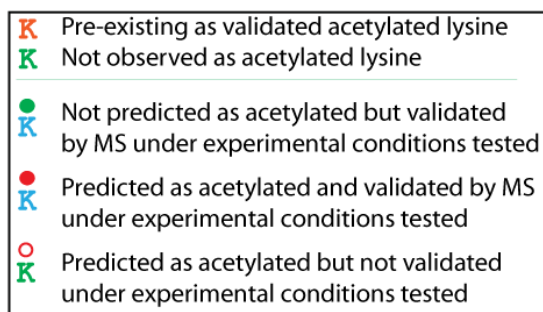
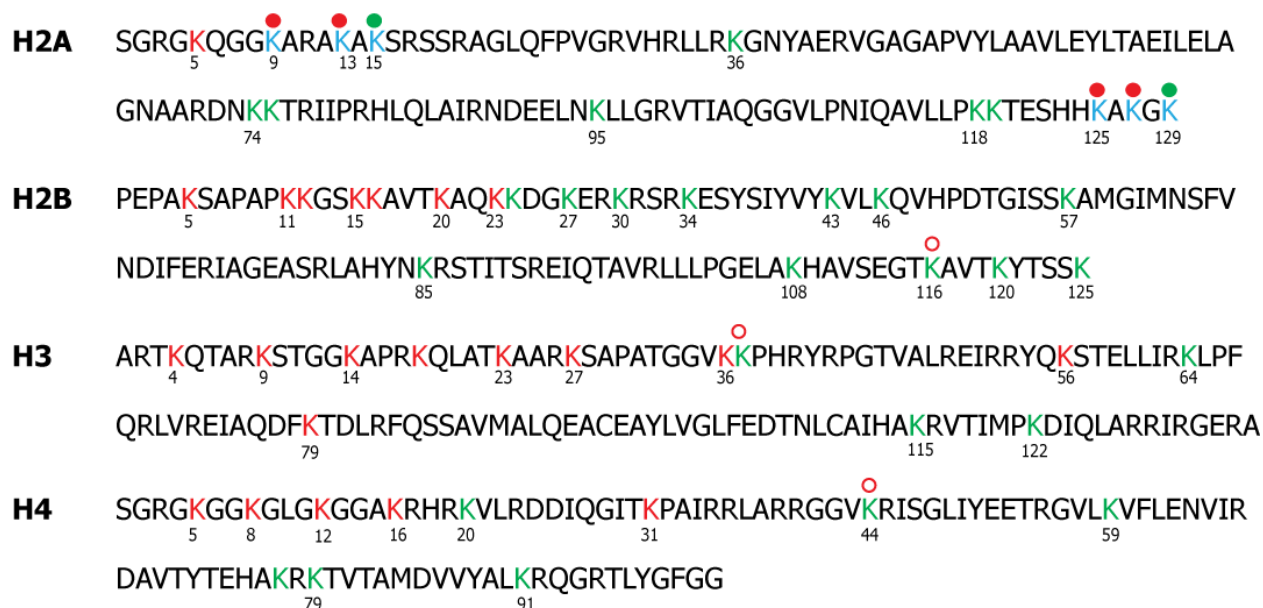
### ***In vivo* validation by mass spectrometry**

In Chapter 2, I described the computational method that I used in order to predict acetylation marks in histone and nonhistone proteins. As described in Chapter 2, I predicted seven novel acetylation sites in the human core histones; four in H2A (K9, K13, K125, K127), one in H2B (K116), one in H4 (K44), and one in H3 (K37) (**Figure 2.3**). This large number of predicted sites was unexpected, since histones have been intensely investigated for PTMs in recent years. To test whether these predicted lysines are acetylated *in vivo*, the Don Hunt lab employed an MS-based approach to examine histone peptides from human cell lines (Hela and HL60) that were asynchronously growing and treated without any HDAC inhibitors. HDAC inhibitors were excluded so that I could capture the non-hyper levels of acetylation in the asynchronously growing cells. All peptides containing the predicted lysines were identified, and importantly four of the seven predicted acetyl-lysines were experimentally validated: H2AK9, H2AK13, H2AK125, H2AK127 (**Figure 3.1**). In collaboration with the Ben Garcia lab, I also looked at histone acetylation marks under sodium butyrate treatment (an HDAC inhibitor), to assess whether this condition would result in an altogether different set of acetylation marks which could potentially alter my training set composition or prediction accuracy on the algorithm.

Under sodium-butyrate treatment, Hela cells displayed H3K37 and H2BK116 acetylation, two of my predicted acetylation sites, (personal communication, Ben Garcia), however since these marks were only observed under these special conditions, I did not count them as bonafide “validated” sites.

Additional acetyl marks observed under sodium butyrate treatment were: K95, K119 in H2A; K34, K46, K108, K116 in H2B; K77, K79, K91 in H4, and K37 in H3. Incorporating these sites into my training set of “validated” lysines did not alter the prediction performance of my algorithm significantly.

In summary, I correctly predicted four of these seven acetyl-lysine sites suggesting that the algorithm is capable to identify acetylation sites in human histone proteins.



**Figure 3.1: Validation on histone predicted lysines**

Human H2A and H2B, H3, and H4 sequences are shown. Literature-validated acetylated lysines are red, and lysines which have not been observed as acetylated are green. Lysines in blue under filled green circle were not predicted as acetylated, but validated under experimental conditions tested using LC/MS/MS. Lysines in blue under filled red circle were predicted as acetylated, and validated under experimental conditions tested using LC/MS/MS. Lysines in green under open red circle were predicted as acetylated, but have not yet been validated under experimental conditions tested.

## **Yeast *in vivo* validation**

The previous predictions were made on human histone and nonhistone datasets, and I therefore wondered whether my algorithm would also be able to predict acetylation sites in proteins from other organisms. Since histone acetylation has been studied extensively in budding yeast, I assessed the performance of my model on a proteome-wide dataset that included acetylated peptides in *S. cerevisiae* (Craig, Cortens et al. 2006) (Chapter 2). In addition, I experimentally validated predicted acetylation sites in candidate yeast nonhistone proteins *in vivo*.

An outline of my overall approach and method is as follows: I) I ran the algorithm against the 30 chromatin-associated proteins (of interest to the Allis Lab) in the *S. cerevisiae* proteome by taking a sequence stretch of six residues N- and C-terminal to the target lysine because of the promising human ROC results. I used the nonhistone lysine traversal method described in Chapter 2 and applied to the human histone-trained tree to find which subgroup of sequences the nonhistone yeast sequence formed the tightest cluster with. Based on the threshold score that was determined by the human ROC analysis, I then selected candidate proteins that had the highest scoring alignments with acetylated histone sequences. The highest ranked predicted lysine acetylation sites were in the proteins Cac2 (component of the chromatin assembly complex) (Enomoto and Berman 1998), Spt6 (a transcriptional elongation factor and nucleosome disassembly factor) (Clark-



Adams and Winston 1987), Sir3 (a silencing factor that establishes a transcriptionally silent chromatin state) (Gasser and Cockell 2001), and Eaf7 (an Esa1 associated factor) (Krogan, Baetz et al. 2004). Additionally, for each of these proteins, I ranked in order of confidence lysines that would be predicted as likely KAT targets. III) I then purified my candidate proteins by performing a TAP purification (**Figure 3.2A**) and subsequently analyzed these via mass spectrometry according to previously published rules (Chen, Kwon et al. 2005).

In conclusion, I validated my predictions on three yeast candidate proteins that had previously not been published to contain acetylated sites (Spt6 (Clark-Adams and Winston 1987), Sir3 (Gasser and Cockell 2001), and Eaf7 (Krogan, Baetz et al. 2004). I expressed and purified the TAP-tagged candidate proteins in *S. cerevisiae* (**Figure 3.2A**) and subsequently subjected them to LC/MS/MS. With PredMod, I predicted 15 sites to be acetylated out of 416 total lysines in three candidate proteins combined. Four of these, within the top six ranked predicted sites (**Figure 3.2B,C,D**), were validated as acetylated by MS and therefore predicted correctly (**Figure 3.2B,C,D**). Since the total number of acetylated lysines in the yeast proteome is approximately 0.6% (discussed in Chapter 2), my *in vivo* hit rate of approximately 25% is of reasonable accuracy.

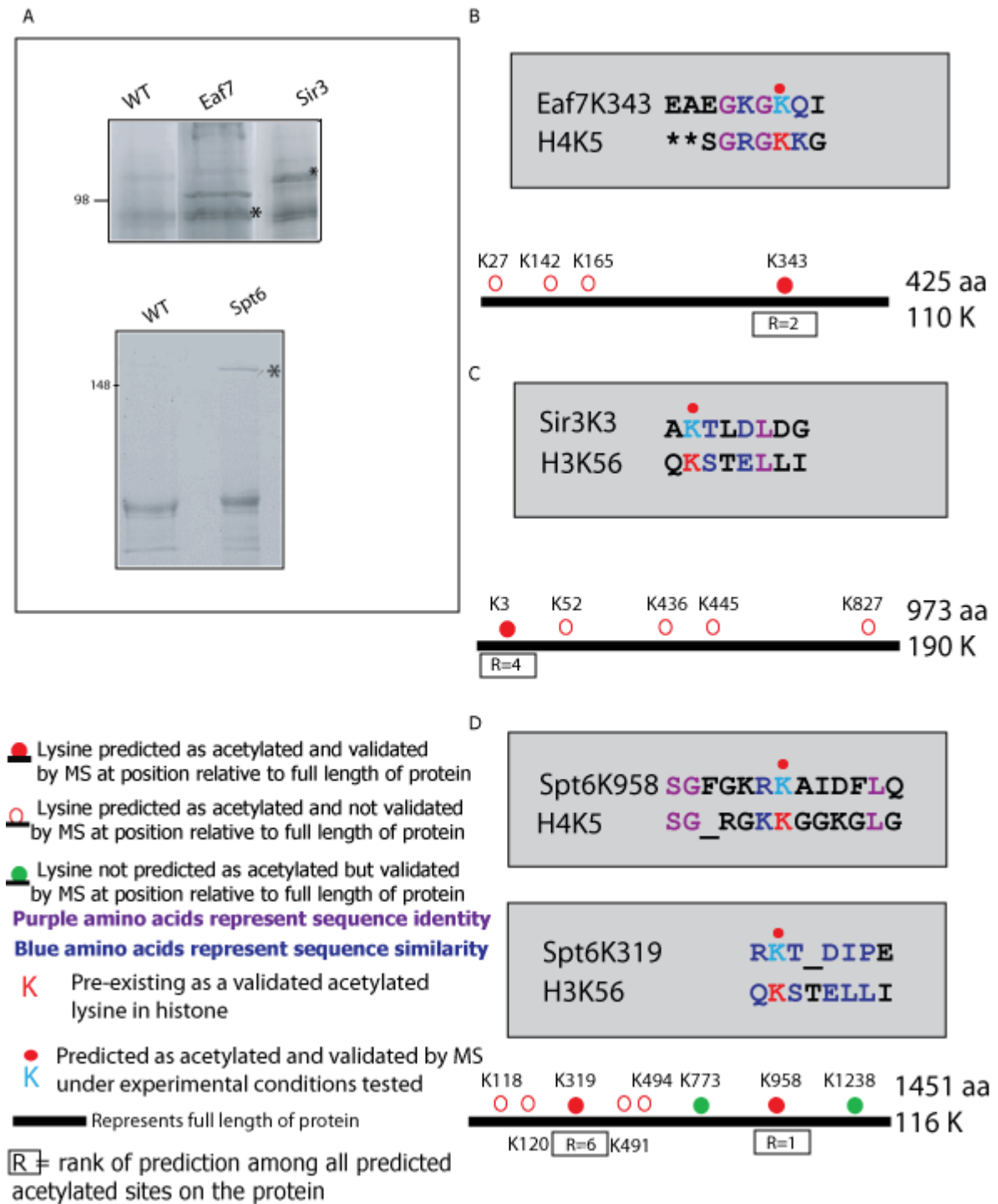
These results demonstrate the power of PredMod for identifying bona fide acetylation sites in nonhistone proteins, and further display the strength of using histone sequences as a useful guide for nonhistone acetylation prediction.

**Figure 3.2: Novel predictions and in-vivo validations in *S. cerevisiae***

**A.** Coomassie-stained gel of TAP pulldown purification of yeast proteins Eaf7, Sir3, and Spt6. Asterisks denote bands that were isolated and inspected for acetylation by MS.

**B-D.** Shown are regions of the sequence alignment where there is sequence identity or similarity. A purple pair of amino acids represents identical residues and a blue pair of amino acids represents residues that can be evolutionarily substitutable in accordance with the BLOSUM matrix. A black bar below the alignment represents the length of nonhistone protein. The numbers to the right represent the total number of amino acids (aa) and the number of lysines (K) in the protein. Red filled circles depict lysines that were predicted correctly on the yeast nonhistone substrate at the specific location relative to the full length of the protein. Red empty circles represent lysines that were predicted, but not confirmed by mass spectrometry under the conditions tested. Green filled circles represent lysines that were not predicted, but validated experimentally by MS. A blue lysine with a red filled circle on top represents those lysines that were correctly predicted. Boxed R represents the rank of prediction among all predicted acetylated sites on the protein. **B.** Eaf7K343 (R=2) sequence alignment with H4K5. **C.** Sir3K3 (R=4) sequence alignment with H3K56. Symbols denoted as in B. **D.** Spt6K958 (R=1) and Spt6K319 (R=6) sequence alignments with H4K5 and H3K56, respectively. Symbols denoted as in B.

Figure 3.2



## ***In vivo* mutagenesis in budding yeast**

My computational analysis in Chapter 2 led me to conduct experiments that could help identify rules, or specific patterns that could potentially be important for acetyltransferase recognition in order to address the question: is immediate sequence context of lysine important in HAT recognition? From my statistical and quantitative analysis methods, I came to the conclusion that sequence context surrounding a target acetylated lysine can dictate KAT substrate preferences. However, to fully test the effectiveness of my approach, I wanted to perform specific wet-bench experiments to formalize the rules necessary for enzyme recognition. Since my computational analysis led to a frequency enrichment analysis of flanking residues surrounding an acetylated lysine, I wanted to mutate these residues *in vivo* to determine what effects these mutations might have on known PTMs. I chose *S. cerevisiae* as the organism to perform site-directed mutagenesis in because of its robust genetic manipulation techniques and easily available reagents. Additionally, yeast H3 and H4 are highly conserved with human H3 and H4 which were a part of the training set the algorithm was based on. My first step was to examine the hierarchical tree in Chapter 2 (**Figure 2.3**), in order to determine whether there were lysine embedded sequences under the same cluster with similar flanking sequence profiles.

Via the clustering analysis, I singled out a cluster containing two sequences that were tightly paired (ranked as the tightest pair of sequences in the tree). As shown in **Fig 3.3A**, these sequences were H3K14 and H3K36. Close examination of the sequence alignment of these two sequences reveals phosphorylatable residues in the -4 position, glycine in the -1 position (small), a proline in the +2 position, and a basic residue in the +3 position (**Figure 3.3A**). Also, H3K14 and H3K36 are both Gcn5-mediated acetyl sites (Howe, Auston et al. 2001; Morris, Rao et al. 2007). Furthermore, H3K36 is a bonafide trimethyl site, where the methylation of this site occurs via the methyltransferase, Set2 (Strahl, Grant et al. 2002). Additionally, both of these acetyl sites are associated with transcriptional activation (Pokholok, Harbison et al. 2005), and H3K36me2 and H3K36me3 are associated with transcriptional elongation, and are needed to prevent cryptic initiation (Kizer, Phatnani et al. 2005; Morris, Shibata et al. 2005). The distinct functional roles of H3K36 acetylation and methylation made me curious as to whether H3K14, a bonafide acetyl site, could also be methylated if its sequence was mutated to resemble H3K36's. I next checked whether there was an existing literature of H3K14 methylation in yeast. To my knowledge, there is no published literature of H3K14 methylation in organisms ranging from mammals to yeast to date (Garcia, Hake et al. 2007).

In order to address the question of whether H3K14 is methylated after mutagenesis within the flanking residues, I mutagenized the flanking residues of H3K14 to resemble H3K36. A key question that I could also attempt to answer in parallel through these experiments was: how do H3K14 acetylation levels fluctuate as a result of specific mutations? The table below (**Table 3.1**) displays the mutations that I chose to make, the rationale behind why selecting these mutations would enable me to learn something about acetylation based on my computationally derived frequency analysis (**Figures 2.5, 2.7**), and a prediction of how acetylation levels would increase or decrease upon making the selected mutations, which were designed to resemble the flanking H3K36 sequence profile (**Figure 3.3B**):

**Table 3.1: Mutations, Rationale, and Acetylation Prediction**

<b>Mutation</b>	<b>Rationale</b>	<b>Acetylation Prediction</b>
G13->V	Valine is beta branched and more conformationally rigid as opposed to glycine with a flexible side chain.	Acetylation will be decreased, as glycines are enriched on the N-terminal side of an acetylated lysine.
A15->V	Control. Since I mutated G13 to V on the left, a parallel mutation on the right is a good control.	No change. Valines are present in the +1 position of acetylated substrates
A15->K	Lysine is a charged amino acid. Do two adjacent charged residues have an effect on lysine acetylation or methylation?	Acetylation could increase or decrease. There are lysines in the +1 positions of acetylated substrates, and in the “not observed as acetylated” substrates.
G13->V, A15->K	Double mutation, identical residues surrounding H3K36	Decrease in acetylation mainly caused by the G13V mutation.

### **Figure 3.3: Mutagenized yeast histones**

**A.** Blast sequence alignment between H3K14 and H3K36. Both sites are acetylated by Gcn5 and H3K36me can be mono, di, or tri-methylated. Colored residues either identical or similar according to BLOSUM matrix(Eddy 2004).

**B.** Mutations that were made in H3 of budding yeast *in vivo*. Effect of methylation and acetylation levels as a result of mutagenesis was the question I tried to answer in study.

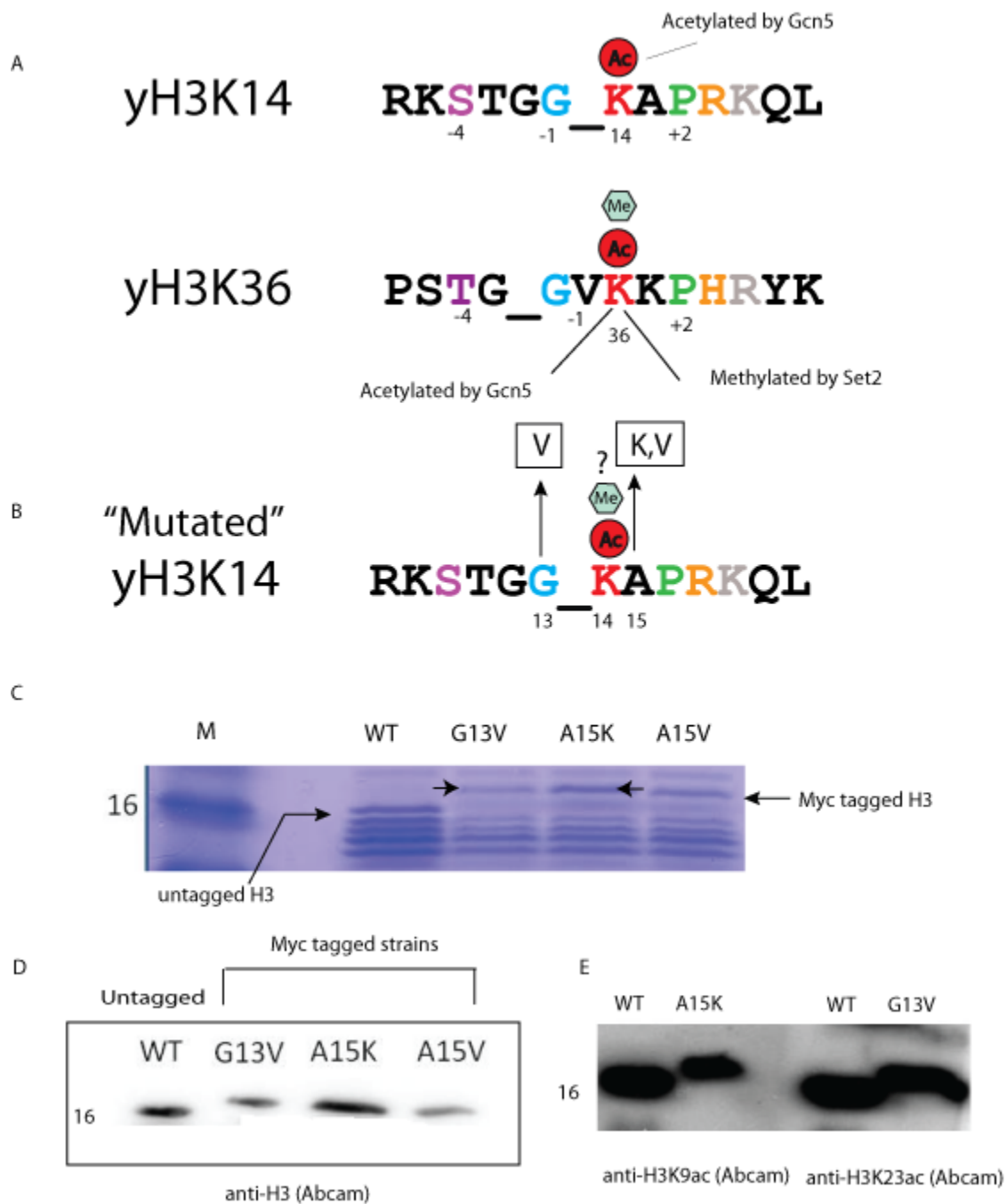
**C.** Acid extracted histones (1 microgram) were electrophoresed in a 15% SDS-PAGE and analyzed by Western blot. Shown is the marker (left) and the three strains that histones were extracted from (right 3 lanes). These strains are WT, H3G13V, H3A15K, and H3A15V. Untagged H3 (shown in second lane), and Myc-tagged H3 is shown with arrow (all other lanes).

**D.** Western Blot using a general H3 antibody (Abcam) with same fractions as in C, electrophoresed in a 15% SDS-PAGE gel.

**E.** Western Blot using H3K9ac (Abcam) and H3K23ac antibodies (Abcam) to display that Myc-tag on the mutant strains did not affect acetylation levels of H3. Strains are H3A15K (left) and H3G13V (right).



Figure 3.3



I used a budding yeast plasmid shuffle strain (detail described in **Methods**) to perform yeast mutagenesis by making the following H3 shuffle plasmid mutations: H3A15K, H3G13V, and H3A15V. In the strain that I wanted to make these mutations in, I shuffled out the WT plasmid and inserted my mutant plasmid containing my desired mutation. To check whether my strains had the correct mutations, I performed a plasmid recovery followed by DNA sequencing. Once I confirmed that I had the correct mutations in my yeast strains, I extracted yeast nuclei through douncing and spheroplasting, and used an acid extraction procedure of histone proteins (**Figure 3.3C, D**). Since the mutant strains were tagged with a N-terminal Myc tag, I was concerned that the Myc-tagging could alter the acetylation status of histone H3. Thus, I checked acetylation levels of my WT versus mutant strains against the general H3, H3K9ac, and H3K23ac antibodies (**Figure 3.3D, E**). In the mutants, acetylation was at a comparable level to WT acetylation levels, suggesting that the Myc tag did not grossly affect levels of acetylation. In collaboration with the Ben Garcia laboratory at Princeton University, mass spectrometry analysis of WT and mutant H3 was performed using LC/MS/MS, and the following results were revealed: there was no methylation visible on H3K14 in either the WT or any of the mutant strains. The double mutant strain (G13V, A15K) displayed a lethality phenotype. While this could be interesting and worth investigating, I chose not to follow up on it further at the moment. In all my

mutant strains excluding my designated control (A15V), H3K14 acetylation was decreased (**Figure 3.4A**); the most significant decrease was under the alanine to lysine mutation in the +1 position as H3K14 decreased by four-fold (**Figure 3.4A**). Under this same mutation, H3K9ac increased by approximately two fold. To my surprise, methylation levels of H3K36 were also affected under the A15K mutation. I did not expect that distorting the sequence immediately surrounding H3K14 would have such a pronounced effect on H3K36 methylation levels, such that the unmodified form of H3K36 increased twenty fold, H3K36 monomethyl increased ten-fold (H3K36me1), while H3K36 dimethyl (H3K36me2) and trimethyl (H3K36me3) levels decreased by 1.5 fold (**Figure 3.4** illustrates the exact values). Errors not shown in figures as the experiments were performed only once. All mass spectrometry figures can be found in the **Appendix**.

**Figure 3.4: PTMs as result of flanking residue mutations surrounding H3K14**

**A.** Table displaying H3K9ac, H3K14ac, H3K36ac, H3K36me0, H3K36me1, H3K36me2, and H3K36me3 levels as a result of making mutations H3A15K, H3G13V, and H3A15V. Numbers represent percentage of the total that is modified. Yellow highlighted numbers represent mutations where there were dramatic increases or decreases.

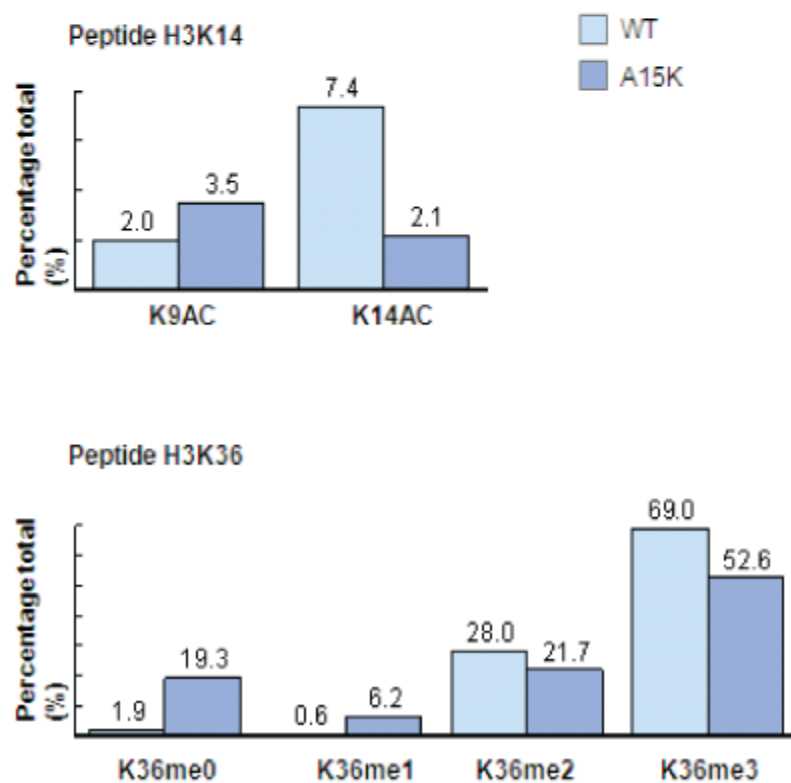
**B.** Bar graphs displaying values as denoted in (A) for both H3K14 and H3K36 peptides. Note the fold percentage differences between H3K36me0, and H3K36me1.

**Figure 3.4**

A

	H3K9 ac	H3K14 ac	H3K36 ac	H3K36 me0	H3K36 me1	H3K36 me2	H3K36 me3
WT H3	2.1%	7.3%	Very low	1.88	.6%	28%	69%
H3G13V	No change	5%	No change	No change	No change	No change	No change
H3A15K	3.5%	2.1%	No change	19.44%	6.15%	21%	52%
H3A15V	2.1%	7.3%	No change	No change	No change	No change	No change

B



### Chapter 3 Discussion

Here I show that I can predict novel acetylation sites in proteins of interest in both human and yeast cells. Furthermore, I have shown that there are residues that may be critical for lysine acetyltransferase recognition and demonstrated that there could be a “crosstalk” occurring between three marks on the histone H3 tail H3K9, H3K14, and H3K36. Together my data can lead to a predictive understanding of acetylation by examining the sequence of the targets themselves.

