Probing the Interfaces of Epigenetic Complexes: Efforts Towards Elucidating and Targeting Critical Protein:Protein and Protein:lncRNA Interactions of Lysine-Specific Demethylase 1 (KDM1A/LSD1)

By

Meghan Frances Lawler

Department of Chemistry

Duke University

Date:_____

Approved:

_____

Dewey G. McCafferty, Supervisor

_____

Emily Derbyshire

_____

Michael Lynch

_____

Qiu Wang

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Chemistry in the Graduate School
of Duke University

2019

ABSTRACT

Probing the Interfaces of Epigenetic Complexes: Efforts Towards Elucidating and Targeting Critical Protein:Protein and Protein:lncRNA Interactions of Lysine-Specific Demethylase 1 (KDM1A/LSD1)

by

Meghan Frances Lawler

Department of Chemistry

Duke University

Date:_____

Approved:

_____

Dewey G. McCafferty, Supervisor

_____

Emily Derbyshire

_____

Michael Lynch

_____

Qiu Wang

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Chemistry in the Graduate School of Duke University

2019

Abstract

The post translational modification (PTM) of histone proteins is a highly dynamic process that is utilized in the control of gene transcription. This epigenetic process involves enzymatic 'writers' and 'erasers' which place or remove chemical modifications to the unstructured tails of histone proteins which protrude out from the nucleosomal core. In a highly dynamic manner, each PTM is spatiotemporally regulated and combinations of PTMs at a gene promotor or enhancer region leads to transcriptional enhancement or repression. The gene targets as well as selectivity and specificity of epigenetic enzymes is regulated by the multimeric complexes each enzyme is co-opted. Each complex contains a unique set of coregulatory proteins with RNA and DNA binding domains and PTM 'reader' domains to direct the catalytic machinery to a specific subset of genes. The coregulatory proteins also affect the specificity and selectivity of the enzyme through mechanisms which are only beginning to be explored.

Our interest is in elucidating the role of coregulatory proteins and lncRNA with respect to lysine-specific demethylase 1 (LSD1/KDM1A). A flavin-dependent mono-and di-demethylase of H3K4me1/2 and H3K9me1/2, KDM1A has been implicated in many different multimeric enzymatic complexes which, in some cases such as the REST and NuRD complexes, function on opposing pathways. This disparity in the downstream outcome being coordinated by the same enzyme highlights the need to understand not

only epigenetic enzymes, but to consider the complexes as a whole towards therapeutic targeting.

The specific aims of my thesis were to (a) interrogate the role of individual and multiple coregulatory partners in enzyme selectivity and specificity (b) establish tools to study the mechanisms of biochemical and biophysical of protein:protein and protein:lncRNA interactions and (c) elucidate key characteristics of protein:protein and protein:lncRNA interfaces towards targeted disruption. To this end, I have utilized cloning and mutagenesis methods towards heterologous expression and purification of coregulatory partners of KDM1A in *E. coli*. I chose coregulatory partners found in a common catalytic core as well as several additional coregulatory proteins from a stable KDM1A-containing 5-mer complex. I have produced multiple constructs for four of these proteins to allow for multiple affinity purification routes as well as for future binding studies. I have further expressed each of these constructs and have made significant efforts towards the purification of each construct based on solubility.

I furthermore established HDX-MS and SELEX protocol in our lab as tools to allow us to explore the dynamics of these epigenetic interactions. I further demonstrated and confirmed that there is no hotspot along the binding interface between KDM1A and CoREST, but that CoREST stabilizes the apical end of the KDM1A tower domain via HDX-MS with the highest change in deuterium uptake, over 20%, long KDM1A TαA residues 440-451.

I also made significant efforts towards elucidating the interaction between KDM1A and HOTAIR. Firstly, I established an RNA radiolabeled EMSA assay for the lab which allowed us to test the binding of HOTAIR to KDM1A. With this assay, we saw that $CoREST_{286-482}$, specifically the linker region (residues 293-380), must be bound to KDM1A for HOTAIR to bind and that the dissociation constant was unchanged at 1.71±0.38 μM and 1.29±0.34 μM, respectively. Further, I confirmed that the first 320 nt of domain 4 of HOTAIR (nt 1500-1820) contain the critical binding and that the dissociation constant was slightly higher at 2.97±0.96 μM.

I have also optimized SHAPE-MaP and crosslinking strategies to explore the binding interface between KDM1A:$CoREST_{286-482}$ and lncRNA. I determined that there were 83 nt that displayed at least a 1.5-fold change in SHAPE reactivity of HOTAIR D4 due to the presence of KDM1A:$CoREST_{286-482}$. I also utilized a free-energy based secondary structure model to establish a secondary structure for HOTAIR D4 based on my SHAPE-MaP data. I noted that 44% of the significant nt were confined to a stretch of RNA (nt 1538-1610, 1779-1844) that is predominantly dsRNA. Further usage of photochemical crosslinking strategies revealed a propensity for G:C paired nt to be crosslinked to KDM1A:$CoREST_{286-482}$. A similar nt sequence around these paired nt suggests a binding motif.

Dedication

These last five years have truly been a journey. I came to graduate school at Duke University ready to learn and partake in exciting and relevant science. So first, and foremost, thank you to my advisor, Dewey McCafferty, for allowing me the space to explore but also to grow as a scientist and person. In addition, I have grown enormously as a person in ways that I could not even have imagined, and there are many people that I would like to take a moment to thank and dedicate this work to.

None of this would have been possible without the support and belief the faculty at Western Carolina University had in me. Dr. William Kwochka, you sparked and fostered my interest in science, and encouraged me every step of the way. Your mentorship and guidance has been unparalleled – I never would have considered graduate school without you. Dr. David Evanoff, thank you for fostering my skills as an independent researcher while at Western, and for all of the advice and guidance during my time at Duke. Brittania Bintz, thank you for pushing me and teaching me to have the confidence to stand up for myself and to push for what I believe in. Bob Buckner, David Starnes, Matt Henley, and Jon Henson: the four of you kept me grounded and reminded me that there will always be time outside of science to follow my other passions. Thank you to the four of you for seeing the potential in me when I couldn't see it in myself.

To my parents, Mary and Kelly Lawler, thank you for your unwavering support and love, it hasn't always been easy, but we've made it work the best way we know

how. Thank you for always granting me opportunities to do what I need to do and not questioning my path. I am so blessed to have such supportive parents in my life, and I will never be able to thank you enough. There truly aren't the words to express how grateful I am to not only have you as parents, but friends.

To my siblings, Jonathan, Grace, and Shannon Lawler, each of you keep me balanced. Jonathan, thank you for all of the Friday evening conversations, you have no idea how much I looked forward to them some weeks. Grace, thank you for being a ray of sunshine and always willing to just talk even when I have nothing to say. Shannon, thanks for always being 100% authentic to yourself, it reminds me to do the same. I am so proud of each of you for pursuing your goals and dreams with drive and purpose, it has been a blessing watching you grow, and I can't wait to keep watching your journeys.

To my friends and Kappa Kappa Psi brothers, thank you for standing by my side and for being my family. Each of you have shaped and changed my perspective and life in ways that I don't think I'll ever be able to fully grasp. I am so grateful for each of you and the lessons you have taught me, the good and the bad, I am a stronger, more compassionate, and more conscientious person because of you. I'd like to specifically thank Brittany Clark, Rachel Caddell, Mary Kate Moore, Katie Durham, Nina Fiore, Jesse Nixon, Alexandra and Andrew Hall, Cameron Holmes, Chuck Hendrick, Christiana Gooden, Katherine and Bumki Kim, Alex Chaconas, and Ken Maksimchuk. Thank you to each of you for keeping me alive, sane, and well-fed over the past years.

You all have each stood by me and stood in my corner through everything, and I can never thank you enough. The confidence each of you hold in me gave me strength when I most needed it. I'd also like to thank those that couldn't walk with me through my entire time here at Duke, you each taught me of my individual strength and courage. I cherish all of the memories I have with each of you and wish you nothing but the best. "

I'd like to also thank my undergrads, Rebecca Haley and Maegan Burns. I know this project was not the easiest or straightforward, but the dedication and drive that both of you have has made me a better person and mentor and that is invaluable to me.

Finally, thank you to my partner-in-crime, Rudy. You have been my guy for the past five years, and I don't think I could have done this without you. Your willingness to stay up during late nights to keep me company kept me moving. Thank you also for keeping me grounded, and reminding me to get outside and take walks, you're the best.

"Courage doesn't always roar. Sometimes it's the quiet voice at the end of the day whispering 'I will try again tomorrow'" –Mary Anne Radmacher

Contents

List of Tables

List of Figures

List of Schemes

Abbreviations

1-M7	1-methyl-7-nitroisatoic anhydride
2-DSC	2-Dimensional Combinatorial Screening system
5-NIA	5-nitroisotoic anhydride
AOD	Amine oxidase domain
ASO	Antisense oligonucleotide
BAP	Bacterial alkaline phosphotase
BME	Betamercaptoethanol
bp	Basepair
cDNA	Complementary DNA
Conc	Concentration
CoREST	REST corepressor 1
CtBP	C-terminal binding protein
DMS	Dimethyl sulfate
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
ds	Double stranded
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EMT	Epithelial-to-mesenchymal

H3K4	Histone H3 Lysine 4
H3K9	Histone H4 Lysine 9
HDAC1/2	Histone deacetylase 1 and 2
HDX-MS	Hydrogen deuterium exchange with mass spectrometry
HOTAIR	Hox antisense transcript RNA
IPTG	Isopropyl β-D-1-thiogalactopyranoside
KDM1A	Lysine-specific demethylase 1
LB	Lennox broth
lncRNA	Long non-coding RNA
MaP	Mutational analysis profiling
me	Methylation
miRNA	Micro RNA
mRNA	Messenger RNA
NaCl	Sodium Chloride
ncRNA	non-coding RNA
NGS	Next generation sequencing
nm	nanometer
nt	Nucleotide
NuRD	Nucleosome remodeling and deacetylase complex
OD600	Optical density 600

pM	Picomolar
pmol	Picomoles
PNA	Peptide nucleic acid
PRC2	Polycomb repressive complex 2
PTM	Post-translational modification
RNA	Ribonucleic acid
rpm	Revolutions per minute
rRNA	Ribosomal RNA
RT	Reverse transcriptase
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
SELEX	Systematic evolution of ligands by exponential enrichment evolution
SHAPE	Selective 2'-hydroxyl acylation analyzed by primer extension
SOC	Super optimal broth
ss	Single stranded
SSII	Superscript II
START	Structure Activity Relationships Through Sequencing
TB	Terrific broth
tRNA	Transfer RNA
Vt	Total volume

WB	Western Blot
x g	Gravitational acceleration
μM	Micromolar
μmol	Micromoles

Acknowledgements

First, and foremost, I would like to acknowledge my advisor, Dewey McCafferty. Thank you for believing in my abilities and allowing me the opportunity to work and learn in your group. I have grown incredibly as a scientist under your guidance and direction. I also appreciate your willingness to allow me to step into the RNA world and your trust in my abilities that I would be able to work without direct guidance from you on this project.

Thank you to my committee, each of you have provided me with invaluable insight and direction over the past years both professionally, scientifically, and personally. The different background and perspective each of you hold has allowed me to gain insight and expand my knowledge-base.

I would like to acknowledge not only the Chemistry department at Duke University, but also the Pharmacology and Cancer Biology department as well as those that headed and mentored me during my time in the PSTP and GAANN training programs. Graduate school may be designed to aid in our growth as scientists, but each of these programs and departments enabled and encouraged my interests in translational science as well as teaching and mentorship. My access to these opportunities has substantially aided my decision-making process in career paths.

Finally, I would also like to specifically thank Dr. Katherine Alser Kim, Dr. Jennifer Link Schwabe, Dr. Jonathan Burg, Dr. Ken Maksimchuk, Dr. Chase Weidmann,

Dr. Matthew Disney, Dr. Kevin Weeks, Dr. Erik Soderblom, Dr. Will Thompson, Samuel Olson, Maegan Burns, Rebecca Haley, and all other members of the McCafferty lab. Each of you contributed and supported my research goals during my time at Duke and I am extremely appreciative. I am fortunate to have been able to work with such great minds, and learn how to approach scientific design, optimization, and data analysis. This dissertation is the culmination of everything I have learned from each of you and would not have been possible without you.

1. Introduction

1.1 Epigenetics

Meaning "above genetics" epigenetics encompasses the disparity between the number of available genes and the transcription and translation of only those required for a specific cell's function. The first utilization of the term epigenetics was in 1942 by Conrad Hal Waddington to discuss the mechanisms employed to enable a zygote, which is "more or less equivalent to the genotype" to a fully developed organism *(1)*. He employed this term in order encompass the disparity between genotype and a phenotype *(2,3)*.

Waddington's ideas regarding regulatory mechanisms between genetic information and what would now be considered transcription and translation of that information still hold true at the basis of what epigenetics encompasses. However, it was David Nanney who first proposed that cellular genetic material was maintained via a "template replicating mechanism" wherein control of which genes would be expressed or repressed was controlled by an "auxiliary mechanism" *(4)*.

While the current field of epigenetics has a firmer grasp on the molecular auxiliary mechanisms to affect the expression of a gene, these initial attempts to depict and describe the variation and differentiation between not only cell types within an organism, but more broadly the apparent influence of the environment on genotype and phenotype are still the basis for epigenetics. The current field of epigenetics

encompasses the regulation of gene transcription and translatin. Currently, the 'central dogma' describes the process through a gene is transcribed into an mRNA product which is then processed, shuttled to the cytoplasm and then translated to protein.

Our interest lies in the ability of a cell to express and repress genes in a spatiotemporal manner across various cell types. There are many mechanisms through which this occurs, but they can broadly be characterized into: (a) DNA modification, (b) RNA modification, and (c) histone modification. Our focus for the remainder of this dissertation will be (c) the post-translational modification (PTM) of histone proteins and we direct the readers to several comprehensive reviews regarding epigenetic DNA and RNA modifications *(5–8)*.

1.2. Post-translational Modification of Histones

1.2.1. DNA packaging

With over 2 meters of genomic DNA per normally dividing somatic cell, there is a clear and present need for packaging of DNA within the nucleus. The nucleosome core particle consists of four histone proteins form an octameric core which roughly 146 bp of DNA is wound around as shown by the original structure solved by Richmond and Luger in 1997 *(9)*. Each of the four-histone proteins contain a high level of basic lysine and arginine residues which contribute to interactions with the phosphate backbone of the DNA and stabilization of the nucleosome (Figure 1) *(10–12)*.

The assembly of the nucleosome occurs via placement of an $(H3\text{-}H4)_2$ tetramer followed by subsequent H2A-H2B dimers to cap the tetramer to form the globular structured histone octamer core *(11,13–18)*. *In vitro,* the protein NAP1 is able to interact with and deposit first the H3/H4 tetramer and then subsequently each H2A/H2B dimer in succession; however, *in vivo* there are a series of proteins believed to be utilized through this process *(14,16)*. The placement of nucleosomes in regularly spaced increments allows for higher order condensation into 30-nanometer chromatin fibers, which loop averaging 300 nm in length. These loops can be further compressed to a 250 nm width and coiled into an x-shaped 'chromosome.' (Figure 2) *(19)*.

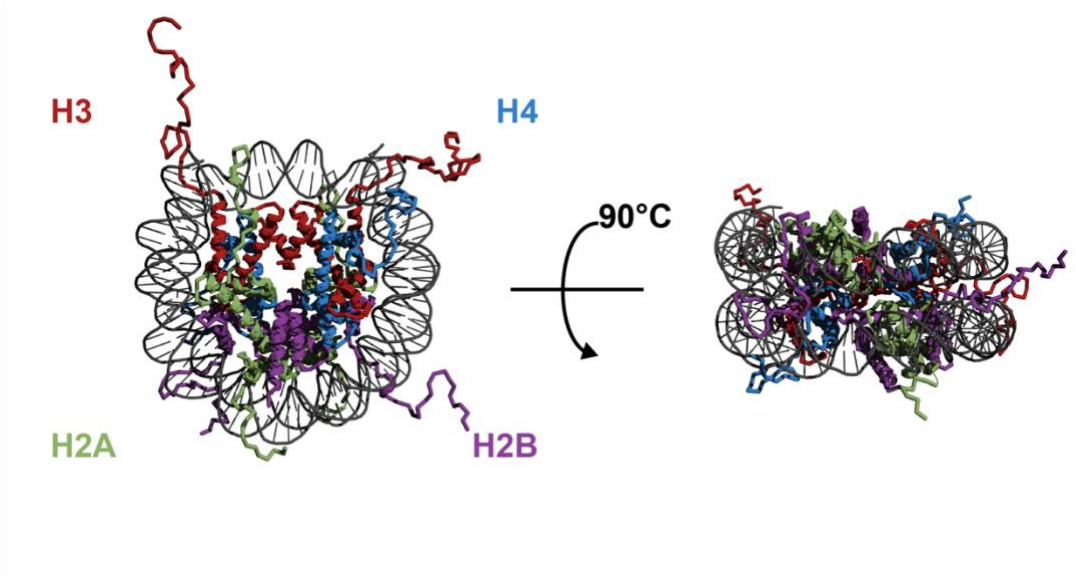


Figure 1: Overall nucleosome structure. The nucleosome consists of roughly 146 bp of DNA (grey) wrapped 1.7 times around an octameric core of two copies of histone proteins: H2A (green), H2B (purple), H3 (red), H4 (blue). The assembly of the nucleosome occurs via placement of an (H3-H4)2 tetramer followed by subsequent H2A-H2B dimers to cap the tetramer to form the globular structured histone octamer core. (PDB 1AOI) (9)

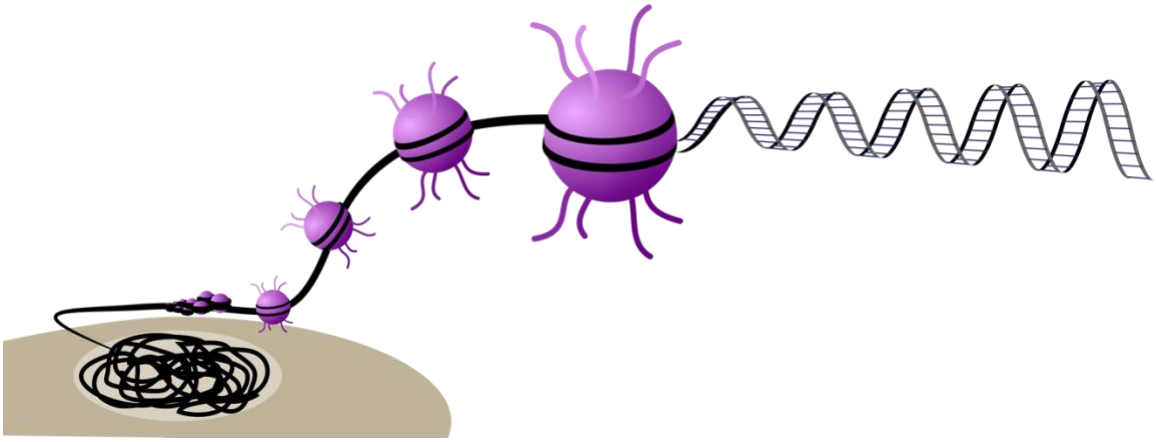


Figure 2: Basic scheme of condensation of DNA within the nucleus. As DNA (black) is wound around core histone octamers (purple), this creates what is referred to as 'beads on a string' which can be further condensed via additional contacts between DNA and the basic histone residues within the nucleus (light tan) of the cell (dark tan).

The condensation of DNA around histone proteins is highly compact and requires local structural remodeling to allow for gene transcription to occur. *(11,16,20,21)*. In cases of heterochromatin, the nucleosomes are more closely spaced, and the region of DNA is more condensed resulting in transcriptionally silenced genes. In the case of euchromatin, the additional spacing between nucleosomes allows for genes to be transcriptionally activated (Figure 3) *(20,22)*. To initiate the removal of histone octamers to allow for gene transcription, or subsequent replacement of histone octamers to silence a gene, multimeric nucleosome remodeling as well as RNA polymerase protein complexes must be signaled to a gene promoter site *(23,24)*. The signaling cascade of gene activation and repression is regulated via post-translational modification (PTM) of the histone proteins.