Using a combined experimental/computational approach, I identified several sites in human histone and nonhistone proteins that were correctly predicted. There were, however, a class of lysines that were predicted by my algorithm, yet have not yet been detected experimentally. Several possibilities can be envisioned for this case: first, the MS approach has limited detection and sensitivity capabilities and cannot recover peptides that are acetylated at only low levels. Second, lysines could be modified only in distinct environmental conditions, cell cycle stages, cell types, and are therefore undetectable in the cell extracts I used. Here it should be noted that additional novel histones acetylation sites were detected by MS/MS when Hela cells were pretreated with HDAC inhibitors (personal communication, B. Garcia), and I retrained the algorithm with this data. Preliminary results from this analysis display that the predictive power of my

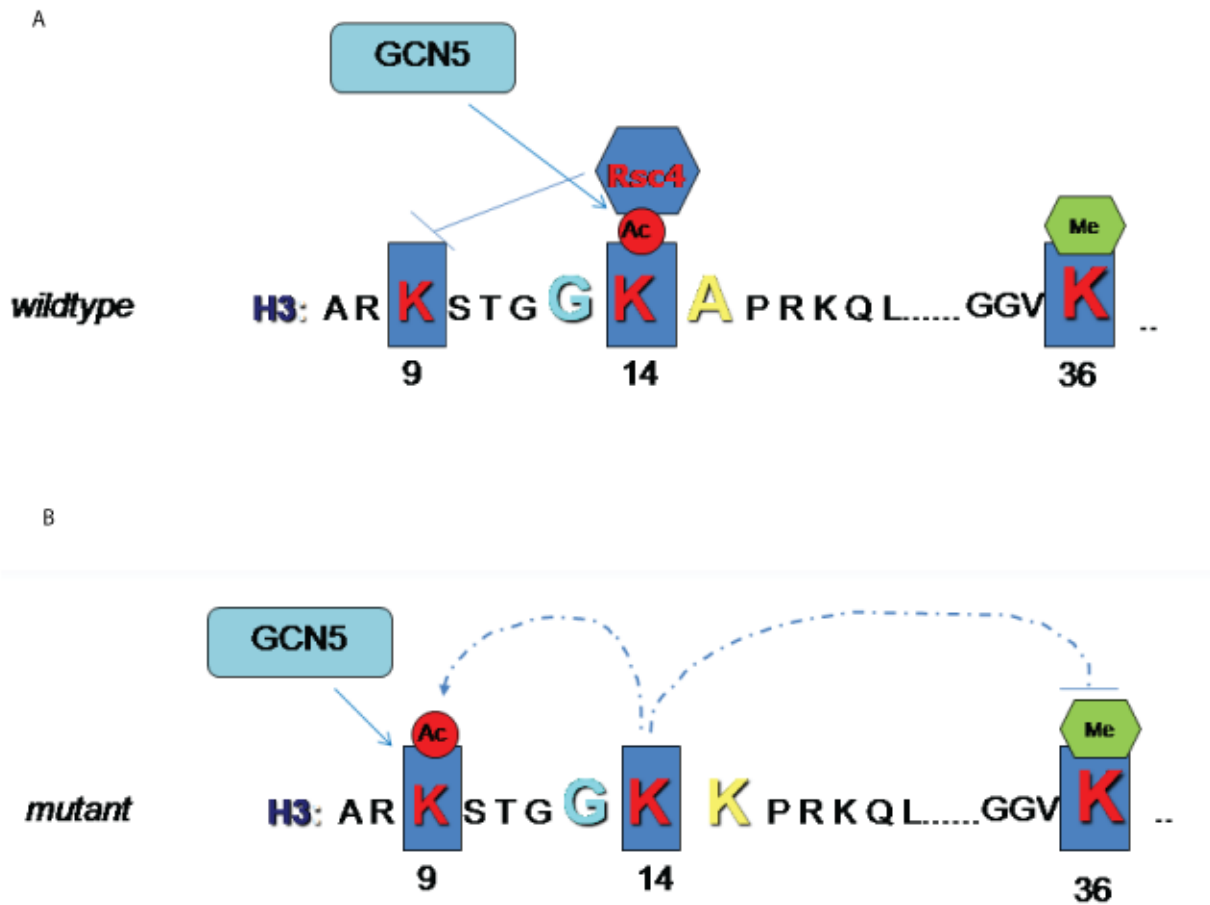
overall approach is not altered significantly, thereby increasing further confidence in the power of my approach. Third, acetylation might be inhibited by adjacent PTMs (negative crosstalk), and therefore the responsible KAT might be prevented from binding to or accessing its target site. Finally, acetylation is a dynamic, transient modification, and thus MS results may depend on a time-specific acetylation state whose kinetic properties have not been adequately captured by the experimental parameters. Of interest is the class of lysines in histones that were not predicted by the algorithm, yet were detected by MS, and which might indicate a different class of KATs that need special sequence surroundings. I observed that one of these lysines, H2AK129, did not contain a full set of flanking residues C-terminal to the lysine, which put this lysine at a “disadvantage” compared to other lysines with a full set of flanking residues. H2AK15 failed to overall strongly resemble the other acetylated histone lysines in my training set by sequence, and therefore was not predicted. This could be due to the training set containing an underrepresented number of HATs, assuming that clusters of sequences under a single node represent similar enzyme recognition.

In this chapter, I also addressed whether sequence surrounding an acetylated lysine can drives HAT recognition by performing site directed mutagenesis *in vivo*. My results display that in the H3A15K background, H3K14ac decreases 3.5 fold, while H3K9ac increases almost two fold.

H3K36me levels are dramatically affected with a twenty fold increase of the unmodified form of H3K36. Thus, H3K14 disruption could be promoting H3K9ac due to the steric binding of the H3K14 acetylation effector, Rsc4 (VanDemark, Kasten et al. 2007) (**Figure 3.5**). Similarly, H3K14 disruption could be acting in an inhibitory manner to H3K36 di and tri methylation, suggesting that H3K14ac is required for normal levels H3K36 methylation (**Figure 3.5**). To my knowledge so far, acetylation of H3K14 by the NuA3 KAT requires prior methylation of H3K36, although the mechanism remains to be determined (Latham and Dent 2007). My model suggests that perhaps H3K14ac promotes H3K36 methylation by inhibition of demethylase activity, or by post-translational modifications in the vicinity of these two marks that are also involved in crosstalk (**Figure 3.5**). Further mutagenesis experiments would have to be performed in order to fully explore this hypothesis. Some preliminary experiments that I suggest are: (1) In budding yeast, design a strain with the H3K14R mutation, and observe whether H3K9ac levels decrease, and H3K36me0 and H3K36me1 levels increase using H3K9ac and H3K36me1 antibodies via western blotting. The result would reveal the distinct nature of the H3A15K mutation against the H3K14R effects. (2) Make Rsc4 bromodomain mutations in WT cells, so that H3K14 binding to Rsc4 is abolished. Observe H3K9ac and H3K36me levels as a result; my above stated hypothesis suggests that H3K9ac would increase as a result. These preliminary experiments could help us learn



more about the crosstalk biology and can guide us to better understand the orchestration of specific H3 acetylation, deacetylation, methylation, and demethylation events.



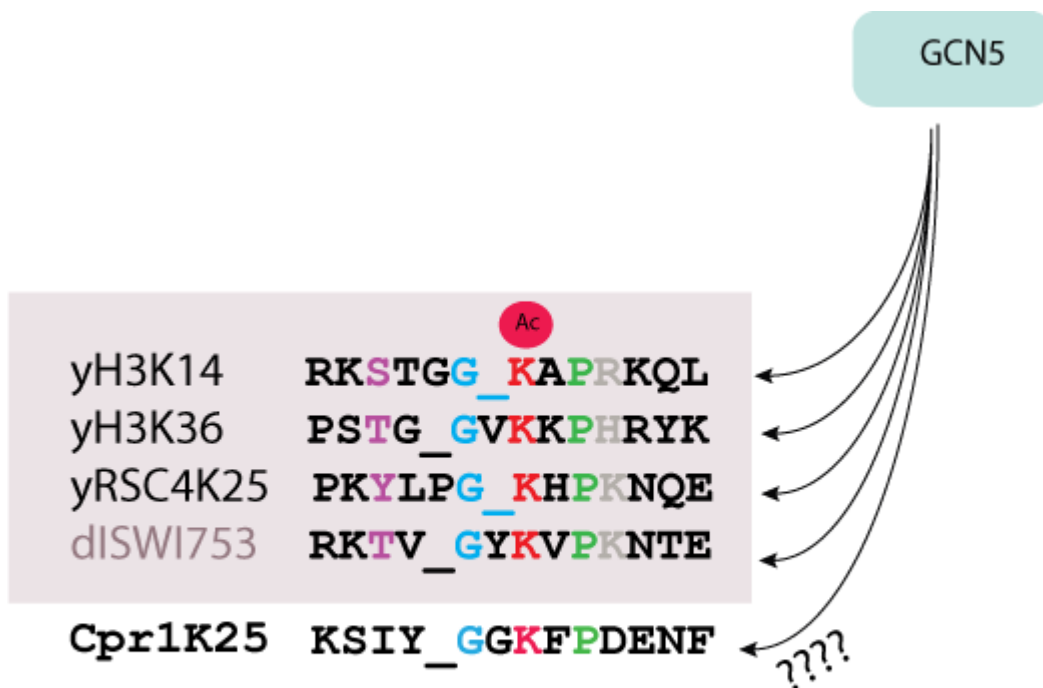
**Figure 3.5: Crosstalk Model**

**A.** In the wildtype background, Gcn5 acetylates H3K9ac. The Rsc4 ( bromodomain ) binds H3K14ac causing a steric occlusion of Gcn5 to acetylate H3K9.

**B.** In the mutant background, Gcn5 does not recognize its preferred site, and acetylates H3K9. An inhibitory mechanism causes disruption of H3K36 di and trimethyl. Dashed lines indicate possible crosstalk.

One outstanding question that remains is what the algorithm presents regarding enzyme specificity within the hierarchical tree. Does the alignment of acetyl sites actually predict the HAT for those sites, and do KATs have characteristic lysine containing motifs that they will target? Through my method, I suggest that there may be an intrinsic substrate specificity for acetyltransferases in general. Whether computational rules can then be subgrouped into rules specific for each enzyme is a question that I would like to address. In order to achieve this, I revisited my tight cluster (H3K14 and H3K36), both Gcn5 dependent marks. I then ran my algorithm against validated nonhistone acetylated proteins and traversed these lysines through the histone-trained tree. I observed that there were nonhistone proteins, particularly those that were Gcn5-mediated, that clustered very tightly within a specific region of the histone trained tree (**Figure 3.6**). Recently, an essential chromatin remodeling factor Rsc4 was uncovered as acetylated at K25 in *S. cerevisiae* (VanDemark, Kasten et al. 2007). When I ran the Rsc4 protein against my algorithm, I noticed that K25 clustered tightly with cluster, and thus were able to predict this correctly (**Figure 3.6**) ((Howe, Auston et al. 2001), (Morris, Rao et al. 2007; VanDemark, Kasten et al. 2007)). Interestingly, a recent study of the ISWI protein in *Drosophila* reveals that it's acetylated at K753, dependent on Gcn5, and also has a G and P in its sequence (Ferreira, Eberhardter et al. 2007)(**Figure 3.6**). Close analysis of the sequences under this cluster

revealed that there was sequence identity and similarity in the flanking residues of the lysine, and thus I was interested in whether acetylation of a lysine is programmed by its surrounding residues and whether there were particular residues that were critical for KAT recognition.



**Figure 3.6: Sequence alignment of Gcn5 mediated substrates**

Sequence alignment displaying H3K14, H3K36, ISWI753, Cpr1K25, and Rsc4K25. I predict that Cpr1K25 is acetylated by Gcn5, the primary enzyme for all the other substrates in the purple box. Similar or identical residues are colored.

Alignment of the flanking residues of the surrounding lysines within the cluster (**Figure 3.6**; grey box) revealed that all four substrates have either a S/T/Y residue and a glycine flanking the left side of the lysine nearly at the same position. I also noticed that all four substrates have a proline in the +2 position and a basic residue H/R/K in the +3 position (**Figure 3.6**). Next, I ran all yeast proteins through my algorithm, and detected proteins that contained a G-K-X-P sequence. One such protein was Cpr1, a peptidyl prolyl isomerase (Arevalo-Rodriguez, Wu et al. 2004) that clustered with this particular group of G-K-X-P sequences. Whether or not this substrate is Gcn5 mediated will be an interesting question to answer and is something I hope to follow up on in the future.

## Chapter 4: Domain prediction of chromatin-associated proteins

### Summary

Protein domain prediction is important for protein structure determination, functional annotation, and mutagenesis among other things. Most eukaryotic proteins receive and process signals which are constructed in a modular fashion from a combination of interaction and catalytic domains (Zarrinpar, Bhattacharyya et al. 2003). These interaction domains mediate the formation of multiprotein complexes that restrict signaling proteins to appropriate subcellular locations and help determine the specificity of enzyme-substrate interactions. Thus, being able to identify these domains using computational and structure based approaches is an important precursor for a range of methods. While Chapters 2 and 3 were focused on “writer” recognition, in this chapter I discuss “readers” or effector domains that specifically bind PTMs, and describe computational approaches used to predict additional unannotated or cryptic domains. I focus on the chromatin-associated Polycomb Group (PcG) proteins, since these proteins contain highly conserved key domains throughout evolution, yet their substrate specificities are highly dissimilar contributing in part to their protein complexity. Being able to predict additional domains and motifs outside of the key domains could help gain insight into their functional versatility. Since PcG proteins have diverged significantly from their *Drosophila*

counterparts and from their paralogs, I use my domain detection approach to infer co-evolution and expansion of these proteins in closely related and distant organisms. In the first part of the chapter, I present the diversification of PcG homolog genes in the context of development and cellular differentiation in species ranging from plant to human, as part of a collaboration with two of my Allis lab colleagues, Sarah Whitcomb and Emily Bernstein (former member of Allis Lab). In the second part of this chapter, I use existing bioinformatic methods to discover additional chromatin-associated effector proteins that bind PTMs. These effectors can potentially be important for downstream signaling and recruitment events.

## **Results**

### **Conservation and diversification of PcG homologs**

I developed a bioinformatics method to discover key domains of PcG homologs. Key domains of PcG homologs are highly conserved between evolutionarily distant organisms (orthologs) and among paralogs in a given organism (typically >75% amino acid similarity). However, outside of key domains, PcG proteins have diverged significantly from their *Drosophila* counterparts and from their paralogs. An important challenge is to understand the functional significance of developmentally regulated expression of orthologs and how this impacts PRC composition, genomic targeting and/or mechanism of transcriptional repression. To illustrate the

functional diversification of PcG paralogs, I focused on homologs of *Drosophila* Pc, Psc, dRing and E(z).

The bioinformatics algorithm that I developed consisted of the following steps: (1) Selection of E(z), Pc, dRing and Psc as the PcG reference set: the “writer” (E(z)) and the “reader” (Pc) of the H3K27me mark, the only other known catalytic member of the PcG (dRing), and dRing’s associated factor (Psc). Sequence and domain information was retrieved for *Drosophila*, mouse, and human PcG proteins from the Swissprot, Uniprot, Ensembl, and SMART (Ponting, Schultz et al. 1999) databases. (2) In order to identify putative homologs of these proteins in other organisms, I performed BLAST searches using the *Drosophila*, mouse and human proteins as queries (Altschul, Gish et al. 1990; Altschul, Madden et al. 1997). The Blastp searches, using the NCBI web server (<http://www.ncbi.nlm.nih.gov/BLAST>) were performed at low, medium, and high stringencies to obtain all possible proteins in the following organisms: *Arabidopsis thaliana*, *Caenorhabditis elegans* (worm), *Strongylocentrotus purpuratus* (sea urchin), *Danio rerio* (zebrafish), *Xenopus laevis* (frog), *Gallus Gallus* (chicken), *Canis familiaris* (dog), *Mus musculus* (mouse), and *Homo sapiens* (human). Since homologous domains between species were very well conserved, variation of BLAST parameters such as gap costs and substitution matrices did not significantly alter my results. (3) I developed a custom made algorithm to query motifs less than 15 amino acids in length

such as the Pc box. Sequence similarity between homologous domains and full-length sequences was used to compute a weighted score by multiplying a fixed weight to the BLAST score and computing an overall weighted sum. Key domains for PcG proteins were identified including: the chromodomain and Pc-box for Pc homologs, the RING domain for Psc and dRing homologs, and the SET and SANT domains for E(z) homologs (**Figure 4.1**). A list of putative homologs containing the highest scores was compiled for each class of PcG proteins, and homologs in different species were selected by manual inspection. (4) Finally, the ClustalW (Thompson, Higgins et al. 1994) program was used to cluster putative homologs to ensure that proteins with similar key domains with minor amino acid differences such as Ring1A and Ring1B were categorized correctly. In addition, a reverse BLAST of all the putative homologs was also performed against the *Drosophila* genome to ensure that the forward BLAST hit was accurate. Many databases of translated cDNA libraries are inherently incomplete, and so gene sequences of those PcG homologs that were found to be absent in a given organism were also queried against the predicted ORFs from that organism's genome. Finally, putative PcG homologs were mapped onto a phylogenetic tree of organisms (**Figure 4.2**). Additional potential paralogs of dPsc were found in *Drosophila* through the algorithm. These proteins are Su(z)2 and I(3)-73Ah; however as they are not well characterized as functional Psc homologs, they have been excluded these proteins from the phylogenetic tree. Sequence

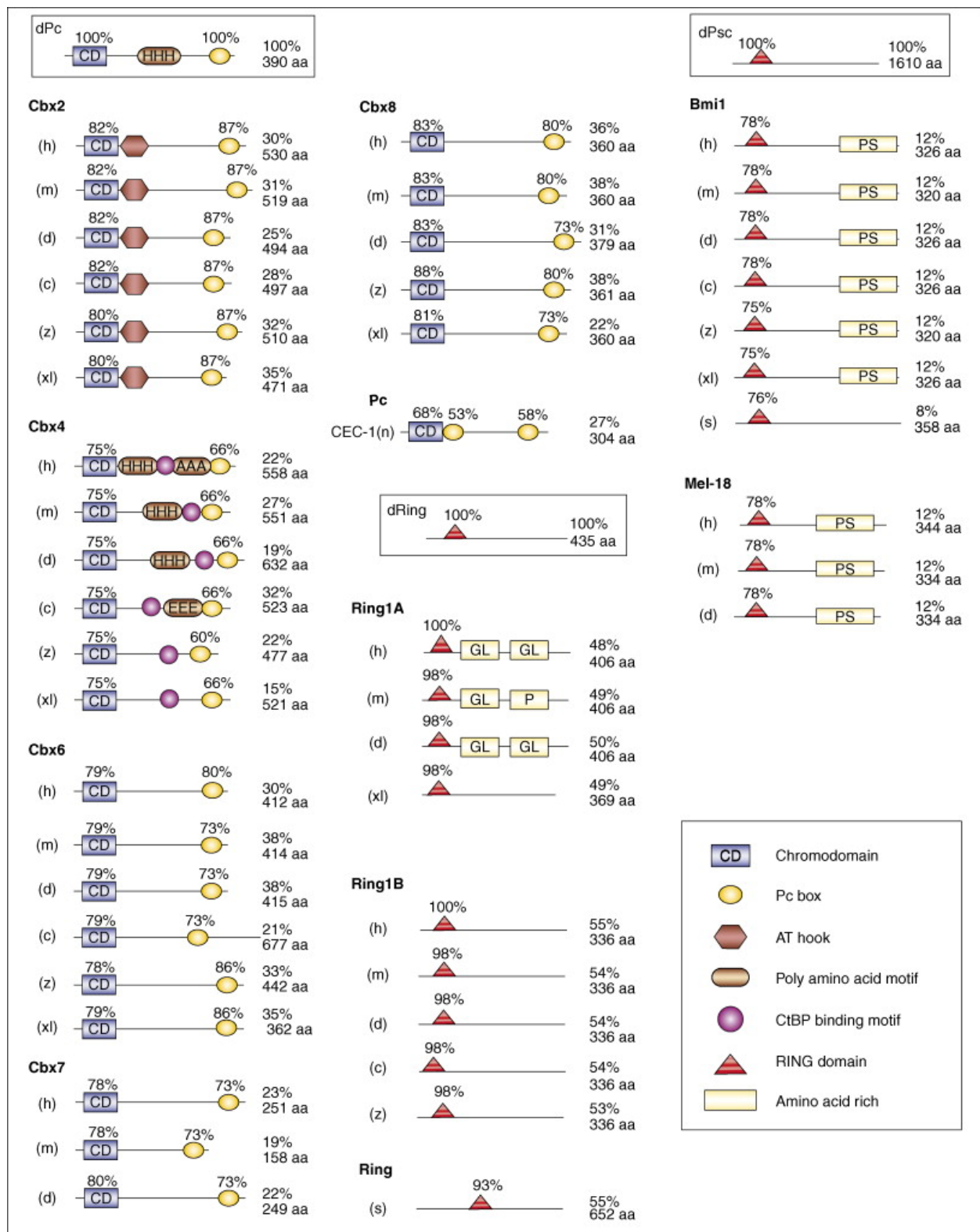


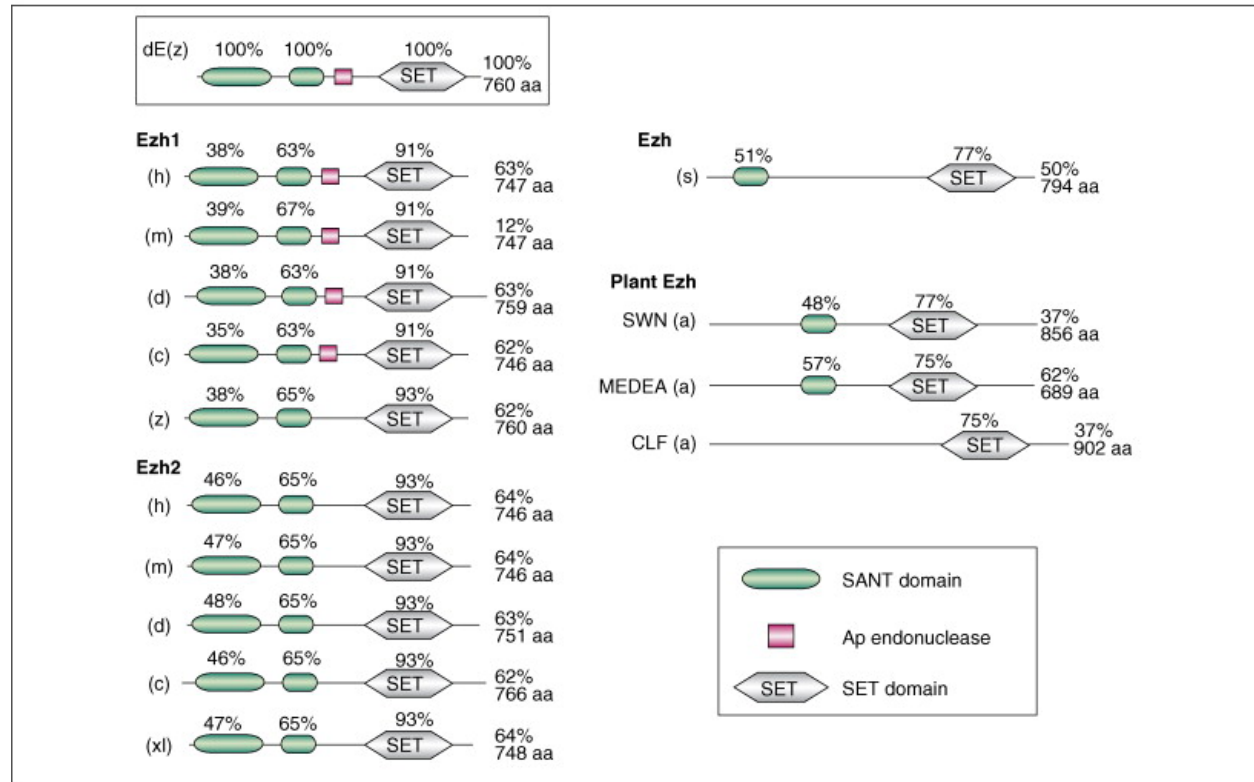
alignments in **Figure 4.3** were constructed using the Emboss program and CEC-1 was predicted as a putative Polycomb homolog because of its high sequence similarity with *Drosophila* and mouse Cbx8 chromodomains. Two stretches of amino acids within the N-terminus and C-terminus of CEC-1 also shared a strong sequence similarity with the *Drosophila* and mouse Cbx8 Pc boxes. In **Figure 4.2**, the ClustalW program was used to create a multiple sequence alignment of Cbx proteins and subsequently fed into the PHYLIP software program (Felsenstein 2008). The protdist and neighbor joining programs within PHYLIP were used to construct the Cbx paralog tree (**Figure 4.3**) and the Emboss local alignment program was used to determine sequence similarity and identity percentages.

#### Figure 4.1: Domain and motif structure of selected PRC1 proteins

Comparison of domain and motif structure of selected PRC1 proteins in *Drosophila* (boxed), human, mouse, dog, chicken, zebrafish, frog, nematode and sea urchin. Protein lengths are scaled exclusively within homolog groups, not among paralog groups, and are represented by a black line. Numbers shown above domains are percentage similarity to the domain in the *Drosophila* homolog. Numbers to the right of proteins represent percentage similarity to the full-length *Drosophila* sequence and the number of amino acids in the protein. Note the high percentage similarity between domains, but low percentage similarity of full-length sequence, between the *Drosophila* protein and its homologs in other organisms. Also note different amino acid lengths of paralog groups (e.g. Cbx4 versus Cbx7) and different domain structure (e.g. Cbx2 versus Cbx4). See text for details. Pc homologs are grouped based on their sequence similarity to mouse Cbx2, Cbx4, Cbx6, Cbx7 and Cbx8. Putative *C. elegans* Pc protein is shown at bottom (CEC-1). Psc homologs are grouped based on their sequence similarity to mouse Bmi1 and Mel-18. Homologs of dRing are grouped based on their sequence similarity to mouse Ring1A and Ring1B. Sea urchin Ring is equally similar to mouse Ring1A and Ring1B, and is listed separately at the bottom. Abbreviations: c, chicken; d, dog; h, human; m, mouse; n, nematode; s, sea urchin; xl, frog; z, zebrafish. Image directly from (Whitcomb, Basu et al. 2007).

Figure 4.1





**Figure 4.2: Domain and motif structure of selected PRC2 proteins.**

Comparison of domain and motif structure of selected PRC2 proteins; see 4.1 for details. E(z) homologs are grouped based on their sequence similarity to mouse Ezh1 and Ezh2. *Arabidopsis* E(z) homologs are listed separately on the right. Abbreviations: a, plant; c, chicken; d, dog; h, human; m, mouse; s, sea urchin; xl, frog; z, zebrafish. Image directly from (Whitcomb, Basu et al. 2007).

## Pc homologs

Results indicate that: vertebrate model organisms have between three and five Pc homologs [known as Chromobox (Cbx)], which all have highly conserved CDs and Pc boxes (**Figures 4.1, 4.2**). However, paralogs differ greatly in length, such as Cbx7 with a protein length of approximately 200 amino acids; these factors might contribute to differential function (**Figures 4.1, 4.2**). Cbx proteins also specifically interact with non-PcG proteins. Cbx4 is the only member of the family that binds the transcriptional co-repressor C-terminal binding protein (CtBP) (Sewalt, Gunster et al. 1999). Cbx4 is also unique among Pc homologs as an E3 SUMO ligase (Kagey, Melhuish et al. 2003). The full range of Cbx4 SUMO targets is unknown, but the sumoylation of several transcriptional regulators, including CtBP, is enhanced by Cbx4 (Kagey, Melhuish et al. 2003), (Roscic, Moller et al. 2006). Additionally, recent biochemical data suggest that the five mammalian Cbx proteins have different histone-binding preferences: the Cbx CDs bind differentially to H3K27me3 and H3K9me3, unlike *Drosophila* Pc CD, which prefers H3K27me3 (Bernstein, Duncan et al. 2006).

## Psc homologs

Mel-18 and Bmi1 (two of six Psc homologs in mammals; **Figure 4.2**) are also likely to be non-redundant paralogs, despite their 63% amino acid sequence identity. These proteins overall have a low sequence similarity to

their *Drosophila Psc* homolog (12%). Bmi1- and Mel-18-deficient mice display similar but distinct phenotypes (Akasaka, Kanno et al. 1996) and (van der Lugt, Domen et al. 1994), and only ~30% of Bmi1-regulated genes were found to be co-regulated by Mel-18 and vice-versa (Wiederschain, Chen et al. 2007). Interestingly, sea urchin does not have a serine-proline rich domain.