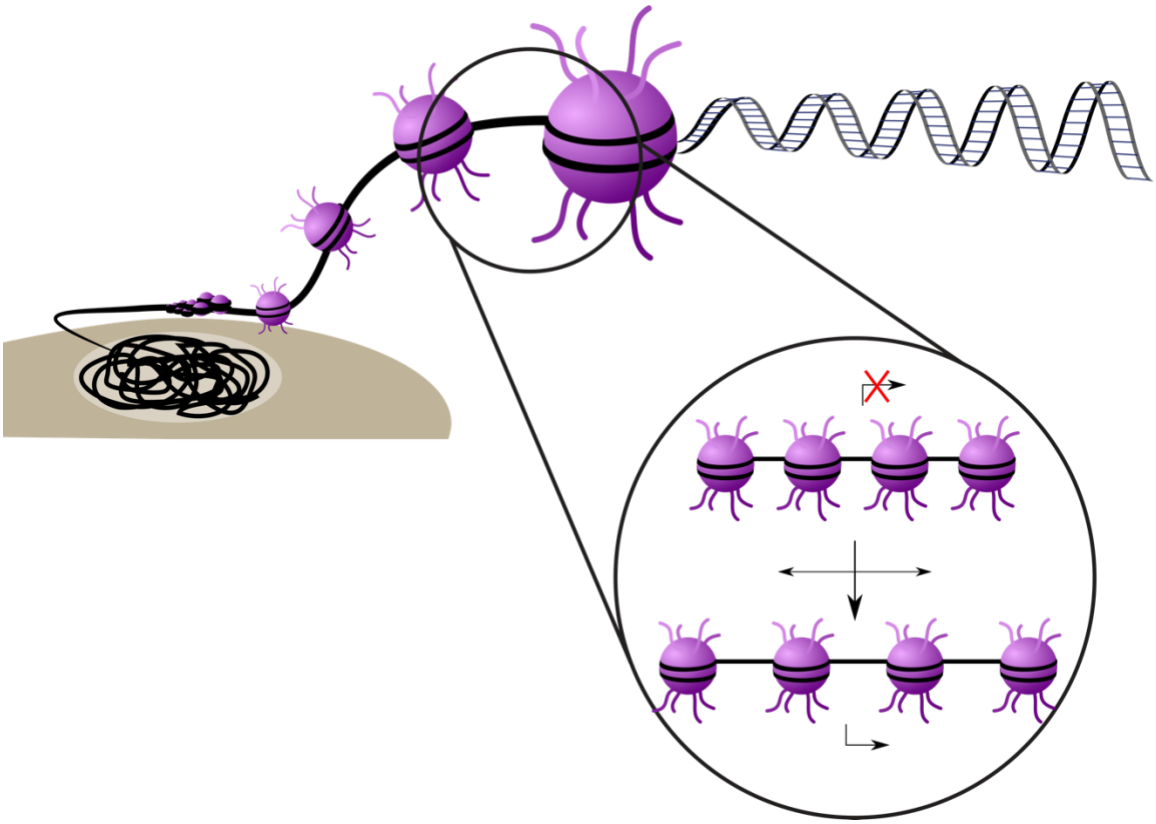


Figure 3: Representation of hetero- and euchromatin. In cases of heterochromatin, the nucleosomes are more closely spaced, and the region of DNA is more condensed resulting in transcriptionally silenced genes. In the case of euchromatin, the additional spacing between nucleosomes allows for genes to be transcriptionally activated.

1.2.2 Histone post-translational modification (PTM)

The majority of the histone proteins are involved in the formation of the core octamer, the unstructured amino-terminus of each histone protein protrudes out from the nucleosome, as shown by H3 and H2B histones in Figure 1. Each histone N-terminal 'tail' is extensively post-translationally modified (PTM) via methylation, acetylation, phosphorylation, ubiquitination, and SUMOylation via epigenetic enzymes in a site selective and specific manner (Figure 4).

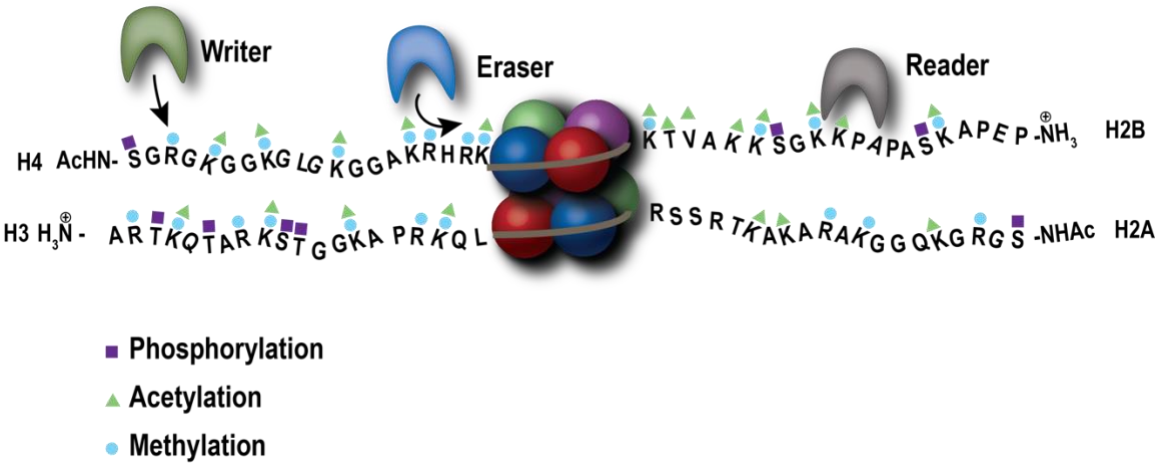


Figure 4: PTM of histone tails via epigenetic enzymes referred to as 'writers' and 'erasers.' Phosphorylation, acetylation, and methylation are only three of many reported modifications to the tails, each with spatio- and temporal-specific qualities and impact on gene regulation

PTM of histone tails affect transcriptional regulation through targeting necessary remodeling and other enzymatic complexes or by altering DNA:histone interactions. Acetylation of lysine residues histones is generally regarded as a gene activating mark, and deacetylation as a gene repressing mark. This is due to the positively charged lysine and arginine residues contributing to interactions with the negatively charged DNA phosphate backbone. Experimental studies and computational models have shown that charge loss at one lysine residue can destabilize the nucleosome by up to 15 kcal/mol, loosening the association of protein to DNA. This enables remodeling complexes to more easily transition the local chromatin structure to euchromatin for activation of gene transcription *(25–29)*. Thus, acetylation of histones is generally regarded as a gene activating mark, and subsequent deacetylation as a gene repressing mark.

Unlike acetylation, methylation of lysine residues proves to be less straightforward and has been demonstrated to be both activating and repressive depending on the residue *(30–32)*. For example, H3K4me2 has been linked to gene activation and only five residues away, H3K9me2 is associated with gene repression *(30–32)*. It is likely that methylation acts in concert with additional PTMs on a nucleosome to target and recruit specific remodeling complexes or transcription factors.

The presence of one mark may preclude the addition of a secondary mark due to steric hindrance *(33–35)*. For example, the repressive H3K9me2 mark does not allow for the addition of a potentially activating H3S10ph mark. If the two marks are in opposition with respect to gene activation or repression, the first mark would need to be removed before the second could be deposited. Not only is this an additional layer of regulation, it speaks to the high level of balance that occurs on the nucleosomes and that transitioning from an active to repressed gene state is highly controlled.

In addition to steric hinderance playing a role in the positioning and placement of PTMs, and potentially gene regulation, there is an increasing number of studies to suggest a high level of interplay between specific marks at a single histone tail. This is referred to the 'histone code hypothesis' wherein it is the combination of multiple marks which leads to an overall change in gene activation or repression *(36)*. On a broader scale, this suggests an interplay between the enzymes which catalyze the addition and removal of each PTM.

1.3 Epigenetic 'readers,' 'writers,' and 'erasers'

The enzymes which catalyze the reactions for histone PTMs can be classified as 'writers' which deposit the PTM, or 'erasers' which subsequently remove the PTM. Each epigenetic enzyme is catalytically specific, and most are specific for a select number of histone residues (Table 1). The regulation of these enzymes is what drives the correct placement of each PTM at not only the correct residue but also in a temporal manner with respect to cellular response. Each epigenetic enzyme function on specific residues at subsets of gene loci. The origin of this selectivity and specificity remains unclear in many cases. An enzyme's active site structure may provide some substrate selectivity whereas additional gene targeting, and refined substrate selectivity arises from protein-protein interactions.

There have been many cases where the dysregulation of an epigenetic enzyme has been demonstrated to be a driver in cancer cell proliferation, extensively covered in these reviews *(37–41)*. Historically, efforts to restore the balance of gene expression in these cancerous cells have been focused on inhibiting the active site of the enzyme. There are several drug therapeutics currently approved for clinical use that target a specific epigenetic enzyme. One example is the histone deacetylase (HDAC) inhibitors SAHA and depsipeptide (vorinostat and romidepsin, respectively) in several different cancers where an increased level of HDAC activity has been demonstrated to drive cancer progression and inhibition decreases proliferation *(42–45)*. In addition to HDACs,

demethylase lysine-specific demethylase 1 (LSD1/KDM1A) is commonly reported to be overexpressed in cancer phenotypes and is of interest for inhibition *(41,46–50)*.

Table 1: Histone posttranslational modifications, positions, and representative associated enzymes (adapted from *(10,51,52)*

Histone Modifications	*Residues Modified*	*Positions Modified*	*Enzymes*
Acetylation (ac)	**K**ac	H3K9	KAT2A
		H3K14	KAT2A/B, KAT8
		H3K18	KAT2A/B
		H3K23	KAT2A/B, KAT8
		H3K27	KAT2A
		H4K5	KAT2B, KAT7, KAT8, HAT1
		H4K8	KAT2B, KAT7, KAT8
		H4K12	KAT7, KAT8, HAT1
		H4K16	KAT2A, KAT8
		H2AK5	KAT2B, KAT8
		H2BK12	KAT2B, ATF2
		H2BK15	KAT2B, ATF3
Deacetylation (ac)	**K**ac	H3K9	Class III HDACs
		H3K14	Class III HDACs
		Multiple positions	Class I HDACs (1, 2, 3, and 8)
		Multiple positions	Class II HDACs (4, 5, 6, 7, and 9)
Methylation (me) (K)	**K**me1/2/3	H3K4	MLL, Set7/9, KMT2D, Set1
		H3K9	SUV38H1, SUV39H2, G9a, SETDB1, EHMT1, KMT1F
		H3K27	EZH2
		H3K36	SETD2, SMYD2, NSD1
		H3K70	KMT7
		H4K20	KMT5B, SETD8
Demethylation (me) (K)	**K**me1/2/3	H3K4	KDM1A, KDM1B, KDM5A-D
		H3K9	KDM1A, KDM3A, KDM4A-D
		H3K27	KDM6A, KDM6B

		H3K36	KDM2A, KDM2B, KDM4A-C
Methylation (me) (R)	**R**me1, **R**me2	H3R2/R17/R26	PRMT4
		H4R3	PRMT1
Phosphorylation (ph)	**S**ph, **T**ph	H3S10	AURKB
		H3T3	GSG2
		H3S28	RPS6K5A, RPS6KA4
		H4S1	CSNK2A1

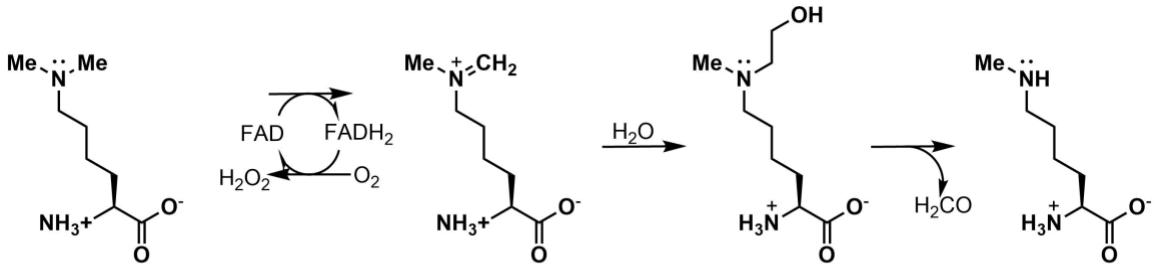
1.3.1 Lysine-specific demethylase 1 is an eraser with important biological functions

1.3.1.1 KDM1A is comprised of three distinct domains which mediate catalysis and coregulatory protein interactions

Lysine-specific demethylase 1 (BHC110/LSD1/KDM1A) is an FAD-dependent mono- and di-demethylase comprised of three domains: a SWIRM domain (residues 172-272), an amine oxidase catalytic domain (AOD) (residues 272-415, 515-852), and a novel 'tower' domain (residues 415-515) which disrupts the AOD (Figure 5a) *(53,54)*. KDM1A has been demonstrated to demethylate at H3K4 and H3K9 and additionally has non-histone substrates such as p53 *(55–59)*. The three-dimensional structure of KDM1A has been solved via X-ray crystallography in unliganded as well as in complex with peptide ligands and inhibitors (Figure 5b) *(53,54,60–63)*. The AOD of KDM1A contains a single, non-covalently bound FAD molecule and the AOD can be separated into substrate binding (357-415, 515-558, and 658-769) and co-factor binding (residues 272-356, 559-657, and 770-852) lobes.

Demethylation of a lysine residue by KDM1A occurs via direct hydride transfer by the flavin to oxidate the substrate. This step is proceeded by hydrolysis by bulk solvent of the iminium intermediate. This produces a hemiaminal that decomposes into formaldehyde (Scheme 1) *(32,64)*. A molar equivalent of H_2O_2 is produced upon reoxidation of the flavin *(65)*.

Scheme 1: Catalytic mechanism of KDM1A



The SWIRM domain is comprised primarily of α-helices and creates a hydrophobic interface with the C-terminus of the AOD. Decreased catalytic activity and stability of KDM1A were demonstrated when mutations were introduced to this interface *(53)*. Found commonly in chromatin associated proteins, SWIRM domains have been demonstrated to bind DNA *(66)*. The conserved DNA-binding residues (α6, residues 247-259) are occluded by the AOD domain and accordingly, KDM1A has not been demonstrated to bind DNA *(67,68)*.

The Tower domain of KDM1A is comprised of two antiparallel α-helices, Tαa and TαB, bridged by the five residue sequence KPPRD (Figure 5b) *(54)*. Previously believed to be unique amongst demethylases, recent hydrogen-deuterium exchange

with mass spectrometry (HDX-MS) and small angle X-ray scattering (SAXS) evidence suggests that KDM5B may also contain a Tower-like domain *(69)*. The Tower-like domain of KDM5B may be important for nucleosomal binding, though the exact role has yet to be established *(70)*. In KDM1A, the functional role of the Tower domain is as the site of association with proteins such as REST co-repressor 1 (CoREST1) (Figure 5a) and metastasis associated tumor protein (MTA1) *(32,53,71,72)*. The structure of KDM1A with CoREST has been established with X-ray crystallography *(53,56)*. It is the linker region of CoREST (Figure 5b) which makes the most extensive contacts with KDM1A along the tower domain forming a coiled-coil motif (Figure 5c).

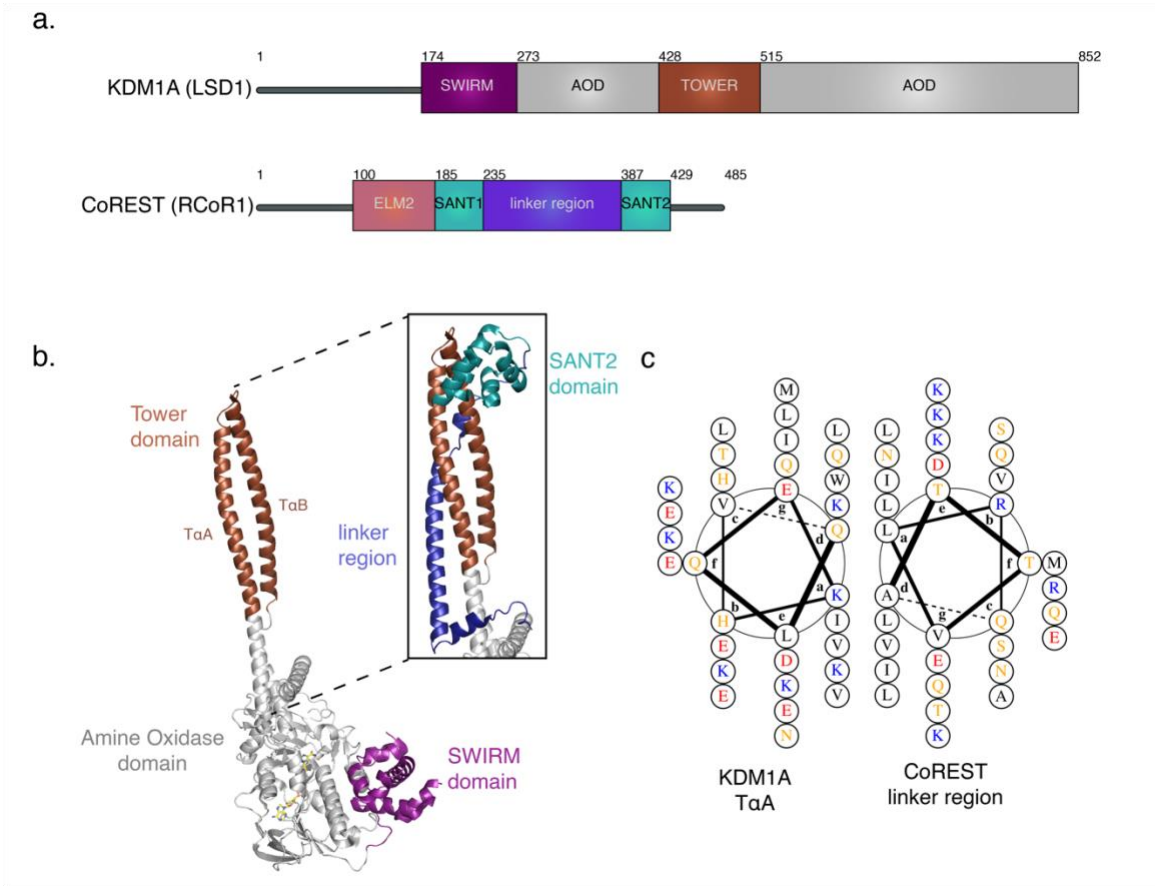


Figure 5: Overview of KDM1A interaction with binding partner CoREST. KDM1A is a FAD-dependent mono- and di-demethylase and specifically demethylates histone residues H3K4 and H3K9 (a) KDM1A contains a SWIRM, AOD, and Tower domain. Binding partner CoREST contains an ELM2, SANT1 and SANT2 domains and in intervening linker region. (b) the structure of KDM1A and KDM1A with CoREST have been solved to reveal a long-binding interface along the tower domain. (c) The interactions between KDM1A and CoREST α-helices represent a coiled-coil motif.

1.3.1.2 Functional roles of KDM1A in cells through the complexes which utilize the demethylase

The role of KDM1A in repression of neuronal genes in non-neuronal cells is one of the first reported functions of KDM1A *(73,74)*. Since then, KDM1A has been further linked to the cellular differentiation of various cell types including pituitary, osteogenic,

and embryonic *(75–78)*. In pluripotent stem cells, regions of chromatin around genes linked to cellular lineage are held in a 'poised' state containing both the repressive H3K27me3 mark as well as the activating H3K4me3 mark *(79)*. KDM1A is a component of cellular machinery that act to rapidly activate or silence long-term the lineage-specific genes depending upon upstream differentiation cues *(80)*. KDM1A is additionally integral in maturation of several hematopoietic cell lines *(76,81–85)* and is involved as well in the monocyte-to-macrophage transition *(86)*.

In addition to its role in cellular differentiation, KDM1A has also been linked to the epithelial-to-mesenchymal transition (EMT) *(87,88)*. In this process, cells transition from adhesive epithelial cells a to a mobile mesenchymal phenotype . In addition to regulating upstream promoters of the EMT such as Wnt, NF-kB and Notch, KDM1A is also an established partner of SNAI1 which functions to silence epithelial genes to promote the EMT pathway *(50,89–92)*.

The EMT is an essential pathway for development; however, a transition to a mesenchymal state is a key rate of metastatic and aggressive cancers *(88,92)*. Inhibition or knockdown of KDM1A in several cancer types such as esophageal squamous cell carcinoma, small cell lung carcinoma, lung cancer, and colon cancer decreases motility of the metastatic cells *(93–96)*.

KDM1A has been widely implicated in the development and progression of many cancer types and is well-documented in neuroblastoma *(97–100)*, colorectal cancer

(101–103), leukemia *(85,104–107)*, and hormone-dependent cancers such as breast and prostate cancer *(50,57,58,108–111)*. Similar to the EMT state, in many cases there are reports that KDM1A knockdown or inhibition decreases the viability of the cancer cell survival *(47,93,104,112)*.

There are a series of active site inhibitors of KDM1A, as cited above; however, to date there are no FDA approved KDM1A-inhibitors on the market. This may be due in part to the wide-range of functionally important interactions that KDM1A makes within each cell. A pan-inhibitor of KDM1A will non-selectively target KDM1A. The aberrantly functioning KDM1A will be inhibited; however, so will the fraction of enzyme which is still functioning normally. Each function of KDM1A is dictated by recruitment by coregulatory proteins. Establishing the mechanisms for the pathobiological functions of KDM1A will enable a targeted approach through specific disruption of protein interactions over active site inhibition.

CoREST is the most well-established binding partner of KDM1A. This protein functions as a scaffolding protein in several complexes containing both KDM1A and HDAC1/2 wherein the linker region and SANT 2 domain contribute to the binding interface with KDM1A and binding to HDAC1/2 occurs via the N-terminal ELM/SANT1 domains of CoREST (Figure 6) *(74,113–116)*. Our lab has previously reported that the linker region is predominantly responsible for the low nanomolar affinity to KDM1A (K_d =7.79 nM) with no significant change of binding affinity upon inclusion of the SANT2

domain *(117)*. The SANT2 domain of CoREST facilitates demethylation of nucleosomes and was demonstrated by the Mattevi lab to bind nucleosomal DNA *(118)*. CoREST has been demonstrated to be required for KDM1A activity on a nucleosomal substrate. It may be that CoREST is required to bind and direct KDM1A to its substrate *(118)*. The interaction with CoREST is at the core of several KDM1A-containing multimeric complexes (Figure 7) and the role of these complexes are outlined below.

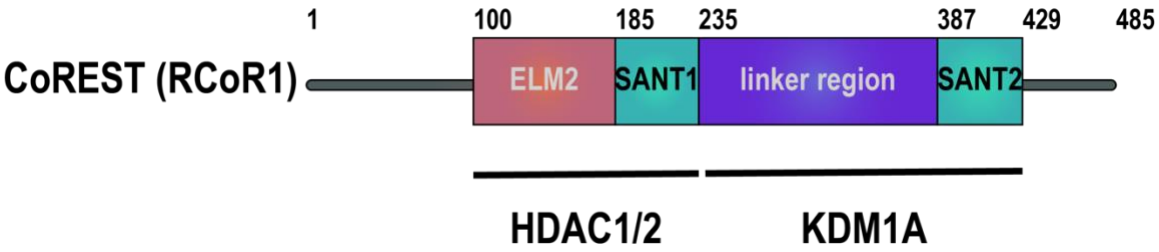


Figure 6: REST Co-repressor 1 (RCoR1/CoREST) domain map. CoREST contains two SANT domains, which historically bind DNA, an ELM2 domain involved in protein dimerization as well as a 'linker region' which contributes a majority of the binding affinity between KDM1A and CoREST.