### **dRing homologs**

Vertebrate homologs of dRing, Ring1A and Ring1B, also exhibit some functional divergence (**Figure 4.2**). Although they share long stretches of high conservation (approximately 100%), Ring1A- and Ring1B-deficient mice have drastically different phenotypes (Madireddi, Coyne et al. 1996).

### **E(z) homologs**

The mammalian organisms that I focused on have two E(z) homologs: Ezh1 and Ezh2 (**Figures 4.1, 4.2**). Little is known about the functional differences between these paralogs in mammals, but the ancestral *E(z)* gene also expanded in plant lineages (**Figure 4.1**). *Arabidopsis* has three *E(z)* homologs: *MEDEA (MEA)*, *CURLY LEAF (CLF)*, and *SWINGER (SWN)* with largely non-overlapping patterns of expression (Goodrich, Puangsomlee et al. 1997). CLF is the only protein among *Arabidopsis* that does not contain an E(z) domain.

## Vertebrates vs. Invertebrate Evolution

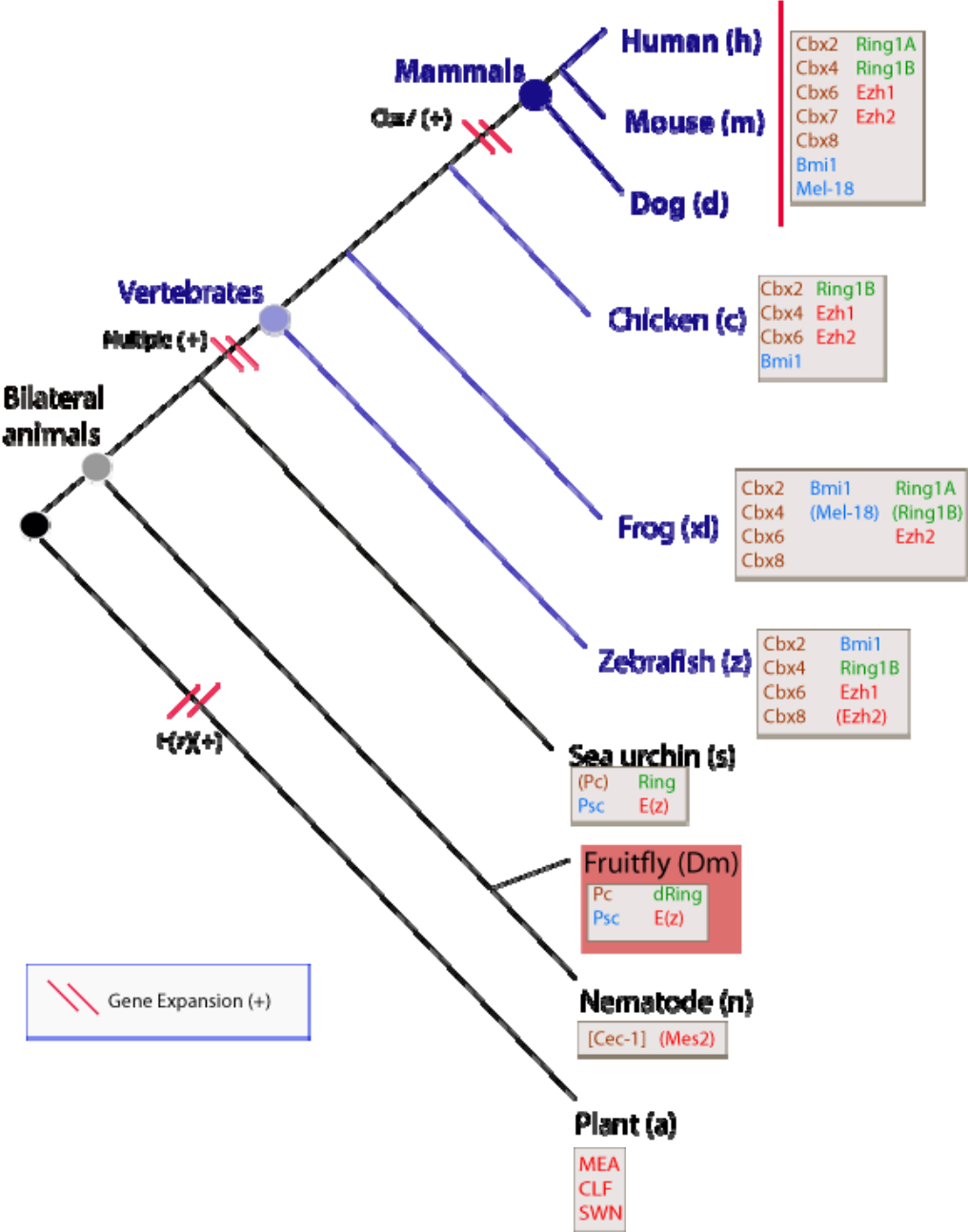
A common mechanism of evolution is gene duplication and subsequent divergence of coding sequences or regulatory elements. Based on the bioinformatics analysis, PcG genes are likely to have undergone multiple duplication events in their evolutionary history (**Figure 4.2**). Perhaps the most dynamic period was during the evolution of vertebrates from invertebrate ancestors. The extant invertebrates, *Drosophila* and sea urchin have single copies of the PcG proteins in the reference set, with the exception of Psc (**Figures 4.1, 4.2**). By contrast, vertebrate species have multiple paralogs of most PcG members (**Figure 4.2**). One striking example of PcG expansion is the Pc family. Represented by a single gene in invertebrates, there are up to five Pc homologs in vertebrates with differences in domain structure and biochemical properties (see below; **Figures 4.1, 4.2**). This could be due to massive gene duplication and exon shuffling.

**Figure 4.3: Phylogenetic representation of selected organisms and their PcG homologs.**

A phylogenetic tree of selected model organisms from plants to humans is shown (adapted from <http://www.tolweb.org/tree/>). This tree illustrates that PcG-encoding genes have undergone multiple duplication events through evolution; the most dynamic period appears to be during the evolution of vertebrates from invertebrates. PRC1 components seem to have been lost in *C. elegans*. However, CEC-1 might be a functional PRC1 homolog (see text for details). *Drosophila* proteins, used (here and in the text) as the PcG reference set, are highlighted in the red box. Shaded boxes next to each organism display homologs of E(z) (red), Pc (orange), Psc (blue) and dRing (green) proteins in each organism. Red slash marks represent probable gene expansion events. The black and grey nodes represent the common ancestor of all selected model organisms and extant bilateral animals, respectively. The light blue and dark blue nodes denote the common ancestor of extant vertebrate and mammalian species, respectively. Parentheses denote proteins with biochemically or genetically defined PcG activity but lacking sufficient sequence conservation with the *Drosophila*, mouse or human proteins to be predicted as homologs by methods used. Brackets indicate putative PcG proteins that were identified by sequence similarity but that need to be confirmed functionally. Asterisks represent proteins that might have multiple (putative) paralogs within a given organism. Note that branch lengths do not represent evolutionary distance between organisms.



Figure 4.3



## Prediction of a putative Pc gene in *C. elegans*

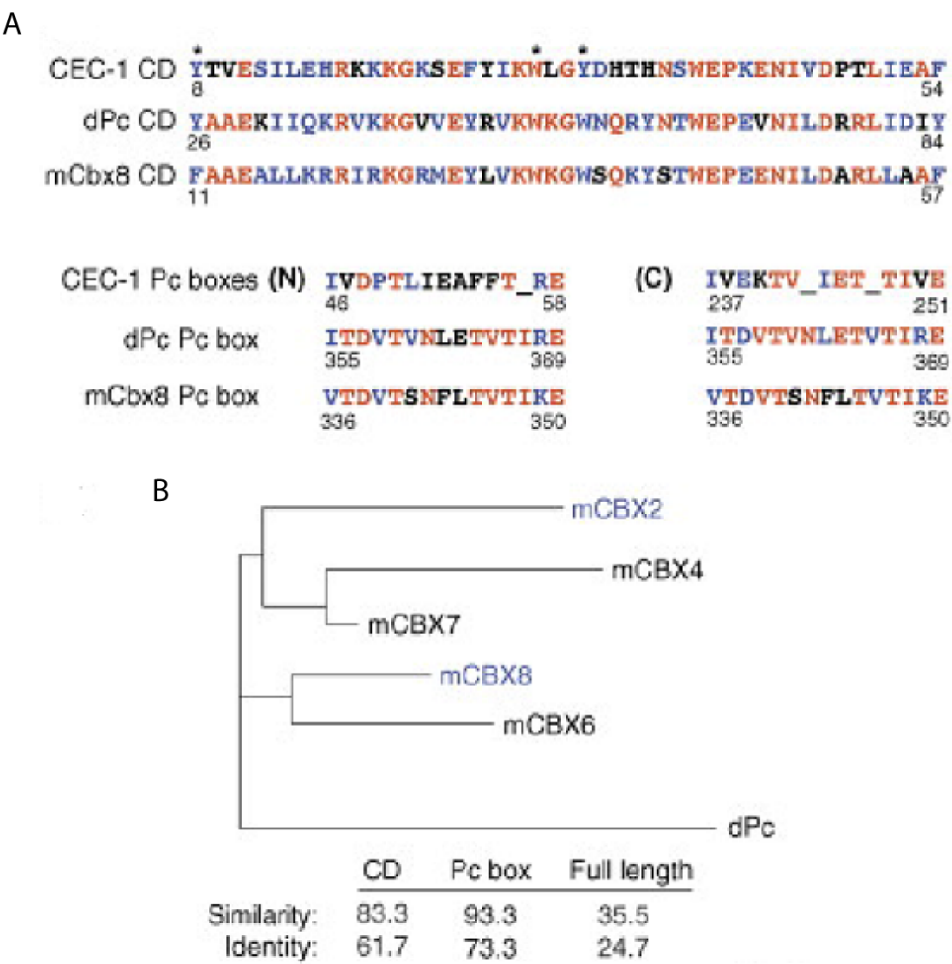
My analysis uncovered a putative Pc homolog in *C. elegans*, previously identified as *C. elegans* chromobox 1 (CEC-1) (Agostoni, Albertson et al. 1996). Little is known about CEC-1 except that it localizes exclusively to somatic nuclei and dissociates from chromosomes at mitosis (Agostoni, Albertson et al. 1996). The domain structure of CEC-1 supports its classification as a Pc homolog: an N-terminal CD and a C-terminal Pc box (as well as a second putative Pc box after the CD) (**Figures 4.1,4.3**). Although the sequence similarity between full-length CEC-1 and dPc or dHP1 are equal (27%), CEC-1 lacks two important sequence characteristics of HP1 proteins: a chromo-shadow domain and a stretch of glutamic acid residues N-terminal of the CD (**Figure 4.3**). The possibility that CEC-1 might regulate H3K27me-dependent gene repression in the worm soma is intriguing but requires further investigation.

**Figure 4.4: CEC-1 sequence alignment and phylogeny of mouse Cbx proteins.**

**A.** Sequence alignment is shown between the chromodomains (CD) and Pc boxes of dPc, mouse Cbx8 and CEC-1 proteins. An asterisk represents aromatic residues within the chromodomains that are required for histone methyllysine binding. Pc-box-like features are found in both the N-terminal and C-terminal regions of CEC-1 and are aligned individually to the Pc boxes of dPc and mouse Cbx8. (N) and (C) represent the N-terminal and C-terminal Pc-box-like features of CEC-1, respectively. Amino acids highlighted in red represent residues that are identical to each other, and those highlighted in blue represent residues that are evolutionarily similar. Overall, the bioinformatics analysis suggests that CEC-1 might represent a Pc homolog, rather than a HP1 homolog. However, this remains to be rigorously tested.

**B.** Pairwise sequence similarities were calculated between all mouse Pc proteins (Cbx 2, Cbx4 and Cbx6–Cbx8) and an unrooted neighbor-joining tree was constructed using the PHYLIP software program (<http://evolution.genetics.washington.edu/phylip.html>). Evolutionary distance between paralogs is represented by the tree branches, which are drawn to scale. The table below shows percentage sequence identity and percentage sequence similarity between the chromodomains, Pc boxes and the full length sequences of Cbx2 and Cbx8 proteins. These proteins were selected for comparison because of their comparable length and evolutionary distance. Note the high similarity and identity between key domains, but significantly less similarity and identity in the full length sequences.

Figure 4.4



## **Predicting chromatin–associated effector proteins**

Chapter 2 illustrates how acetylation has shown to be important for cellular function, and why being able to predict and map these marks is an essential first step in the discovery process of higher level regulatory pathways. Of paramount importance is the ability to uncover effector proteins that bind PTMs, as they serve as another layer of regulation in many cellular processes. Examples of additional effector proteins that bind to post-translationally modified amino acids are Plant Homeo Domain (PHD) finger, bromodomains (BD) (acetyl-lysine binding), chromodomains (CD) (methyl-lysine binding), and Src-homology 2 (SH2) domains (phospho-tyrosine binding), all which have been studied thoroughly. However, there likely exist a number of unidentified proteins containing cryptic versions of these domains or similar domains that have eluded detection by conventional bioinformatic analysis, thus additional studies of effector domains are required to identify these proteins and their functional relevance.

My interest in this project developed as a result of Allis lab colleagues and I extensively using the SMART/PFAM (Bateman, Coin et al. 2004) domain prediction programs and recognizing that certain domains that we considered as putative effectors by sequence/structure-gazing were not contained in either the SMART/PFAM databases, for example p300, a lysine acetyltransferase (personal communication, A. Ruthenburg) containing a

putative PHD finger domain. Manual screening of effector proteins that were not predicted by PFAM/Smart, together with domain querying of PcG proteins with limited results, led us to invest in a project aimed at discovering additional effector proteins.

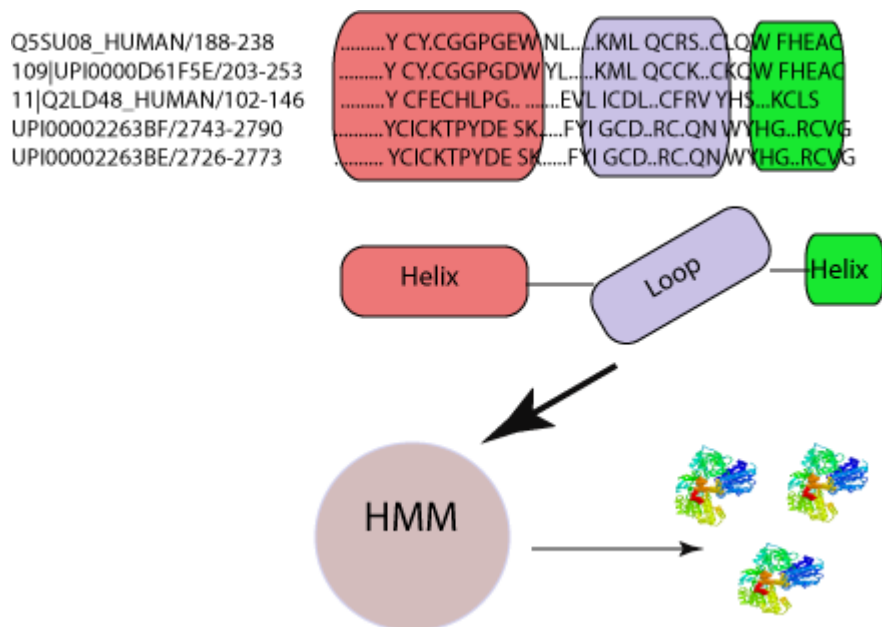
I hypothesized that software programs such as SMART and PFAM were often limited in their domain predictive ability due to default parameters usage in within all features of the protein. For example, loop structure, although more flexible than other regions a protein fold, such as the  $\alpha$ -helix, or  $\beta$ -strand, obtains an identical gap penalty score as the other structured elements of the protein. A tight conservation between loops within an alignment is less essential than conservation between the structured regions of the protein, since the binding pocket frequently resides on the surface of the protein ensconced within well-ordered secondary structural motifs. Moreover, gap penalties contribute to the overall score of alignments, and therefore, the size of the gap penalty relative to the entries affects the alignment that is finally selected. Though previous bioinformatic/structural studies have used a similar loop adjustment method in protein finding analyses (Qiu and Elber 2006), I wanted to focus on applying this method to PTM effector proteins for the benefit of our lab and the chromatin community at large.

## Training set, key assumptions, and method

In order to begin my computational analysis, I collected all sequences of known effector proteins BD, CD, SH2, and PHD finger containing sequences found in the SMART database. The SMART database contains more than 500 domain families found in signaling, extracellular and chromatin-associated proteins. These domains are extensively annotated with respect to phyletic distributions, functional class, tertiary structures and functionally important residues. The number of human training sequences in each of these respective domains was: BD (99), CD (70), SH2 (192), and PHD (256). I next performed a multiple sequence alignment (ClustalW) on all training sequences through three step process (**Figure 4.5**): 1) I used JPred, a secondary structure prediction program (built from a consensus of multiple prediction methods) (Cuff, Clamp et al. 1998) in order to label the loop element(s) in my sequences. 2) Applied a less rigid gap penalty exclusively in the loop region of protein, while applying a harsher penalty ( default PFAM parameters ) on the more rigid areas 3) Apply the same gap penalty to the entire length of the proteins in the alignment. Next, I fed the loop-adjusted and unadjusted version of the sequence alignment into the Hidden Markov Model 2.0 (HMM), a statistical model in which the system being modeled is assumed to be a Markov process with unobserved state. Running the HMM against MSA resulted in an output of additional effector proteins with an accompanying E-value. The resulting alignment is the

alignment between the consensus sequence which is derived from highest residue frequency at each position in the MSA and the predicted effector domain. Based on PFAM's threshold of selection, the E-value is the number of hits that would be expected to have a score equal or better than this by chance alone. A good E-value is much lower than 1, and measures statistical significance. In my analysis, I used an E-value  $\leq .01$  to threshold my boundary. After running my model against the human proteome, I manually scanned the list of proteins that were not annotated in literature with the desired domain, but which contained a maximum number of residues that were critical annotated for PTM binding. As a final step of verification, and work which is in progress, I analyze the known structures of the predicted effectors in order to ensure sure that the predicted effectors are homologous to the canonical structure.





**Figure 4.5: Schematic of domain prediction method**

Proteins with known effector domains (chromodomain, bromodomain, etc.) are separated by their secondary structure elements such as helix, loop, and beta strand. Loops within the proteins are given a less stringent gap penalty than the highly conserved regions in the protein such as the helix within the multiple sequence alignment. The alignment is fed into a Hidden Markov model (HMM) and then subsequently, the HMM produces an output of additional effector proteins.

Via my method, I was able to predict additional effector proteins containing all the domains I have specified above. I compiled a list of 7 putative candidates in the human and *S. cerevisiae* proteomes that contain previously unannotated BDs, CDs, PHD fingers, or SH2 domains, respectively (**shown in figures below**). Capital letters in the alignment represent residues that are aligned to the consensus. Small letters represent residues that HMM arbitrarily assigns based on a probabilistic function of residues in the alignment. For validation purposes, I used the few literature validated effector proteins that were excluded from my training set to assess whether my method was promising, particularly for PHD predictions.

### **PHD finger predictions**

At the time of my analysis, several of my predicted PHD finger proteins were not annotated in literature as PHD finger containing proteins, but have now been experimentally validated, and published. These include DNMT3 (Ooi, Qiu et al. 2007), RAG2 (Matthews, Kuo et al. 2007), ATRX (Baker, Allis et al. 2008) and others. In some cases the PHD finger domains appeared atypical by amino acid sequence, particularly for RAG2, which has an atypical aromatic cage, and an extremely divergent loop region (**Figure 1.7B**). In the case of RAG2, I achieved an E value of 0.5 before loop adjustment, and an E value of  $10e^{-12}$  after loop adjustment, which is much lower and of higher statistical significance. My finding is consistent with the finding that

RAG2 PHD finger behaves like a canonical H3K4me3 effector (Matthews, Kuo et al. 2007). **Figure 4.6** displays the alignment between the consensus HMM sequence (see **Methods** for determination) and the novel predicted PHD finger sequences (EP300, SNF2), those that are currently not in the database. Aromatic cage residues are highlighted as these are H3K4me3 recognition residues, as well zinc- coordinating residues. Interestingly, many of the predicted PHD finger proteins also contained bromodomains (shown below the alignment of the proteins in **Figure 4.6**).

### **Bromodomains**

In humans, I predicted two BD containing proteins, MLL2 and SP110 and in yeast I predicted Arp8, an actin related protein involved in chromatin remodeling (**Figure 4.6**). Among all of these, a promising bromodomain candidate protein was Arp8 (E-value=0.01) (**Figure 4.6**). This was also a compelling candidate because it was a yeast protein, where experiments could be facilitated by commercially available strains. Intriguingly, it also contained an important asparagine residue in a conserved histone binding acetyl-lysine position (in red). To determine whether the predicted BD in Arp8 could bind to acetylated lysine, I performed a yeast peptide-pulldown experiment using a commercially available tap-tagged strain of Arp8. The preliminary experiment shown in **Figure 4.7B** suggests that Arp8 binds selectively to acetyl peptides H3K4me3K14ac and H3K4me3K9ac, but not to

H3K56ac. Next, I cloned and recombinantly expressed full length Arp8 to determine whether Arp8 itself, or a possible complex member was responsible for peptide binding (data not shown). I was successfully able to achieve cloning of full length Arp8, however further experiments would be required to complete the experiment by cloning the bromodomain itself, and then subsequently repeating the peptide binding assay using additional unmodified and acetyl peptides.

## **Chromodomains**

The next group of proteins that I predicted are CD-containing proteins, a family containing different subgroups classes of proteins. The first class includes proteins having an N-terminal chromo domain followed by a region termed the chromo shadow domain, eg. *Drosophila* and human heterochromatin protein Su(var)205 (HP1). The second class includes proteins with a single chromo domain, eg. *Drosophila* protein Polycomb (Pc); mammalian modifier 3; human Mi-2 autoantigen and several yeast and *C.elegans* hypothetical proteins. In the third class, paired tandem chromo domains are found, eg. in mammalian DNA-binding/helicase proteins CHD-1 to CHD-4 and yeast protein CHD1. The training set included a conflation of all the CDs found in PFAM, and were not separated individually by family. I found two candidates (SMRC1 and MRG1), with the most promising predicted candidate, SMRC1, which appears most similar to the

chromodomain protein Cbx2 ( $E = 5.3 \times 10^{-7}$ ) among all other CD containing proteins in the training set. Also, SMRC1's sequence similarity to is HP1 ( $E = 5 \times 10^{-6}$ ), which is also statistically significant.

## **SH2**

SH2 domains are modules of  $\sim 100$  amino acids that bind to specific phospho-tyrosine (pY)-containing peptide motifs. In yeast, the only known protein with a SH2 domain is SPT6, a transcriptional elongation protein. However, there is no known function associated with the SH2 domain. My analysis led me to discover a SH2 domain containing protein, Ste6, a mating pathway protein in yeast ( $E = 5.2 \times 10^{-7}$ ).

**Figure 4.6: Effector Protein Predictions**

An alignment is shown between the consensus domain sequence and the predicted effector protein. Domain predictions displayed are the following: PHD, CD, BD, and SH2. Residues that are identical between sequences are shown. Underlined residues are those that are critical for histone binding. E value is shown for the predicted proteins as well as the residue positions in the predicted protein. Capital letters in the consensus sequence show that this residue is highly conserved, whereas the small letters represent residues that have been inserted via probabilistic method. Sequence similarity and identity (%) between the consensus and predicted protein is shown in the red box. Shown below alignment are the other domains in the protein as annotated in the SMART database.

Figure 4.6

PHD Predictions

hSNF2 (gi|51476310). E=0.13.residues 549-698

EC-ICGELDQIDRKPRVQCLK-CHLWQHAKCV---N-YDEKNLKI KPFYCPHCLVA  
yCsvC.....llqCDgaCdrwfHlaClqlstpeplte..Fe. WyCpeCkpk

Sequence similarity: 23.7  
Sequence identity: 18.6

EP300 (gi|3024341). E=1.residues 1200-1278

FCEKCFNEIQGESVSLGDDPSQPQTTINKEQFSKRKNDTLDPELFVECTECGRKMHCICVLHH.EIIPAGFVCDGCLKK  
yCsvC.....llqCDgaCdrwfHlaClqlstpeplte..Fe.Wy.CpeCkpk

Domains: bromo, 1048-1158, TAZ

Sequence similarity: 20  
Sequence identity:10

Chromodomain Predictions

hSMRC1 (gi|57012964). E= 5.3e-07.residues 215-262

vEkildvrwgvsdkdvekGevleiveeYLVKWkG.wsyshdtWepees  
DEEWLRP---VMRKE---KQV-----LVHW-GfYFDSYDTWVHSN  
eLelekrideklikeFkrkk  
DVDAEI-EDPP-IPF---KP

Sequence similarity: 40  
Sequence identity: 22

hMRG1 (gi|45767884). E=0.013.residues 27-78

vEkildvrwgvsdkdvekGevleiveeYLVKWkGwsyshdtWepees  
EAKCVK---VAIKD---KQV---KYFIHYSGWKNKNDWVVPESR  
LelekrideklikeFk.rkk  
V-LKY-VDTN-LQKQReLQK

Sequence similarity: 43  
Sequence identity: 23

Figure 4.6 (continued)

#### Bromodomain Predictions

hMLL2 (gi|37999860), E=4.2\*10e-20, residues 2941-3046

tellqeklerlllaikshkdswpflepvdapvvdeeeesGRllSelf  
SGALQGGLRQVLQGLLSSKVVGPLLLCTQCGPDG-----  
ielPskkeaPdYydiIkkPMDLsTikeklengkYrsveeFvaDvrLmFsN  
-----KQLHPGPCG-----LQAVSQRFEDGHYKSVHSFMEDMVGILMR  
aktYNepnpdSeiykaakkLekffeeeklkelep

Sequence similarity: 30  
Sequence identity: 17

HSE--EG--ETPDRRAGGQMKGLLLKLLESAPG

Domains: PhD finger, SET

hSp110 (gi|22256925), E=5.1e-17, residues 586-676

tellqeklerlllaikshkdswpflepvdapvvdeeeesGRllSelf  
QPQDQLKCEFLLLKAYCHPQSS-FFTGIIPFN-----  
ielPskkeaPdYydiIkkPMDLsTikeklengkYrsveeFvaDvrLmFsN  
-----IRDYGEPPQEAMWLDLVKERLITEMYT-VAWFVRDMRLMFRN  
aktYNepnpdSeiykaakkLekffeeeklkelep  
HKTFYK---ASDFGQVGLDLEAEFEKDLKDVLG

Sequence similarity: 38  
Sequence identity: 28

yArp8 (gi|6324715), E=.01, residues 374-468

leeilellqsyckdlswpFleFVdpkeyPkskerlpDYydiIkkPM.  
IEWIFDDSKL--YYGSDA-LRCVDEKFV-----IRKPFr  
..DLsTikeklenkSgkYrsveeFvaDvrLmFs...NaktYNepnpdSei  
ggSF-----NVKS--PYYS LAELISDVTKLLEhalNSETLV-KPTK-F  
..ykaakkLedvferqklkeiea  
ngYKVVLVIPDIFKKSHVETFIR

Sequence similarity: 35  
Sequence identity: 20

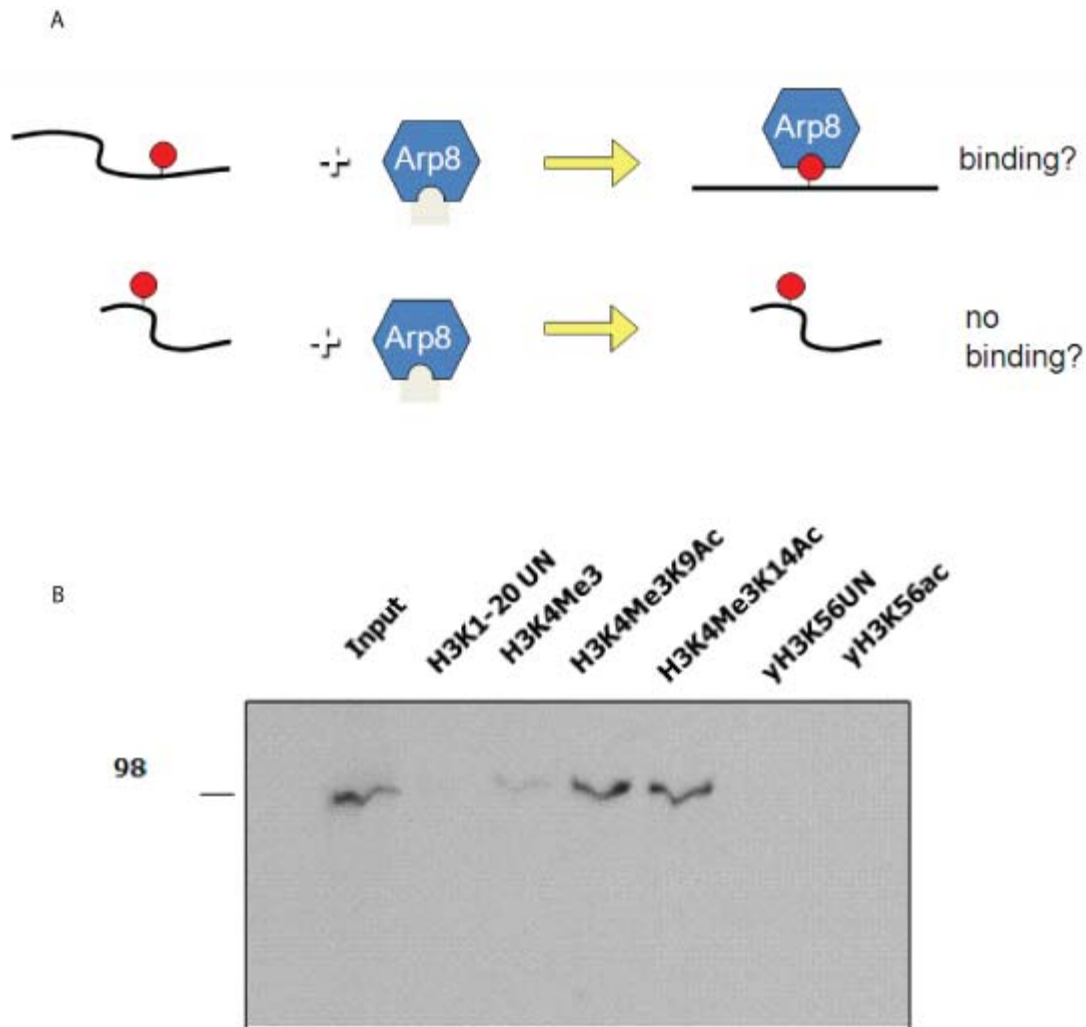
#### SH2 Predictions

ySTE4(gi|6324786), E=5.2e-7, residues 99-205

pWyhGklnisesRekkeAEeLLqaggkdGsFLvRdSeskkspG...  
RW-----SR---DSKRILSA-SQDGFMLIWDSAS----G1kqn  
.....dyvLSvrwkddahgqkk.tg...kGkpnlvkHyrIepvrtddg  
aipldsQWVLSAIS---PSSTLvASaglnN-----NCTIYR-VSKEN  
ks.k.....lyylgergrtreeyFdSLpeLvehYsknplg  
RVaQnvasifkghTCYISDIE-----FTDNAHILTASGDMTCA

Sequence similarity: 17.8  
Sequence identity: 10.7





**Figure 4.7: Peptide binding assay using Arp8 whole cell extract.**

**A.** Schematic of a peptide pulldown assay. The peptide is shown with either an acetylated mark or as unmodified. When incubated with Arp8, the question is whether binding occurs.

**B.** Shown are peptides H31-20 unmodified, H31-20 with H3K4Me3, H31-20 with K4Me3K9ac, H31-20 with H3K4Me3K14ac, H348-63 with K56 unmodified, and H348-63 with K56ac. Pulldown assay was performed with a DALK protein-A antibody and bands represent binding between the whole Arp8-TAP tagged and the respective peptides.

## Chapter 4 Discussion

The phylogenetic tree (**Figure 4.2**) argues that PcG genes underwent multiple duplication events in the evolution of plants and animals. One possibility for this is that these extra genomic copies diverged in sequence resulting in differential functions conferring fitness advantages. Here, the focus was on PcG gene expansion events, but I suspect that PcG proteins were also lost from genomes. A thorough comparative genomic analysis of loss and expansion from genomes can help us delve further into which genes were deleted through mechanisms of translocation or chromosomal inversion.

An outstanding question in the Polycomb/trithorax field is how these transcriptional regulator complexes are localized to specific genes in a tissue specific manner. In *Drosophila*, DNA elements known as PRE/TREs (Polycomb/trithorax response elements) have been mapped near the TSS of a handful of PcG/trxG target genes, and have been shown to be essential for proper PcG/trxG localization and expression of these genes. Attempts have been made to discover and map functional PRE/TREs in *Drosophila* using genomic, microarray, ChIP, and bioinformatic techniques. These studies identified a total of 167 novel PRE/TREs, some of which mapped to genes involved in development and cell proliferation (Ringrose, Rehmsmeier et al. 2003).

Unfortunately, the sequence characteristics of putative PRE/TREs in mammals have remained elusive. However, new technology such as ChIP coupled to Solexa sequencing has produced large, genome wide data sets relating to PcG occupancy and target gene expression. I present an extension of the Ringrose study that I propose in order to predict PREs in the human genome (Ringrose, Rehmsmeier et al. 2003). A brief outline of this is as follows: (1) use the current *Drosophila* annotated PRE data (167 sequences) to create a computational model for training purposes. Utilize empirical methods (supervised or unsupervised clustering) to learn about sequence constitution of known PRE elements. There could be multiple methods, but training and testing on the 167 elements themselves (using cross validation) could be an initial pilot to assess baseline accuracy. (2) Run the algorithm against the mammalian genome and make high quality predictions that can be tested in-vivo. (3) Perform ChIP experiments with PcG antibodies, and/or mine datasets already published to validate target DNA elements. It would be interesting to observe if validated mammalian PRE/TRE map to promoter regions as Ringrose and others observed in *Drosophila*. Identifying mammalian DNA elements necessary for PcG/trxG localization, if they indeed exist, would be an important step towards understanding tissue-specific epigenetic maintenance of both activated and silenced states.

## Unmodified H3K4me0 vs. H3K4me Effectors

Several papers have recently shown PHD finger-containing proteins binding H3K4me0. Recent studies display that DNMT3L, BHC80, TAF3, and AIRE among others bind H3K4me0 (Lan, Collins et al. 2007; Ooi, Qiu et al. 2007; Koh, Kuo et al. 2008; van Ingen, van Schaik et al. 2008). Co-crystal structures of these proteins with the H3 tail reveal crucial residues mediating this recognition. For example, the basis for binding of H3me0 to DNMT3L is the steric occlusion of the aspartic acid 90 in DNMT3L and H3K4me2/3 (**Figure 4.8A**). Other studies have revealed that the salt bridge between the unmodified lysine on H3 and acidic residues on the effector are crucial for favorable enthalpic contributions to binding free energy. Recently, the Yang Shi and Xiodong Cheng's groups showed that substrate specificity of the BHC80 PHD finger is determined primarily through the recognition of the H3 amino terminus, H3K4 and H3R8, and the three main chain carbonyl oxygen atoms (residues 523, 524 and 525) on BHC80 that form a hydrogen bond 'cage' that recognizes the N terminus of H3 (Lan, Collins et al. 2007). Molecular recognition of the unmodified lysine is primarily through bonds to the unmodified epsilon amino group and steric exclusion of appended methyl groups, where a second or third methyl group would engender steric clashes with the D489 carboxylate, the amide carbonyl of E488, and the  $\beta$ -carbon of the H487 side chain (van Ingen, van Schaik et al. 2008). Additional motifs include a Proline-(x)-Glycine -x - Tryptophan motif before the last pair of

Cysteines (**Figure 4.8B** in blue, c-terminal). **Figure 4.8** is an alignment of the current list of H3K4me0 binders.

In order to predict additional effectors using my method, I use the training method described above to predict new unmodified H3K4 binders, and use the known H3K4me0 as my training set of sequences (**Figure 4.8B**). Even though the training set is small (only five sequences), there are a number of conserved residues in these proteins, and for further screening of H3K4me0 binders, I manually scanned for proteins with these residues in the conserved positions. In my list (**Figure 4.8C**), many of the predicted proteins are either published or are being worked on by other groups, such as DNMT3, JARID1A, and AIRE (Ooi, Qiu et al. 2007; Koh, Kuo et al. 2008; Wang, Song et al. 2009). The published list of proteins gives me confidence that using more sophisticated methods, additional effectors can be discovered which could be regulatory in nature.

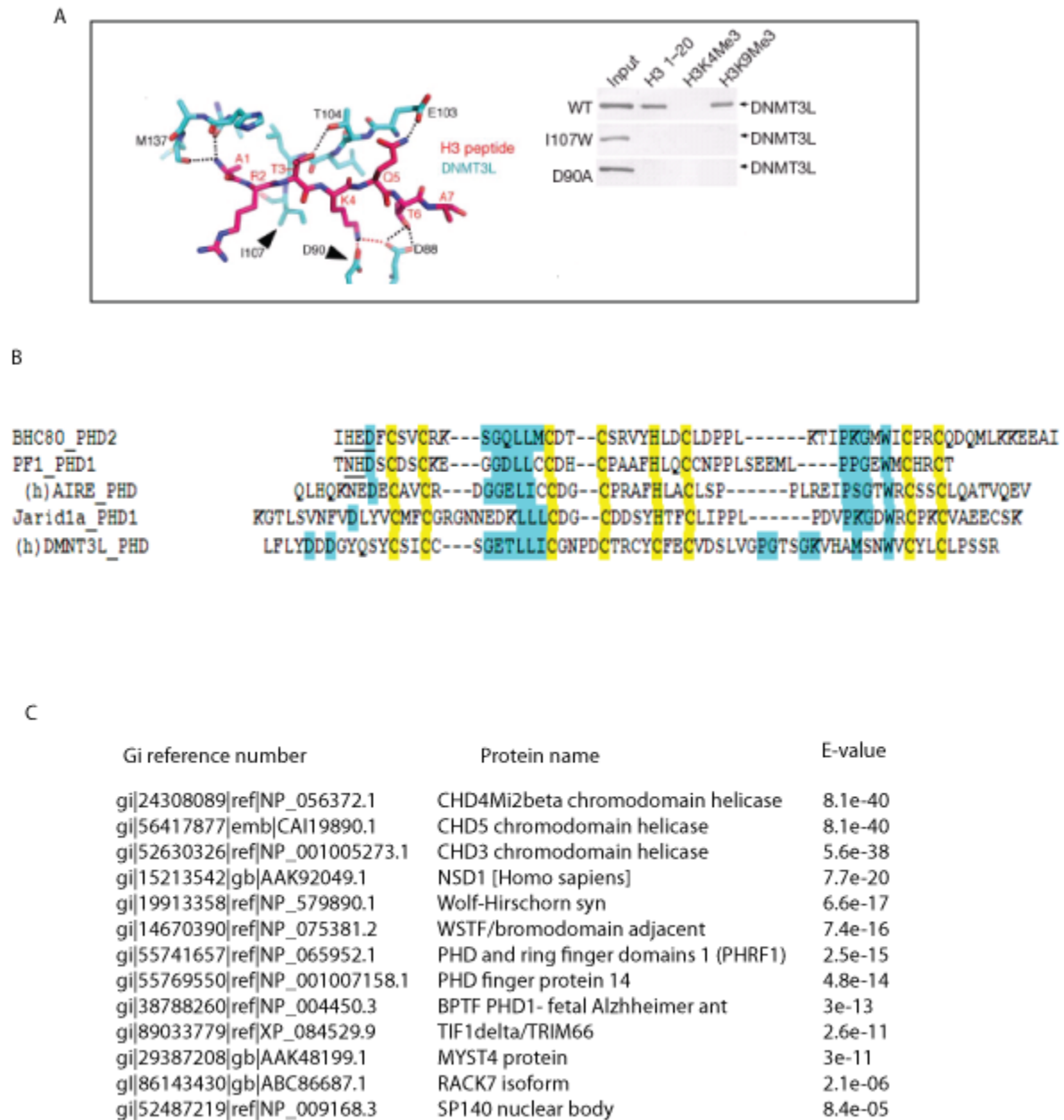
**Figure 4.8: H3K4MeO Known and Putative Effectors**

**A.** (Left) Interaction between the H3 N terminus (amino acids in magenta) and DNMT3L (in black). Dashed lines indicate potential interactions between amino-acid side chains. H3K4 makes contacts with DNMT3L. D90 and D88 and methylation of K4 will occlude these interactions. (Right) Mutagenesis of residues (tick marks on left) of DNMT3L abolished binding. Image adapted from (Ooi, Qiu et al. 2007).

**B.** Alignment of all known H3K4MeO binding proteins. Yellow represents identical residues; blue represents similar patches of residues. Alignment constructed using ClustalW.

**C.** Predicted H3K4meO binders (gi annotation to the left) with E-value resulting from Hidden Markov Model output are shown the right of the predicted protein.

Figure 4.8



Taken together, the studies I performed in this chapter reveal that there are still a number of undiscovered protein domains. Improved algorithms, more accurate prediction modeling, and higher throughput evolutionary analysis can benefit multi-domain classification problems, and potentially mitigate the gap between the various domain definitions.



## Chapter 5: General Discussion

A rapidly growing literature suggests that acetylation in histone and nonhistone proteins is important for a number of biological processes, including cellular differentiation, DNA binding, and other chromatin-templated processes. In the previous chapters, I have discussed my work developing a computational model that will enable the scientific community to characterize and map putative acetylation sites in any protein of interest. I have also shown using my model that we can predict acetylation sites in both histone and nonhistone proteins. Furthermore, I have shown that there are potentially key residues that are essential for acetyltransferase recognition, and that acetyltransferase activity might be regulated via crosstalk between modifications on the H3 tail. My computational model and results represent a step towards gaining a framework for predicting lysine acetylation sites in both human and yeast proteomes. It will be of interest in future studies to see whether our algorithm is capable of predicting lysine acetylation sites in other organisms. Below I will discuss some of the implications of my work.

## Enzyme prediction and discovery

Elegant structural studies by the Marmorstein group have shown that specific residues within KATs (Gcn5, Esa1) that form contacts between the enzyme and substrate are conserved among organisms (Marmorstein and Roth 2001). My algorithm agrees with the importance of these conserved residues, yet I was limited by the quantity of structural information available. Further structural analysis of various KAT-substrate relationships would allow us to determine whether the rules that we have observed are general acetyltransferase rules or enzyme specific principles. However, structural studies remain experimentally challenging and low-throughput by nature.

Alternatively, my prediction algorithm might be improved through additional experimental determination of KAT-substrate specificity, this in turn leading to an improved training dataset. To achieve a deeper understanding of KAT enzymes and their substrate specificity proteome-wide, I could perform the following experiments: first, on a small scale to establish the assay, and then-genome wide, I would design two identical proteome arrays containing histones H3, H4, and bovine serum albumin (BSA) as a negative control. Each array would then be incubated with radiolabeled acetyl-CoA and either of the histone acetyltransferases Gcn5 or Esa1, respectively. The array would then be subjected to fluorography, and

a positive result would entail Gcn5 labeling H3, and Esa1 labeling H4 (**Figure 5.1A**). I would not expect to see a signal in the BSA lane. If my pilot arrays confirm the expected results, I could carry out this experiment on a proteome-wide scale.

Further testing on magnified scale would involve protein microarrays containing 5800 yeast proteins incubated individually with all known yeast HATs (currently there are seven). Detection of incorporated radiolabeled acetyl-CoA would result in a number of possible KAT substrates. These substrates would then be validated from individually TAP-tagged lines. First, all acetylated proteins would be immunoprecipitated (IP) using a monoclonal pan-acetyl-lysine antibody (there are several commercially available). The presence of the potential HAT substrate would then be probed by immunoblotting with an anti-TAP antibody. Mass spectrometry would then be used to map the position of the acetylated lysine on the protein. A similar study was published recently by the Shelley Berger lab revealing NuA4's substrate specificity (Krishnamoorthy, Chen et al. 2006) (**Figure 5.1B**). A follow-up gene ontology (GO) analysis could indicate pathways or subset of pathways that the candidate acetyl proteins are enriched for.

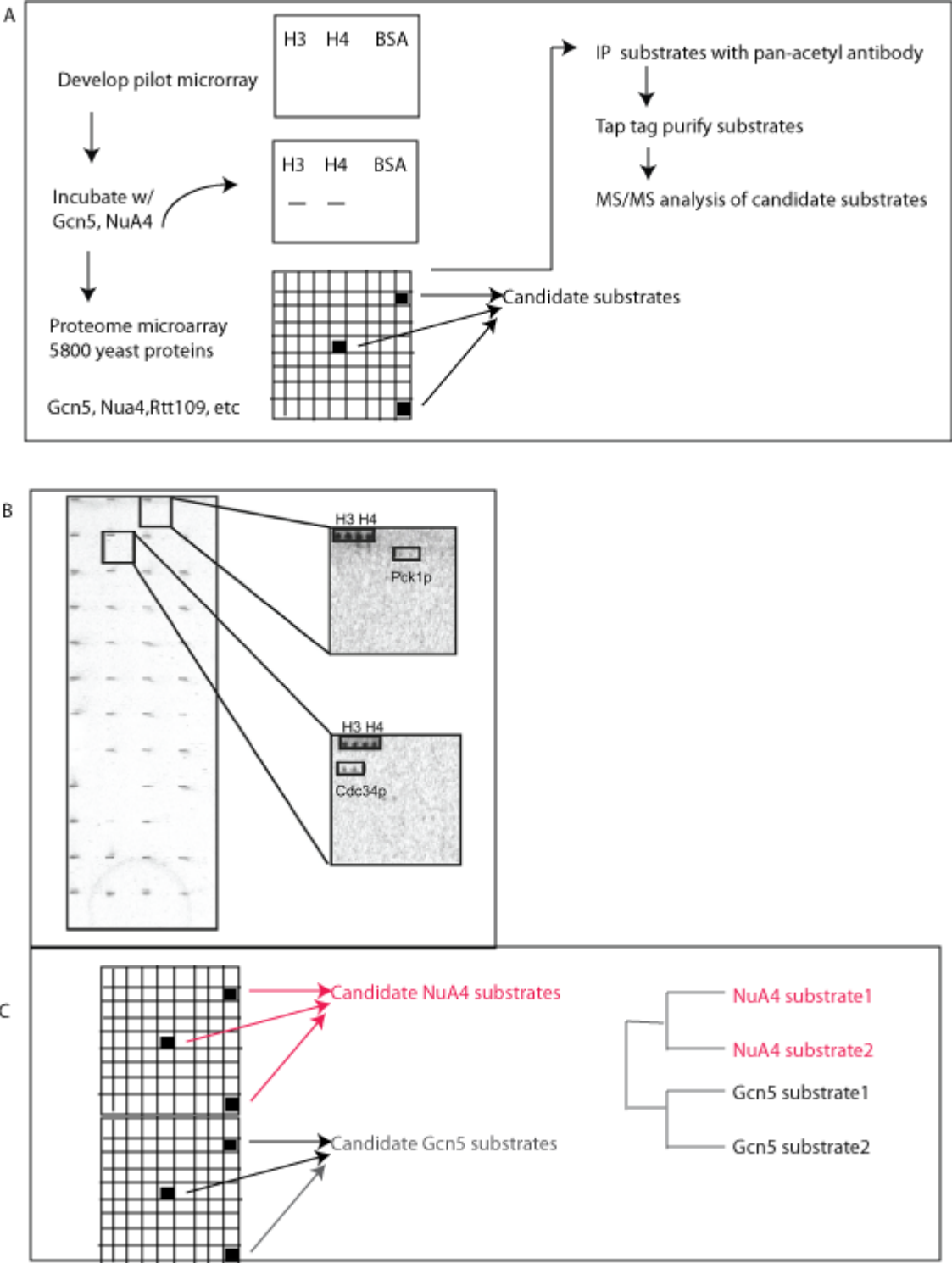
**Figure 5.1: Proposed discovery of novel KAT substrates**

**A.** Proteome-wide microarray of enzymatic discovery of substrates. In step 1, develop pilot protein array with H3 and H4 on the array. Incubate array with Gcn5 and Esa1. Once controls have been tested, include all 5800 proteins per array, and incubate array with each enzyme separately. Black boxes represent putative acetylated proteins as result of this assay, which become candidate substrates. An immunoprecipitation of substrates with pan-acetyl antibody followed by a TAP tag purification and subsequent MS/MS analysis could be performed to map the acetyl site.

**B.** Example of an array which includes 5800 proteins on an array. Acetylated substrates that were detected were Pck and Cdc34 among others (Lin, Lu et al. 2009).

**C.** Proposed computational model. Substrates with identical enzyme could potentially cluster together, and a separate tree could be built for all 8 KATs in yeast.

Figure 5.1



The results of these experiments could allow me to add enzymatic information onto the hierarchical tree (**Figure 2.2**), which could add predictive power to my algorithm. Potentially, these results would help uncover enzyme specificity for nonhistone acetyl substrates. I would expect substrates of the same KAT to cluster in my analysis, therefore if a substrate of an unknown HAT clusters tightly with a substrate of known HAT activity, I would expect these two to be substrates of the same HAT. This procedure could help narrow down whether there are specific clusters designated by enzyme, or whether there are acetylation rules in general, so that we can gain a further predictive understanding of our targets (**Figure 5.1C**). Furthermore, this experiment would allow me to gain insight into residues that are critical for specific enzyme recognition, which could ultimately feed back into the algorithm by changing the weight function of flanking residues.

In Chapter 3, I discussed my aim to induce methylation on H3K14 by mutagenizing residues *in vivo* to emulate the flanking residue profile of H3K36, a famous di and tri-methyl site. However, little is known about methyltransferase specificity and enzyme recognition. My initial computational analysis of methylation substrates displayed that a strong consensus sequence or signal was not present among the flanking residues of methyl-lysines, possibly due to limited experimental data. This could be due to the fact that methyltransferases have different selectivity mechanisms and that we may be underfitting the data by studying a

conflation of these substrates. Using the same protein microarray as for acetylation, methylation could also be studied by incubating the microarrays with radiolabeled S-adenosylmethionine (SAM) and various methyltransferases. Further computational analysis and more defined datasets could allow us to explore methylation rules as well.

### **Crosstalk of modifications on the same histone tail domain**

To my knowledge thus far, there have been few studies that have rigorously computationally or experimentally tested the necessity of specific flanking residues in maintenance of an acetyl mark. My results in Chapter 3 demonstrate that the flanking residue mutations of H3K14 can initiate crosstalk that can potentially occur between modifications on the same tail. It has been previously shown that acetylation of a particular lysine can be inhibited by adjacent PTMs (negative crosstalk), with the implication that the responsible KAT might be prevented from binding to or accessing its target site (Yang and Seto 2008). My mutagenesis studies on the flanking residues of H3K14 suggested that there could be a longer range crosstalk occurring between H3K14, H3K9, and H3K36, and perhaps these modifications are inhibitory to each other in some manner. Briefly, I observed that upon mutation of H3A15 to a lysine, H3K14ac levels decreased approximately 4 fold, while H3K9ac levels increased by two fold. H3K36me0 and H3K36me1

levels soared, while H3K36me2 and H3K36me3 decreased by approximately 1.5-2 fold.

To further study whether the lysine A15 is an enzyme specific residue, I could perform the experiment in reverse: select a lysine that is low in abundance and then mutate the “next door” residue to an alanine to determine whether the target residue acetylation levels are stimulated. An ideal lysine for this experiment would be H3K36, since it is of low abundance in yeast (Morris, Rao et al. 2007), and has a lysine (K37) adjacently positioned. Thereafter, I would extract histones as I previously did, and analyze their PTM levels by mass spectrometry. First, I could address whether the acetylation of K36 increases in intensity as a result of this mutation, and whether H3K37 is also modified (**Figure 5.2B,C**). Perhaps an altered acetylation state of H3K36 would recruit an enzymatic complex which could stimulate H3K37 acetylation or methylation states. Second, it would be interesting to see whether H3K9ac, H3K14ac, or any of the H3K36me0, me1, me2, and me3 marks are affected by the K37A mutation, since crosstalk has been shown to occur previously on histone tails (Latham and Dent 2007). Perhaps a rise in H3K36ac levels would prevent methylation to occur on H3K36 (**Figure 5.2C**), and H3K37A would carry a similar profile to H3A15K (**Figure 5.2C**) (Latham and Dent 2007). Another related question that could be asked is whether mutations in the flanking residues of H3K14 and H3K36 would alter transcriptional profiles genome-wide since all three



marks, H3K9ac, K14ac, and K36me have a link to transcriptional activation (Kurdistani, Tavazoie et al. 2004). My finding suggests that there might be a distinctive role for each acetylation, and it is now possible to examine the cross-talk of adjacent acetylation and methylation marks in vivo.

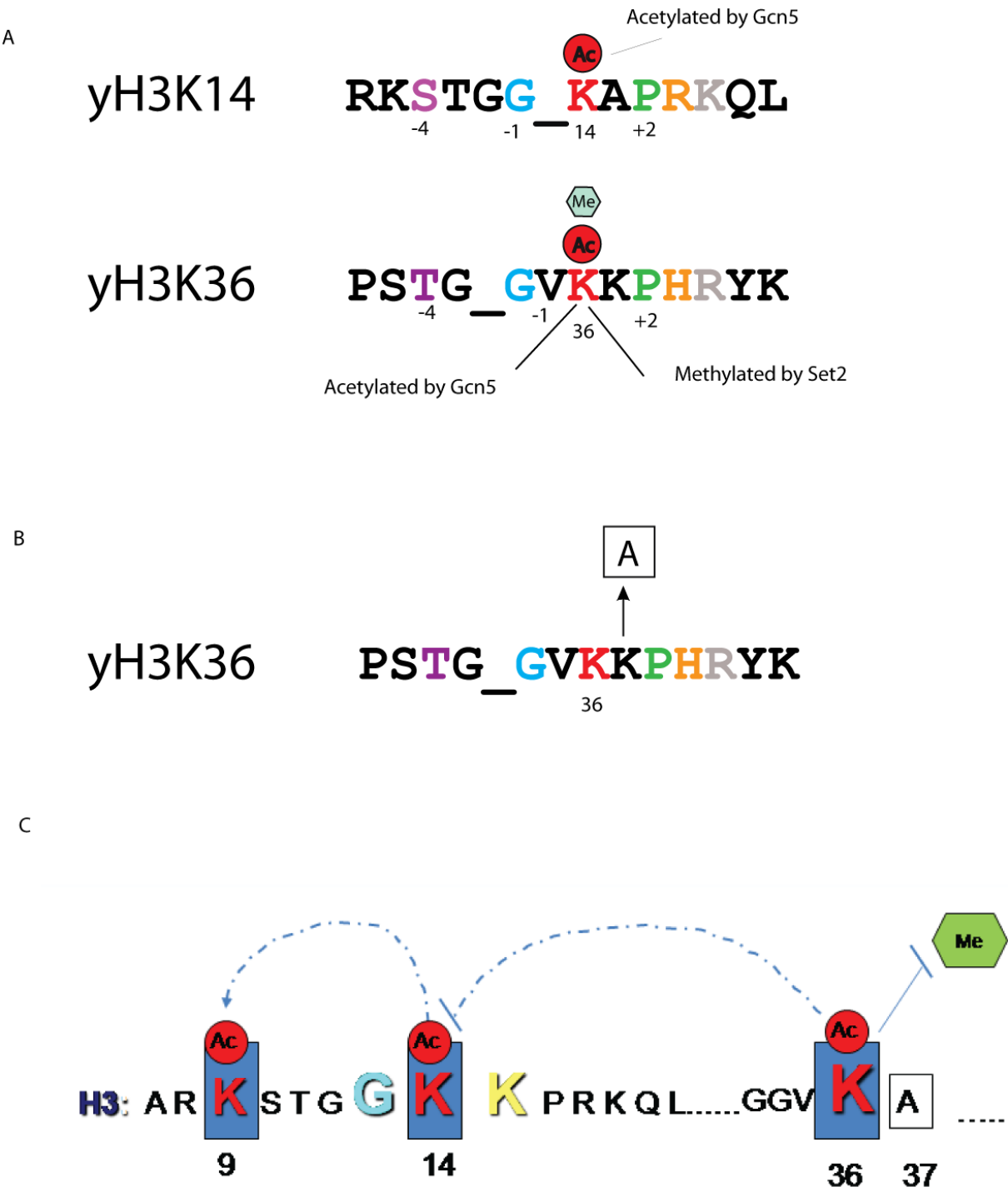
**Figure 5.2: Proposed future mutagenesis of H3K36**

**A.** Sequence alignment of H3K14 and H3K36. Colors as in Figure 3.3.

**B.** Proposed mutation of H3K37 to an alanine.

**C.** Proposed cross-talk effect of H3K37A mutation. H3K36 acetylation levels could increase, which could cause H3K36 methylation to decrease. H3K36 effects could regulate H3K14ac, by inhibiting its acetylation, which could allow for Gcn5 (H3K9 KAT) to acetylate H3K9.