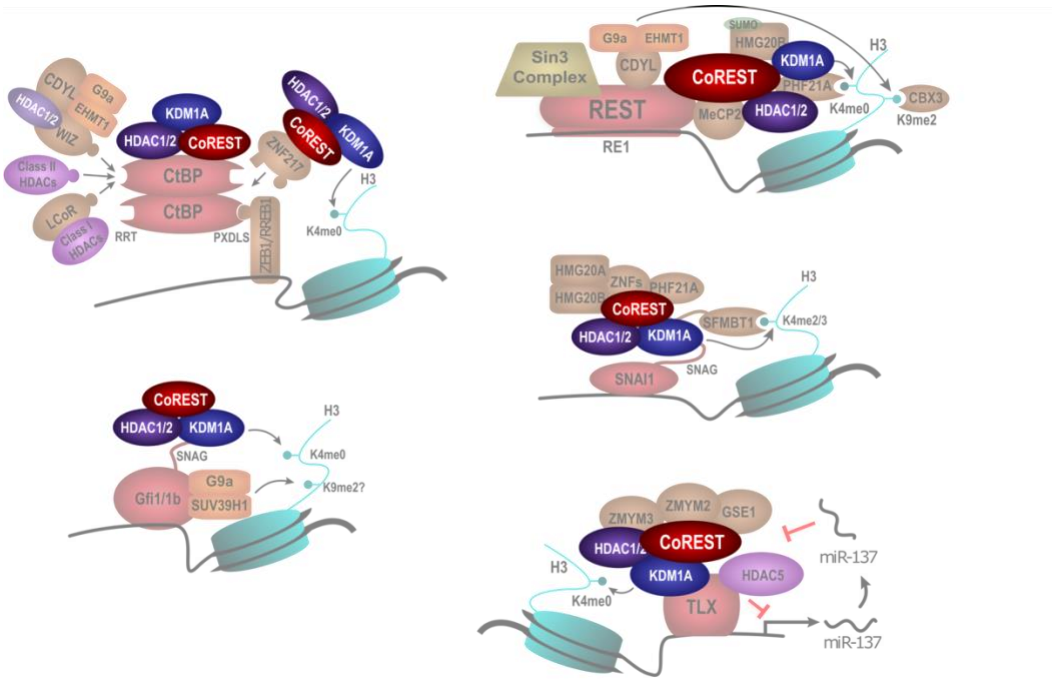


Figure 7: KDM1A has been established, with CoREST and HDAC1/2, in a series of complexes including REST, CtBP and NuRD.

The REST complex

RE1-silencing transcription factor (REST) is involved in neuronal differentiation as a transcriptional repressor in neuron-specific genes *(119)*. In a pluripotent cell state, these genes are held in a 'poised' state and REST recruits CoREST and mSin3 to its C- and N-terminal domains respectively. Post-differentiation of non-neuronal cells, the CoREST complex persists to alter the local chromatin structure via KDM1A-driven H3K4 demethylation *(119)*. Further stabilization of REST-mediated repression is imparted by additional REST-associated coregulatory proteins such as BHC80 which stabilizes KDM1A on chromatin via it's H3K4me0 reader domain *(31)*. Another coregulatory protein, Braf35 (HMG20B) is sumoylated and recruited specifically to the

CoREST complex *(113,120)*. The specific mechanism of Braf35 is not clear, however, dimerization with iBraf (HMG20A) disrupts Braf35 sumoylation and binding leading to an suppression of the REST complex activity *(121)*

The CtBP complex

The C-terminal binding proteins (CtBPs) are generally considered to be repressors of mammalian gene expression driven to specific contexts by recruitment of different sets of coregulatory proteins *(30,122)*. KDM1A has been associated with several CtBP functions such as suppression of the tumor suppressor BRCA1, development of the pituitary gland, activation of gastrointestinal endocrine-specific genes, and suppression of the epithelial phenotype-associated gene E-cadherin *(123–126)*.

CtBPs interact with coregulatory proteins through a hydrophobic pocket with the sequence motif PXDLS as well as a surface structure which binds to a corresponding RRT motif *(127,128)*. CoREST does not contain a PXDLS motif but has still been reported to bind directly to CtBP1 to direct KDM1A to the complex *(129)*. Separately, KDM1A recruitment to the CtBP complex may be driven by an interaction with ZNF217 which localizes to the complex for transcriptional repression of several genes including E-cadherin and BRCA1 *(125,130)*. Furthermore, ZEB1 and RREB1, both DNA-binding proteins, may separately recruit KDM1A towards repression of pituitary cell differentiation as well as CtBP-mediated gene activation, respectively *(126,131)*. KDM1A's activity within the CtBP complex has far-reaching biological implications,

though the exact mechanisms of recruitment and gene loci targeting still remain largely unknown.

The SNAI1/2 complexes

Regulators of EMT, SNAI1 and SNAI2 bind to E-box motifs of epithelial genes in order to repress transcription and induce a mesenchymal phenotype *(132,133)*. KDM1A is recruited by the SNAG domains of these proteins *(134,135)*. The SNAG domain mimics the H3 tail and binds the active site of KDM1A in order for these complexes to co-opt its demethylation activity at H3K4me2. SNAI1 has been coprecipitated with the KDM1A:CoREST:HDAC1/2 catalytic core as well as Braf35 and BHC80 *(136)*. It is unclear whether these coregulatory proteins play an additional role at these gene loci in regulation of gene expression.

The Gfi1/1b complexes

Also containing SNAG-domains are Gfi1 and Gfi1b which play a role in hematopoiesis of lymphocytes, granuloctes, macrophages (Gfi1) as well as megakaryocytes and erythrocytes (Gfi1b) *(137–141)*. KDM1A is recruited by these proteins to promote transcriptional representation through H3K4me2 demethylation *(142)*. It remains unclear if additional coregulatory proteins or complexes are associated with KDM1A when it is recruited by Gfi1/1b complexes.

The NuRD complex

The nucleosome remodeling and histone deacetylase complex (NuRD) has been implicated in a variety of cellular processes including development and cancer progression *(131)*. KDM1A was first linked to the NuRD complex through repression of

the TGFβ signaling cascade to repress the EMT pathway and inhibit the metastatic phenotype of a breast cancer cell line *(70,72)*. KDM1A was demonstrated to bind MTA1, which contains similarities with CoREST. The three MTA isoforms each possess a BAH, ELM2, SANT, and zinc finger domains, and the ELM2/SANT domains have been demonstrated to recruit HDAC1/2 *(73,143)*. It is unclear if MTA proteins recruit KDM1A in a manner analogous to CoREST. Furthermore, KDM1A has not been demonstrated to be a unit of the NuRD complex in HeLa cell extracts, which suggests that the interaction is context-specific *(144)*.

The androgen and estrogen receptors

A nuclear hormone receptor (NHR), the androgen receptor (AR) is tightly associated with normal tissue development to the initiation and development of prostate cancers *(145)*. KDM1A was first reported to be associated with the AR by Metzger and colleagues wherein they also demonstrated that the demethylase's substrate in this context was H3K9me2 *(146)*, activating gene expression. Interestingly, there is evidence that in the absence of the AR, KDM1A resides at the target genes of the AR to repress these genes *(58,146)*. There is also evidence that the AR recruits KDM1A to the AR gene loci to initiate a negative feedback loop through H3K4 demethylation *(147)*.

Another NHR, the estrogen receptor, the estrogen receptor alpha (ERα) is the main estrogen regulator in estrogen-responsive tissue types and dysregulation results in the initiation and progression of several cancer types *(148)*. Similar to the AR, there is evidence that KDM1A demethylates H3K4 at ERα target genes when not recruited by

and directly interaction with the ERα *(149,150)*. Upon estrogen addition, ERα is recruited to these promoters and subsequent gene activation occurs, likely via demethylation of H3K9me2 by KDM1A *(150,151)*.

For both androgen and estrogen receptors, it remains unknown how the substrate targeting of KDM1A is shifted from H3K4 to H3K9. It may be that upon recruitment by the coregulatory NHRs, KDM1A is repositioned. It is unclear whether there is a conformational shift in the active site of KDM1A to redirect substrate specificity.

lncRNA and KDM1A

In addition to being directed by protein complexes, KDM1A may be recruited to specific gene loci via non-coding RNA (ncRNA) such as HOTAIR and TERRA. First reported by the Chang lab, HOTAIR is a long-noncoding RNA (lncRNA) that is dramatically upregulated in a wide variety of cancer phenotypes including ovarian, breast, oesophageal, colon, and gastric cancers *(152–162)*. HOTAIR recruits the PRC2 complex and KDM1A at its 5′ and 3′ ends, respectively, to *HOXD* gene loci for gene repression *(163)*. Targeted disruption of the interaction between PRC2 and HOTAIR in ovarian and prostate cell models has demonstrated the decreased viability of the cancer cell type as well as xenograft tumor growth *(164,165)*. Little work has been reported on the molecular mechanism of the HOTAIR:KDM1A interaction, though it is likely based on the PRC2 results, that disruption of this interaction would provide similar results.

Another lncRNA, TERRA, is associated with the shortening of telomers in the absence of telomere repeat-binding factor 2 (TRF2). TERRA recruits KDM1A to telomeric DNA via a G-quadruplex repeat UUAGGG which binds the SWIRM domain. KDM1A in turn recruits and enhances the activity of the nuclease MRE1, promoting telomere shortening *(166,167)*. The histone methylation levels in these regions do not change, so it is unclear if KDM1A has a catalytic role, or functions to recruit MRE1 to the telomeres.

1.3.1.3 Current gap in knowledge

In each of these complexes, KDM1A is able to selectively and specifically target a specific substrate and subset of gene loci to affect cellular function and disease progression. Due to the widespread nature of KDM1A's function, it is important to establish the specific role of KDM1A within each complex. Thus, there is a clear and present need to produce disrupters of complex-specific interactions with KDM1A. Disruption of KDM1A from being recruited by only one complex will allow us to elucidate how the cell is affected by KDM1A in that specific context.

In addition to producing targeted inhibitors, it is important to understand the basic molecular mechanisms of recruitment and substrate specificity and selectivity. As we saw with the NHR, KDM1A can function on both H3K4 as well as H3K9 residues at the same gene loci and it is unclear what dictates this change in substrate.

There are likely allosteric conformational changes to the enzyme, as well as additional substrate intractions by coregulatory proteins that have yet to be fully

elucidated in KDM1A. There are examples in other epigenetic complexes and enzymes which demonstrate allosteric regulation by coregulatory proteins.

1.4 Epigenetic complexes provide selectivity and specificity to writer and eraser enzymes

Similar to KDM1A, each enzyme functions as catalytic machinery within the context of large multimeric complexes composed of varying combinations of coregulatory proteins. Each coregulatory protein contains histone PTM recognition 'reader' domains, DNA binding domains, or RNA binding domains. Varying permutations of coregulatory proteins enable multiple complexes to co-opt the same enzyme while targeting genes specific to each complex. The coregulatory proteins involved in many complexes have been established by pull-down assays. The manner these complexes assemble have broadly been characterized by domain mapping to regionally assign protein-protein interactions. However, precise interfaces between proteins within the complex remain widely unknown. It is also not known how the complexes are able to selectively target gene loci.

Within an epigenetic complex, the interfaces formed between coregulatory proteins and enzyme can alter the catalytic efficiency of the enzyme (Figure 8) *(168,169)*. In extreme cases, enzymes exist in an off-state where the conformation of the enzyme occludes access to the active site and only upon binding coregulatory proteins is the active site available for cofactor or substrate binding *(170–174)*. This acts as a mechanism for regulation ensuring that the enzymes are not aberrantly affecting the epigenetic

landscape *(68)*. Other enzymes exist in an on-state where the catalytic efficiency further enhanced through coregulatory proteins. The tunable nature of the catalytic efficiencies could arise from minor conformational shifts of the active site, or the by-product of additional substrate binding interactions by the additional complex partners. In order to best understand the current knowledge in the field, we use EZH2, KDM1A, and UCH37 to further explore how catalytic efficiency is affected by non-catalytic coregulatory proteins.

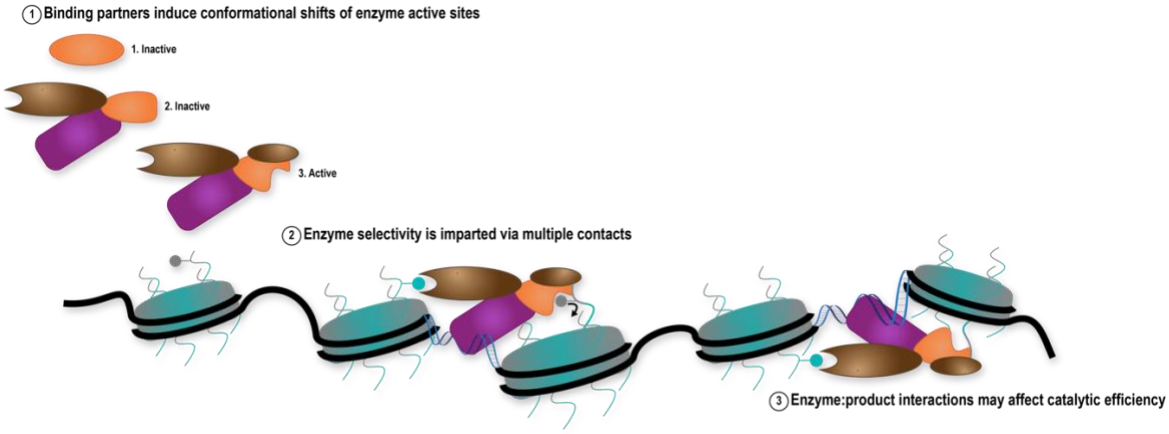


Figure 8: A general scheme of the alteration of enzyme catalysis and kinetic rates. (a) There are a series of epigenetic enzymes that must be activated by coregulatory proteins for the active site to be accessible for cofactor or substrate. This activation is due to a conformational shift in the active site. (b) Each complex is formed by coregulatory proteins each with domains which impart specificity and selectivity via a series of reader, RNA/DNA/protein binding domains. (c) In some cases, the product binds more tightly than the substrate resulting in a decreased off-rate altering the catalytic efficiency of the enzyme.

1.4.1 Coregulatory complex proteins control enzyme activity

1.4.1.1 EED and SUZ12 are necessary for EZH2 activity

Enhancer of zeste homolog 2 (EZH2) is a SET-domain histone methyl transferase (HMT) most commonly associated with the polycomb repressive complex 2 (PRC2) wherein the enzyme trimethylated histone 3 lysine 27 (H3K27me3) *in vivo (175–179).* EZH2 is autoinhibited unless bound to coregulatory proteins SUZ12 and EED which, together, form the PRC2 catalytic core (Figure 9a,e) *(178,180–182)*. Suppressor of zeste 12 (SUZ12) is a zinc-finger containing protein which is essential for EZH2 to be catalytically active based on kinetic and knockdown studies (Figure 9d) *(183–185)*. Embryonic ectoderm development protein (EED) is a reader protein critical for EZH2 activity. The seven WD40 repeats of EED assemble into a β-propeller structure forming an aromatic cage at the center of the protein and utilization of electrostatic and cation-π interactions by the aromatic cage allow for recognition of trimethyllysine residues (Figure 9b,c) *(179,186–189)*. Confirming initial activity assays, crystal structures of an incomplete PRC2 complex, both human and orthogonal proteins in different organisms, revealed that EED and SUZ12 make extensive contacts with EZH2 (Figure 9a), and upon binding induce a conformational shift to expose the EZH2 SET domain to cofactor and histone substrate allowing for catalysis to occur *(173,190–192)*.

Most recently, the Nogales lab published a cryo-EM assembly overlaid with crystal structures to reveal a complete PRC2 complex (Figure 9a) *(193)*. Congruent with previously reported structures, the α-helical N-terminus of EZH2 wraps around the

front face of EED (Figure 9b) which extends EZH2 residues 106-125 to form the EZH2 'activation loop', outlined in green in Figure 9d, these residues have been reported to play a key role activating the EZH2 SET domain *(192)*. While more extensive dynamic biophysical experiments are required, it is currently believed that, as a SET domain methyltransferase (MTAse), EZH2 likely exists in an off-state wherein the activation loop and overall SET-domain conformation sterically preclude cofactor or substrate binding. Upon EED binding, the N-terminus of EZH2 clamps around EED, extending the activation loop, which can be engaged and stabilized by EED resulting in a major conformational shift to reveal an activated binding pocket for cofactor and substrate. While the role of SUZ12 is less established compared to that of EED, mutation of a residue found at the interface of SUZ12 and the EZH2 activation loop, W591, has been reported in patients with cancers including AML *(194–196)*.

Kinetic studies of EZH2 have revealed catalytic activity depends on the binding of EED to a trimethyllysine and a small molecule binding the aromatic cage of EED has been shown to act as an allosteric inhibitor (Figure 9c) *(187,197–199)*. The EED aromatic cage can accommodate trimethyllysine residues with various sequences; however, binding affinity of a peptide is not directly related to the activity of EZH2 *(187)*. While EED may be promiscuous in certain *in vitro* assays, EED has not been shown to be promiscuous when other coregulatory proteins are present. This suggests that the coregulatory proteins function as a whole to position EZH2 at the correct substrate. Two

EZH2 α-helices, SRM and SET-I, span the gap between the EED reader domain and the SET domain forming several distinct interactions, such as a salt bridge between D142(SRM) and K660(SET-I), (Figure 9d) which may allow further communication between EED and EZH2 *(173,191,193)*. There are likely additional conformational changes that occur upon trimethyllysine:EED binding that have yet to be elucidated; however, it is clear that an intact EED bound to a trimethyllysine-containing peptide is crucial for EZH2 activity and requirement of EED binding a trimethyllysine adds an additional layer of control over the selectivity of EZH2.

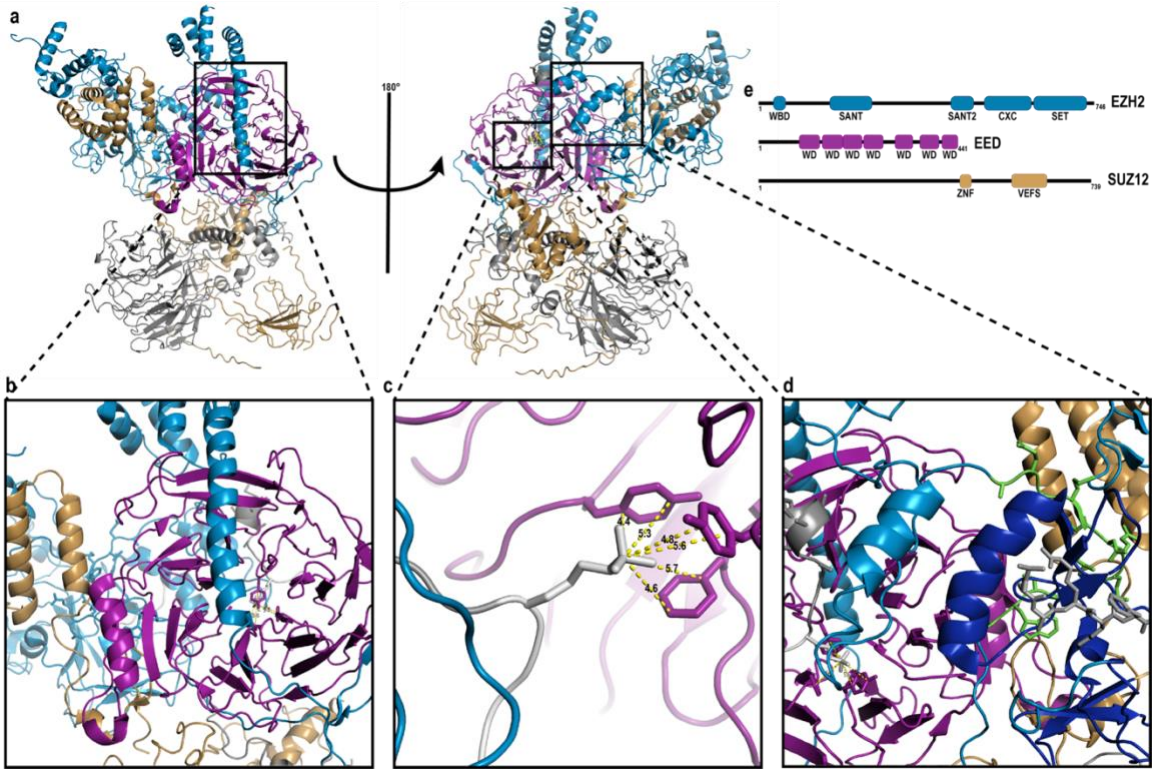


Figure 9: The catalytic core of the PRC2 complex. (a) Two orientations of the PRC2 complex with EZH2 (blue), EED (purple), and SUZ12 (old gold) highlighted over the additional coregulatory proteins in grey (AEBP2, RBBP4, JARID2). (b) C-terminal residues of EZH2 (39-62) extend around the WD40 protein, EED. (c)The seven WD40 domains of EED form a hydrophobic cage which utilizes cation-π interactions to recognize trimethylated lysine residues. (d) The catalytic SET domain of EZH2 (dark blue) is in front of an activation loop which allows for substrate to bind (dark grey sticks, JARID2). EED and EZH2 may also communicate via two α-helices (SRM is light blue, SET-I is dark blue). (e) Domain maps of EZH2, EED, and SUZ12. (PDB 6C23)

1.4.1.2 UCH37 activation and inactivation is dependent on coregulatory proteins

One of the few known cases of an enzyme being activated or inhibited by its coregulatory proteins is the deubiquitinase Ubiquitin carboxyl-terminal hydrolase isozyme L5 (UCH37). UCH37 exists in an off-state that can be activated by RPN13 *(200)*. *In vitro* binding studies and crystal structures illustrate an active site crossover loop

(ASCL) which extends over the active site to preclude substrate binding *(201)*. A C-terminal domain of UCH37 can fold over the ASCL to expose the active site for substrate binding. The interaction between the C-terminal domain and ASCL is stabilized upon RPN13 binding. This stabilization catalytically actives UCH37.