Figure 5.2



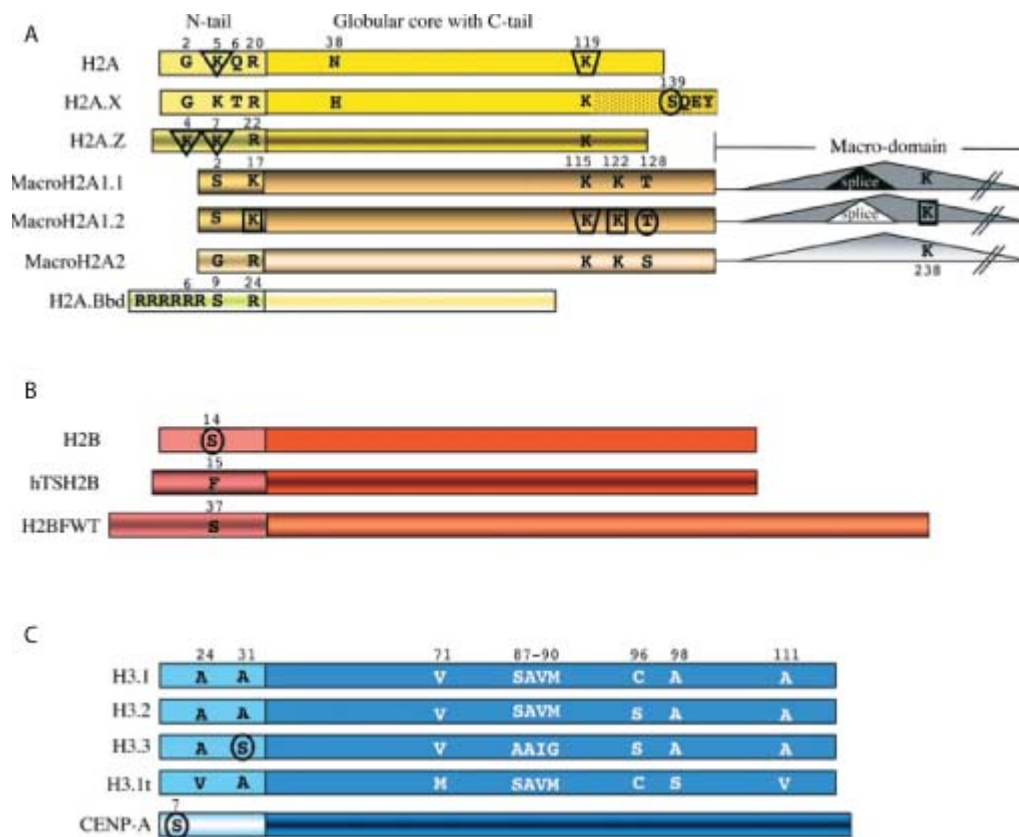
## Histone variants and PTMs

Changes in the chromatin template can occur through various interrelated mechanisms: post-translational modifications of histones, ATP-dependent chromatin remodeling, and the incorporation (or replacement) of specialized histone variants into chromatin (Bernstein and Hake 2006). These variants have specialized functions and in some cases are synthesized through the cell cycle (Bernstein and Hake 2006). While the two major human histone variants of H3, CENP-A and H3.3, have been extensively studied, the larger number of human H2A variants, including H2A.Z, H2A.X, and others, has been studied relatively less. There are at present three testis-specific H2B variants, and to date, no variants from the H4 family have been uncovered (Bernstein and Hake 2006) (**Figure 5.3**, courtesy Sandra B. Hake).

Since many variants have specialized functions and unique sequences, PTM profiles of these variants would potentially help gain a mechanistic insight into variant function. For example, phosphorylation on the serine in the H2A.X motif "SQ(E/D)" upon DNA damage is a phenomenon which is conserved across species (Rogakou, Pilch et al. 1998; Downs, Lowndes et al. 2000; Stiff, O'Driscoll et al. 2004; Xiao, Li et al. 2009). Predicting additional modifications on these proteins might help unravel further distinct regulatory functions or pathways. I attempted to predict modification patterns on H3 variants, however, as H3 variants are dissimilar to the canonical H3 by only

four or five amino acids, as shown in the alignment, the algorithm did not predict any additional modifications on this variant.

Human histone variants have been studied relatively in depth, however yeast histone variants are still an area that has not been as intensely explored. When I ran the yeast variant Cse4 (homolog of human CENP-A) against my algorithm, to my surprise, a protein which seems dissimilar by “eye” to the canonical H3, had a strong similarity (80%) to H3K56 in yeast (**Figure 5.4A,B**(red arrow)). Since the enzyme responsible for yeast H3K56ac has been discovered (Rtt109/p300) (Tang, Holbert et al. 2008) and functionally deciphered, Cse4 might also be acetylated by similar HATs, which would point to a sequence conservation of K56-type substrates. If Cse4 is indeed acetylated, one could then perform knockouts of the HATs themselves to determine whether Rtt109/p300 is responsible for Cse4 acetylation. This would also allow one to pinpoint residues that may be critical for Rtt109/p300 recognition.

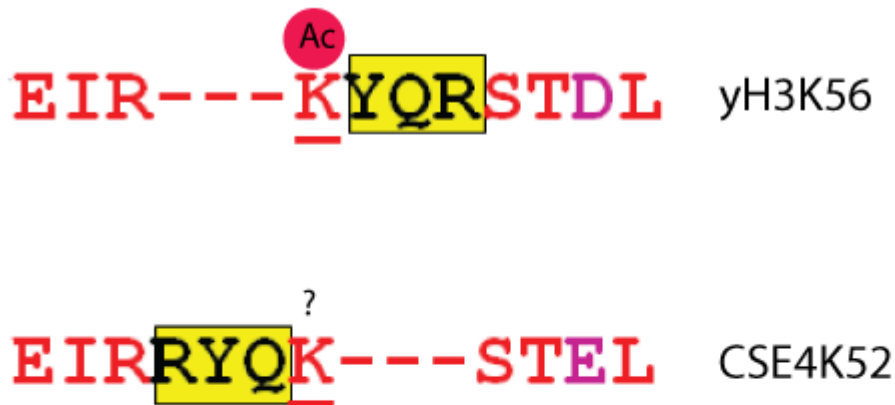


**Figure 5.3: Schematic of mammalian histone variant proteins**

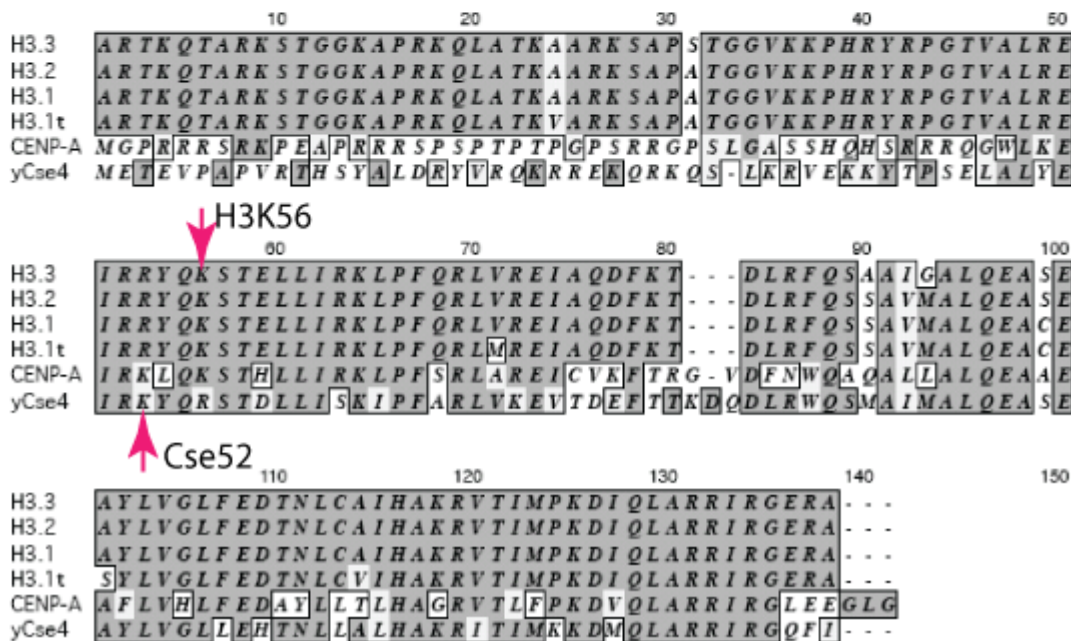
Schematic representation of the mammalian histone variant proteins containing the N-terminal tails and the globular core with the C-terminal tails: (A) H2A variants (yellow), (B) H2B variants (red), (C) H3 variants (blue), (D) H4 variants (green). Protein sequences that are highly divergent between the conventional histone and its variants, histones are depicted in different color shades without highlighting sequence differences. Specific amino acids are depicted when only a few key differences are found among variants, or when these amino acids are post-translationally modified. Residues that have been found to be post-translationally modified are marked in the following manner: circle, phosphorylation; square, methylation; triangle, acetylation; trapezoid, ubiquitination. The macrodomains of macroH2A histones are not drawn to scale and are shown as triangles to highlight that these domains are not histone-like sequences. Image adapted from (Bernstein and Hake 2006).

A

## Alignment of H3K56 and Cse4K52



B



**Figure 5.4: H3K56 and Cse4**

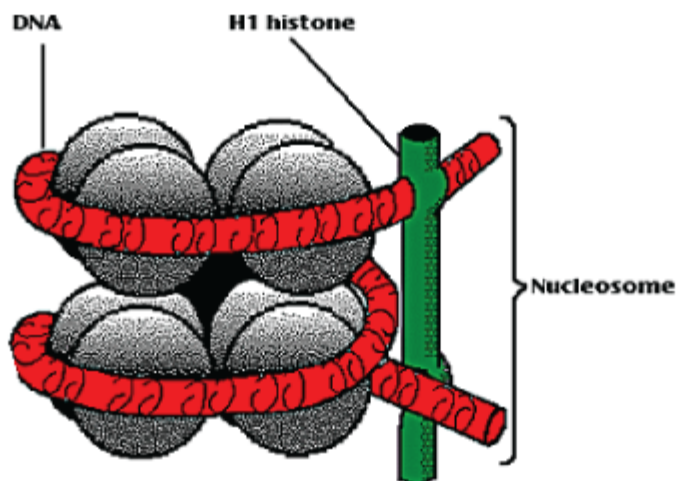
**A.** Alignment of H3K56 and Cse4 in budding yeast. Yellow highlighted residues represent patch of similar residues in a different order in H3K56 vs. Cse4. Purple residue represents similar amino acids, whereas red denotes identical residues.

**B.** Alignment of all H3 variants, hCENP-A, and yCse4. Arrows indicate H3K56 which is conserved across variants, and Cse4K52.

In addition to the four core histones, there is a fifth histone, linker histone H1, involved in nucleosome structure (**Figure 5.5A**). These proteins are crucial to nucleosome-nucleosome contact and higher order chromatin structure. It is possible that post-translational modifications could regulate their contact, causing dramatic chromatin structure changes, including decreased global nucleosome spacing, reduced local chromatin compaction, and decreases in certain core histone modifications. A multiple sequence alignment of the H1 variants is shown **Figure 5.5B**, and as displayed these variants are conserved in the core region, but divergent within the tail region. To predict additional modifications as well as experimentally validated modifications, I ran the yeast and human histone variants including H1 against my algorithm. Interestingly, out of 5 known H1 validated acetyl sites, I was able to predict 3 with my algorithm (these included K90, K146, and K22) (**Figure 5.5B**; red arrows) (Wisniewski, Zougman et al. 2007). Notably, these acetylation sites all contained a flanking glycine and basic residues in the vicinity as well. Further validation of the predicted unknown modifications might lead to an important understanding of the role of acetylation in H1 function. Finally, genomic and proteomic studies, evolutionary analysis using bioinformatics, and experimental approaches in model organisms will provide new insights into the biological roles of these variant proteins.



A



B

	10	20	30	40	50			
H1.1	MSETVP	PAPAA	SAAPEK	PLAGKKK	K - - - - KPAKAAAASKKKPA	GSPVSE		
H1.2	MSETAPA	APAA	APPAEK	APVKKKA	- - - - - AKKAGGTP	- RKASGPPVSE		
H1.3	MSETAPL	AP	TI	PAPAEK	TPVKKK	- - - - - AKKAGATAG	KRKASGPPVSE	
H1.4	MSETAPA	APAA	PAPAEK	TPVKKKA	- - - - - RKASAGAAK	- RKASGPPVSE		
H1.5	MSETAPA	ETATP	APVEK	SPAKKKA	T - - - - KKAAGAGAA	KRKATGPPVSE		
H1.X	MSVLE	EEAL	LPV	TTAEGMA	KKVTKA	GGSAA	LSPSKKRKNSKKKNQPGKY	SQ
H1.0	MTE	NSTS	APAA	KP	- - - - - KRAKA	SKKS	- - - - - TDHPKY	SD

	60	70	80	90	100				
H1.1	LIVQ	AA	SS	SKER	G	GVSLAALK	-	KALAAAGYDVEKNNSRIKLG	LKSLVSKG
H1.2	LITKAVA	AASKER	S	GVSLAALK	-	KALAAAGYDVEKNNSRIKLG	LKSLVSKG		
H1.3	LITKAVA	AASKER	S	GVSLAALK	-	KALAAAGYDVEKNNSRIKLG	LKSLVSKG		
H1.4	LITKAVA	AASKER	S	GVSLAALK	-	KALAAAGYDVEKNNSRIKLG	LKSLVSKG		
H1.5	LITKAVA	AASKER	N	GLSLAALK	-	KALAAAGYDVEKNNSRIKLG	LKSLVSKG		
H1.X	LVVETI	IRRLGER	NG	SSLA	KI	YTEAKKV	PWF	DQONGRTY	LKYSIKALVQND
H1.0	MIVAI	QAEKN	RAG	S	SRQS	IQ	-	KYIKSHYKV	GENADSQIKLSIKRLVTTG

	110	120	130	140	150											
H1.1	TLVQ	TKGTGAS	SGSFKLNKKA	S	SVE	TKPGAS	SKV	-	-	ATK	TKA	TGAS	KKL	KKA		
H1.2	TLVQ	TKGTGAS	SGSFKLNKKA	AS	SGEAK	PKVKKAGG	TKP	K	K	K	P	V	GAA	KPKKA		
H1.3	TLVQ	TKGTGAS	SGSFKLNKKA	AS	SGEGE	PKAKKAGAA	K	P	R	K	P	A	GAA	KPKKV		
H1.4	TLVQ	TKGTGAS	SGSFKLNKKA	AS	SGEAK	PKAKKAGAA	K	A	K	K	P	A	GAA	KPKKA		
H1.5	TLVQ	TKGTGAS	SGSFKLNKKA	AS	SGEAK	PKAKKAGAA	K	A	K	K	P	A	GAT	-	PKKA	
H1.X	TLLQ	VKGTGAS	NGSFKLNRE	K	LEG	-	-	-	-	GGER	RGA	P	AA	TAPAPT	AHAKKA	
H1.0	V	LKQ	TKGTGAS	SGSFR	LAK	SDEPK	KS	VAFK	K	TKKE	I	K	VAT	PKKA	S	KPKKA

**Figure 5.5: H1 Schematic and Predictions**

- A. Nucleosome made up of DNA (red), Histone octamer (black) and Histone H1 (green).
- B. Alignment of all H1 variants, CENP-A, and yCse4. Arrows indicate acetylated lysines that I predicted correctly in histone H1.

Many questions remain as to whether acetylation, like phosphorylation, is a key signaling mechanism for multiple cellular pathways. Mann's study reveals that a striking feature of acetylation is that it tends to occur in large macromolecular complexes involved in diverse cellular processes (Choudhary, Kumar et al. 2009). With higher throughput proteomic datasets becoming available, we can begin to dissect the role of distinct acetylation marks across multiple cell lines and analyze their importance in cellular pathways, signaling, and gene regulation. Answering these questions can help us achieve state of the art epigenetic drug therapy targeted towards specific acetylation marks, however mapping these marks is a critical first step. Understanding the crosstalk between acetylation and other PTMs such as methylation and phosphorylation remains to be answered will potentially be the subject of future research endeavors.

## Chapter 6: Materials and Methods

### **Acetylation (computational)**

#### **Datasets**

Training set: 56 human and *S. cerevisiae* core histone lysine sequences were collected from the Swissprot database <http://ca.expasy.org/sprot/>.

Test set: Source of nuclear protein and pan-acetyl antibody datasets are described in Chapter 2. The budding yeast proteome-wide dataset of observed peptides with acetyl modifications was derived from the publicly available GPM Annotated Spectrum Library (Craig, Cortens et al. 2006) for *S. cerevisiae* (v. 2008.10.1). Peptide sequences from the library corresponding to lysine modifications (nominal modification mass = 42 Da.) were curated and mapped onto the appropriate set of protein sequences. Given the mass resolution of many of the spectra used to create the library, it was not possible to distinguish a priori between acetyl-lysine versus trimethyl-lysine on the basis of the tandem spectra alone.

#### **Hierarchical Clustering Analysis**

I performed hierarchical clustering on the sequences surrounding each of the 56 histone lysines. All 56 sequences were aligned to each other creating a matrix of pairwise alignment scores; the metric was based on these pairwise scores. Sequence alignment scores were computed by performing

BLAST local alignments using the NCBI BLAST 2.0 server. A standard BLOSUM62 evolutionary substitution matrix was applied (Eddy 2004). The hierarchical clustering works in an iterative process with the sequence alignment score representing the metric value: it begins with each protein sequence as a singleton cluster; during each iteration, it finds two clusters with the lowest metric value, then joins these two clusters into a new cluster, and updates the metric value between this new cluster and all others (see text for details). An average alignment score is calculated when there are multiple leaves under a node, thereby assigning a single metric value to each node. As an example, H2AK127, a lysine not observed as acetylated, and H4K12, an acetylated lysine, clustered tightly via sequence alignment. This is illustrated by their shared node placing them to right of the threshold line (**Figure 2.2A**; blue box). As a result, H2AK127 was predicted to be acetylated via my approach. In contrast, H2BK27 clustered with H2BK11 more weakly, and thus their shared node was positioned on the left side of the threshold line. Hence, this lysine H2BK27 was not predicted as acetylated.

## Statistical Analysis

ROC calculations are described in main text. Hypergeometric probability calculation:  $Pr = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}$ ; (N = all lysines in human proteome; K = number of times the particular residue is seen flanking in each position in

human proteome;  $n$  = total number of lysines in each independent validation dataset;  $m$  = number of times the particular residue is seen flanking in each position in validation dataset). Sensitivity ( $S_n$ ) was calculated as the total number of correctly identified acetylation sites from the positive dataset divided by the total positive dataset. Specificity ( $S_p$ ) was calculated as the total number of negative sites that were not predicted to be acetylated divided by the total negative dataset size. Accuracy calculated as the number lysines correctly predicted as acetylated/total number of acetylated lysines in dataset.

### **Additional Methods of Classification**

When I first began the project, I used a Support Vector Machine Classifier (SVM Light) to analyze my data (Yu, Joachims et al. 2008). I used a linear kernel function, which measures the similarity between a pair of inputs, and defines an inner product in the feature space. My feature space consisted of  $k$ -mers, where  $k = (2..6)$ . These  $k$ -mers were extracted from the residues surrounding the flanking lysine. The results of this method (75% accuracy on histone lysines, where a lysine was either classified as “validated” or “not observed” as acetylated) encouraged me to proceed and use additional machine learning methods empirically. Hence, I also used a mismatch kernel (Leslie, Eskin et al. 2004), which consists of a class of string kernels for use with support vector machines (SVMs) to measure sequence similarity

allowing for mutations between patterns. Thus, if two protein sequences contain many k-length subsequences that contain m, mismatches, the inner product one would expect is large. The mismatch kernel did not alter my results significantly as compared to the linear kernel. Other types of kernels I used were the profile and quadratic kernels, however the best results were achieved using the linear and mismatch kernels.

### **Weighting**

As described in Chapter 2, I weighted residues to the left and to the right of target lysines within the sequence alignment. The weight was added to the raw alignment score such that the  $T_s = \text{Raw alignment score} + \sum w_i$  of each identical or evolutionarily similar flanking residue, where  $T_s$  = total score, and  $w_i = 1/d$ ;  $d$  = position of flanking residue with respect to lysine. I also grouped chemically similar residues together, such that residues that may not be evolutionarily similar according to the BLOSUM matrix, but within the same chemical group, such as small, aromatic, hydrophobic, etc. are also given weight to. Using chemical similarity improved the algorithm performance by a five to eight percent margin (accuracy measure) for both histones and non-histone proteins, and thus I included chemical similarity in my optimal tree analysis.

## **Sequence logos**

Sequence logos for displaying the flanking residue distribution of all lysines in my training and test datasets were created according to (Crooks, Hon et al. 2004).

## **Software URL**

The acetylation prediction software, PredMod can be found at

<http://www.cs.cornell.edu/w8/~amrita/predmod.html>

## **Cross Validation and Evaluation of Test Sets**

I used Leave One out cross Validation (LOV) of histone lysines on the training set and performed a ROC analysis. The predictive performance was monitored using the AUC metric (Dodd and Pepe 2003) on the test lysines. I applied the procedure to 1000 random permutations of the labels of the observed and not observed lysines. The independent datasets were also measured by a ROC analysis, as described previously.

## **Human Nonhistone Analysis Method**

In order to measure whether the prediction method also works on nonhistone lysines, I first applied my algorithm to a list of validated lysine acetylation sites in human nonhistone proteins. I used the following two independent validation datasets for my analysis: Firstly, a proteomics

survey of published cytosolic and nuclear protein fractions from Hela cells that were subjected to immunoaffinity purification using a pan-acetyl lysine antibody was applied. Isolated peptides were analyzed by HPLC-MS/MS and the final output contained 51 acetylated lysines in 38 proteins (Kim, Sprung et al. 2006). Secondly, I screened the literature for acetylated proteins and found 32 molecules containing 73 acetylated lysines that were reported as acetylated both *in vitro* and/or *in vivo* using mass spectrometry and immunoblotting detection methods. The test data sets were composed of lysines that were identified as acetylated and represent my positive set ("validated"), and lysines that were not observed as acetylated within these investigated substrates represented my negative dataset ("not observed"). By comparing validated acetyl marks to my computationally predicted acetylation marks, we tested whether sites are dictated by the surrounding amino acid sequences to validate my stated assumption that there is an intrinsic substrate specificity for KATs true for both histone and nonhistone proteins.

## **Acetylation (Experimental)**

### **Cell Lines**

Mammalian cell lines were grown in Iscove's DMEM supplemented with 10% fetal calf serum and penicillin/streptomycin at 37°C and 5% CO<sub>2</sub>.



## **Human Histone Purification**

Cell nuclei were isolated by hypotonic lysis in buffer containing 10 mM Tris-HCl (pH 8.0), 1 mM KCl, 1.5 mM MgCl<sub>2</sub>, 1 mM DTT, 0.4 mM PMSF, and protease and phosphatase inhibitors. Pelleted nuclei were extracted using 0.4 M sulfuric acid. The acid-soluble histones were precipitated with trichloroacetic acid and resuspended in water. Histones were separated by reverse phase HPLC using a C8 column (220 by 4.6 mm Aquapore RP-300, PerkinElmer Life Sciences) with a linear ascending gradient of 35–60% solvent B (solvent A: 5% acetonitrile, 0.1% trifluoroacetic acid, solvent B: 90% acetonitrile) over 75 min at 1.0 ml/min on a Beckman Coulter System Gold 126 Pump Module and 166/168 Detector. For the additional data that was obtained from sodium butyrate treated samples, histones were purified from HeLa S3 cells and treated with 10mM sodium butyrate for 15 hmys prior to harvesting (STable 6). H3 was also purified from HEK293 cells treated with sodium butyrate (see below).

## **Mass Spectrometric Analysis of Human Histones H2A and H3**

Histones were isolated and H2A was purified as described previously (Shechter, Dormann et al. 2007) from HL60 cells. The HPLC fraction containing purified H2A was resuspended in water, and an aliquot was diluted with 100 mM ammonium bicarbonate, pH 8. The diluted aliquot was divided in half. One half was digested with GluC protease (Princeton

Separations, Inc, Adelphia, NJ) at a substrate-to-enzyme ratio of 20:1 for 4 h at RT, and glacial acetic acid was used to quench the digest. An aliquot of the GluC-digested H2A containing approximately 1 picomole of H2A was diluted 3-fold with ammonium bicarbonate to increase the pH to 8.0. H2A peptides were then derivatized by treatment with deuterio succinimido acetate. This reagent was created by adding deuterio acetic anhydride and triethylamine to N-Hydroxysuccinimide, and the white precipitate was collected after rinsing with hexanes. 0.35mg ( $\approx$  2 micromoles) of deuterio succinimido acetate was added to the H2A peptide mixture, containing an estimated 20 picomoles of free amino groups. The deuterio-acetylation reagent and H2A peptides were allowed to react for 2 h at 4 °C. Deuterio-acetylation of the N-termini of peptides and the epsilon amino groups of lysine residues increases the hydrophobicity of the H2A peptides, and allows the smaller, hydrophilic peptides of H2A to be retained on a C18 column. Instead of adding a 42-Da acetyl group (C<sub>2</sub>H<sub>3</sub>O), this reagent adds a 45-Da acetyl group (C<sub>2</sub>D<sub>3</sub>O) since it contains 3 deuteriums instead of 3 hydrogens. Therefore, an *in vivo* lysine acetylation is distinguished from a deuterio-acetylation because of their difference in mass.

After performing this reaction, the mixture was acidified with glacial acetic acid and loaded onto a capillary precolumn (360  $\mu$ m O.D. x 75  $\mu$ m I.D. fused silica, Polymicro Technologies, Phoenix, AZ) packed with irregular 5–20  $\mu$ m C18 resin (YMC Inc., Wilmington, NC). The precolumn was then connected to

an analytical column packed with regular 5  $\mu$ m C18 resin equipped with an electrospray tip. H2A peptides were separated using nanoflow HPLC on an 1100 series binary HPLC pump (Agilent Technologies, Palo Alto, CA) coupled to a micro-electrospray ionization source on a Finnigan LTQ Orbitrap mass spectrometer (Thermo Scientific, San Jose, CA). The HPLC gradient consisted of 0-60 %B in 50 min and 60-100 %B in 10 min (solvent A: 0.1 M acetic acid, solvent B: 70 % acetonitrile, 0.1 M acetic acid), with a flow rate of 60 nL/min. Full mass spectra were acquired with the Orbitrap as the analyzer, and MS/MS spectra were acquired in the LTQ ion trap. After each full MS scan, m/z values of 535.6, 537.1, 513.1, 514.6, 534.1, and 511.6 were targeted sequentially for isolation and fragmentation. The last scan event in the cycle (prior to acquisition of the next full MS scan) was a data-dependent MS/MS scan of the most abundant ion in the previously acquired full MS scan. MS/MS spectra were manually interpreted. This approach employing GluC digestion and deuterio-acetylation was used to examine the C-terminal H2A peptide, residues 122-129.