Conversely, the binding of NFRKB to UCH37 has been shown to decrease the k_{cat} of UCH37 by over 200-fold *(169,202)*. Crystal structures between UCH37 and either RPN13 or NFRKB demonstrate that each binds BAP1 along the same region *(169)*. However, upon NFRKB binding, there is a 4-5Å shift of the C-terminal domain of UCH37 which occludes the active site from ubiquitin binding. An α-helix of NFRKB also binds around the globular region of UCH37 which may act as a clamp increasing the stability of the protein's binding interaction *(169)*. Elucidating additional interactions with each coregulatory protein will require expression and purification of the full-length proteins for utilization in further x-ray crystallography or dynamic interaction studies. As both proteins bind in a similar region, it would be of potential interest to produce chemical biology probe, such as an aptamer or peptide mimetic able to either increase or decrease stability of one coregulatory protein for differential outcomes.

1.4.2 Coregulatory proteins tune enzyme activity

Both of the examples above, and others referenced in Table 2, demonstrate that coregulatory proteins can induce major conformational shifts to affect active site

availability. The effects of coregulatory proteins can also be limited to a change in catalytic efficiency without a major conformational shift as we demonstrate below. These changes in activity are likely due to either uncharacterized long-range contacts between coregulatory protein and the substrate or an increased stabilization of the substrate through additional electrostatic interactions.

1.4.2.1 Coregulatory proteins of KDM1A and LSD2 enhance enzyme activity

The manner through which CoREST affects KDM1A's activity has been studied by our lab and others. *In vitro*, LSD1 requires at least a 16-mer peptide to be active, but is incapable of demethylating full nucleosomes without the presence of the minimal SANT2 domain/'linker' of CoREST *(203)*. Likely, CoREST makes additional contacts, such as those to nucleosomal DNA with the SANT2 domain, with the full nucleosome to direct KDM1A to its substrate *(118)*. This, however, does not explain the observed 2-fold increase in k_{cat} of a peptide substrate when CoREST binds KDM1A *(117)*. An increase in k_{cat} could result from a change in the activation energy due to a conformational shift of the active site. As outlined in Figure 10b, KDM1A α-helix bridges interactions between CoREST and the H3 peptide via electrostatic and cation-π interactions *(56)*. Crystal structures only provide a static view of a protein-protein interaction and dynamic studies such as HDX-MS would establish the dynamic movement of KDM1A's active site due to the presence of CoREST. We acknowledge that while 2-fold is not a dramatic change in k_{cat}, no current kinetic studies or crystal structures utilize full-length CoREST,

which is due, in part to an inability to express soluble CoREST from *E. coli* (Figure 10c). It is possible that full-length CoREST makes additional contacts further tuning the catalytic rate of KDM1A in response to CoREST binding. Other KDM1A binding partners may also alter the activity of the enzyme but there is little data currently in the literature regarding this.

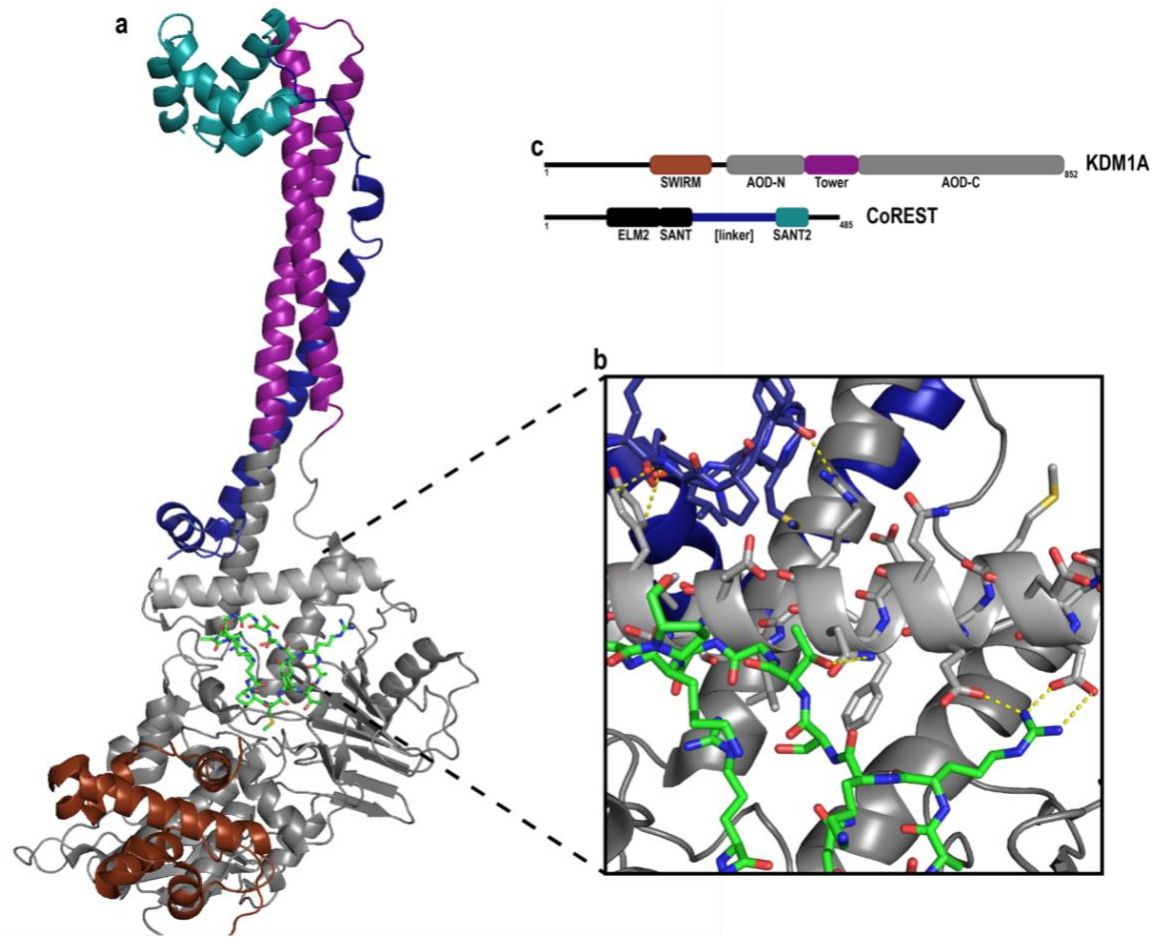


Figure 10: Interactions between KDM1A:CoREST interface bridge histone substrate interactions. (a) KDM1A bound to the linker-SANT2 region of CoREST (residues 308-440). The tower domain (deep magenta) consists of a pair of α-helices which create a handle through which coregulatory proteins such as CoREST can bind. The linker region (dark blue) is critical for KDM1A:CoREST and (b) the C-terminal region of the linker domain forms contacts with the amine oxidase domain (AOD) of KDM1A which in turn interacts with the H3 peptide substrate (bright green, residues 1-16). (c) Domain map of KDM1A and CoREST. (PDB 2V1D)

It is worth noting that LSD2 (KDM1B) is an ortholog of KDM1A (Table 2). Similar to KDM1A, LSD2 targets H3K4me2 and H3K9me2; however, each enzyme targets different genes and are recruited by different coregulatory proteins *(48)*. In fact,

an overlay of LSD2 and KDM1A reveals that the structures of the two are very similar despite the marked absence of a tower domain in LSD2 *(204)*.

While less is known about LSD2, the enzyme is not active *in vitro* towards nucleosome substrates without a coregulatory protein, NPAC/GLYR1, which is a putative H3K36me3 reader protein with a potential DNA binding domain *(205,206)*. A 2013 study produced a crystal structure with NPAC residues 214-225 structured and shown to bind in a groove between the AOD and SWIRM of LSD2 *(68)*. Further assays revealed that these 12 residues were the minimal functional unit with F217 being critical for activity within the dodecapeptide *(68)*.

It is currently unclear how the minimal unit of NPAC enables enzyme activity against a nucleosome. In the case of both CoREST and NPAC, there is not a significant change in active site due to binding, suggesting that changes in activity are due to long-range contacts that each coregulatory protein makes with the substrate.

1.4.2.2 JARID and AEPB2 enhance EZH2 activity

Binding selectivity and catalytic activity of EZH2 is mediated by additional coregulatory partners AEBP2 and JARID2 (Figure 11d) *(207–210)*. Adipocyte enhancer-binding protein 2 (AEBP2) and Jumonji/ARID domain-containing protein 2 (JARID2) each bind to DNA and enhance gene targeting of the PRC2 complex *(207–210)*. AEBP2 and JARID2 have additionally been shown to enhance EZH2 activity above the catalytic efficiency of EZH2:EED:SUZ12 as well as stabilize the PRC2 complex *(192,211,212)*.

The Nogales lab's crystal structure reveals that AEBP2 and JARID2, are focused on the bottom half of the PRC2 complex (Figure 11a,b) *(193)*. The placement of AEBP2 and JARID2 is much farther from the active site than EED; however, both of these proteins, along with another reader protein, RBB4/7 (RbAp46/48) make stable interactions with the central SUZ12 (Figure 11c) *(192,213)*. We hypothesize that the AEBP2 and JARID2 interactions with SUZ12 may function to stabilize and rigidify SUZ12 which traverses through the entire PRC2 complex (Figure 11b). The stabilization of SUZ12 might then allow for the active site to experience an increased catalytic activity.

Interestingly, JARID2 has been postulated to be a substrate of PRC2 as well as a coregulatory protein *(210)*. The Nogales lab utilized a minimal binding unit of JARID2 (106-450) trimethylated at 116 revealing JARID2 at two positions within the structure. The first is bound to SUZ12 and AEBP2 (Figure 11c) and the second is involved in the EED aromatic cage (Figure 11b). Whether PRC2 utilizes JARID2 *in vivo* for activation has yet to be established.

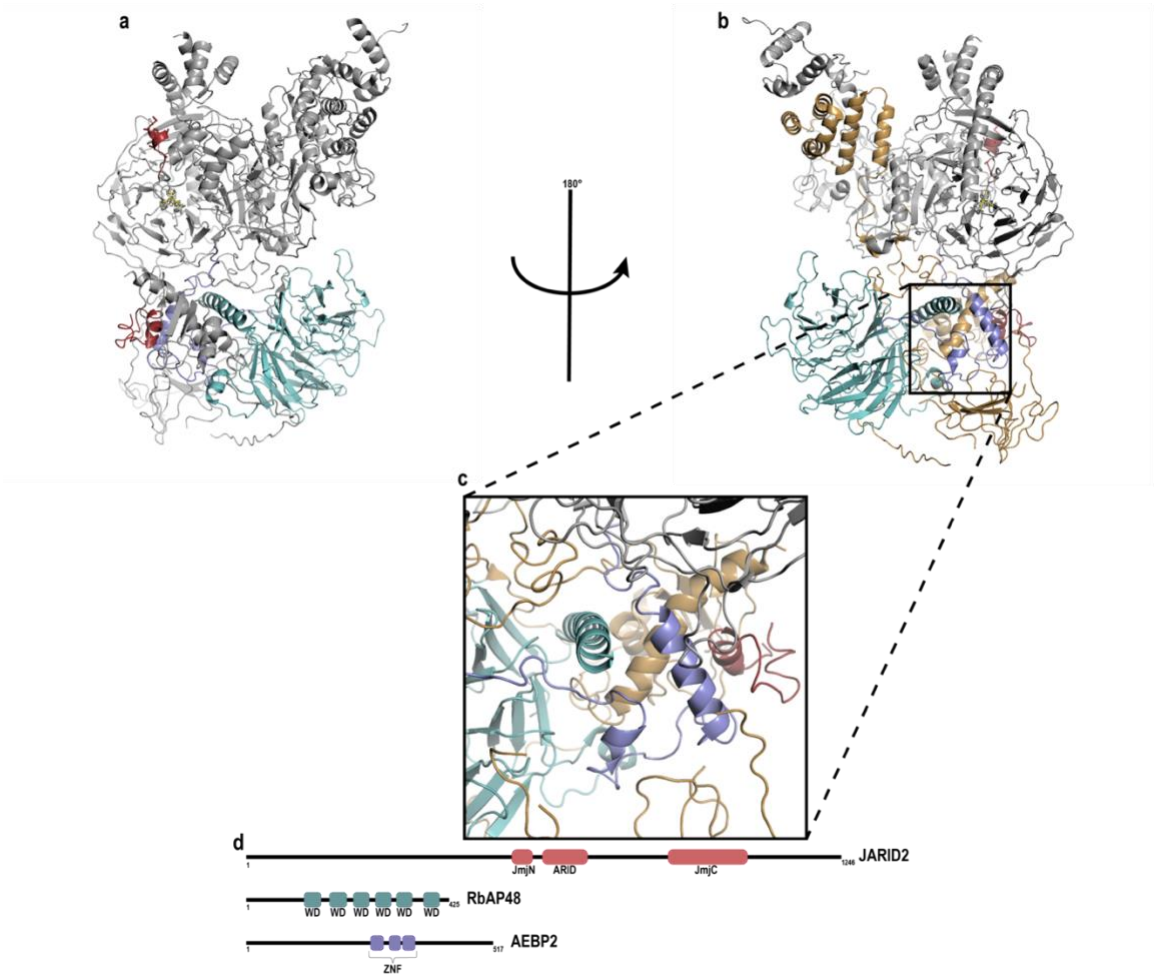


Figure 11: PRC2 coregulatory proteins can modulate EZH2 catalytic activity via long-range contacts. (a) JARID2 (red), RbAP48 (teal), and AEBP2 (light purple) are positioned around the catalytic core of PRC2. (b),(c) Rotating the complex 180° on the y-axis highlights that SUZ12 spans through PRC2 and makes contacts with each of the three additional coregulatory proteins. (d) Domain maps of JARID2, RbAP48, AEBP2. (PBD 6C23)

1.4.3 Enzyme products can modulate catalytic activity

Just as coregulatory proteins can affect protein activity via conformational shift, the product of each enzyme could alter the enzymatic activity. If an enzyme has a slow off-rate from its product, the catalytic efficiency of the protein will decrease. This

protects the newly modified nucleosome from another enzyme reversing this and provides a handle for additional complexes or transcription factors to bind to the promoter region.

1.4.3.1 A full-length H3 protein can alter the catalytic efficiency of KDM1A

Our lab has shown that a full length unmodified H3 protein, the product of KDM1A demethylation, can act as a competitive inhibitor over a 21-mer dimethylated substrate *(214)*. This is likely via long-range contacts increasing the binding affinity of the histone to the enzyme which suggests that the k_{off} between LSD1 and its product could be biologically relevant.

There is no crystal structure or dynamic interaction binding data between LSD1 and a full length H3 histone protein; however, there are several crystal structures with an H3 peptide *(60,63)*. These structures depict the peptide exiting the active site towards a negatively charged grove between the AOD and SWIRM domain of KDM1A (Figure 12). The electronegative charge of this cavity would provide the interactions necessary to observe an increased binding with the positively-charged full-length H3 histone. Whether this diminished off-rate persists in the more biologically relevant tetramer, octamer, or nucleosome has yet to be established. In order to establish the orientation of KDM1A with respect to its substrate, it is also important to utilize techniques such as cross-linking, x-ray crystallography, or HDX-MS.

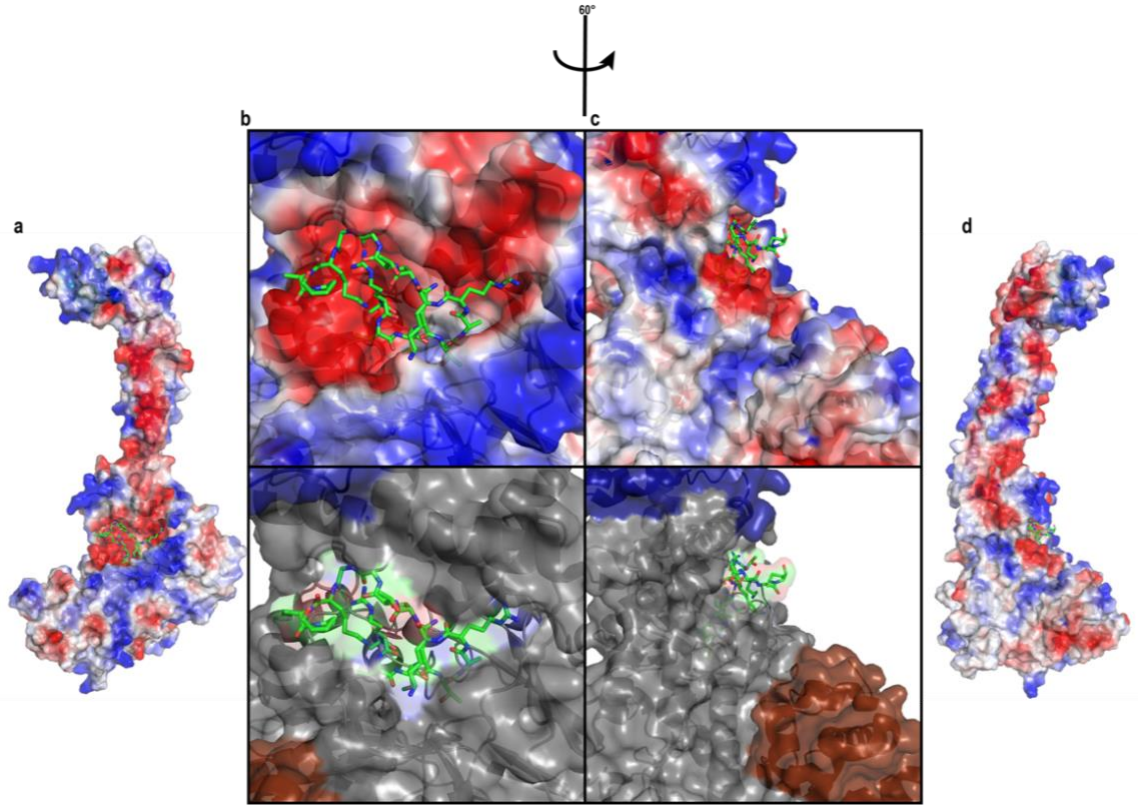


Figure 12: Electrostatics of the surface of KDM1A provide insight into potential binding interface of H3 substrate. (a) KDM1A:CoREST:H3 in the same perspective as Figure 10a with electrostatic surface as predicted by PyMOL v 1.7.4.5 Edu. (b) A zoomed in view of the KDM1A active site with the H3 peptide1-16 with (top) the electrostatic surface or (bottom) a space-filling surface model. (c) A side view of KDM1A revealing a potential interface for the H3 substrate to bind with (top) the electrostatic surface or (bottom) a space-filling surface model. (d) A full view of the KDM1A:CoREST:H3 structure at the side angle. (Electrostatics of PDB 2WI5 were calculated with PDB 2WI5 imposed upon PDB 2VID)

1.4.3.2 Long-range HDAC8 interactions affect specificity

Histone deacetylases 1 and 2 (HDAC1/2) are Class I histone deacetylases and are of particular interest due to their upregulation in various cancers *(40,43,44,215)*. Suberoylanilide hydroxamic acid (SAHA), also known as vorinostat, can inhibit Class I HDACs and is approved in treatment with cutaneous T-cell lymphoma (CTCL) *(216)*.

However, there is still much about the activity of these enzymes that is unclear due to difficulty obtaining large quantities required for many biochemical characterizations. HDAC1/2 require phosphorylation and inositol phosphate for catalytic activity making them much less accessible via *E. coli* expression *(217,218)*. Another Class I HDAC enzyme, HDAC8, does not require phosphorylation to be active and has been actively expressed via recombinant *E. coli* expression *(219)*. While HDAC8 does not reside in the same complexes as HDAC1/2 it does provide insight on how Class I HDACs interact with their substrate as the globular active-sites of HDAC1 and HDAC8 are virtually indistinguishable. The Fierke lab has recombinantly expressed HDAC8 and performed extensive kinetic studies to establish metal dependency, likely Fe (II) *in* vivo, as well as to assess the overall activity of the enzyme *(220–223)*.

In 2017, the Fierke lab published a paper to establish the selectivity of HDAC8 from 7-mer peptides to full nucleosomes *(224)*. They demonstrated that the active site of the enzyme displays a local sequence preference surrounding the PTM. This study also established that between 7-mer and 17-mer peptides, the rates were within a 3-fold change, with the slight increase in k_{cat}/k_m towards the longer peptides. This trend in increasing $k_{1/2}/k_m$ continued for the tetramer and octamer due to a much lower K_m. This implies that larger HDAC8 substrates have an increase affinity due to long range contacts critical for enzyme binding and selectivity. The $k_{1/2}/k_m$ of HDAC8 against a nucleosome falls back to roughly the rate of the 7-mer peptide. It may be that, like

KDM1A, there are coregulatory proteins required for HDAC8 to bind the nucleosome over the other substrates. Overall, we observe a generally amplified affinity of HDAC8 towards increasingly biologically relevant substrates which results in a decreased catalytic efficiency.

1.4.4 Long range complex contacts mediate selectivity

Each coregulatory protein in a complex contains reader domains, DNA binding domains, RNA binding domains, or protein binding domains which target specific PTMs, DNA sequences, or long non-coding RNA (lncRNA). As evidenced by EED, each coregulatory protein may obtain some level of promiscuity and it is only through the formation of a complete complex is the true selectivity revealed. We hypothesize that each coregulatory protein adds a level of selectivity and binding affinity and only when most, if not all, of the coregulatory proteins bind is the enzyme positioned properly at its substrate for catalysis.

1.4.4.1 PRC2 coregulatory proteins impart EZH2 selectivity

Looking first at the PRC2 complex, PRC2 substrate specificity is towards H3K27me3 *in vivo* suggesting that the observed EED promiscuity is an artifact of *in vitro* studies not utilizing the full PRC2 complex *(187,191)*. It is thus the culmination of the additional coregulatory proteins which target genes and position EZH2 and EED.

The 5-mer PRC2 complex (EZH2, EED, SUZ12, RBBP4/7, and AEBP2) binds dinucleosomes 23±9 nM over 280±19 nM (K_d^{app}) mononucleosomes *in vitro (212,225)*.

Further investigation revealed that the 5-mer preferentially binds to dinucleotides with over 50 bp length between each nucleosome with H3K27methylation significantly increased at a 50 bp linker over 100 or 10 bp linker length *(212)*. This suggests that the PRC2 complex is more likely to interact with euchromatin over heterochromatin. This preference could be due to the ability of AEBP2 to bind DNA. The binding motif of AEBP2 has been determined to be CTT$(N)_{15\text{-}23}$CAGGCC with a co-occupancy rate with SUZ12 of 70-80% *(207)*. This motif would enable AEBP2 to aid in PRC2 targeting specific genes. Based on current evidence it is unclear if AEBP2 binds to linker DNA over DNA bound to a histone octamer, though the PRC2 preference for long DNA linkers could provide a clue to this.

While JARID2 is commonly associated with the PRC2 complex, it is not always associated with PRC2 *in vivo (226)*. PRC2 has been implicated in binding several lncRNA *(227–229)*. While there is currently a debate about the selectivity PRC2 binding lncRNA, it may be that the lncRNA serve to add an additional method of gene targeting for PRC2 with a mechanism similar to JARID2.

Overall, there are several models and hypotheses that PRC2 spans two nucleosomes with EED interacting with a pre-existing H3K27me3 at an adjacent nucleosome to PRC2 and additional lncRNA or coregulatory proteins aiding in gene targeting of PRC2.

1.4.4.2 KDM1A-containing complexes target different histone residues and gene promoters

KDM1A is found within the C-terminal binding protein (CtBP) and nuclear remodeling complex (NuRD) which target opposing pathways *(32,48,72,230)*. The CtBP complex promotes the epithelial-to-mesenchymal transition (EMT) via repression of E-cadherin and related proteins; whereas the NuRD complex functions, in part, to repress the EMT *(72,231–233)*. Elucidation of the role of KDM1A within each complex will require selective inhibition of one complex, not pan-inhibition of KDM1A's active site. Chemical biology probes such as aptamers or peptide inhibitors would allow for disruption of key interfaces within a complex to establish its cellular role.

The residue selectivity of KDM1A can also be altered from H3K4me2 to H3K9me2. When associated with the estrogen receptor, KDM1A targets H3K4me2 in complex with CoREST, whereas there is evidence that the substrate selectivity of KDM1A is H3K9me2 when in complex with the androgen receptor *(32,56,58,71)*. While it is possible that this change in residue selectivity is due to conformational shift, there is no evidence to support this. Based on KDM1A's inability to demethylate nucleosomes *in vitro*, we hypothesize that KDM1A exists in an active state and is directed and positioned to a substrate by its coregulatory proteins.

KDM1A was first reported within a stable 5-mer subunit of the REST complex. This 5-mer complex contained CoREST, HDAC1/2, DNA binding Braf35, and reader domain protein BHC80 *(113)*. While the only crystal structures that exist are of the

aforementioned KDM1A:CoREST$_{308-440}$, and few kinetic or binding assays have been performed to establish the role of coregulatory proteins, there are several indications as to the orientation of the complex *in vivo*. First, based on the potential orientation of the H3 tail to KDM1A, a molecular dynamics simulation was produced that models KDM1A and CoREST wrapping around a mononucleosome on the same face as the DNA *(123)*. Secondly, BHC80 interacts with unmethylated H3K4 residues which could be critical for this CoREST complex to span multiple nucleosomes, similar to PRC2, or to allow for an increased residence time at a nucleosome post KDM1A demethylation *(31)*. Each would provide increased stability and gene targeting specificity of the complex.

Additionally, KDM1A has been reported to bind lncRNA which, similar to PRC2, may aid the complex in targeting specific genes and pathways *(163,166)*. However, there is little evidence as to how lncRNA bind KDM1A. Overall, elucidation of the orientation of KDM1A, and KDM1A-containing complexes and the importance of each coregulatory protein in affecting not only KDM1A activity but selectivity is currently lacking compared to PRC2 and requires significant work. Establishing these interfaces and dynamic protein interactions could aid in the discovery of chemical biology tools to examine the role of KDM1A within disease states.

Table 2: Comprehensive overview of epigenetic enzymes and their coregulatory interactions

Enzyme	*Enzyme Activity*	*Common Residue(s) of Interest*	*Product of enzyme affects catalytic efficiency*	*Selectivity via multiple contacts*	*Coreg. protein alters active site*	
SIRT2	HAT	H3, H4 peptides	Yes			*(234)*
KDM7A	KDM	H3K4me3, H3K9me2 H3K27me2			Yes	*(235)*
PHF8	KDM	H3K4me3, H3K9me2 H3K27me2			Yes	*(235)*
BAP1	DUB				Yes	*(236–238)*
G9a	MTase	H3K9	Yes	Yes		*(239)*
Rtt109	ACT	H3K56	Yes	Yes		*(240)*
KMD6A	KDM	H3K27		Yes	Yes	*(241)*
CBP	HAT	H3, H4 peptides		Yes		*(242,243)*
Gcn5	HAT	H3K9/14/23		Yes		*(244)*
HBO1	HAT	H3K14		Yes		*(245,246)*
KMD4A	KDM	H3K23me3		Yes		*(247)*
KMD4B	KDM	H3K23me3		Yes		*(247)*
KMD5 family	KDM	H3K4me3		Yes		*(248)*
KMD6B	KDM	H3K27		Yes		*(241)*
LSD2	KDM	H3K4me1/2		Yes		*(68)*
PRMT5	MTase	H3/4R3, H2AR3		Yes		*(249)*
Pr-Set7	MTase	H4K20		Yes		*(171)*
p300	ACT	H3, H4 peptides		Yes		*(242,243)*

1.4.4.3 Long non-coding RNA can additionally impart selectivity on epigenetic complexes

While each PTM may be associated with either transcriptional activation or repression for a gene, it is in fact, a concert of multiple PTMs on a gene's enhancer and promoter regions which sum together to either promote or repress transcription. This 'histone code hypothesis' relies on rapid and precision addition and removal of each PTM to match the ever-changing landscape of the cellular environment requiring multiple epigenetic enzymes to function together. As has been discussed above, one method the cells have adopted to increase the efficient turnover of multiple PTMs is through multimeric complexes which contain multiple enzymes. Thus, a singular complex can interact with a histone and lead to changes at multiple histone residues, potentially even spanning multiple nucleosomes *(250)*.

We will provide a more comprehensive overview of the role of non-coding RNA (ncRNA) as well as various roles of long-noncoding RNA (lncRNA) in Chapter 4; however, lncRNA have been shown to be important in gene regulation through several mechanisms including scaffolding whereby multiple epigenetic complexes with RNA binding regions interact with different regions of the same lncRNA which is capable of targeting a specific subset of genes.

One of the best studied examples of this is also one of the first examples of not only a trans-acting lncRNA was of HOX transcript antisense RNA (HOTAIR) which is a

2146-nt lncRNA transcribed from the HOXC locus which functions along HOXD loci targeting both PRC2 and KDM1A-containing complexes. The Chang lab at Harvard utilized pull-down assays with segments of HOTAIR to establish that nt 1-300 were critical for PRC2 binding and that nt 1500-2146 were important for binding KDM1A-containing complexes *(163)*. They further demonstrated that nt 1820-2146 could not interact with KDM1A-containing complexes alone, suggesting that the majority of the binding affinity lies in nt 1500-1820 *(163)*. The Pyle lab utilized several RNA footprinting techniques which will be further discussed in Chapter 5 and established that HOTAIR contained four distinct domains wherein Domains 1 and 4 interacted with the PRC2 and KDM1A-containing complexes, respectively *(251)*.

To date, there has been significant work to explore the interface of the HOTAIR:PRC2 complex but not that of HOTAIR:KDM1A. EZH2 alone can bind to $HOTAIR_{1\text{-}300}$ at a K_d of 755 ± 45 nM this was lower than the minimal active 3-mer PRC2 unit of EZH2:EED:SUZ12 at 165 ± 16 nM *(252)*. With HOTAIR:PRC2, nt 212-300 have been established to be the minimal nucleic acid sequence required for a comparable binding affinity to the EZH2:SUZ12:EED 3-mer *(252)*. From this work, there have been several studies producing inhibitors of the PRC2:HOTAIR interface which have demonstrated both *in vivo* and in mouse xenograft models that disruption of this interface hinders cancer cell survival rate and tumor growth *(165,253)*. Toward the HOTAIR:KDM1A interaction, only the Chang lab's pull-down assay demonstrates the

interaction between KDM1A-containing complexes and HOTAIR. It is unclear what the affinity is, nor what the minimal nt or protein contributions are necessary for the interaction to be formed.

Even though the work performed demonstrates a utility for the disruption of the PRC2:HOTAIR interface, there has been significant debate as to how selective PRC2 is for binding RNA as the complex has been demonstrated to bind a significant number of other ncRNA non-selectively *(212,254–257)*. However, even if PRC2 isn't selective for HOTAIR, it is clear that there is a significant effect of disrupting their interaction. To mitigate the disparity between these two sets of results, the 'junk mail' hypothesis has been suggested wherein lncRNA rest along chromatin and sample epigenetic complexes as they interact along the chromatin. When the local chromatin landscape is compatible with the complex interacting with the lncRNA is able to impart additional contacts with reader and DNA-binding domains in the complex to increase the binding affinity and stability of the complex at a gene locus. When this occurs, the catalytic machinery of the complex is able to initiate its chemistry and alter the epigenetic landscape as necessary. Conversely, when the epigenetic landscape is not compatible, the binding between complex and lncRNA alone is not enough and the complex will continue past that specific gene loci. This is an attractive hypothesis, particularly when looking at the μM binding affinity of PRC2 to HOTAIR, that multiple lower affinity interactions are

required to fully target a specific gene locus and imparts the potential importance of looking at these complexes and substrates in more biologically relevant contexts.

1.4.5 Allosteric modulation by epigenetic enzymes

Overall, we begin to see a model emerge from these individual studies. While the catalytic core of each epigenetic complex drives the change of the PTM landscape at a gene locus, this is predicated by the involvement of the additional coregulatory partners within each complex. If we envision a specific gene locus with a specific set of PTM and a lncRNA, even before reaching the chromatic landscape, coregulatory proteins within the complex affect the overall conformation of the enzymes leading to important changes in the substrate specificity and catalytic activity. Each conformational shift can affect the binding and efficacy of a small molecule inhibitor *in vivo*, leading to a decreased viability of a therapeutic compared to in an isolated *in vitro* system. We envision that the complexes travel along gene loci, testing different gene loci to find those which enable (a) a high-binding affinity through multiple low-affinity interactions and (b) the proper positioning of the catalytic machinery to its substrate. Once multiple coregulatory partners within the complex are able to bind to their ligand, such as a PTM, DNA, or a lncRNA, the additive effect of the affinities enables the complex to have a longer residence time at that gene loci. At this time, the catalytic machinery should be able to function upon its correct substrate. Post-catalysis the enzyme:product as well as complex:product interactions dictate the residence time and the overall enzyme activity

and selectivity. We hope to gain a more complete understanding of the molecular mechanisms within this potential model system as well as elucidate if there are critical interactions within a complex which may dictate complex formation, or the targeting of a specific gene set. This information will drive future discovery of targeted therapeutics.

1.5 Dissertation overview

As we begin to grasp the full complexity of the interplay between epigenetic enzymes and their coregulatory partners, it is clear there is still much about the molecular mechanisms between these interactions that have yet to be elucidated. Within the context of this dissertation, we have made progress towards establishing protein:protein and protein:lncRNA interactions utilizing the demethylase KDM1A. Herein we report efforts towards (a) the expression and purification of a core KDM1A-containing complex, (b) utilization of techniques including HDX-MS and SELEX to probe and potentially disrupt protein interfaces, (c) establishing and disrupting a KDM1A:lncRNA interface.

First, to fully understand how enzyme activity and conformation was affected via each coregulatory partner, we chose a bottom-up approach wherein each coregulatory partner would be expressed and purified separately as opposed to whole-complex pull-down from mammalian cells. We chose a well-established core complex of 5-subunits which was stable over a series of six affinity purification columns *(113)*. We

have been able to express each member of the core complex via recombinant *E. coli* expression and will discuss the progress towards purification in Chapter 2.

Our next goal was to establish methods through which we could systematically disrupt specific protein:protein interactions within a complex. As protein:protein interfaces have long been deemed 'undruggable' our first aim within this goal was to utilize biophysical and biochemical characterization of these protein:protein interactions between coregulatory proteins. Our rationale was that if the binding affinity was spread along the interface as opposed to localized to a 'hotspot' of 1-2 residues, this would help inform whether a peptide or RNA aptamer would be better suited for targeted disruption over a small molecule that might be considered for a 'hotspot.' Thus, we sought to establish a model system through which to assess our approach. KDM1A has a well-defined interface with CoREST, which had been previously explored in our lab by Sunhee Hwang to establish if any hotspots resided along the long interface between the two proteins. I furthered her research via hydrogen-deuterium exchange mass spectrometry to demonstrate a lack of hotspots along the interface. This led me to establish a protocol for utilization of an RNA aptamer to disrupt protein:protein interfaces.

Finally, in addition to changes to specificity and selectivity by coregulatory partners, it is becoming increasingly clear that lncRNA are able to scaffold and direct epigenetic complexes to specific gene loci. KDM1A had been demonstrated to interact

with HOTAIR; however, there have been no efforts to our knowledge to elucidate this interaction at a molecular level nor disrupt it. I have utilized a variety of techniques including binding assays, RNA footprinting, photochemical crosslinking, and proteomics to establish the molecular basis for the KDM1A:HOTAIR interface towards a targeted disruption of the interface. I have defined the interface on both a nucleotide and peptide level which will inform future studies involving this interface.

While not included in the main text, I also report systematic proteomic analysis of a *Chlamydia trachomatis* (*C. trachomatis*) proteomic data set by myself and Dr. Katherine Alser Kim in Appendix b. The protease chlamydial protease-like activity factor (CPAF) has been of interest to our lab for its potential small molecule inhibition. Utilizing(+/- CPAF) isogenic strains, we were able to establish proteins which changed due to the presence of CPAF, implicating them in host co-option pathogenic pathways in which CPAF has been previously implicated. As the targets of CPAF have been under debate, my work allows for a more defined set of pathways and systems to probe for CPAF involvement. In future work by members of our lab and others, we will begin to assess CPAF-dependent cleavages based on the established pathways and proteins by this proteomic dataset analysis.

2. Expression and purification of a core demethylase complex

Epigenetic enzymes regulate the PTM of histones, DNA, and RNA which results in changes to gene expression and mRNA translation. In addition to interest in understanding the molecular mechanisms of epigenetic regulation, there are many reports of enzymatic dysregulation being a critical factor in pathobiological progression. There has been significant progress in establishing the kinetics of many of these enzyme classes as well as focus on producing small molecules towards catalytic inhibition.

As I outlined in Chapter 1, there are many cases wherein specificity and selectivity of an enzyme is impacted by coregulatory protein and lncRNA interactions. The widespread impact of coregulatory subunits on an enzyme are only now being fully realized in complexes such as PRC2 and is likely true for most epigenetic complexes. My goal was to express and purify coregulatory proteins which could be used in an increasingly complex system to assess communication and regulation of epigenetic enzymes.

I envisioned being able to assess: binding and timing events for complex formation, effect on enzymatic activity, and changes to substrate specificity. This could be accomplished via pre-established activity assays for KDM1A in our lab as well as a readily available HDAC1/2 activity assay. Stoichiometric binding could be assessed via pull-down assays and gel-filtration assays and biophysically probed with hydrogen-deuterium exchange with mass spectrometry (HDX-MS), surface plasmon resonance

(SPR), and isothermal calorimetry (ITC). Understanding the interfaces between the subunits of the complex, as well as gaining and appreciation for critical interfaces will potentially enable us to produce targeted probes to disrupt or enhance a specific protein:protein interaction to gain a clearer understanding of the specific function of a complex within a cellular environment. Thus, I also envisioned establishing a versatile and readily accessible method to produce high-affinity targeted interaction inhibitors.

The attractiveness of each of these techniques and this bottom-up approach overall is that it enables us to test each protein interaction individually as well as assess an increasingly complex system. Furthermore, I believe this approach will provide an unparalleled method to assess protein:protein interfaces (PPI) via biochemical and biophysical techniques.

KDM1A is a biologically important histone demethylase towards H3K4me1/2 and H3K9me1/2. KDM1A is overexpressed in a series of cancer types including breast, prostate, ovarian, leukemia, and has inhibitors under clinical investigation for treatment of small cell lung carcinoma. This may be due, in part to the various complexes KDM1A is co-opted by within even a single cellular environment. Each complex which utilizes KDM1A targets a specific gene set and, in some cases, opposing biological roles such as the C-terminal binding protein (CtBP) and the nucleosome remodeling deacetylase (NURD) complexes *(32,48,72,230)*. CtBP complex acts to repress E-cadherin expression, promoting the epithelial-to-mesenchymal transition (EMT); whereas the NuRD complex

represses the EMT pathway. KDM1A is overexpressed in various cancers and other disease states and has been targeted for active-site inhibition; however, as an enzyme in both pathways, pan-inhibition of KDM1A would target both complexes and mitigate the potential benefit of the NuRD complex over the CtBP complex *(46,88,95,258,259)*.

The expression and purification of KDM1A and the region of CoREST critical for binding (residues 286-493) have been established, our first goal was to establish procedures through which to express and purify additional coregulatory proteins of KDM1A. With CoREST as an initial coregulatory protein having been well-established, we sought to utilize other coregulatory proteins which were (a) involved in CoREST-containing complexes and (b) formed stable interactions.

In addition to KDM1A:CoREST, there is a common demethylase-scaffold-deacetylase core catalytic complex with histone deacetylases 1 and 2 (HDAC1/2) that exists within several complexes that have been extensively reviewed previously *(48)*. This core demethylase-scaffold protein-deacetylase has been subject to inhibition studies and there are reports that a dual-inhibition strategy of both a demethylase and deacetylase promotes a synergistic effect on gene expression *(258,260,261)*.While this core has been expressed and purified from Sf9 and HEK293F cells, I desired to develop a protocol for heterologous expression and purify this core in *E. coli (261,262)*. There are several benefits to an *E. coli* expression system which I found attractive including a much quicker growth time which would result in a higher yield of cells and protein in a

more rapid amount of time. Additionally, the HDAC proteins have been demonstrated to dimerize and in mammalian cell purifications, both are purified even if only HDAC1 or HDAC2 is expressed *(217,263)*. Assessing the effect of HDAC12 dimerization on demethylase and deacetylase activity was of interest and would require the expression and purification of each HDAC individually.

Expanding from the catalytic core, the first complex that reported the presence of KDM1A established that a 5-mer complex was able to stably pass over six chromatography columns *(113)*. This 5-mer was comprised of the catalytic core of KDM1A, CoREST, and HDAC1/2. In addition to the catalytic core of the 5-mer, there were two additional proteins, BRCA associated factor 2/High mobility group 20B/SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1-related (Braf35, HMB20B, SMARCE1-related protein), and BHC80/PHD finger domain 21A (BHC80/PHF21A), which may impart additional selectivity or specificity to the catalytic core due to their reader domains *(31,113)*. BHC80 and Braf35 fit our two criteria for being involved in a CoREST-containing complex as well as a clear interaction stability as they were able to stay bound over a non-trivial number of purification steps.

I envisioned working towards expression and purification of each of these components. The interaction between each of these proteins and KDM1A has not yet been fully elucidated. Expression and purification of each individual protein would

provide unprecedented access towards utilization of these proteins in the activity assays as well as the biochemical and biophysical assays I outlined above. With each additional protein, I would be able to assess the changes to an increasingly complex KDM1A-containing epigenetic structure.

Herein I report my efforts in cloning, expression, and purification of HDAC1/2, BHC80, Braf35, and MTA1. I have outlined two methods to obtain a KDM1A:CoREST heterodimer as well as multiple methods to access HDAC1/2. I have additionally cloned and expressed both BHC80 and Braf35 in multiple expression constructs. I have additionally expressed and purified an additional scaffold protein, MTA1, to assess the commonality of scaffold proteins in KDM1A binding. Overall, I have made significant progress towards our expression and purification goals to will enable us to assess how KDM1A is affected by coregulatory proteins.

2.1 Co-expression and purification of KDM1A and full-length CoREST

REST co-repressor 1 (RCOR1,CoREST) a scaffolding protein in several complexes containing both KDM1A and HDAC1/2 wherein the linker region and SANT 2 domain contribute to the binding interface with KDM1A and binding to HDAC1/2 occurs via the ELM/SANT1 domains of CoREST (Figure 13) *(74,113–116)*. The SANT2 domain of CoREST was demonstrated by the Mattevi lab to bind nucleosomal DNA *(118)*. CoREST has been demonstrated to be required for KDM1A activity on a nucleosomal substrate. It may be that CoREST is required to bind and direct KDM1A to its substrate. *(118)*.