The second half of the H2A aliquot was treated with propionic anhydride to derivatize endogenously monomethylated and unmodified epsilon amino groups of lysine residues. Chemical derivatization with propionic anhydride converts amino groups of lysines to their corresponding propionyl amides and has been detailed previously (Garcia, Mollah et al. 2007). Briefly, equal volumes of propionylation reagent and protein were reacted, and

derivatization was repeated twice to ensure full conversion of amino groups. The sample was vacuum-dried after each derivatization. H2A was then digested with trypsin (Promega, Madison, WI) at a substrate-to-enzyme ratio of 20:1 for 7 h at 37 °C. Derivatization blocks lysine residues from cleavage, and thus trypsin cleaves C-terminal to arginine residues only. The resulting H2A peptide mixture was acidified with glacial acetic acid and loaded onto a reverse phase capillary column for LC-MS/MS analysis as described above. H2A peptides were analyzed using a hybrid quadrupole linear ion trap Fourier transform (LTQ-FT) mass spectrometer (Thermo Scientific, San Jose, CA). The LTQ-FT instrument was operated in data-dependent mode with dynamic exclusion enabled. The data-dependent method consisted of acquisition of a full scan mass spectrum using the FT as analyzer followed by ten MS/MS scans of the ten most abundant ions in the initial full scan. MS/MS scans were acquired using the ion trap as the analyzer and spectra manually interpreted. This approach employing propionylation and trypsin digestion was used to examine N-terminal H2A peptides, 4-11 and 12-17.

Histone H3 was purified as described previously (Garcia, Barber et al. 2005; Hake, Garcia et al. 2006) from sodium butyrate-treated HEK293 cells. H3 was treated with propionic anhydride and digested with trypsin, similar to the procedure described for H2A. Following digestion, the samples were again reacted with propionic anhydride to derivatize the amino-termini of the

trypsin-generated H3 peptides. The peptide mixture was dried in a speed-vac concentrator, and subsequently reconstituted in 0.1% acetic acid. H3 peptides were loaded onto a capillary reverse phase precolumn, and the precolumn was connected to an analytical column as described above. Peptides were gradient-eluted into an LTQ Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose, CA). The instrument was operated in data dependent mode and cycled through acquisition of a full mass scan followed by MS/MS scans of the ten most abundant peptide cations in the initial full scan. This approach was used to examine the N-terminal H3 peptide, 27-40. All MS/MS spectra were manually interpreted.

### **Mass spectrometric analysis of nonhistone lysine acetylation sites**

Tagged cells of my nonhistone proteins were lysed under cryogenic conditions. Tandem Tap-tag purification was performed on candidate yeast proteins as described (Puig, Caspary et al. 2001) and eluates run on SDS-PAGE gels and stained with coomassie. Protein bands were in-gel digested with trypsin or chymotrypsin, and peptides extracted.

Each of the protein bands were cut into two pieces with similar size and washed with 10% acidic acid in 50% ethanol (acetic acid:ethanol:water = 10:50:40 (v/v/v)) three times and then overnight. After destaining three times with 25 mM ammonium bicarbonate in ethanol buffer (ethanol:water = 50:50 (v/v)), the gel bands were swollen in water twice and then cut into

small pieces. They were dehydrated in acetonitrile and then dried in a SpeedVac (Thermo electron, Waltham, MA). Overnight digestion was performed at 37 °C with about 200 ng of modified porcine trypsin (Promega, Madison, WI) or chymotrypsin (Roche, Indianapolis, IN) in 50 mM ammonium bicarbonate. The resulting peptides were extracted sequentially with 5% TFA/50% acetonitrile/45% water (v/v/v), and 0.1% TFA/75% acetonitrile/24.9% water (v/v/v), and 100% acetonitrile. The extracts were combined and dried in a SpeedVac. The resulting peptides were cleaned with C18 ZipTips (Millipore, Bedford, MA) according to the manufacturer's instructions, prior to nano-HPLC/mass spectrometric analysis.

The extracted peptides were separated using a capillary HPLC column (11 mm length  $\times$  75  $\mu$ m I.D., 4  $\mu$ m particle size, 90 Å pore diameter) packed in-house with Jupiter C12 resin (Phenomenex, Torrance, CA). LC-MS/MS analysis was performed in an integrated system that includes an Agilent 1100 series nanoflow LC system (Agilent, Palo Alto, CA) and a LTQ 2D trap mass spectrometer (Thermo Electron, Waltham, MA) equipped with a nanoelectrospray ionization (NSI) source. The gradient-eluted peptides were electrosprayed directly into the LTQ mass spectrometer, which was operated in a data-dependent mode. Mascot (version 2.1, Matrix Science, London, U.K.) was used for database searching. Acetylated lysine containing peptides identified with a Mascot score of 20 or above were manually verified by a method described previously (Chen, Kwon et al. 2005).

## **Yeast Strains and Plasmids.**

Strain MSY421 from M. M. Smith (University of Virginia) [*MAT* a,  $\Delta(hht1-hhf1)$   $\Delta(hht2-hhf2)$  *leu2-3, 112, ura3-62, trp1, his3*, pMS329 (*HHT1-HHF1, URA3, CEN*)] was used to shuffle in plasmids containing histone H3 A15K, H3G13V, and H3A15V mutations using 5-FOA as a counterselecting agent for the *URA3* plasmid. The mutant plasmids were generated by PCR mutagenesis (H3A15K, H3A15V, H3G13V) and confirmed by sequencing. The final yeast strain was confirmed by PCR amplification of the *HHT2-HHF2* locus and DNA sequencing. Yeast growth, plasmid and DNA fragment transformation of yeast cells were done according to standard yeast protocols (Hieter et al, Methods in Yeast Genetics).

## **Extraction of Histones**

Histones were extracted from nuclei made from 1 liter of cells grown to an  $OD_{600}$  of 0.8 in yeast-rich media. Nuclei were prepared in the presence of protease inhibitors, and histones were obtained by acid extraction as described in (Hsu, Sun et al. 2000). One-tenth of each fraction was used to confirm the presence of purified histones by SDS/PAGE separation and Coomassie blue staining. The remainder of the fraction was used for MS analysis.

## **Mass Spectrometry of Yeast Histones**

Histones were prepared using propionic anhydride reagent and digested with trypsin as described previously (Garcia, Mollah et al. 2007). For mass spectrometry analysis, histones were loaded on to C18 packed columns and separated using an Agilent 1200 series HPLC system (Agilent Technologies), and the LC gradient used was 5% B to 45% B over 60 min. All data was acquired on an LTQ-Orbitrap as previously described (Leroy, Toubreau et al. 2006).

## **Domain Prediction (computational)**

I performed the domain prediction analysis by creating a multiple sequence alignment of known effector proteins of my domain of interest using the software program ClustalW. Sequences that were fed into my MSA were based on a concatenated sequence, ie. If the domain consisted of a helix, loop, helix, then I ensured that the loop region of the sequence had a lower gap penalty, while the helix part of the sequence had a more stringent gap penalty (JPred) (Cole, Barber et al. 2008). The MSA was then fed into a Hidden Markov Model 2.0 which calculated the probability of the output sequence based on a probabilistic weight analysis of input sequences. The final result I obtained was an E-value of the statistical significance of my prediction given the selected parameters. I compared this value against an



unadjusted loop, and so I obtained a list of predictions for the unadjusted and adjusted loops.

### **PcG domain detection**

I selected E(z), Pc, dRing and Psc as my PcG reference set: the “writer” (E(z)) and the “reader” (Pc) of the H3K27me mark, the only other known catalytic member of the PcG (dRing), and dRing’s associated factor (Psc). Sequence and domain information was retrieved for *Drosophila*, mouse, and human PcG proteins from the Swissprot, Uniprot, Ensembl, and SMART databases. In order to identify putative homologs of these proteins in other organisms, I performed BLAST searches using the *Drosophila*, mouse and human proteins as queries<sup>2, 3</sup>. The Blastp searches, using the NCBI web server (<http://www.ncbi.nlm.nih.gov/BLAST>) were performed at low, medium, and high stringencies to obtain all possible proteins in the following organisms: *Arabidopsis thaliana*, *Caenorhabditis elegans* (worm), *Strongylocentrotus purpuratus* (sea urchin), *Danio rerio* (zebrafish), *Xenopus laevis* (frog), *Gallus Gallus* (chicken), *Canis familiaris* (dog), *Mus musculus* (mouse), and *Homo sapiens* (human). Since homologous domains between my species are very well conserved, variation of BLAST parameters, such as gap costs and substitution matrices did not significantly alter my results. In addition, I developed a program (Motif Search) to query motifs less than 15 amino acids in length such as the Pc box. Sequence similarity between

homologous domains and full-length sequences was used to compute a weighted score by multiplying a fixed weight to the BLAST score and computing an overall weighted sum. Key domains for PcG proteins were identified including: the chromodomain and Pc-box for Pc homologs, the RING domain for Psc and dRing homologs, and the SET and SANT domains for E(z) homologs. A list of putative homologs containing the highest scores was compiled for each class of PcG proteins, and homologs in different species were selected by manual inspection. Finally, the ClustalW program was used to cluster putative homologs to ensure that proteins with similar key domains with minor amino acid differences such as Ring1A and Ring1B were categorized correctly. In addition, a reverse BLAST of all the putative homologs was also performed against the *Drosophila* genome to ensure that the forward BLAST hit was accurate. Many databases of translated cDNA libraries are inherently incomplete, and so gene sequences of those PcG homologs that were found to be absent in a given organism were also queried against the organism's genome. Finally, putative PcG homologs were mapped onto a phylogenetic tree of organisms. Additional potential paralogs of dPsc were found in *Drosophila* through my algorithm. These proteins are Su(z)2 and l(3)-73Ah; however as they are not well characterized as functional Psc homologs, these proteins have been excluded these proteins from my phylogenetic tree.

Sequence alignments were constructed using the Emboss program and CEC-1 was predicted as a putative Polycomb homolog because of its high sequence similarity with *Drosophila* and mouse Cbx8 chromodomains. Two stretches of amino acids within the N-terminus and C-terminus of CEC-1 also shared a strong sequence similarity with the *Drosophila* and mouse Cbx8 Pc boxes. In **Figure 4.5B**, the ClustalW program was used to create a multiple sequence alignment of Cbx proteins and subsequently fed into the PHYLIP software program<sup>5</sup>. The protdist and neighbor joining programs within PHYLIP were used to construct the Cbx paralog tree (**Figure 4.5B**) and the Emboss local alignment program was used to determine sequence similarity and identity percentages.

### **Domain Prediction (Experimental)**

#### **Peptide Pulldown Assay**

Extracts were made by taking frozen, lysed cells (1 g/ pull-down condition) and extracting them in 500mM extraction buffer (500 mM NaCl, 20 mM HEPES pH 7.9, 25% glycerol, 1.5 mM MgCl<sub>2</sub>, 0.2 mM EDTA, 1 mM PMSF, Complete Mini EDTA-free [Roche], 0.2% Triton X-100) for 1 hour at 4°C. Extracts were then diluted to 150 mM NaCl with 'no-salt' extraction buffer, mixed with 2.5 µg of biotinylated histone peptide-linked Dynabeads (M-280 Streptavidin, Dynal) at the ratio of 2.5 µg /100 µL beads and nutated for 30 minutes at 4°C. Peptide-linked Dynabeads and associated proteins were

then washed five times in 300 mM KCl wash buffer (300 mM KCl, 20 mM HEPES pH 7.9, 0.2% Triton X-100), and one time in a buffer containing 4 mM Hepes pH 7.5 and 10 mM NaCl. Peptide-bound proteins were eluted in boiling SDS-PAGE loading buffer, resolved on Novex 4-20% gradient gels, and probed with antibodies recognizing the PrA tag (DAKOP0450). Peptides were synthesized by Upstate Biotech (UBI) and Proteomics Resource Center of The Rockefeller University.

## **Appendix**

### **Collaborations**

During my PhD, I had a chance to collaborate with a variety of groups on and off the Rockefeller campus. A group that I collaborated with was Dr. Dmitri Krainc's group, at Harvard Medical School (Cambridge, MA). His group was interested in looking at human Htt (a Huntington disease related protein), and was interested in studying PTMs on this specific protein. I contributed to this project by using my software tool, PredMod, to predict several acetylation sites on this protein. My top ranked predicted lysine was K444, which eventually was shown to facilitate trafficking into autophagosomes, and had an effect on neurodegeneration in cultured neurons and in mouse brain (Jeong, Then et al. 2009).

A second collaboration that I was involved in with the lab of Dr. Elliott Hertzberg at Albert Einstein College of Medicine, who was involved in looking at a series of Connexin proteins (gap junction proteins) in rat cells. My top predictions aligned with their primary detected acetyl-sites in Connexin23. They are in now the process of conducting further experiments to determine the acetyltransferase of the lysine mark.

## Software Tools

During my graduate studies, I developed PredMod as described in the computational section of my thesis. The usage of this tool is as follows:

**Usage of PredMod.** After you have inserted the sequence of your protein of interest, press "Submit" and the following output will appear: The **first** section displays the entire sequence with each lysine colored red or blue. A red colored lysine indicates that the lysine is predicted as acetylated, and a blue colored lysine suggests that the lysine is not acetylated. Listed below the protein sequence are the total number of predicted acetylated lysines, not predicted as acetylated lysines, and a total lysine count. In the **second** section, you will see that there are two tables of proteins, one table shaded in yellow and one table shaded in blue. Table columns are as follows: Position of lysine in protein, confidence (calculation described below), and the flanking sequence with the target lysine bolded. The yellow shaded list contains **predicted** acetylated lysines, and the blue shaded list lysines that are **not predicted** as acetylated. Note that the confidences are in ranked order. The top-most yellow shaded lysine is the lysine predicted as acetylated with the highest confidence. The bottom-most blue shaded lysine is least likely to be acetylated. Note that this program runs at a lower stringency threshold than results reported in the text (high stringency), and thus the output contains a higher number of positive predictions.

Confidence levels are calculated by the difference between the output score and the optimal calculated threshold. Thus, if a lysine scores above the designated threshold, it is predicted as an acetylated lysine. Confidence values range between 0.5 and 67, where 67 most likely represents an identical match between a histone sequence existing in the training set and an identical input sequence.

Another tool, useful for the lab was Motif Search, a software program which enables a user to enter any length motif, and then this motif is searched for in the entire proteome of all organisms. The output received is something like this:

Together, PredMod along with Motif Search can potentially help the scientific community obtain additional information on their “favorite protein.”

## All Tables

Table represents lysines in human histones H3, H4, H2B, H2A and their acetylation status as noted in literature (references are in rightmost column). Asterisk (\*) represents lysines that were predicted by the algorithm and validated experimentally by mass spectrometry.

### Human Histone H3.1 (NP\_003520.1)

<u>Lysine residue</u>	<u>Validated as Acetylated</u>	<u>Not Observed as Acetylated</u>	<u>Reference(s)</u>
K4	X		(Zhang, Tang et al. 2002; Zhang and Tang 2003; Zhang, Eugeni et al. 2003; Garcia, Barber et al. 2005; Hake, Garcia et al. 2006; Garcia, Hake et al. 2007)
K9	X		(Cocklin and Wang 2003; Beck, Nielsen et al. 2006)
K14	X		(Cocklin and Wang 2003; Beck, Nielsen et al. 2006; Kim, Sprung et al. 2006)
K18	X		(Garcia, Hake et al. 2007),(Hake, Garcia et al. 2006),(Zhang and Tang 2003),(Beck, Nielsen et al. 2006),(Kim, Sprung et al. 2006)
K23	X		(Garcia, Hake et al. 2007),(Hake, Garcia et al. 2006),(Zhang and Tang 2003),(Beck, Nielsen et al. 2006),(Kim, Sprung et al. 2006)
K27	X		(Cocklin and Wang 2003; Beck, Nielsen et al.



			2006)
K36	X		(Morris, Rao et al. 2007)
K37			*(sod but)
K56	X		(Xie, Song et al. 2009),(Das, Lucia et al. 2009)
K79	X		(Garcia, Hake et al. 2007)
K64		X	
K115		X	
K122		X	

#### **Human Histone H4 (NP\_003529.1)**

<b><u>Lysine residue</u></b>	<b><u>Validated as Acetylated</u></b>	<b><u>Not Observed as Acetylated</u></b>	<b><u>References</u></b>
K5	X		(Zhang, Williams et al. 2002; Smith, Gafken et al. 2003; Pesavento, Kim et al. 2004)
K8	X		(Zhang, Williams et al. 2002; Smith, Gafken et al. 2003; Pesavento, Kim et al. 2004)
K12	X		(Zhang, Williams et al. 2002; Smith, Gafken et al. 2003; Pesavento, Kim et al. 2004)
K16	X		(Zhang, Williams et al. 2002; Smith, Gafken et al. 2003; Pesavento,

			Kim et al. 2004)
K20		X	
K31	X		(1)
K44		X	
K59		X	
K77		X	
K79		X	
K91		X	

#### **Human Histone H2B.2(Q) (NP\_003519.1)**

<b><u>Lysine residue</u></b>	<b><u>Validated as Acetylated</u></b>	<b><u>Not Observed as Acetylated</u></b>	<b><u>References</u></b>
K5	X		(Beck, Nielsen et al. 2006),(Kim, Sprung et al. 2006)
K11	X		(Beck, Nielsen et al. 2006),(Kim, Sprung et al. 2006)
K12	X		(Bonenfant, Coulot et al. 2006)
K15	X		(Bonenfant, Coulot et al. 2006)
K16	X		(Beck, Nielsen et al. 2006),(Kim, Sprung et al. 2006)
K20	X		(Beck, Nielsen et al. 2006),(Kim, Sprung et al. 2006),(Bonenfant, Coulot et al. 2006)
K23	X		(Kim, Sprung et al. 2006)
K24	X		(9)
K37		X	
K30		X	
K34		X	
K43		X	
K46		X	
K57		X	
K85		X	
K108		X	
K116			*(sod but)
K120		X	
K125		X	

# Human Histone H2A.C (NP\_003503.1)

<u>Lysine residue</u>	<u>Validated as Acetylated</u>	<u>Not observed as Acetylated</u>	<u>References</u>
K5	X		(Bonenfant, Coulot et al. 2006)
K9			*
K13			*
K15			*
K36		X	
K75		X	
K76		X	
K95 (R)		X	
K118		X	
K119		X	
K125 (R)			*
K127(R)			*
K129			*

Note (R) stands for reversed sequences that were used in the alignment

H2AK125	KTESHH <b>K</b> AKGK	(R)KGKA <b>K</b> HHSETK
H2AK125	ESHHKA <b>K</b> GK	(R)KGKA <b>K</b> HHSE
H2AK96	NDEELN <b>K</b> LLGRVT	(R)TVRGLLL <b>K</b> NLEEDN

(sod but) =This modification observed under sodium butyrate treatment

## Human pan-acetyl IP substrates

Table includes published proteins that were immunoprecipitated with a pan-acetyl antibody (Kim, Sprung et al. 2006). Asterisk (\*) represents lysines that were not predicted by the algorithm but experimentally identified by mass spectrometry.

Protein/accession number	Total residues	Lysines	Predicted/validated (position on protein)	Predicted but not validated experimentally by algorithm (position on protein) at threshold	*
Set-translocation (gi:4506891)	277	18		7,10,159	
HrNP (gi:14043070)	372	18	350	144,145,52	
SMARCA (gi 55958983)	1570	121	1531,1533, 1535	999, 1003 , 472, 1366, 984	
P29ING4 (gi 18873723)	249	28	130	129, 156, 160	127*
CREB binding (peptide info only known)			6		
DNMT1 (gi:62088406)	1601	137	1100, 1098, 1102,1104	135,1589,160,944	
SON DNA BASS1 (gi 5737751)	344	26	206	144, 157, 98	117*
HMG1 (gi 96888)	215	43		7,8,114,177	28*
MYST3 (gi 150378493)	2004	139		1165, 365,1014,1154	815*
MLL3(insufficient support for the transcript and the protein)					
EF1alpha (gi 62896589)	462	47		385, 386	318*
PHD15 (gi 18676594)	639	39		605, 617, 604	109*
Heat shock 90kd Protein 1 Beta (gi 34304590)	724	75	624	148, 505	
Heat shock 27 KD (gi 4504517)	205	6	123	171, 198, 112	
Chaperonin (gi 1137641)	539	40	20	509, 364, 528	
Peptidyl-prolyl isomerase(cyclophilin) (gi 10863927)	165	14	125	44,49,76	
Cofilin(gi 30582531)	166	25	132	164,31	
Profilin1 (gi 30582841)	140				105*
Tropomyosin3(gi 55665780)	232	22		100	212*
Actin (gi 16924319)	363	19	49		

LMNA (gi 21619981)	465	32	418	316, 90, 180	
ANXA5(gi 12655149)	320	22	101	29, 26 , 212	
RHO gdi alpha (gi 30582607)	204	19	141		
Phospholipase Cbeta 1 (gi 9438229)	1210	120		966, 1070, 1079	971*
Phosphoglycerate mutase 1	254	18	106	39	
PEST (gi 9966827)	179	22			152*
Cyclin T1 (gi 2981196)	726	51		93,219, 356	492*
VegF (gi 3712671)	254	21	171, 169	174, 44	176*
B23 nucleoplasmin (gi 825671)	280	66	198	40,188	136*
Thierodoxin (gi 55957946)	85	11		65	74*
BCell (gi 32698936)	1494	41	36	49,275, 234	
GAPD(gi 230868)	334	26		2,112,116,65	60*
LOC84081 (gi 62020992)	557	70	499	410, 413, 56	
HYPK (gi 27692116)	129	9	35	87	
Rbbp7 (gi 57209887)	285	16		79, 9, 21	39*
UBL4A (gi 57284174)	180	10	48		54*
Transketotase (gi 31417921)	457	33	66	94, 78	431*
Aldolase A (gi 28595)	108	7	13, 42	28	
P300(gi 50345997)	2414	108	1555 ,1554	1549,1550	1560 *, 1558 *

## Literature validated human proteins.

Table includes lysines that were reported as acetylated in literature. Asterisk (\*) represents lysines that were not predicted by the algorithm but identified experimentally by MS.

Protein/accession number	Total residues	Lysines	Predicted/validated (position on protein)	Predicted but not validated experimentally by our algorithm (position on protein) at threshold	*
TFIIB (gi 135629)	290	31	238	267, 136, 14	52*
P53(gi 23491729)	393	19	320, 370 321,305	357, 139, 120	372*, 386*, 373*, 382*
Rch1 (gi:791185)	529	28	22	102, 42	
Myc (gi:71774083)	439	24	317, 143, 157, 371, 275, 323	422,289,355	
Smad 7 (gi 18418630)	426	20	64,70	5,185	
Stat1 (gi 2507413)	749	55		679, 511	410*, 413*
HIV1 – integrase <a href="http://ca.expasy.org/uniprot/A6YEJ0">http://ca.expasy.org/uniprot/A6YEJ0</a>	288	26	266,273	219, 71	264*
MyoD (gi 34862)	319	11	102	133	99*, 104*
HIV-tat (gi:114842145)	101	13		53, 12	50*
Alpha-Tubulin (gi 55977864 )	451	19		60, 338, 112	40*
ACTR (gi:2707770)	1412	55		691, 316 , 87	630*, 629*
HMG14(Chen, Lin et al. 1999) (gi:184251)	216	43	11 , 2	8, 87 , 172	
CDK9 (gi 8099630)	372	29	44	24, 272, 178	48*
P65 (gi 5689767)	551	17		123, 343	221*
HMG-A1 (gi 123377 )	107	16	65, 67,71	55, 15	
Stat3 ( gi 48429227)	770	47	685	631, 548, 153	
ANDR (gi 113830)	919	39	630,632,633	299,316, 590	
Esr1 (gi 544257)	595	28	299	401,266	302*, 303*
Gata3 (gi 120962)	443	21	302	250, 159, 254	305*
P73 (gi:2370177)	500	17	331, 321	346, 138, 157, 192	327*
RB (gi 132164  )	928	76		417, 432,	873*,

				265,8	874*
Beta catenin (gi 860988)	781	26	49	270,672	345*
HIF-1 (gi 4504385)	826	49		297, 159, 629	532*
SATB1 (gi 417747)	763	44	136	241, 11, 518, 129	
FOXO1 (gi 116241368)	655	25	245, 248, 265	272, 354, 179	
INAR2 (gi 1352466)	515	21		180, 182, 243	399*
HSP90alpha (gi 92090606)	732	80		277, 513, 276	294*
Ku70 (gi 125729)	609	59		253,249,358	539*, 542*
PGRC1 (gi 6647589)	798	47		646, 656, 194, 254	278*, 271*, 146*
HTT (gi 1170192)	3144	278	444	178,1168,92, 410,700	
RIP140(gi 57232746)	1158	92		724, 249,1105, 127, 97	111*, 158*, 286*, 310*, 481*, 446*
CDT1(gi 224471822)	546	23	49	429,218, 166, 100	24*

**Literature validated human proteins and the references of acetylation sites.**

<b><u>Protein</u></b>	<b><u>References</u></b>
TFIIB	(Imhof, Yang et al. 1997)
HMG-A1	(Munshi, Merika et al. 1998)
P53	(Gu and Roeder 1997)
Gata1	(Boyes, Byfield et al. 1998)
Rch1	(Bannister, Miska et al. 2000)
HMG-14	(Bergel, Herrera et al. 2000)
C-Myb	(Tomita, Towatari et al. 2000)
E2F1	(Martinez-Balbas, Bauer et al. 2000)
C-myc	(Zhang, Faiola et al. 2005)
Smad7	(Simonsson, Heldin et al. 2005)
Stat1	(Kramer, Baus et al. 2006)
HIV1-integrase	(Kiernan, Vanhulle et al. 1999)
MyoD	(Poleskaya, Duquet et al. 2000)
HIV-tat	(Dormeyer, Dorr et al. 2003)
ACTR	(Chen, Lin et al. 1999)
CDK9	(Fu, Yoon et al. 2007)
P65	(Ishinaga, Jono et al. 2007)
SATB1	(Pavan Kumar, Purbey et al. 2006)
STAT3	(Glozak, Sengupta et al. 2005)
HMGB-1	(Glozak, Sengupta et al. 2005)
Beta-catenin	(Wolf, Rodova et al. 2002; Levy, Wei et al. 2004)
Alpha-tubulin	(Glozak, Sengupta et al. 2005)
GATA3	(33)
P73	(33)
SMC3	(Zhang, Shi et al. 2008)
HIF-1	(Jeong, Bae et al. 2002)
RB	(Markham, Munro et al. 2006)
ESR1	(Fu, Wang et al. 2004)
Ku70	(35)
Hsp90	(35)



Fox01	(Greer and Brunet 2005)
PGC1	(35)
INF	(35)
AR	(Fu, Wang et al. 2004)
HTT	(Jeong, Then et al. 2009)
CDT1	(Glozak and Seto 2009)
RIP140	(Yang and Seto 2008)

## P-values in Chapter 2

To test whether the observed enrichment of small residues (G/A), K, S was statistically significant in the histone and nonhistone datasets, I determined the frequency of these same residues flanking a lysine in the entire human proteome. I employed the hypergeometric test to measure the statistical relevance of this observation (STable 4). "LIT" represents those lysines validated in literature, "HIST" represents lysines in histones, and "PAN" represents pan-acetyl antibody substrate lysines. Top row shows the position of the residue with respect to the target lysine (position 0, not shown). List below the table displays the values for each of the variables in the hypergeometric distribution: all lysines in the human proteome = N; number of times the particular residue is seen flanking in each position in human proteome= K; total number of lysines in each independent validation dataset = n; number of times particular residue is seen flanking in each position in validation dataset = m.