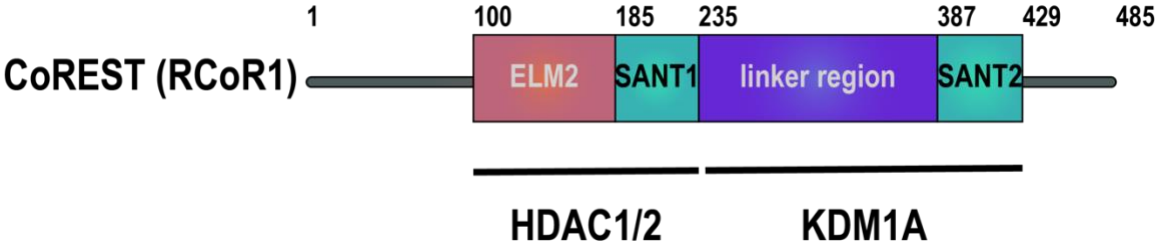


Figure 13: REST Co-repressor 1 (RCoR1/CoREST) domain map with protein binding regions depicted. CoREST contains two SANT domains, which historically bind DNA, an ELM2 domain involved in protein dimerization as well as a 'linker region' which contributes a majority of the binding affinity between KDM1A and CoREST.

All of the current knowledge in the literature regarding the binding of KDM1A to its scaffold partner, CoREST, is based on the linker region and SANT2 domain (residues 286-493). The major contribution of binding affinity to KDM1A is made through the linker domain as previously established by our lab and others *(53,117,203)*. Previously unidentified regions of CoREST may make additional contact with KDM1A to impart conformational changes that are not yet established they may still make additional contacts with KDM1A . Additionally, the ELM2/SANT1 domains of CoREST interact with HDAC1/2 and accessing these regions of CoREST would allow us to probe the CoREST:HDAC1/2 interface *(115,217)*.

The expression and purification of KDM1A has been well established both in our lab and by several other labs *(53,117)*. However, there are currently no published methods for heterologous expression and purification of CoREST. Attempts by previous lab members resulted in expression of protein, but none of the expression or purification methods resulted in soluble protein.

Previous lab member Jonathan Burg in hand generated two pET-DUET1 constructs containing KDM1A$_{N\Delta 179}$ and CoREST$_{N\Delta 59}$. Each construct utilized tandem affinity purification (TAP) tags wherein each protein was orthogonally tagged. The underlying principle behind TAP is that each protein has a separate affinity handle enabling for the complex to be run over multiple columns to increase purity but also ensure a 1:1 ratio between each of the construct components. This method of co-expression enables KDM1A to function as a stabilizing protein for CoREST to bind upon translation.

The second-generation construct that he produced utilized a hexahistidine (6xHis) tag on KDM1A and a maltose binding protein (MBP) tag on CoREST, each with a tobacco-etch virus (TEV) protease cleavage site between the tag and the protein. We utilized expression conditions established by Dr. Burg to optimize a purification strategy which results in 0.126 mg/L KDM1A:CoREST to 90% purity. However, this strategy is lengthy and did not produce high yields of protein. The third-generation construct that Dr. Burg produced utilized, again, a 6xHis tag on KDM1A; however, CoREST was preceded by a glutathione s-transferase (GST) tag in this construct. I optimized expression conditions for this construct as well as a purification strategy.

2.1.1. Expression and purification of a second generation KDM1A:CoREST complex

A pET-DUET1 vector with ampicillin resistance, a T7 promoter, *lac* operon, KDM1A$_{N\Delta 179}$ and CoREST$_{N\Delta 59}$ were each tagged N-terminally with a 6xHis and MBP tag,

respectively. Each protein was full length except for the nuclear localization signal at the N-terminus of KDM1A (1-179) and the first 60 residues of CoREST which were predicted to be unstructured.

While the construct could be induced via Dr. Burg reported a leaky expression of protein from a BL21(DE3) cell line (Novagen) when incubated at isopropyl β-D-1-thiogalactopyranoside (IPTG) for 40 hours at 30 °C in a shaking incubator at 200 rpm. I have worked to optimize an initial purification strategy run on a small-scale by Dr. Burg and have accessed low yields of KDM1A/CoREST at 0.126 mg/L and >90% purity.

2.1.1.1 Heterologous expression of a second generation KDM1A:CoREST complex in *E. coli*

The second-generation pET-DUET1 vector was electrochemically transformed into a BL21 Star(DE3) (ThermoFisher) cell line and grown on an ampicillin resistant lennox broth (LB, GeneseeScientific 11-125) medium overnight. A swipe of colonies was utilized to inoculate a 100 mL starter culture. Upon reaching confluency, the large-scale culture flasks were inoculated to a final 0.1% starter culture and then incubated in a shaking incubator (200 rpm, 30 °C, 40 hrs.). The resulting cells were harvested via SLA-3000 rotor (5000 rpm, 4 °C, 10 min), transferred to a 50 mL conical and stored in the -20 °C freezer until use.

2.1.1.2 Purification of a second generation KDM1A:CoREST complex

In an initial attempt, a thawed pellet was resuspended in 50 mM Tris-HCl buffer system at pH 7.4 containing 350 mM NaCl, 10 mM imidazole, 5% glycerol, 1X Halt

protease inhibitor cocktail (ThermoFisher 78429), lysozyme (1.0 mg/L, Sigma L6876), and PMSF (0.4 mM, Sigma P7626) and then mechanically lysed in a high-pressure cell disruption system. The resulting cell lysate was subjected to ultracentrifugation to remove insoluble material (44,000 *x g*, 4 °C, 45 min). The supernatant was then loaded onto a 25 mL IDA chelating sepharose column previously charged with nickel at a flow rate of 2.0 mL/min via benchtop peristaltic pump. After completion of loading, the column was washed extensively with a 50 mM Tris-HCl buffer system at pH 7.4 containing 350 mM NaCl, 10 mM imidazole, and 5% glycerol utilizing a GE Healthcare fast protein liquid chromatography (FPLC) ÄKTA purification system (ÄKTA). An imidazole gradient was then implemented up to 300 mM imidazole over a 15 CV to elute bound protein from the column (Figure 14a).

To remove the high levels of imidazole and glycerol from the samples to ensure compatibility with the amylose resin (NEB), the fractions containing KDM1A and CoREST were concentrated and buffer exchanged to a buffer containing 20 mM Tris-HCl (pH 7.4), 200 mM NaCl, 1 mM EDTA, and 1 mM DTT with an Amicon Stirred Cell system. The concentrated protein was loaded onto a 10 mL amylose resin below 1 mL/min via peristaltic pump. Upon completion of loading, the column was washed extensively on an ÄKTA for 7 CV and then eluted with 10 mM maltose (Figure 14b) over 4 CV. The pooled fractions containing a 1:1 ratio of KDM1:CoREST were incubated overnight with 6xHis TEV protease to cleave the affinity tags from the proteins.

The resulting cleaved mixture was loaded onto a 25 mL IDA chelating sepharose column charged with nickel at a flow rate of 1 mL/min via peristaltic pump. The column was washed on an ÄKTA for 4 CV and bound protein was eluted over an imidazole gradient up to 300 mM over 6 CV. We hypothesized that KDM1A:CoREST would be in the flow through, but SDS-PAGE revealed that the complex required a low level of imidazole to be eluted, which has been noted previously with KDM1A alone. We noted that imidazole was required for the cleaved heterodimer to be eluted from the column (Figure 14c). The requirement for imidazole to elute KDM1A:CoREST resulted in 6xHis-TEV protease contamination.

The pooled fractions were concentrated using an EMD Millipore centrifugal filter unit (10 kDa MWCO, 2000 *x g*, 4 °C, 20 min) to a volume less than 1 mL and then loaded onto a 16/600 S200 Sephacryl column that eluted with 1.5 CV (20 mM HEPES (pH 7.8), 200 mM NaCl, 1 mM DTT) at 0.7 mL/min. The resulting gel had extremely faint bands corresponding to the correct MW at two different elution points (Figure 14d). The two MW bands corresponding to KDM1A and CoREST in the green box of Figure 14d appeared to be in a 1:1 ratio. However, the MW bands corresponding to KDM1A and CoREST in the second peak (Figure 14d, blue box) did not appear at a 1:1 ratio.

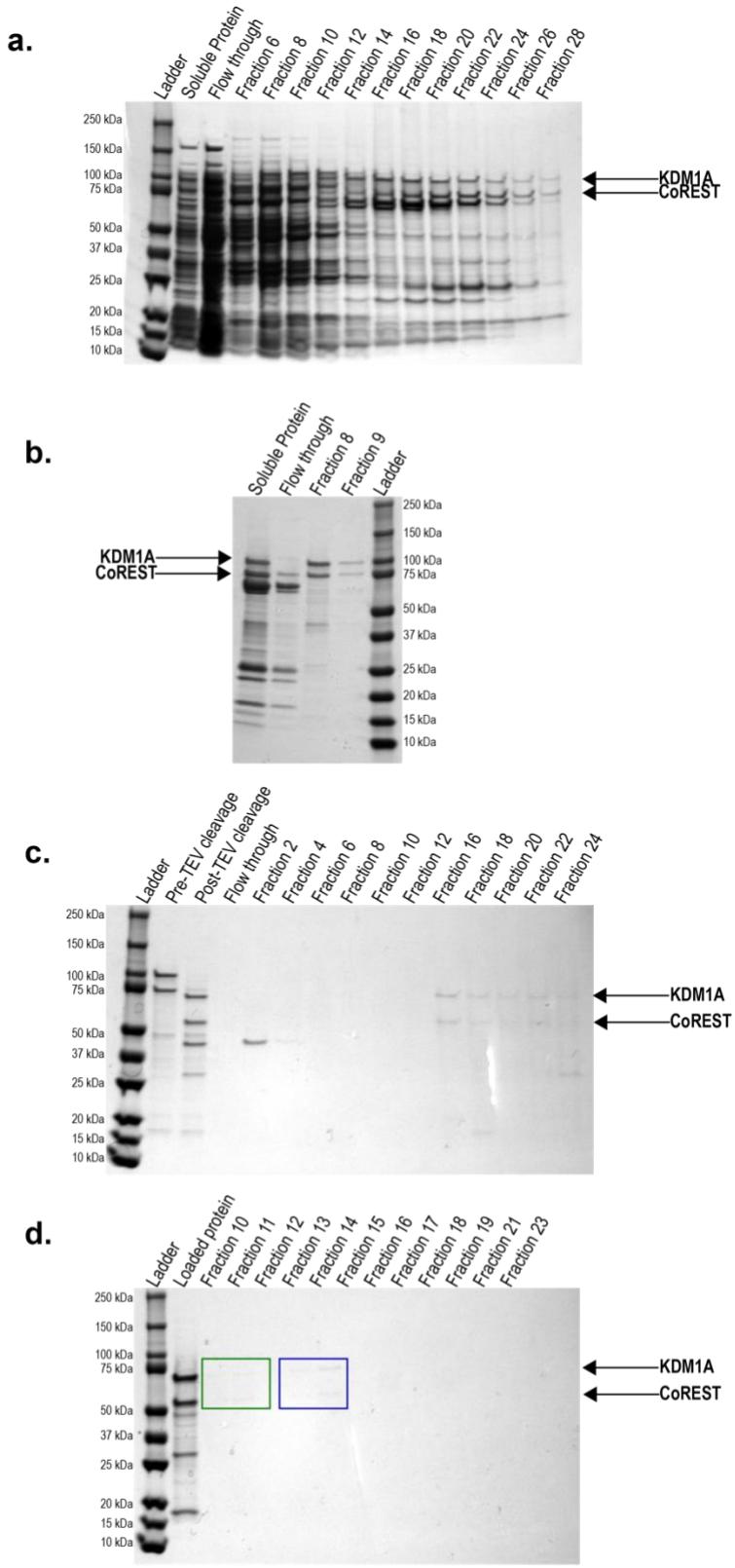

a.
Ladder
Soluble Protein
Flow through
Fraction 6
Fraction 8
Fraction 10
Fraction 12
Fraction 14
Fraction 16
Fraction 18
Fraction 20
Fraction 22
Fraction 24
Fraction 26
Fraction 28
250 kDa
150 kDa
100 kDa
75 kDa
50 kDa
37 kDa
25 kDa
20 kDa
15 kDa
10 kDa
KDM1A
CoREST
b.
Soluble Protein
Flow through
Fraction 8
Fraction 9
Ladder
KDM1A
CoREST
c.
Ladder
Pre-TEV cleavage
Post-TEV cleavage
Flow through
Fraction 2
Fraction 4
Fraction 6
Fraction 8
Fraction 10
Fraction 12
Fraction 16
Fraction 18
Fraction 20
Fraction 22
Fraction 24
KDM1A
CoREST
d.
Ladder
Loaded protein
Fraction 10
Fraction 11
Fraction 12
Fraction 13
Fraction 14
Fraction 15
Fraction 16
Fraction 17
Fraction 18
Fraction 19
Fraction 21
Fraction 23
KDM1A
CoREST

Figure 14: Initial strategy to purify KDM1A:CoREST from a second-generation construct. (a) IDA chelating sepharose (50 mM Tris HCl (pH 7.4), 350 mM NaCl, 5% glycerol, 10 mM imidazole) run on an AKTA with a 10 CV wash and 15 CV elution. (b) Amylose resin (20 mM Tris HCl (pH 7.4), 200 mM NaCl, 1 mM EDTA, 1 mM DTT) run on an AKTA with a 7 CV wash and 4 CV elution of 10 mM maltose. (b) Post-overnight TEV protease cleavage, the protein mixture was run over an IDA chelating sepharose column on an AKTA with a 4 CV wash and 6 CV gradient (same buffer system as A). (d) Hi-Load 16/600 S200 Sephacryl column over 1.5 CV (20 mM HEPES (pH 7.8), 200 mM NaCl, 1 mM DTT)

We next sought to assess if lengthening the elution of the second IDA chelating sepharose column would allow for separation between KDM1A:CoREST and TEV protease. We performed the purification utilizing the same protocol as above. At the second IDA chelating sepharose column, post-TEV cleavage, we performed a 20% 4CV step-wise elution to a concentration of imidazole at 300 mM. SDS-PAGE analysis revealed apparent separation between KDM1A:CoREST and TEV protease (Figure 16a) . We concentrated the protein complex with an EMD Millipore centrifugal filter unit (30 kDa MWCO, 2500 *x g*, 4 °C, 20 min) to a concentration of 8.248 μM with a yield of 0.126 mg/L of protein (Figure 16b).

While this is not an unreasonable level of protein to obtain over a purification strategy, overall this construct carries several inherent issues. Firstly, the length of the growth period makes this strategy less attractive over a more standard IPTG expression timescale and secondly, amylose resin is not able to be regenerated consistently after 5 purifications due to amylase in the cell lysate. Thus we sought to utilize a resin that retained a higher binding capacity. Interestingly, a recent study reported utilization of

rice flour in their purification of an MBP-tagged protein due to the high levels of amylose in rice flour *(264)*. We sought to test if we could utilize a homemade amylose resin.

A protocol reported by Kellerman and Ferenci in 1982 depicts their method of crosslinking amylose to produce a homemade amylose resin for purification of MBP tagged proteins wherein they utilized amylose heated and then crosslinked using strong base and epichlorohydrin *(265)*. We were curious if we could utilize this crosslinking strategy with readily accessible rice flour, potato starch, and corn starch. Crosslinking the material would allow us to more readily filter the resin through mechanical rigidity.

For the crosslinked samples, we heated 10 grams of each sample in 40 mL of milliQH_2O to 50 °C upon which time 5M NaOH was added to a final volume of 100 mL followed by 30 mL epichlorohydrin with stirring. The samples were then cooled to room temperature on the benchtop, the gels broken to pieces and washed 3x with a 50 mM glycine, 0.5 M NaCl buffer utilizing a fritted filter funnel. The gels were then washed 2x further with 10 mM Tris-HCl (pH 7.2). The washed gel pieces were carried forward into a binding test.

In the binding test, equal weights of crosslinked and uncrosslinked homemade resins were equilibrated in water and then a standard amylose resin binding buffer as described above to the equivalent of 5 CV. An *E. coli* cell lysate expressed with 6xHisKDM1A/MBPCoREST was divided amongst the samples and all were incubated

overnight with rotation at 4 °C. The samples were spun using a swinging bucket rotor (2500 rpm, 4 °C, 20 min), the supernatants removed. The resins were washed to 15 CV and then incubated with buffer containing 10 mM maltose for 1 hour, spun down, and the supernatant removed. SDS-PAGE analysis revealed that all of the test resins, except for uncrosslinked rice flour, were able to bind MBP-CoREST (Figure 15). The uncrosslinked corn starch and potato starch revealed binding of a retained a high level of MBP-CoREST. This suggested to us that future attempts to purify MBP-tagged proteins should utilize a batch elution strategy with corn starch or potato starch.

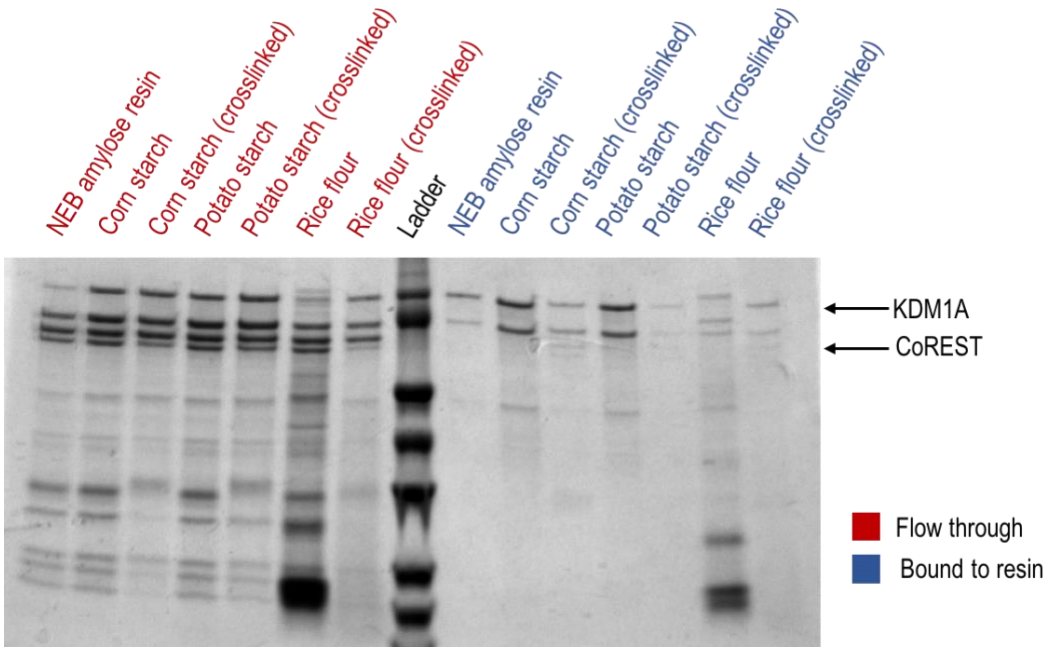


Figure 15: Assessment of the ability of crosslinked and non-crosslinked potato starch, corn starch, and rice flour to bind an MBP-tagged CoREST protein with KDM1A in complex as compared to the NEB amylose standard

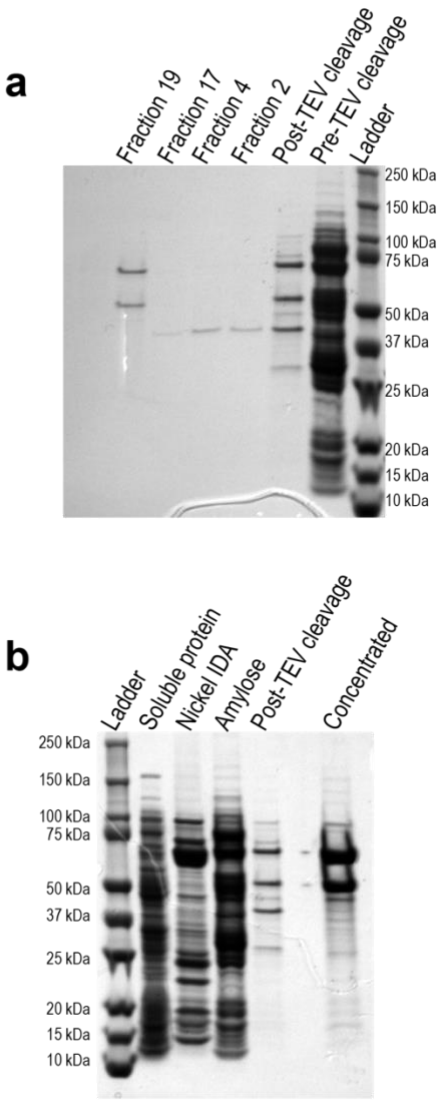


Figure 16: Second Ni-IDA column post-TEV cleavage and overall purification strategy. (a) We utilized a 20% step gradient to allow for the separation between TEV protease and KDM1A:CoREST. (b) Overall purification scheme resulting in 90& pure protein at 0.126 mg/L yields.

2.1.2. Expression and purification of a third generation KDM1A:CoREST complex

Even with the uncrosslinked resin being successful, we still felt there was benefit towards utilizing the third-generation KDM1A:CoREST construct due to the benefits of the expression strategy. Previous lab member, Jonathan Burg, designed and constructed what we will refer to as a third-generation construct. Containing kanamycin resistance, a T7 promoter, and *lac* operon, KDM1A and CoREST were each tagged N-terminally with a 6xhis and GST tag, respectively. Each protein was full length except for the nuclear localization signal at the N-terminus of KDM1A (1-179) and the first 60 residues of CoREST which were predicted to be unstructured.

In his hands, Dr. Burg described an expression strategy with an overnight starter culture (30 °C, 200 rpm) in LB (GeneseeScientific 11-125) with kanamycin resistance, induction of large-scale cultures to a final inoculum of 0.1%, with 1 mM $MgCl_2$ (30 °C, 220 rpm). Upon reaching an OD_{600} of 0.8, the cells were induced with IPTG to a final concentration of 1 mM and allowed to grow overnight (19 °C, 220 rpm). He noted via SDS-PAGE gel that he was able to visualize two overexpressed bands at the MW corresponding to the expected weights of 6xHis-KDM1A and GST-CoREST.

While I noted these bands as well during expression, upon purification there was no binding to the glutathione sepharose column. I have thus worked through expression testing and a purification strategy through which we have been able to obtain KDM1A:CoREST through this strategy.

2.1.2.1 Expression of a third generation KDM1A:CoREST complex

The original expression conditions of the third-generation construct were based on the previously reported expression conditions for CoREST in our lab *(117)*ns

. We decided to expand the expression conditions to include a slightly higher and lower expression temperatures of 21 °C and 15 °C at four different IPTG induction concentrations (1.0 mM, 0.5 mM, 0.25 mM, and 0.1 mM). For these expression tests we inoculated an overnight starter flask with kanamycin and chloramphenicol resistance with a glycerol stock of BL21 Star (DE3) cell line that had been electrochemically transformed and contained the third-generation construct. The induction cultures were grown at 30 °C to an OD_{600} of 0.8, cooled to either 21 °C or 15 °C for 10 minutes, induced with the appropriate amount of IPTG, and allowed to grow overnight to reach confluency.

We collected samples from each of the test induction flasks and ran SDS-PAGE and Western Blot (WB) analysis for both 6xHis and GST (Figure 17). We noted no significant changes between the sample taken before induction and the final time point in most of the samples except for the 0.5 mM and 0.25 mM IPTG at 15 °C. While the 6xHis tag was shown consistently on the WB for all eight of the conditions, there was no GST-tag observed (Figure 17b,c). Even though BL21 Star(DE3) cells are designed to limit proteolysis of proteins, we suspected that GST-CoREST was being degraded overnight.

Thus, I next decided to perform an expression test over shorter time points at higher induction temperatures.

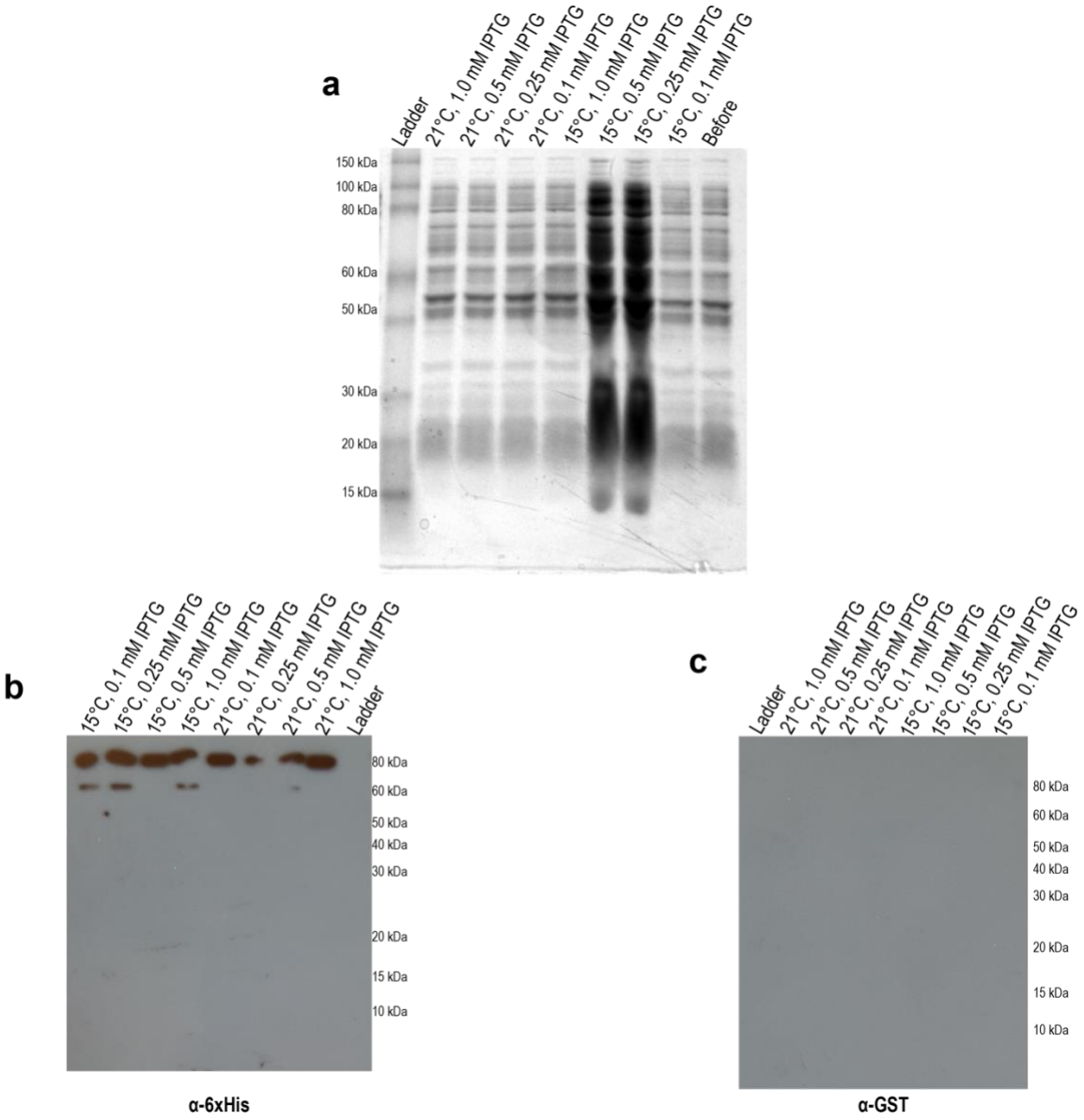


Figure 17: Expression test of third-generation construct. A BL21 Star (DE3) cell line containing the third-generation construct was induced at 15 °C and 21 °C utilizing 4 different IPTG concentrations (1.0 mM, 0.5 mM, 0.25 mM, 0.1 mM). The induced cultures were grown overnight and then subjected to (a) SDS-PAGE (Bio-Rad 4-20%, 120V 65 min), and subsequent WB analysis utilizing an (b) α-6xHis (1:5000) or (c) α-GST(1:1000). We observed expression of 6xHis-KDM1A across all conditions, but no expression of GST-CoREST was observed via WB or visually by SDS-PAGE with comassie staining.

To examine the possibility that CoREST was being degraded, I examined the effects of reducing the incubation time. I chose induction temperatures of 30 °C and 37 °C with IPTG concentrations at 1.0 mM and 0.25 mM. Similar to above, an overnight starter flask inoculated with cells from a glycerol stock was utilized to inoculate the induction flasks. The flasks were grown at 30 °C to an OD_{600} of 0.8 and then induced with the appropriate amount of IPTG. Time points were taken at 0, 1, 2, and 4 hours, samples spun down in a tabletop centrifuge (8000 rpm, 4 °C, 2 min), supernatant removed, and the pellet stored at -20 °C until use. We subsequently ran SDS-PAGE and WB analysis of the samples for both 6xHis and GST (Figure 18). I noted consistent expression across all GST-CoREST samples, and highest expression for 6xHis-KDM1A after 4 hours post-0.25 mM IPTG induction at 30 °C.

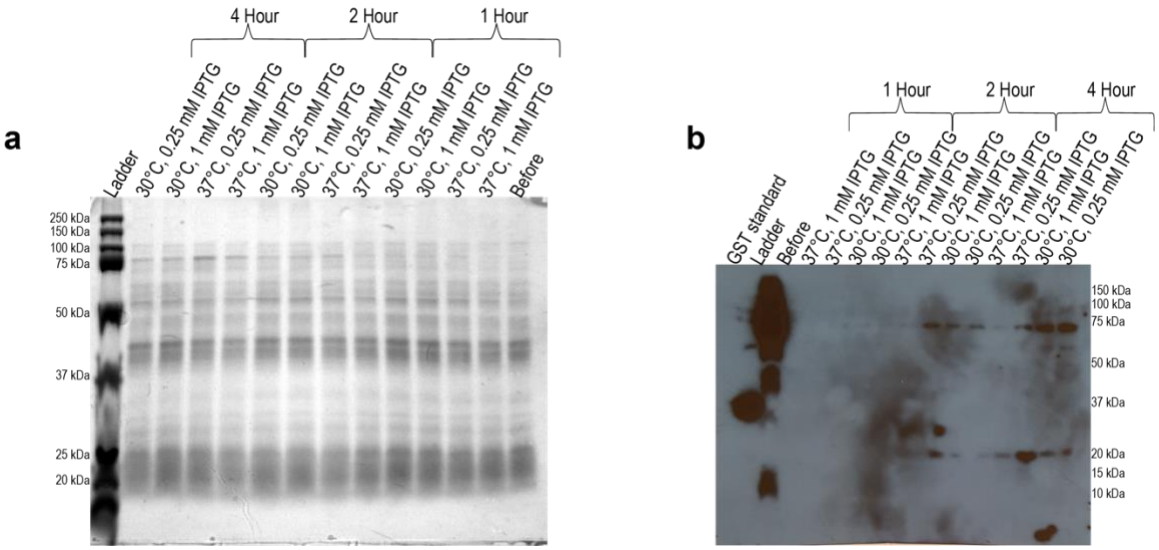


Figure 18: Expression test of third-generation construct. A BL21 Star (DE3) cell line containing the third-generation construct was induced at 30 °C and 27 °C utilizing 2 different IPTG concentrations (1.0 mM, 0.25 mM). Time points were taken at 0, 1, 2, and 4 hours and then subjected to (a) SDS-PAGE (Bio-Rad 4-20%, 120V 65 min), and subsequent WB analysis utilizing an (b) α-6xHis (1:5000) or α-GST(1:1000). We observed expression of GST-CoREST across all conditions, but varied expression of 6xHis-KDM1A, with the highest observed level after 4 hours at 30 °C with a 0.25 mM IPTG induction.

2.1.2.2 Purification of a third generation KDM1A:CoREST complex

The adaptation of shortened induction times facilitated production of GST-CoREST, so I then turned to the development and expression of a purification protocol. I envisioned a strategy wherein the cell lysate was run over a Ni-IDA chelating sepharose column followed by a glutathione agarose column, subsequent overnight TEV-protease cleavage and then a final size exclusion chromatography column.

A thawed pellet was resuspended in 25 mM sodium phosphate buffer system at pH 7.0 containing 300 mM NaCl, 10 mM imidazole, 5% glycerol, 1X protease inhibitor cocktail (Biotool 14001), lysozyme (1.0 mg/L Sigma L6876), and PMSF (0.4 mM, Sigma

P7626) and then mechanically lysed in a high-pressure cell disruption system. The resulting cell lysate was subjected to ultracentrifugation to remove insoluble material (43,000 *x g*, 4 °C, 45 min). The supernatant was then loaded onto a 25 mL IDA chelating sepharose column previously charged with nickel at a flow rate of 2.0 mL/min via benchtop peristaltic pump. After completion of loading, the column was washed extensively with a 25 mM sodium phosphate buffer system at pH 7.0 containing 300 mM NaCl, 10 mM imidazole, 5% glycerol utilizing. An imidazole gradient was then implemented up to 250 mM imidazole over a 20 CV to elute bound protein from the column (Figure 19a).

We next pooled the fractions containing KDM1A:CoREST as assessed via SDS-PAGE and proceed to a 10 mL glutathione agarose column (Genesee Scientific). The column was subsequently eluted with 10 mM reduced glutathione (Figure 19b) over 5 CV. The pooled fractions containing a 1:1 ratio of KDM1:CoREST were incubated overnight with 6xHis TEV protease to cleave the affinity tags from the proteins.

Following concentration of the mixture by centricon (30 kDa MWCO, 2000 *x g*, 4 °C, 20 min), the was then loaded onto a previously equilibrated 16/600 S200 Sephacryl and eluted over 1.5 CV (20 mM HEPES (pH 7.8), 200 mM NaCl, 1 mM BME, 0.7 mL/min) (Figure 19c). The fractions were analyzed by SDS-PAGE (Figure 20a) with Coomassie staining and confirmed the presence of protein in fractions 11 and 12. (Figure 20b-d). Fractions 11-12 were concentrated by a centricon concentrator (30 kDa MWCO 2000x g, 4

°C, 20 min) to a volume of 1 mL and assessed concentration to be 2.088 uM with a yield of 0.2588 mg/L via Bradford assay.

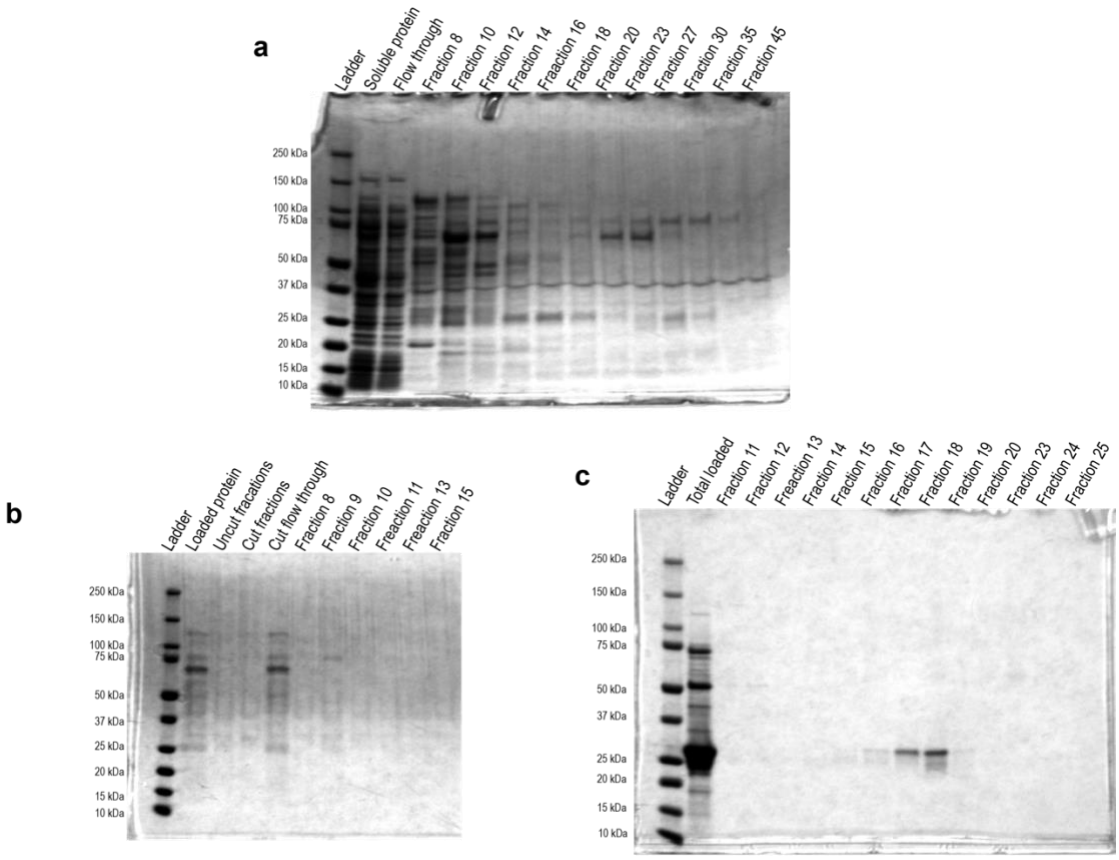


Figure 19: Strategy to purify KDM1A:CoREST from a third-generation construct. (a) IDA chelating sepharose (25 mM sodium phosphate (pH 7.0), 300 mM NaCl, 5% glycerol, 10 mM imidazole) run on an AKTA with a 15 CV wash and 20 CV elution. (b) Glutathione agarose resin (PBS with 2 mM BME to wash, 50 mM Tris-HCl (pH 7.4) with 150 mM NaCl, 10 mM reduced glutathione, 2 mM BME for elution) run on an AKTA with a 15 CV wash and 5 CV elution. (c) Post-overnight TEV protease cleavage, the protein mixture was concentrated and run over an SEC 16/600 S200 Sephacryl column (25 mM HEPES (pH 7.8), 200 mM NaCl, 1 mM BME) for 1.5 CV

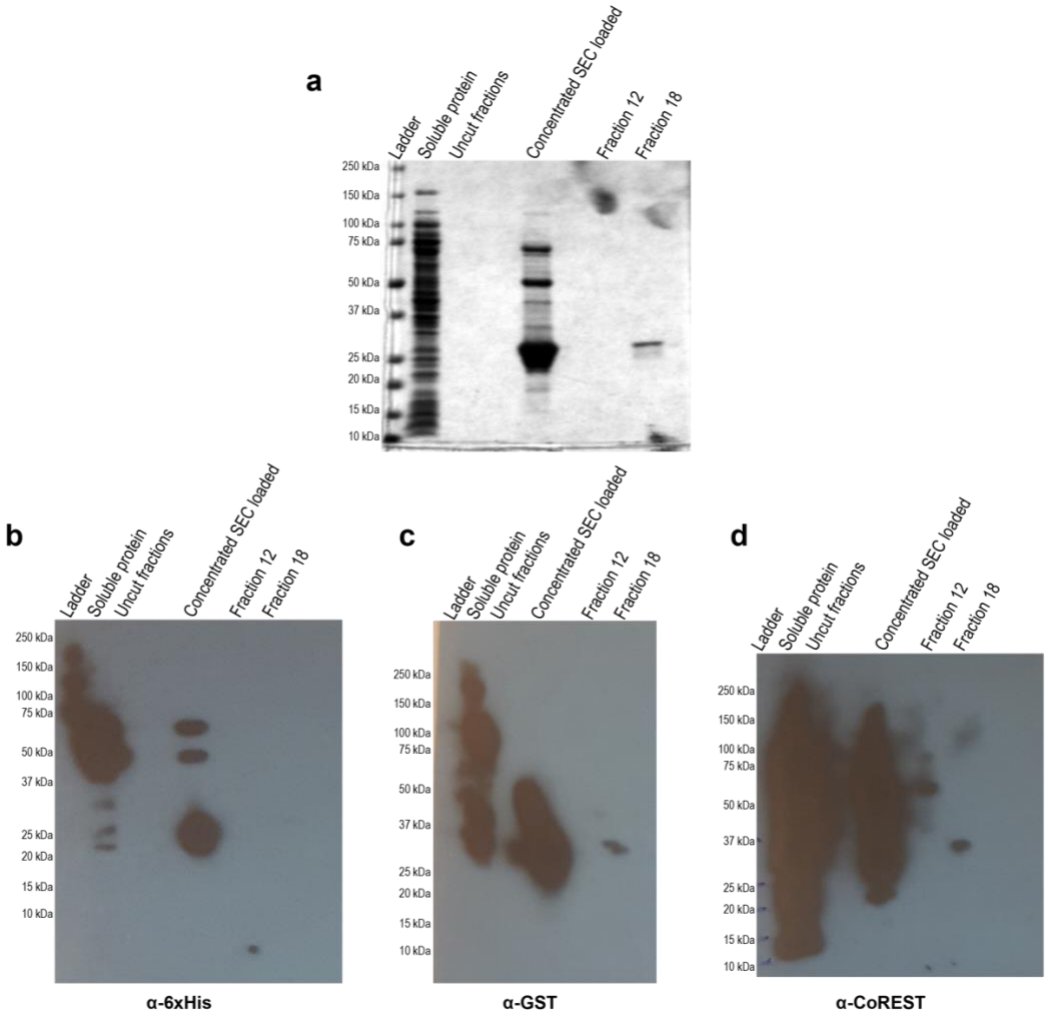


Figure 20: Assessment of the overall purification strategy for the third-generation construct. (a) SDS-PAGE (Bio-Rad 4-20%, 120V 65 min), and subsequent WB analysis utilizing an (b) α-6xHis (1:5000), (b) α-GST(1:1000), or (c) α-CoREST (1:8000). We observed presence of CoREST, and no 6xHis or GST indicating complete cleavage of tags.

In conclusion, I developed and optimized a method to produce KDM1A:Corest via recombinant co-expression in the heterologous host *E. coli.* have provided to

methods through which KDM1A:CoREST can be obtained to sufficient purity for subsequent interaction studies such as ITC and HDX-MS as well as in enxymatic assays with additional coregulatory proteins of not only the REST complex, but any complex which contains both KDM1A and CoREST.

2.2 Expression and purification of Braf35

BRCA associated factor 2/High mobility group 20B/SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily E member 1-related (Braf35, HMB20B, SMARCE1-related protein) is one of the proteins within the 5-mer complex. It was first noted to be found in complexes containing CoREST and HDACs as part of a parallel study which was assessing the interaction between BRCA2 and Braf35. This lab noted that a distinct subpopulation of Braf35 did not interact with BRCA2 and instead was found with CoREST-containing complexes *(120)*. They further assessed the DNA binding ability of Braf35, presumably through its HMG domain (70-138, Figure 21a) , and found it specifically interacted with 4-way DNA junctions (Figure 21b).