	-6	-5	-4	-3	-2	-1	1	2	3	4	5	6
<b>small</b>												
LIT	0.004	1	0.9	2e-1	9e-4	0.003	0.004	0.4	0.32	0.47	7e-6	0.5
HIST	0.01	0.01	0.7	1.9e-8	9.2e-4	1.6e-5	2e-5	1e-2	2e-3	3e-1	2e-1	1e-3
PAN	0.05	1	0.44	0.09	0.03	8.6e-5	0.004	0.85	0.18	0.17	5.91e-4	0.7
<b>K</b>												
LIT	0.06	0.7	0.03	0.13	0.36	0.6	0.01	0.3	0.02	0.3	0.1	0.2
HIST	0.33	0.02	1.86e-4	0.7	0.5	0.47	0.4	0.5	0.4	0.3	0.08	0.6
PAN	7.3e-9	5.6e-4	0.2	0.0099	0.11	0.07	0.7	0.01	0.3	0.16	0.007	0.02
<b>S</b>												
LIT	0.39	0.12	0.005	0.05	0.12	0.09	0.99	0.8	0.15	0.3	0.5	0.8
HIST	0.45	0.5	0.2	0.8	0.8	1	0.01	0.9	0.3	0.8	0.5	0.46
PAN	0.2	0.4	0.04	0.11	0.11	0.6	0.09	0.8	0.07	0.15	0.2	0.2

**N** in human dataset = 5127294

**K** in each position for small, serine, and lysine residues dataset containing all human lysines.

Position 0: GA:535591 S:291208 K:389280  
Position 1: GA:621026 S:315489 K:323314  
Position 2: GA:677578 S:307757 K:416567  
Position 3: GA:663899 S:312774 K:468003  
Position 4: GA:632418 S:310841 K:537247  
Position 5: GA:584283 S:289536 K:486015  
Position 7: GA:597697 S:296377 K:486758  
Position 8: GA:610849 S:404638 K:537885  
Position 9: GA:593694 S:362154 K:462707  
Position 10: GA:586526 S:336826 K:418813  
Position 11: GA:591499 S:303442 K:325808  
Position 12: GA:531503 S:292427 K:387452

**m** in each position for small, serine, and lysine residues: Human Histone Dataset

Position 0: GA:7 S:2 K:3  
Position 1: GA:8 S:2 K:5  
Position 2: GA:3 S:3 K:9  
Position 3: GA:16 S:1 K:2  
Position 4: GA:10 S:1 K:3  
Position 5: GA:12 S:0 K:3  
Position 7: GA:12 S:5 K:3  
Position 8: GA:8 S:1 K:3  
Position 9: GA:9 S:3 K:3  
Position 10: GA:4 S:1 K:13  
Position 11: GA:5 S:2 K:4  
Position 12: GA:9 S:2 K:2

**m** in each position for small, serine, and lysine residues: Pan-acetyl (PA) Dataset

Position 0: GA:10 S:5 K:12  
Position 1: GA:11 S:4 K:11  
Position 2: GA:8 S:7 K:6  
Position 3: GA:11 S:6 K:11  
Position 4: GA:12 S:3 K:9  
Position 5: GA:17 S:6 K:9  
Position 7: GA:11 S:6 K:4

Position 8: GA:5 S:3 K:12  
Position 9: GA:9 S:7 K:6  
Position 10: GA:9 S:6 K:7  
Position 11: GA:11 S:5 K:9  
Position 12: GA:5 S:5 K:9

**m in each position for small, serine, and lysine residues:** Literature  
(Lit) Dataset

Position 0: GA:13 S:4 K:8  
Position 1: GA:7 S:6 K:3  
Position 2: GA:4 S:9 K:9  
Position 3: GA:9 S:7 K:8  
Position 4: GA:16 S:6 K:7  
Position 5: GA:14 S:6 K:5  
Position 6: GA:0 S:0 K:57  
Position 7: GA:8 S:2 K:11  
Position 8: GA:8 S:7 K:7  
Position 9: GA:8 S:5 K:10  
Position 10: GA:7 S:4 K:6  
Position 11: GA:14 S:2 K:6  
Position 12: GA:6 S:13 K:6

$n(\text{hist}) = 21$  ;  $n(\text{lit}) = 49$   $n(\text{pan}) = 67$

### **S. cerevisiae chromatin-associated protein predictions**

yEAF7

Predictions: K27, K343, K165, K142

ySPT6

Predictions: K958, K491, K494, K118, K120, K319

ySir3

Predictions: K52, K436, K445, K3, K827

### **Sodium butyrate (10mM) treated samples**

Additional acetylation sites detected:

H2A: K95, K119

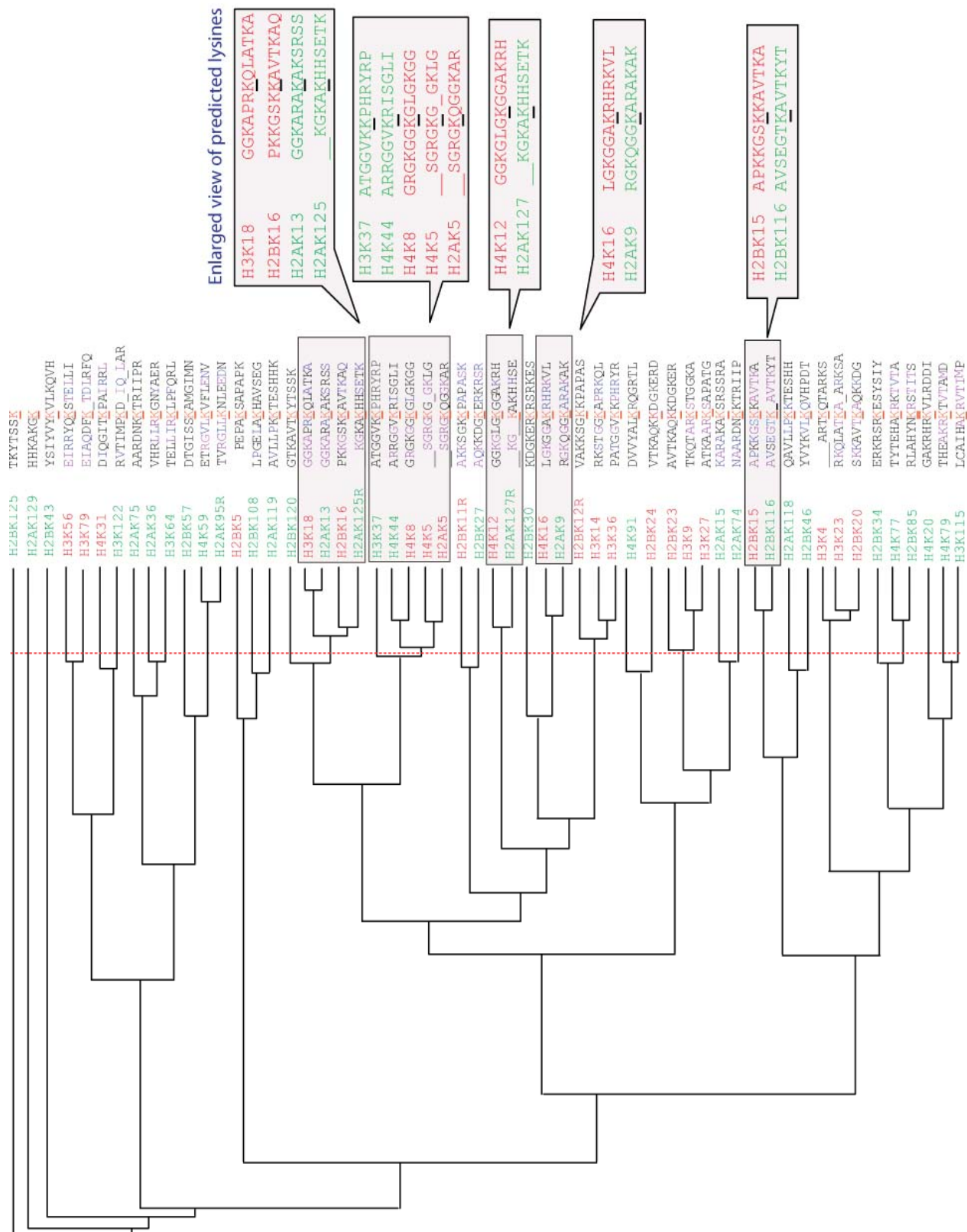
H2B: K34, K46, K108, K116

H4: K77, K79, K91

H3:K37

## **Appendix Figures**

**Appendix Figure 1: Enlarged tree from Figure 2.3.**  
Legend same as in 2.3.



### Appendix Figure 2. Experimental Mass Spectrometry validation in human cells

**A.** CAD (Collisionally Activated Dissociation) mass spectrum of the  $[M+2H]^{2+}$  histone H2A 4-11 peptide. Shown is a mixture of two single-acetylated species with the same sequence but with different sites of acetylation. The MS/MS data indicate acetylation on K5 and K9. The amino acid sequence is shown above the spectrum, and the masses associated with the sequence correspond to the expected b- and y-type fragment ions for the propionylated H2A 4-11 peptide. The masses of the observed b- and y-type fragment ions are assigned to their corresponding  $m/z$  peaks in the spectrum and are also underlined with the sequence. The asterisks denote fragment ions that are shifted lower in mass by 14 amu, indicating the presence of *in vivo* singly-acetylated species. For this analysis, histone H2A was purified from HL60 cells, lysines were derivatized with propionic anhydride, and H2A peptides were generated with trypsin. H2A peptides were gradient-eluted via nanoflow-HPLC and mass analyzed with an LTQ-FT mass spectrometer.

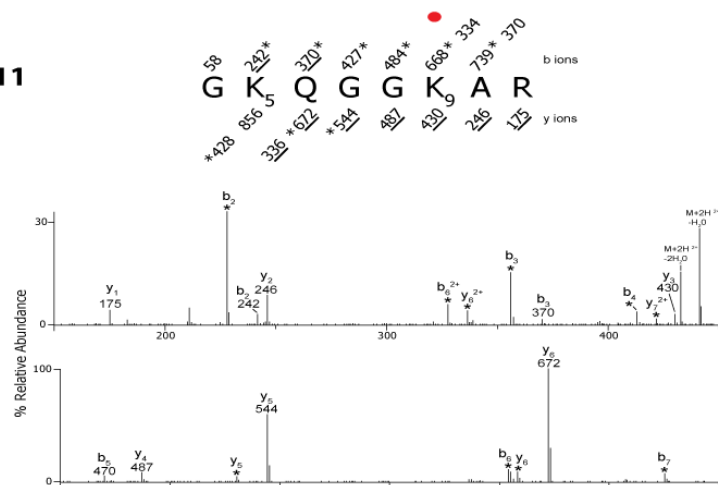
**B.** CAD mass spectrum of the  $[M+2H]^{2+}$  histone H2A 12-17 peptide. Shown is a mixed spectrum of two single-acetylated species with the same sequence but with different sites of acetylation. The MS/MS data indicate acetylation on K13 and K15. The amino acid sequence is shown above the spectrum, and the masses associated with the sequence correspond to the expected b- and y-type fragment ions for the propionylated H2A 12-17 peptide. The masses of the observed b- and y-type fragment ions are assigned as in (A). For this analysis, histone H2A was purified, derivatized, trypsin digested, and gradient-eluted as in C.

**C.** CAD mass spectrum of the  $[M+2H]^{2+}$  histone H2A 122-129 peptide. Shown is a mixed spectrum of three mono-acetylated species with the same sequence but with different sites of acetylation. The MS/MS data indicate acetylation on K125, K127, and K129. The amino acid sequence is shown above the spectrum, and the masses associated with the sequence correspond to the expected singly-charged b- and y-type fragment ions for the fully deuterio-acetylated H2A 122-129 peptide. The masses of the observed b- and y-type fragment ions are assigned to their corresponding  $m/z$  peaks in the spectrum and are also underlined with the sequence. The asterisks denote fragment ions that are shifted lower in mass by 3 amu, indicating the presence of *in vivo* (non-deutero) mono-acetylated species. For this analysis, histone H2A was purified from HL60 cells, digested with GluC, derivatized with deuterio succinimido acetate to increase hydrophobicity of the 122-129 H2A peptide, and H2A peptides were gradient-eluted via nanoflow-HPLC and mass analyzed with an LTQ Orbitrap mass spectrometer. To obtain this spectrum, the precursor ion,  $m/z$  534.1, was targeted for isolation with a 3 amu window and fragmented via CAD. Image and text provided by Kristie Rose, Donald Hunt lab, UVA (Charlottesville,VA)



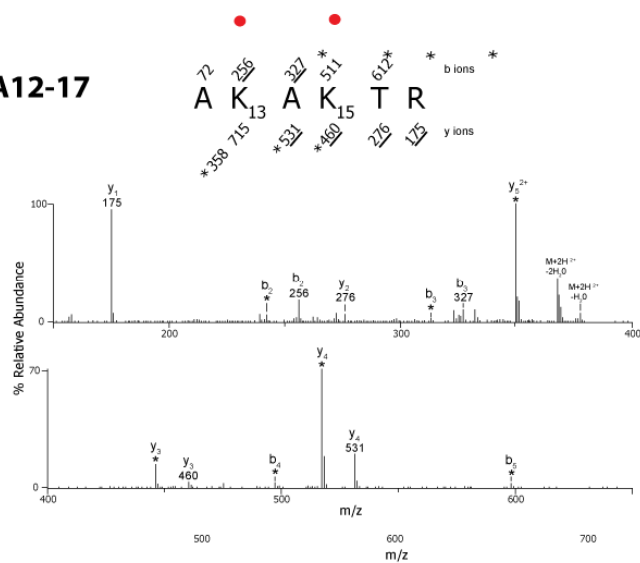
A

H2A4-11



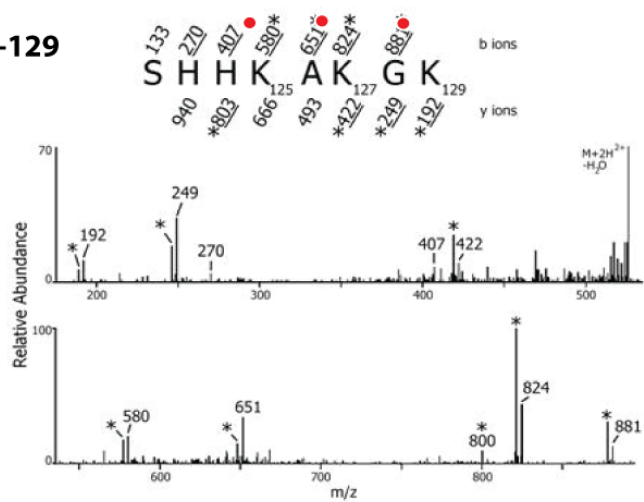
B

H2A12-17



C

H2A122-129



**Appendix Figure 3. MS validation of acetylation sites in budding yeast proteins Eaf7, Sir3, and Spt6.**

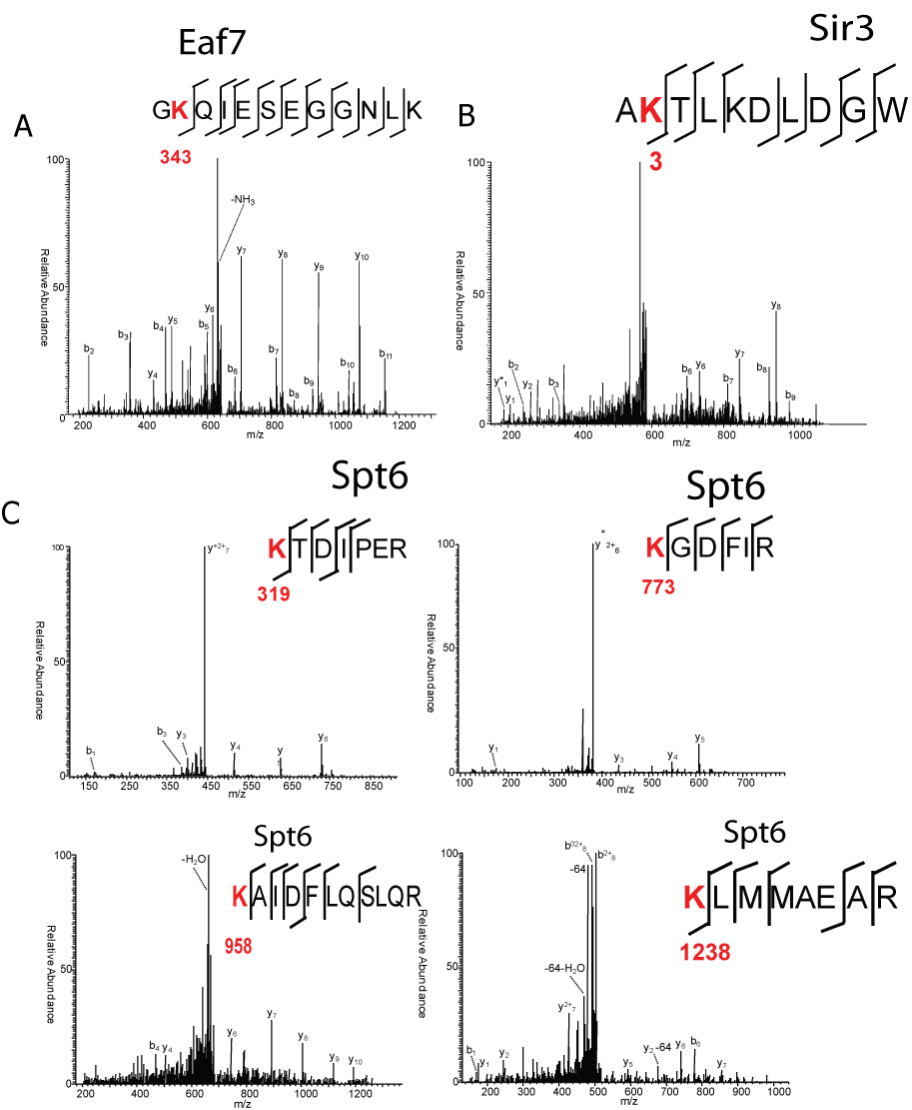
Identification of lysine acetylated peptides in Eaf7, Sir3, and Spt6 by LC/MS/MS. The labels b and y designate the N- and C-terminal fragments, respectively, of the peptide produced by breakage at the peptide bond in the mass spectrometer. The number represents the number of N- or C-terminal residues present in the peptide fragment. The superscripts 0, \* mean water or ammonia loss, respectively

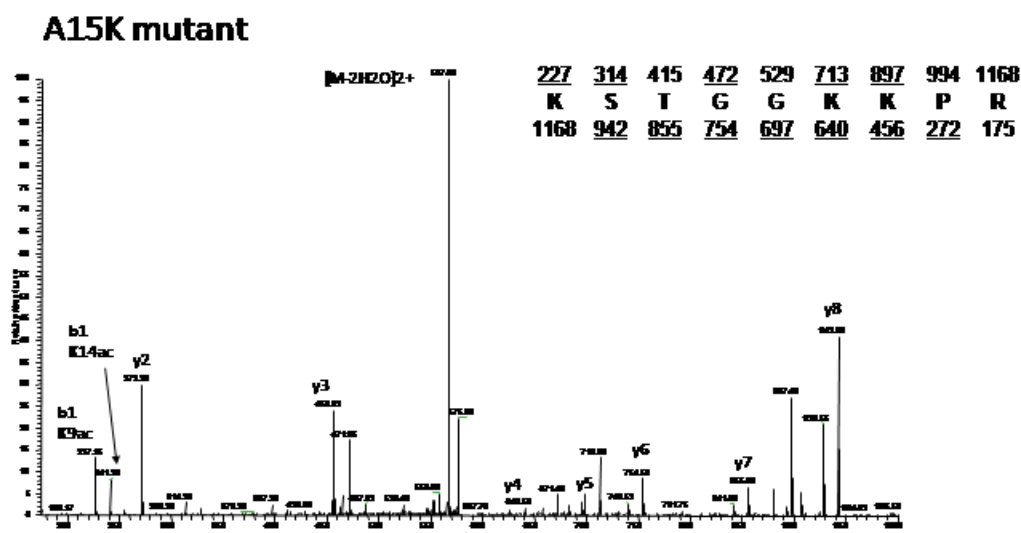
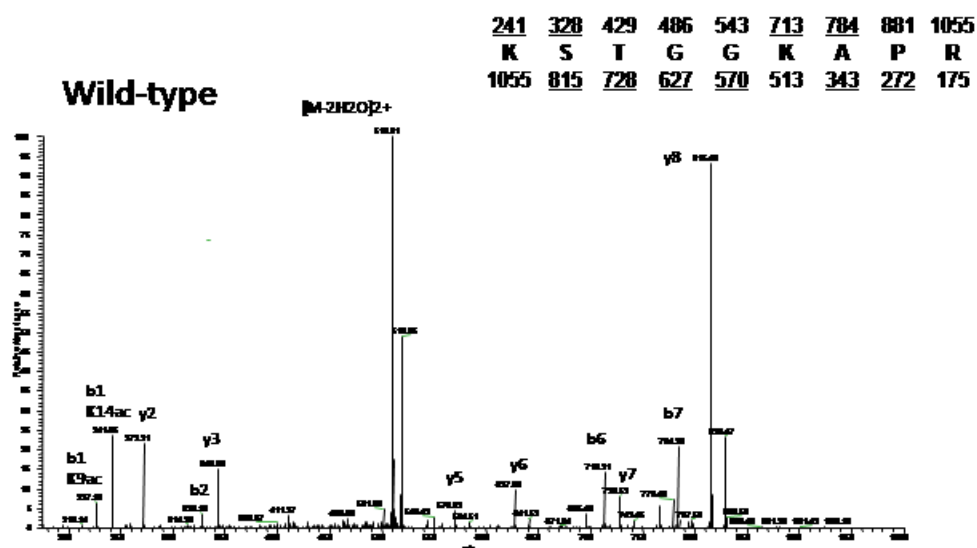
**A.**Eaf7K343

**B.**Sir3K3

**C.**Spt6K319, Spt6K773, Spt6K958, and Spt6K1238

Images and data provided by Yingming Zhao, UT Southwestern Medical Center (Dallas, TX)





**Appendix Figure 4 MS validation of acetylation sites in WT and A15K mutant strain**

LC/MS/MS spectra displaying H3K9ac and H3K14 ac. Figures supplied by Ben Garcia displaying the WT (H3K9-17) peptide and H3A15K (H3K9-17) spectrum. (Images and data provided by Ben Garcia, Princeton University (Princeton, NJ). H3K36 methylation data under A15K mutation and G13V mass spectra image is presently unavailable, but quantitative data presented in Chapter 3.

## References

- Agostoni, E., D. Albertson, et al. (1996). "cec-1, a soma-specific chromobox-containing gene in *C. elegans*." Dev Biol **178**(2): 316-326.
- Akasaka, T., M. Kanno, et al. (1996). "A role for mel-18, a Polycomb group-related vertebrate gene, during theanteroposterior specification of the axial skeleton." Development **122**(5): 1513-1522.
- Allfrey, V. G., R. Faulkner, et al. (1964). "Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis." Proc Natl Acad Sci U S A **51**: 786-794.
- Allis, C. D., S. L. Berger, et al. (2007). "New nomenclature for chromatin-modifying enzymes." Cell **131**(4): 633-636.
- Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.
- Altschul, S. F., T. L. Madden, et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.
- Arevalo-Rodriguez, M., X. Wu, et al. (2004). "Prolyl isomerases in yeast." Front Biosci **9**: 2420-2446.
- Baker, L. A., C. D. Allis, et al. (2008). "PHD fingers in human diseases: disorders arising from misinterpreting epigenetic marks." Mutat Res **647**(1-2): 3-12.
- Bannister, A. J., E. A. Miska, et al. (2000). "Acetylation of importin-alpha nuclear import factors by CBP/p300." Curr Biol **10**(8): 467-470.
- Basu, A., K. L. Rose, et al. (2009). "Proteome-wide prediction of acetylation substrates." Proc Natl Acad Sci U S A **106**(33): 13785-13790.

- Bateman, A., L. Coin, et al. (2004). "The Pfam protein families database." Nucleic Acids Res **32**(Database issue): D138-141.
- Beck, H. C., E. C. Nielsen, et al. (2006). "Quantitative proteomic analysis of post-translational modifications of human histones." Mol Cell Proteomics **5**(7): 1314-1325.
- Beisel, C., A. Imhof, et al. (2002). "Histone methylation by the Drosophila epigenetic transcriptional regulator Ash1." Nature **419**(6909): 857-862.
- Bergel, M., J. E. Herrera, et al. (2000). "Acetylation of novel sites in the nucleosomal binding domain of chromosomal protein HMG-14 by p300 alters its interaction with nucleosomes." J Biol Chem **275**(15): 11514-11520.
- Bernstein, E., E. M. Duncan, et al. (2006). "Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin." Mol Cell Biol **26**(7): 2560-2569.
- Bernstein, E. and S. B. Hake (2006). "The nucleosome: a little variation goes a long way." Biochem Cell Biol **84**(4): 505-517.
- Bird, A. W., D. Y. Yu, et al. (2002). "Acetylation of histone H4 by Esa1 is required for DNA double-strand break repair." Nature **419**(6905): 411-415.
- Blom, N., T. Sicheritz-Ponten, et al. (2004). "Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence." Proteomics **4**(6): 1633-1649.
- Bonenfant, D., M. Coulot, et al. (2006). "Characterization of histone H2A and H2B variants and their post-translational modifications by mass spectrometry." Mol Cell Proteomics **5**(3): 541-552.
- Boyes, J., P. Byfield, et al. (1998). "Regulation of activity of the transcription factor GATA-1 by acetylation." Nature **396**(6711): 594-598.

- Brownell, J. E., J. Zhou, et al. (1996). "Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation." Cell **84**(6): 843-851.
- Burnett, G. and E. P. Kennedy (1954). "The enzymatic phosphorylation of proteins." J Biol Chem **211**(2): 969-980.
- Carrozza, M. J., B. Li, et al. (2005). "Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription." Cell **123**(4): 581-592.
- Cereseto, A., L. Manganaro, et al. (2005). "Acetylation of HIV-1 integrase by p300 regulates viral integration." EMBO J **24**(17): 3070-3081.
- Chen, H., R. J. Lin, et al. (1999). "Regulation of hormone-induced histone hyperacetylation and gene activation via acetylation of an acetylase." Cell **98**(5): 675-686.
- Chen, Y., S. W. Kwon, et al. (2005). "Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra." J Proteome Res **4**(3): 998-1005.
- Choudhary, C., C. Kumar, et al. (2009). "Lysine acetylation targets protein complexes and co-regulates major cellular functions." Science **325**(5942): 834-840.
- Clark-Adams, C. D. and F. Winston (1987). "The SPT6 gene is essential for growth and is required for delta-mediated transcription in *Saccharomyces cerevisiae*." Mol Cell Biol **7**(2): 679-686.
- Clarke, D. J., L. P. O'Neill, et al. (1993). "Selective use of H4 acetylation sites in the yeast *Saccharomyces cerevisiae*." Biochem J **294** ( Pt 2): 557-561.

- Cocklin, R. R. and M. Wang (2003). "Identification of methylation and acetylation sites on mouse histone H3 using matrix-assisted laser desorption/ionization time-of-flight and nanoelectrospray ionization tandem mass spectrometry." J Protein Chem **22**(4): 327-334.
- Cole, C., J. D. Barber, et al. (2008). "The Jpred 3 secondary structure prediction server." Nucleic Acids Res **36**(Web Server issue): W197-201.
- Cooper, G. F., C. F. Aliferis, et al. (1997). "An evaluation of machine-learning methods for predicting pneumonia mortality." Artif Intell Med **9**(2): 107-138.
- Craig, R., J. C. Cortens, et al. (2006). "Using annotated peptide mass spectrum libraries for protein identification." J Proteome Res **5**(8): 1843-1849.
- Crooks, G. E., G. Hon, et al. (2004). "WebLogo: a sequence logo generator." Genome Res **14**(6): 1188-1190.
- Cuff, J. A., M. E. Clamp, et al. (1998). "JPred: a consensus secondary structure prediction server." Bioinformatics **14**(10): 892-893.
- Dai, J., E. M. Hyland, et al. (2008). "Probing nucleosome function: a highly versatile library of synthetic histone H3 and H4 mutants." Cell **134**(6): 1066-1078.
- Dang, W., K. K. Steffen, et al. (2009). "Histone H4 lysine 16 acetylation regulates cellular lifespan." Nature **459**(7248): 802-807.
- Das, C., M. S. Lucia, et al. (2009). "CBP/p300-mediated acetylation of histone H3 on lysine 56." Nature.
- de Napoles, M., J. E. Mermoud, et al. (2004). "Polycomb group proteins Ring1A/B link ubiquitylation of histone H2A to heritable gene silencing and X inactivation." Dev Cell **7**(5): 663-676.



- Dhalluin, C., J. E. Carlson, et al. (1999). "Structure and ligand of a histone acetyltransferase bromodomain." Nature **399**(6735): 491-496.
- Dodd, L. E. and M. S. Pepe (2003). "Partial AUC estimation and regression." Biometrics **59**(3): 614-623.
- Dormeyer, W., A. Dorr, et al. (2003). "Acetylation of the HIV-1 Tat protein: an in vitro study." Anal Bioanal Chem **376**(7): 994-1005.
- Downs, J. A., N. F. Lowndes, et al. (2000). "A role for *Saccharomyces cerevisiae* histone H2A in DNA repair." Nature **408**(6815): 1001-1004.
- Driscoll, R., A. Hudson, et al. (2007). "Yeast Rtt109 promotes genome stability by acetylating histone H3 on lysine 56." Science **315**(5812): 649-652.
- Eddy, S. R. (2004). "Where did the BLOSUM62 alignment score matrix come from?" Nat Biotechnol **22**(8): 1035-1036.
- Eissenberg, J. C. (2001). "Molecular biology of the chromo domain: an ancient chromatin module comes of age." Gene **275**(1): 19-29.
- Enomoto, S. and J. Berman (1998). "Chromatin assembly factor I contributes to the maintenance, but not the re-establishment, of silencing at the yeast silent mating loci." Genes Dev **12**(2): 219-232.
- Felsenstein, J. (2008). "Comparative methods with sampling error and within-species variation: contrasts revisited and revised." Am Nat **171**(6): 713-725.
- Ferreira, R., A. Eberhardter, et al. (2007). "Site-specific acetylation of ISWI by GCN5." BMC Mol Biol **8**: 73.
- Finn, R. D., J. Tate, et al. (2008). "The Pfam protein families database." Nucleic Acids Res **36**(Database issue): D281-288.

- Fischle, W., Y. Wang, et al. (2003). "Molecular basis for the discrimination of repressive methyl-lysine marks in histone H3 by Polycomb and HP1 chromodomains." Genes Dev **17**(15): 1870-1881.
- Francis, N. J., R. E. Kingston, et al. (2004). "Chromatin compaction by a polycomb group protein complex." Science **306**(5701): 1574-1577.
- Fu, J., H. G. Yoon, et al. (2007). "Regulation of P-TEFb elongation complex activity by CDK9 acetylation." Mol Cell Biol **27**(13): 4641-4651.
- Fu, M., C. Wang, et al. (2004). "Acetylation of nuclear receptors in cellular growth and apoptosis." Biochem Pharmacol **68**(6): 1199-1208.
- Garcia, B. A., C. M. Barber, et al. (2005). "Modifications of human histone H3 variants during mitosis." Biochemistry **44**(39): 13202-13213.
- Garcia, B. A., S. B. Hake, et al. (2007). "Organismal differences in post-translational modifications in histones H3 and H4." J Biol Chem **282**(10): 7641-7655.
- Garcia, B. A., S. Mollah, et al. (2007). "Chemical derivatization of histones for facilitated analysis by mass spectrometry." Nat Protoc **2**(4): 933-938.
- Gasser, S. M. and M. M. Cockell (2001). "The molecular biology of the SIR proteins." Gene **279**(1): 1-16.
- Glozak, M. A., N. Sengupta, et al. (2005). "Acetylation and deacetylation of non-histone proteins." Gene **363**: 15-23.
- Glozak, M. A. and E. Seto (2009). "Acetylation/deacetylation modulates the stability of DNA replication licensing factor Cdt1." J Biol Chem **284**(17): 11446-11453.
- Goodrich, J., P. Puangsomlee, et al. (1997). "A Polycomb-group gene regulates homeotic gene expression in Arabidopsis." Nature **386**(6620): 44-51.

- Grant, P. A., L. Duggan, et al. (1997). "Yeast Gcn5 functions in two multisubunit complexes to acetylate nucleosomal histones: characterization of an Ada complex and the SAGA (Spt/Ada) complex." Genes Dev **11**(13): 1640-1650.
- Greer, E. L. and A. Brunet (2005). "FOXO transcription factors at the interface between longevity and tumor suppression." Oncogene **24**(50): 7410-7425.
- Grewal, S. I. and J. C. Rice (2004). "Regulation of heterochromatin by histone methylation and small RNAs." Curr Opin Cell Biol **16**(3): 230-238.
- Grunstein, M. (1990). "Histone function in transcription." Annu Rev Cell Biol **6**: 643-678.
- Gu, W. and R. G. Roeder (1997). "Activation of p53 sequence-specific DNA binding by acetylation of the p53 C-terminal domain." Cell **90**(4): 595-606.
- Hake, S. B., B. A. Garcia, et al. (2006). "Expression patterns and post-translational modifications associated with mammalian histone H3 variants." J Biol Chem **281**(1): 559-568.
- Heard, E. (2004). "Recent advances in X-chromosome inactivation." Curr Opin Cell Biol **16**(3): 247-255.
- Hirayama, J., S. Sahar, et al. (2007). "CLOCK-mediated acetylation of BMAL1 controls circadian function." Nature **450**(7172): 1086-1090.
- Howe, L., D. Auston, et al. (2001). "Histone H3 specific acetyltransferases are essential for cell cycle progression." Genes Dev **15**(23): 3144-3154.
- Hsu, F., T. H. Pringle, et al. (2005). "The UCSC Proteome Browser." Nucleic Acids Res **33**(Database issue): D454-458.

- Hsu, J. Y., Z. W. Sun, et al. (2000). "Mitotic phosphorylation of histone H3 is governed by Ipl1/aurora kinase and Glc7/PP1 phosphatase in budding yeast and nematodes." Cell **102**(3): 279-291.
- Huh, W. K., J. V. Falvo, et al. (2003). "Global analysis of protein localization in budding yeast." Nature **425**(6959): 686-691.
- Imhof, A., X. J. Yang, et al. (1997). "Acetylation of general transcription factors by histone acetyltransferases." Curr Biol **7**(9): 689-692.
- Ingolfsson, H. and G. Yona (2008). "Protein domain prediction." Methods Mol Biol **426**: 117-143.
- Ishinaga, H., H. Jono, et al. (2007). "TGF-beta induces p65 acetylation to enhance bacteria-induced NF-kappaB activation." EMBO J **26**(4): 1150-1162.
- Jenuwein, T. and C. D. Allis (2001). "Translating the histone code." Science **293**(5532): 1074-1080.
- Jeong, H., F. Then, et al. (2009). "Acetylation targets mutant huntingtin to autophagosomes for degradation." Cell **137**(1): 60-72.
- Jeong, J. W., M. K. Bae, et al. (2002). "Regulation and destabilization of HIF-1alpha by ARD1-mediated acetylation." Cell **111**(5): 709-720.
- Kagey, M. H., T. A. Melhuish, et al. (2003). "The polycomb protein Pc2 is a SUMO E3." Cell **113**(1): 127-137.
- Kaplan, T., C. L. Liu, et al. (2008). "Cell cycle- and chaperone-mediated regulation of H3K56ac incorporation in yeast." PLoS Genet **4**(11): e1000270.
- Karolchik, D., R. Baertsch, et al. (2003). "The UCSC Genome Browser Database." Nucleic Acids Res **31**(1): 51-54.

- Kennison, J. A. (1995). "The Polycomb and trithorax group proteins of *Drosophila*: trans-regulators of homeotic gene function." Annu Rev Genet **29**: 289-303.
- Kiernan, R. E., C. Vanhulle, et al. (1999). "HIV-1 tat transcriptional activity is regulated by acetylation." Embo J **18**(21): 6106-6118.
- Kim, C. A., M. Gingery, et al. (2002). "The SAM domain of polyhomeotic forms a helical polymer." Nat Struct Biol **9**(6): 453-457.
- Kim, S. C., R. Sprung, et al. (2006). "Substrate and functional diversity of lysine acetylation revealed by a proteomics survey." Mol Cell **23**(4): 607-618.
- Kimura, A. and M. Horikoshi (1998). "How do histone acetyltransferases select lysine residues in core histones?" FEBS Lett **431**(2): 131-133.
- Kimura, A. and M. Horikoshi (1998). "Tip60 acetylates six lysines of a specific class in core histones in vitro." Genes Cells **3**(12): 789-800.
- Kizer, K. O., H. P. Phatnani, et al. (2005). "A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation." Mol Cell Biol **25**(8): 3305-3316.
- Koh, A. S., A. J. Kuo, et al. (2008). "Aire employs a histone-binding module to mediate immunological tolerance, linking chromatin regulation with organ-specific autoimmunity." Proc Natl Acad Sci U S A **105**(41): 15878-15883.
- Kramer, O. H., D. Baus, et al. (2006). "Acetylation of Stat1 modulates NF-kappaB activity." Genes Dev **20**(4): 473-485.
- Krishnamoorthy, T., X. Chen, et al. (2006). "Phosphorylation of histone H4 Ser1 regulates sporulation in yeast and is conserved in fly and mouse spermatogenesis." Genes Dev **20**(18): 2580-2592.

- Krogan, N. J., K. Baetz, et al. (2004). "Regulation of chromosome stability by the histone H2A variant Htz1, the Swr1 chromatin remodeling complex, and the histone acetyltransferase NuA4." Proc Natl Acad Sci U S A **101**(37): 13513-13518.
- Kurdistani, S. K., S. Tavazoie, et al. (2004). "Mapping global histone acetylation patterns to gene expression." Cell **117**(6): 721-733.
- Lan, F., R. E. Collins, et al. (2007). "Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression." Nature **448**(7154): 718-722.
- Latham, J. A. and S. Y. Dent (2007). "Cross-regulation of histone modifications." Nat Struct Mol Biol **14**(11): 1017-1024.
- Leroy, B., G. Toubeau, et al. (2006). "Identification and characterization of new protein chemoattractants in the frog skin secretome." Mol Cell Proteomics **5**(11): 2114-2123.
- Leslie, C. S., E. Eskin, et al. (2004). "Mismatch string kernels for discriminative protein classification." Bioinformatics **20**(4): 467-476.
- Levy, L., Y. Wei, et al. (2004). "Acetylation of beta-catenin by p300 regulates beta-catenin-Tcf4 interaction." Mol Cell Biol **24**(8): 3404-3414.
- Li, Q., H. Zhou, et al. (2008). "Acetylation of histone H3 lysine 56 regulates replication-coupled nucleosome assembly." Cell **134**(2): 244-255.
- Lin, Y. Y., J. Y. Lu, et al. (2009). "Protein acetylation microarray reveals that NuA4 controls key metabolic target regulating gluconeogenesis." Cell **136**(6): 1073-1084.
- Ling, X., T. A. Harkness, et al. (1996). "Yeast histone H3 and H4 amino termini are important for nucleosome assembly in vivo and in vitro: redundant and position-independent functions in assembly but not in gene regulation." Genes Dev **10**(6): 686-699.

- Liu, X., L. Wang, et al. (2008). "The structural basis of protein acetylation by the p300/CBP transcriptional coactivator." Nature **451**(7180): 846-850.
- Lo, W. S., R. C. Trievel, et al. (2000). "Phosphorylation of serine 10 in histone H3 is functionally linked in vitro and in vivo to Gcn5-mediated acetylation at lysine 14." Mol Cell **5**(6): 917-926.
- Luger, K., A. W. Mader, et al. (1997). "Crystal structure of the nucleosome core particle at 2.8 Å resolution." Nature **389**(6648): 251-260.
- Madireddi, M. T., R. S. Coyne, et al. (1996). "Pdd1p, a novel chromodomain-containing protein, links heterochromatin assembly and DNA elimination in Tetrahymena." Cell **87**(1): 75-84.
- Markham, D., S. Munro, et al. (2006). "DNA-damage-responsive acetylation of pRb regulates binding to E2F-1." EMBO Rep **7**(2): 192-198.
- Marks, P. A. (2007). "Discovery and development of SAHA as an anticancer agent." Oncogene **26**(9): 1351-1356.
- Marmorstein, R. (2001). "Protein modules that manipulate histone tails for chromatin regulation." Nat Rev Mol Cell Biol **2**(6): 422-432.
- Marmorstein, R. (2001). "Structure and function of histone acetyltransferases." Cell Mol Life Sci **58**(5-6): 693-703.
- Marmorstein, R. (2001). "Structure of histone acetyltransferases." J Mol Biol **311**(3): 433-444.
- Marmorstein, R. and S. Y. Roth (2001). "Histone acetyltransferases: function, structure, and catalysis." Curr Opin Genet Dev **11**(2): 155-161.
- Martinez-Balbas, M. A., U. M. Bauer, et al. (2000). "Regulation of E2F1 activity by acetylation." Embo J **19**(4): 662-671.

- Matthews, A. G., A. J. Kuo, et al. (2007). "RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination." Nature **450**(7172): 1106-1110.
- Megee, P. C., B. A. Morgan, et al. (1990). "Genetic analysis of histone H4: essential role of lysines subject to reversible acetylation." Science **247**(4944): 841-845.
- Morris, S. A., B. Rao, et al. (2007). "Identification of histone H3 lysine 36 acetylation as a highly conserved histone modification." J Biol Chem **282**(10): 7632-7640.
- Morris, S. A., Y. Shibata, et al. (2005). "Histone H3 K36 methylation is associated with transcription elongation in *Schizosaccharomyces pombe*." Eukaryot Cell **4**(8): 1446-1454.
- Mujtaba, S., L. Zeng, et al. (2007). "Structure and acetyl-lysine recognition of the bromodomain." Oncogene **26**(37): 5521-5527.
- Muller, J. (1995). "Transcriptional silencing by the Polycomb protein in *Drosophila* embryos." EMBO J **14**(6): 1209-1220.
- Munshi, N., M. Merika, et al. (1998). "Acetylation of HMG I(Y) by CBP turns off IFN beta expression by disrupting the enhanceosome." Mol Cell **2**(4): 457-467.
- Nakanishi, S., B. W. Sanderson, et al. (2008). "A comprehensive library of histone mutants identifies nucleosomal residues required for H3K4 methylation." Nat Struct Mol Biol **15**(8): 881-888.
- Nakayama, J., J. C. Rice, et al. (2001). "Role of histone H3 lysine 9 methylation in epigenetic control of heterochromatin assembly." Science **292**(5514): 110-113.
- Nelson, C. J., H. Santos-Rosa, et al. (2006). "Proline isomerization of histone H3 regulates lysine methylation and gene expression." Cell **126**(5): 905-916.



- Olsen, J. V., B. Blagoev, et al. (2006). "Global, in vivo, and site-specific phosphorylation dynamics in signaling networks." Cell **127**(3): 635-648.
- Ooi, S. K., C. Qiu, et al. (2007). "DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA." Nature **448**(7154): 714-717.
- Ozaki, T., R. Okoshi, et al. (2009). "Acetylation status of E2F-1 has an important role in the regulation of E2F-1-mediated transactivation of tumor suppressor p73." Biochem Biophys Res Commun **386**(1): 207-211.
- Pavan Kumar, P., P. K. Purbey, et al. (2006). "Phosphorylation of SATB1, a global gene regulator, acts as a molecular switch regulating its transcriptional activity in vivo." Mol Cell **22**(2): 231-243.
- Pesavento, J. J., Y. B. Kim, et al. (2004). "Shotgun annotation of histone modifications: a new approach for streamlined characterization of proteins by top down mass spectrometry." J Am Chem Soc **126**(11): 3386-3387.
- Pflum, M. K., J. K. Tong, et al. (2001). "Histone deacetylase 1 phosphorylation promotes enzymatic activity and complex formation." J Biol Chem **276**(50): 47733-47741.
- Pokholok, D. K., C. T. Harbison, et al. (2005). "Genome-wide map of nucleosome acetylation and methylation in yeast." Cell **122**(4): 517-527.
- Polesskaya, A., A. Duquet, et al. (2000). "CREB-binding protein/p300 activates MyoD by acetylation." J Biol Chem **275**(44): 34359-34364.
- Ponting, C. P., J. Schultz, et al. (1999). "SMART: identification and annotation of domains from signalling and extracellular protein sequences." Nucleic Acids Res **27**(1): 229-232.

- Puig, O., F. Caspary, et al. (2001). "The tandem affinity purification (TAP) method: a general procedure of protein complex purification." Methods **24**(3): 218-229.
- Qiu, J. and R. Elber (2006). "SSALN: an alignment algorithm using structure-dependent substitution matrices and gap penalties learned from structurally aligned protein pairs." Proteins **62**(4): 881-891.
- Recht, J., T. Tsubota, et al. (2006). "Histone chaperone Asf1 is required for histone H3 lysine 56 acetylation, a modification associated with S phase in mitosis and meiosis." Proc Natl Acad Sci U S A **103**(18): 6988-6993.
- Ringrose, L. and R. Paro (2004). "Epigenetic regulation of cellular memory by the Polycomb and Trithorax group proteins." Annu Rev Genet **38**: 413-443.
- Ringrose, L., M. Rehmsmeier, et al. (2003). "Genome-wide prediction of Polycomb/Trithorax response elements in *Drosophila melanogaster*." Dev Cell **5**(5): 759-771.
- Rogakou, E. P., D. R. Pilch, et al. (1998). "DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139." J Biol Chem **273**(10): 5858-5868.
- Rojas, J. R., R. C. Trievel, et al. (1999). "Structure of Tetrahymena GCN5 bound to coenzyme A and a histone H3 peptide." Nature **401**(6748): 93-98.
- Roscic, A., A. Moller, et al. (2006). "Phosphorylation-dependent control of Pc2 SUMO E3 ligase activity by its substrate protein HIPK2." Mol Cell **24**(1): 77-89.
- Ruthenburg, A. J., C. D. Allis, et al. (2007). "Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark." Mol Cell **25**(1): 15-30.

- Sampath, S. C., I. Marazzi, et al. (2007). "Methylation of a histone mimic within the histone methyltransferase G9a regulates protein complex assembly." Mol Cell **27**(4): 596-608.
- Sanchez-Pulido, L., D. Devos, et al. (2008). "RAWUL: a new ubiquitin-like domain in PRC1 ring finger proteins that unveils putative plant and worm PRC1 orthologs." BMC Genomics **9**: 308.
- Saunders, N. F., R. I. Brinkworth, et al. (2008). "Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites." BMC Bioinformatics **9**: 245.
- Schuettengruber, B., D. Chourrout, et al. (2007). "Genome regulation by polycomb and trithorax proteins." Cell **128**(4): 735-745.
- Schultz, J., F. Milpetz, et al. (1998). "SMART, a simple modular architecture research tool: identification of signaling domains." Proc Natl Acad Sci U S A **95**(11): 5857-5864.
- Schuster, T., M. Han, et al. (1986). "Yeast histone H2A and H2B amino termini have interchangeable functions." Cell **45**(3): 445-451.
- Schwartz, D., M. F. Chou, et al. (2009). "Predicting protein post-translational modifications using meta-analysis of proteome scale data sets." Mol Cell Proteomics **8**(2): 365-379.
- Sewalt, R. G., M. J. Gunster, et al. (1999). "C-Terminal binding protein is a transcriptional repressor that interacts with a specific class of vertebrate Polycomb proteins." Mol Cell Biol **19**(1): 777-787.
- Shechter, D., H. L. Dormann, et al. (2007). "Extraction, purification and analysis of histones." Nat Protoc **2**(6): 1445-1457.
- Shi, X., T. Hong, et al. (2006). "ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression." Nature **442**(7098): 96-99.

- Shogren-Knaak, M., H. Ishii, et al. (2006). "Histone H4-K16 acetylation controls chromatin structure and protein interactions." Science **311**(5762): 844-847.
- Simonsson, M., C. H. Heldin, et al. (2005). "The balance between acetylation and deacetylation controls Smad7 stability." J Biol Chem **280**(23): 21797-21803.
- Smith, C. M., P. R. Gafken, et al. (2003). "Mass spectrometric quantification of acetylation at specific lysines within the amino-terminal tail of histone H4." Anal Biochem **316**(1): 23-33.
- Smith, E. R., A. Eisen, et al. (1998). "ESA1 is a histone acetyltransferase that is essential for growth in yeast." Proc Natl Acad Sci U S A **95**(7): 3561-3565.
- Song, N., R. D. Sedgewick, et al. (2007). "Domain architecture comparison for multidomain homology identification." J Comput Biol **14**(4): 496-516.
- Sparmann, A. and M. van Lohuizen (2006). "Polycomb silencers control cell fate, development and cancer." Nat Rev Cancer **6**(11): 846-856.
- Sterner, D. E. and S. L. Berger (2000). "Acetylation of histones and transcription-related factors." Microbiol Mol Biol Rev **64**(2): 435-459.
- Stiff, T., M. O'Driscoll, et al. (2004). "ATM and DNA-PK function redundantly to phosphorylate H2AX after exposure to ionizing radiation." Cancer Res **64**(7): 2390-2396.
- Strahl, B. D. and C. D. Allis (2000). "The language of covalent histone modifications." Nature **403**(6765): 41-45.
- Strahl, B. D., P. A. Grant, et al. (2002). "Set2 is a nucleosomal histone H3-selective methyltransferase that mediates transcriptional repression." Mol Cell Biol **22**(5): 1298-1306.

- Suka, N., Y. Suka, et al. (2001). "Highly specific antibodies determine histone acetylation site usage in yeast heterochromatin and euchromatin." Mol Cell **8**(2): 473-479.
- Sung, S. and R. M. Amasino (2004). "Vernalization and epigenetics: how plants remember winter." Curr Opin Plant Biol **7**(1): 4-10.
- Tang, Y., M. A. Holbert, et al. (2008). "Fungal Rtt109 histone acetyltransferase is an unexpected structural homolog of metazoan p300/CBP." Nat Struct Mol Biol **15**(7): 738-745.
- Taunton, J., C. A. Hassig, et al. (1996). "A mammalian histone deacetylase related to the yeast transcriptional regulator Rpd3p." Science **272**(5260): 408-411.
- Taverna, S. D., H. Li, et al. (2007). "How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers." Nat Struct Mol Biol **14**(11): 1025-1040.
- Thompson, J. D., D. G. Higgins, et al. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice." Nucleic Acids Res **22**(22): 4673-4680.
- Tjeertes, J. V., K. M. Miller, et al. (2009). "Screen for DNA-damage-responsive histone modifications identifies H3K9Ac and H3K56Ac in human cells." EMBO J **28**(13): 1878-1889.
- Tomita, A., M. Towatari, et al. (2000). "c-Myb acetylation at the carboxyl-terminal conserved domain by transcriptional co-activator p300." Oncogene **19**(3): 444-451.
- Tordai, H., A. Nagy, et al. (2005). "Modules, multidomain proteins and organismic complexity." FEBS J **272**(19): 5064-5078.
- Tse, C., T. Sera, et al. (1998). "Disruption of higher-order folding by core histone acetylation dramatically enhances transcription of nucleosomal arrays by RNA polymerase III." Mol Cell Biol **18**(8): 4629-4638.

- Turner, B. M. (2008). "Simplifying a complex code." Nat Struct Mol Biol **15**(6): 542-544.
- van der Lugt, N. M., J. Domen, et al. (1994). "Posterior transformation, neurological abnormalities, and severe hematopoietic defects in mice with a targeted deletion of the bmi-1 proto-oncogene." Genes Dev **8**(7): 757-769.
- van Ingen, H., F. M. van Schaik, et al. (2008). "Structural insight into the recognition of the H3K4me3 mark by the TFIID subunit TAF3." Structure **16**(8): 1245-1256.
- VanDemark, A. P., M. M. Kasten, et al. (2007). "Autoregulation of the rsc4 tandem bromodomain by gcn5 acetylation." Mol Cell **27**(5): 817-828.
- Verreault, A., P. D. Kaufman, et al. (1998). "Nucleosomal DNA regulates the core-histone-binding subunit of the human Hat1 acetyltransferase." Curr Biol **8**(2): 96-108.
- Wang, G. G., J. Song, et al. (2009). "Haematopoietic malignancies caused by dysregulation of a chromatin-binding PHD finger." Nature **459**(7248): 847-851.
- Wang, H., L. Wang, et al. (2004). "Role of histone H2A ubiquitination in Polycomb silencing." Nature **431**(7010): 873-878.
- Wang, L., Y. Tang, et al. (2008). "Structure and chemistry of the p300/CBP and Rtt109 histone acetyltransferases: implications for histone acetyltransferase evolution and function." Curr Opin Struct Biol **18**(6): 741-747.
- Whitcomb, S. J., A. Basu, et al. (2007). "Polycomb Group proteins: an evolutionary perspective." Trends Genet **23**(10): 494-502.
- Wiederschain, D., L. Chen, et al. (2007). "Contribution of polycomb homologues Bmi-1 and Mel-18 to medulloblastoma pathogenesis." Mol Cell Biol **27**(13): 4968-4979.

- Wisniewski, J. R., A. Zougman, et al. (2007). "Mass spectrometric mapping of linker histone H1 variants reveals multiple acetylations, methylations, and phosphorylation as well as differences between cell culture and tissue." Mol Cell Proteomics **6**(1): 72-87.
- Wolf, D., M. Rodova, et al. (2002). "Acetylation of beta-catenin by CREB-binding protein (CBP)." J Biol Chem **277**(28): 25562-25567.
- Wynshaw-Boris, A. (2009). "Elongator bridges tubulin acetylation and neuronal migration." Cell **136**(3): 393-394.
- Wysocka, J., T. Swigut, et al. (2006). "A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling." Nature **442**(7098): 86-90.
- Xiao, A., H. Li, et al. (2009). "WSTF regulates the H2A.X DNA damage response via a novel tyrosine kinase activity." Nature **457**(7225): 57-62.
- Xie, W., C. Song, et al. (2009). "Histone h3 lysine 56 acetylation is linked to the core transcriptional network in human embryonic stem cells." Mol Cell **33**(4): 417-427.
- Xu, F., K. Zhang, et al. (2005). "Acetylation in histone H3 globular domain regulates gene expression in yeast." Cell **121**(3): 375-385.
- Xu, P., D. M. Duong, et al. (2009). "Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation." Cell **137**(1): 133-145.
- Yang, X. J. and E. Seto (2008). "Lysine acetylation: codified crosstalk with other posttranslational modifications." Mol Cell **31**(4): 449-461.
- Youdell, M. L., K. O. Kizer, et al. (2008). "Roles for Ctk1 and Spt6 in regulating the different methylation states of histone H3 lysine 36." Mol Cell Biol **28**(16): 4915-4926.

- Yu, C. N., T. Joachims, et al. (2008). "Support vector training of protein alignment models." J Comput Biol **15**(7): 867-880.
- Yuan, J., M. Pu, et al. (2009). "Histone H3-K56 acetylation is important for genomic stability in mammals." Cell Cycle **8**(11): 1747-1753.
- Zarrinpar, A., R. P. Bhattacharyya, et al. (2003). "The structure and function of proline recognition domains." Sci STKE **2003**(179): RE8.
- Zeng, L., K. L. Yap, et al. (2008). "Structural insights into human KAP1 PHD finger-bromodomain and its role in gene silencing." Nat Struct Mol Biol **15**(6): 626-633.
- Zhang, J., X. Shi, et al. (2008). "Acetylation of Smc3 by Eco1 is required for S phase sister chromatid cohesion in both human and yeast." Mol Cell **31**(1): 143-151.
- Zhang, K., F. Faiola, et al. (2005). "Six lysine residues on c-Myc are direct substrates for acetylation by p300." Biochem Biophys Res Commun **336**(1): 274-280.
- Zhang, K. and H. Tang (2003). "Analysis of core histones by liquid chromatography-mass spectrometry and peptide mapping." J Chromatogr B Analyt Technol Biomed Life Sci **783**(1): 173-179.
- Zhang, K., H. Tang, et al. (2002). "Identification of acetylation and methylation sites of histone H3 from chicken erythrocytes by high-accuracy matrix-assisted laser desorption ionization-time-of-flight, matrix-assisted laser desorption ionization-postsource decay, and nanoelectrospray ionization tandem mass spectrometry." Anal Biochem **306**(2): 259-269.
- Zhang, K., K. E. Williams, et al. (2002). "Histone acetylation and deacetylation: identification of acetylation and methylation sites of HeLa histone H4 by mass spectrometry." Mol Cell Proteomics **1**(7): 500-508.



Zhang, L., E. E. Eugeni, et al. (2003). "Identification of novel histone post-translational modifications by peptide mass fingerprinting." Chromosoma **112**(2): 77-86.