While the exact role of Braf35 has yet to be established, a 2012 study by the Reyes lab established that Braf35 is sumoylated within KDM1A-CoREST containing complexes *(121)*. A paralog of Braf35, iBraf can form heterodimers with Braf35, inhibiting both sumoylation and its interaction with the complexes, affecting the ability of the complexes to function within the cells *(121)*. While it is not clear what the cause and

effect is and what specific role sumoylation plays, it is clear that Braf35 does play a critical role in KDM1A-CoREST complexes *in vivo*.

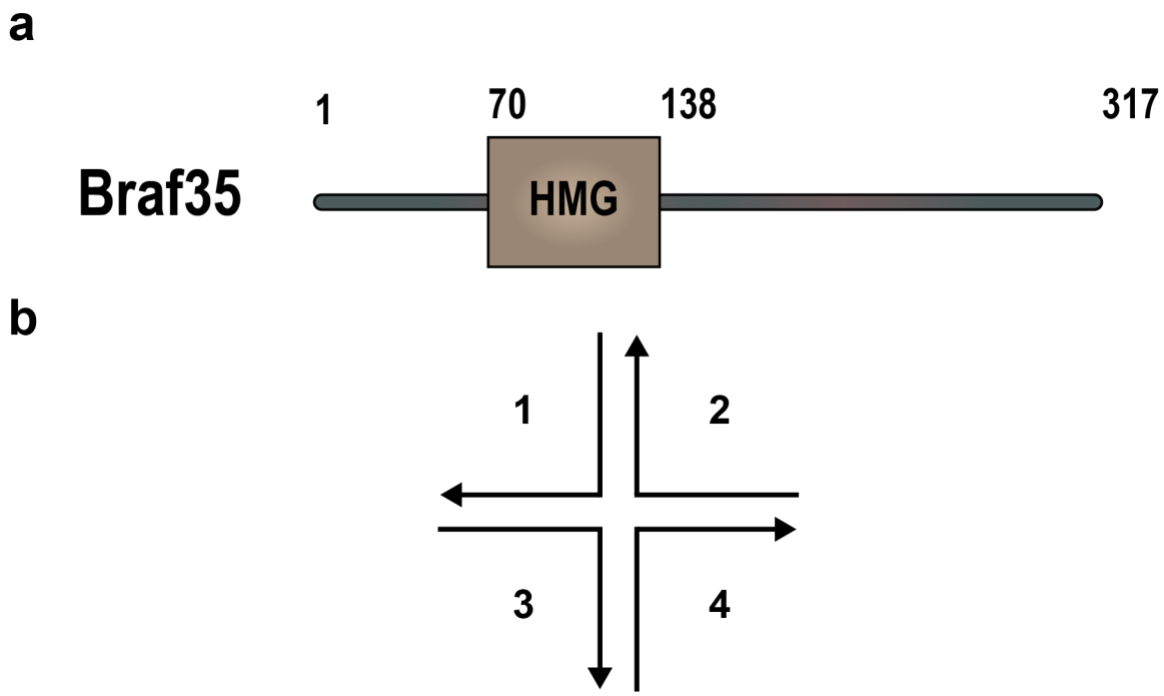


Figure 21: Overview of Braf35 with domain map and representation of a 4-way DNA junction (a) domain map of Braf35 depicting the placement of the HMG domain within the protein and (b) a pictorial representation of a 4-way DNA junction that Braf35 preferentially binds.

Herein, we report our efforts towards cloning, expression, and purification of Braf35 recombinantly in *E. coli*. We were able to produce both a pET15-b construct with an N-terminal 6xHis tag and a TEV protease cut-site preceding the protein as well as a pET28-b(+) construct with an N-terminal 6xHis-GST tandem tag followed by a TEV protease cut site directly before Braf35. I have performed extensive expression testing on both constructs and demonstrated expression of the protein via both plasmids. While I

was unable to obtain soluble protein via the pET15-b vector, I anticipate that a purification protocol for Braf35 from a pET28-b vector will be rapidly accessible.

2.2.1 Cloning of Braf35 into pET expression plasmids

In order to produce a construct containing Braf35, I first obtained a gBlock from IDT containing a codon optimized sequence for *E. coli* flanked by N-terminal NdeI and C-terminal XhoI cleavage sequences. The utilization of different codons varies between strains, and as a result, there are differing levels of tRNAs for each codon between species *(266)*. If a codon is significantly more frequent in the host species, it is much more likely for the ribosome to stall while waiting for the correct tRNA, resulting in a significant number of truncation products.

We chose first to access the pET15-b construct utilizing restriction digestion of the backbone and gBlock synthetic Braf35 linear sequence. Following the IDT protocol for gBlock restriction digests, we incubated 1 μg of both backbone and insert, separately, with CutSmart buffer (NEB), and 20 U of NdeI and XhoI (NEB). An additional 20 U of each enzyme was added halfway through the digestion (37 °C, 3 hrs.). The backbone was dephosphorylated to enhance subsequent ligation (37°C, 1 hour). An agarose gel was run (0.8% agarose, 80V, 30min, 1 kB ladder) to compare the cut backbone to the original vector and allow for separation between the backbone, and the sequence that had been digested from the pET15-b vector (Figure 22). The digested insert and backbone were excised and subsequently purified (Qiagen).

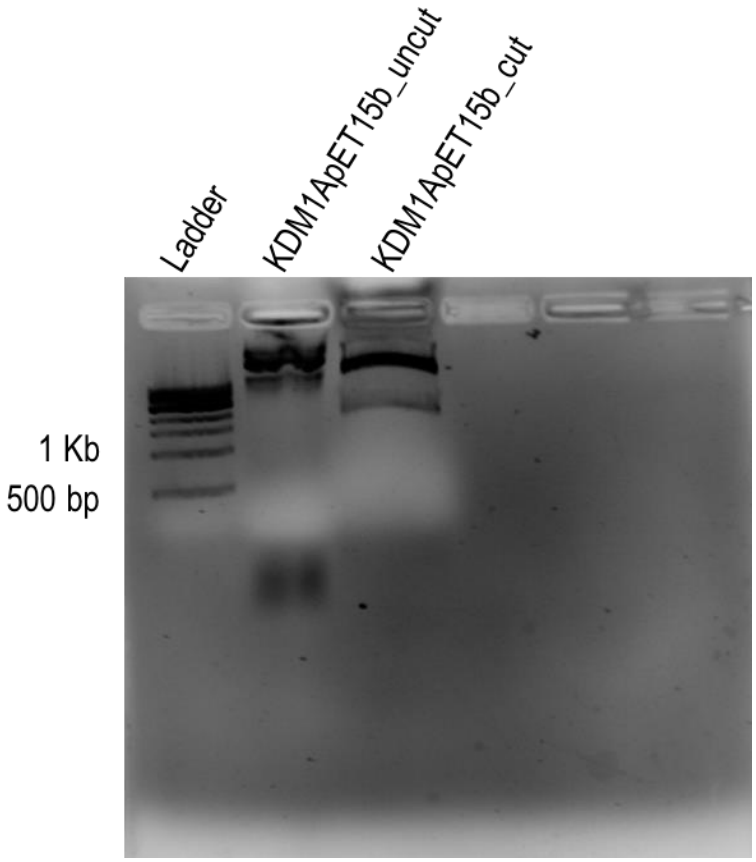


Figure 22: Restriction digest of a pET15-b vector for Braf35 restriction digest. Digest was run over the course of 3 hours utilizing CutSmart buffer and 40 U of both XhoI and NdeI. A 0.8% agarose gel was run (80V, 30 Min, 1kB ladder) of both an uncut vector and the cut vector. In the cut vector lane, we noted a lower molecular weight band corresponding to the original sequence between the two restriction enzymes – KDM1A.

For the ligation reaction, we measured the concentrations of the purified pieces and ran two ligation reactions (NEB) at 3:1 and 4:1 insert:vector molar ratios and incubated overnight at 16 °C. In order to remove excess enzyme and salts that could inhibit an electrochemical transformation, we performed a DNA purification utilizing

EMD Millipore PelletPaint Co-Precipitant. We performed an electrochemical transformation and plated the cells on an ampicillin plate overnight at 37 °C.

Similarly, for the production of a pET28-b(+) vector containing Braf35, we utilized Xho1 and NdeI in our restriction digest of the pET28-b(+) backbone and pET15-b containing the Braf35 insert (37 °C, 2 hrs). I observed cleavage via agarose gel and proceeded to excise and purify the DNA for ligation (1.2% agarose gel, 80V, 30 min,1 kB ladder) (Figure 23). DNA ligation was performed at 3:1, 4:1 and 6:1 insert:vector molar ratios and incubated overnight (16 °C, NEB T4 DNA ligase). The ligation mixture was desalted, transformed into electrocompetent DH10β cells, incubated for 1 hour at 37 °C, and plated on LB agar containing 50 μg/mL kanamycin and incubated overnight. Upon obtaining both vectors with confirmed sequencing from Eton Biosciences, we proceeded to expression tests for both vectors to determine the best method to access a purified Braf35 from *E. coli*.

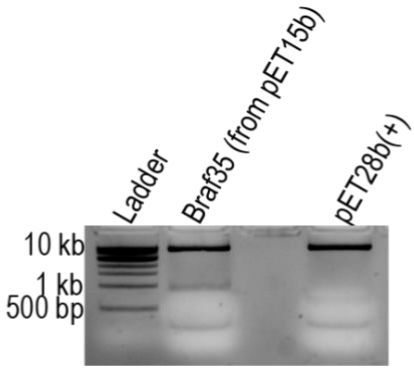


Figure 23: Restriction digest of a pET15-b vector containing Braf35 and a pET28-b(+) vector containing CoREST$_{286-493)}$ for restriction digest. Digest was run over the course of 3 hours utilizing CutSmart buffer and 40 U of both XhoI and NdeI. A 0.8% agarose gel was run (80V, 30 Min, 1kB ladder) of both cut vectors. In both, we noted two bands corresponding to the MW of the backbone (higher) and the original insert between the restriction enzymes.

2.2.2 Expression tests for 6xHis-Braf35 (pET15-b)

I next conducted expression tests of the pET15-b 6xHis-Braf35 construct. Four induction temperatures (37 °C, 30 °C, 21 °C, and 15 °C) were examined at two different IPTG concentrations (1.0 mM and 0.5 mM).

I transformed electrocompetent *E. coli* BL21(DE3) (Novagen) with pET16b-Braf35 plasmid DNA, as above. Several single colonies of *E. coli* BL21(DE3) were used to inoculate an overnight starter flask (37 °C, 200 rpm). Upon reaching confluency, 100 mL flasks of LB media with ampicillin were inoculated to 1% v/v$_t$ and incubated (37 °C, 200 rpm) to an OD_{600} of 0.6 at which time the flasks were moved to their induction temperature for 10 minutes and the induced with either 0.5 or 1.0 mM of IPTG. Time points were taken at 0, 1, 2, and 4 hours for the 30 °C and 37 °C samples whereas I only

took time points for the overnight growth samples of 15 °C and 21 °C due to the slowed growth of the two lower temperature samples.

The time point samples were assessed by SDS-PAGE with Comassie staining (homemade 12% acrylamide gel, 140V, 90 min) (Figure 24). The expected molecular weight was 38 kDa, which we noted only for the overnight time point for 21 °C induction temperature with 0.5 mM IPTG. However, with the 21 °C induction at 0.5 mM IPTG in hand, I decided to expand conditions around this growth to optimize the amount of protein I was able to collect from an expression.

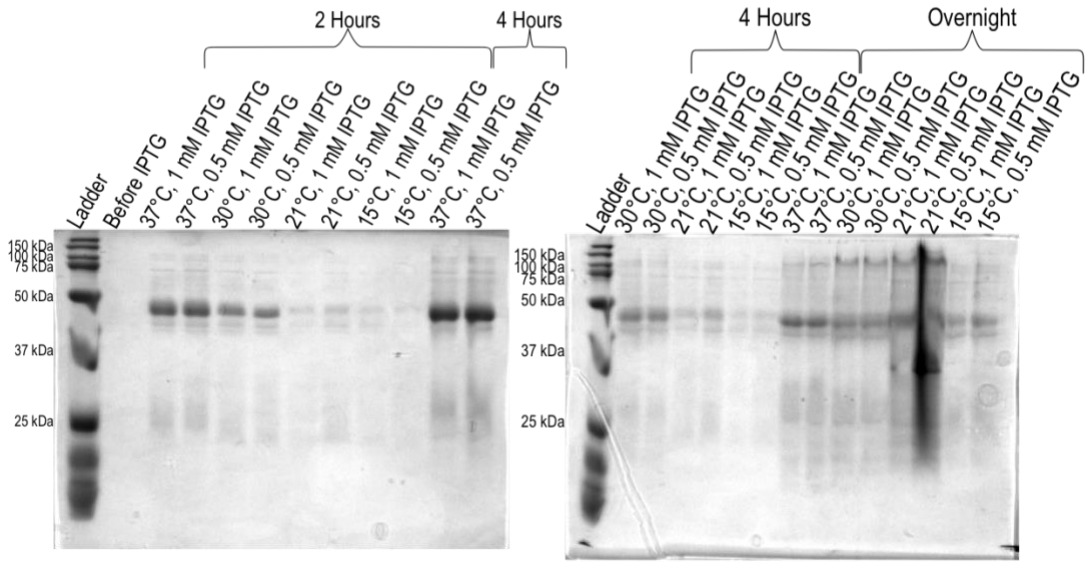


Figure 24: Expression test of 6xHis-Braf35. A BL21 (DE3) cell line containing the 6xHis-Braf35 construct at expression temperatures of 37 °C, 30 °C, 21 °C and 15 °C utilizing 2 different IPTG concentrations (1.0 mM, 0.5 mM). Time points were taken at 0, 1, 2, and 4 hours for the higher two temperatures and overnight for the lower two temperatures and then subjected to SDS-PAGE (homemade 12% acrylamide gel, 140V, 90 min). We noted one band at 21 °C induction with 0.5 mM IPTG that contained the expected MW band at 38 kDa

Additional optimization of pET-15b-Braf35 expression was performed for the 21°C sample with 0.25 and 0.5 mM IPTG at an OD_{600} of both 0.4 and 0.8, and collected time points at 0, and overnight. Subsequently, I ran an SDS-PAGE gel (Bio-Rad 4-20%, 120V, 65 min) and performed a WB analysis with a 6xHis antibody to assess the level of expression in these additional conditions (figure 25).

We detected 6xHis at the correct MW (38 kDa) across all four conditions, indicating the presence of protein. While all four conditions contained an apparent relatively similar amount of protein, we opted for the condition which had a slightly higher apparent level of protein, 0.5 mM IPTG at an OD_{600} of 0.4, for large scale expressions.

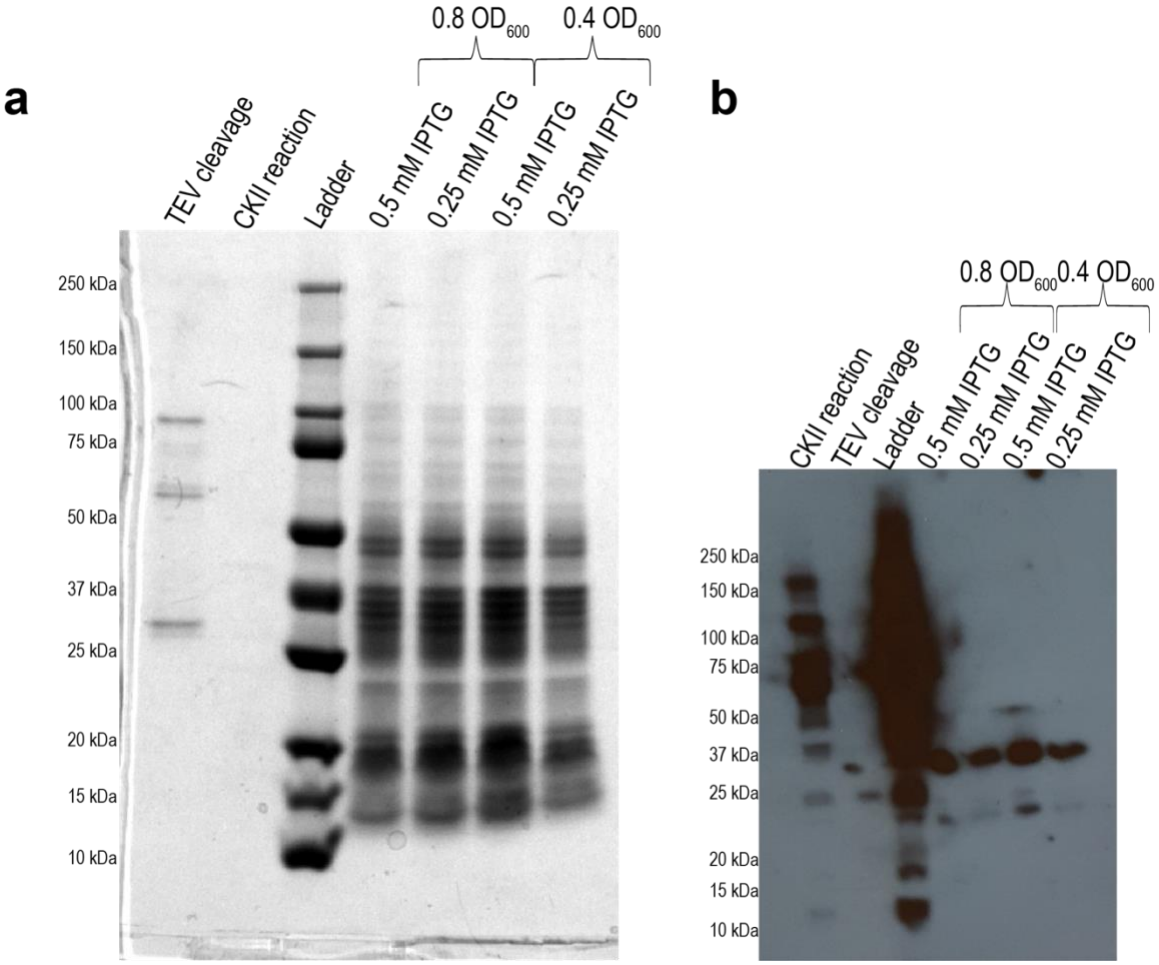


Figure 25: Expression test of 6xHis-Braf35. (a) SDS-PAGE (Bio-Rad 4-20%, 120V 65 min), and subsequent WB analysis utilizing an (b) α-6xHis (1:5000). I observed presence of 6xHis across all four expression conditions, the highest being at 0.5 mM IPTG at an OD_{600} of 0.4.

2.2.3 Purification for 6xHis-Braf35 (pET15-b)

Having established expression conditions in BL21(DE3) cells, I next moved to attempt purification. I envisioned a strategy wherein the cell lysate was run over a Ni-IDA chelating sepharose column. Depending on purity, with a predicted isoelectric point (pI) of 9.36, ion exchange chromatography was a potential second step. A backup

strategy would be an overnight TEV protease cleavage and a second Ni-IDA chelating sepharose column.

A thawed *E. coli* BL21 (DE3) containing pET-15b Braf35 plasmid pellet was resuspended in a 50 mM Tris-HCl buffer at pH 7.5 containing 350 mM NaCl, 10 mM imidazole, 5% glycerol, 1X protease inhibitor cocktail (Biotool 14001), lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626). The cells were then mechanically lysed in a high-pressure Avestin cell disruption system and the lysate was clarified by ultracentrifugation (42,000 *x g*, 4 °C, 45 min). The supernatant was then loaded onto a 25 mL IDA chelating sepharose column previously charged with nickel(II)sulfate at a flow rate of 2.0 mL/min via benchtop peristaltic pump. The column was washed extensively with a 50 mM Tris-HCl buffer system at pH 7.5 containing 350 mM NaCl, 10 mM imidazole, and 5% glycerol utilizing a GE Healthcare FPLC then eluted with an imidazole gradient (0 to 300 mM) over 15 CV (Figure 26a).

At this point, I wished to assess the eluted fractions from this column by SDS-PAGE (Bio-Rad 4-20%, 120V, 65 min) with Coomassie staining and WB analysis (1:5000, α-6xHis) (Figure 26a). A band in fractions 20-26 was present at the expected MW for Braf35, however, there was no protein visualized via WB analysis (Figure 26b).

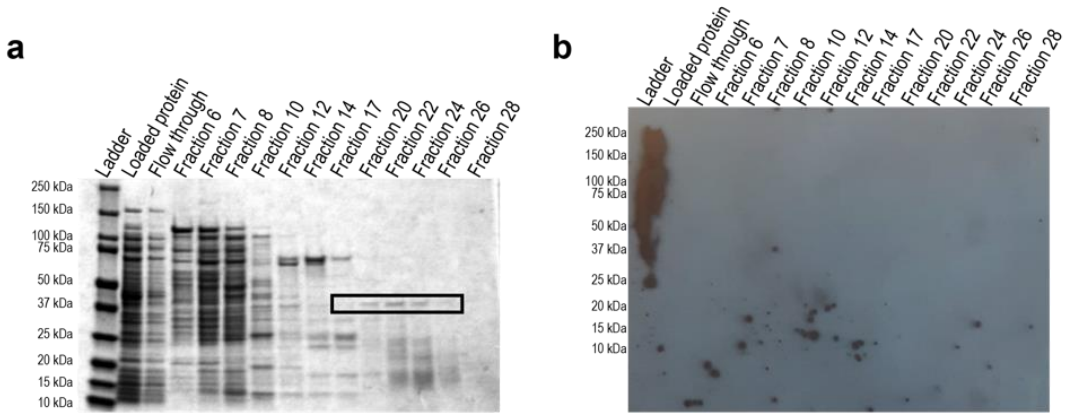


Figure 26: Initial purification attempt with pET15-b 6xHis-Braf35. (a) Utilizing a Ni-IDA chelating sepharose column (50 mM Tris-HCl, 350 mM NaCl, 5% glycerol, 10 mM imidazole) we eluted over a 15 CV gradient to 300 mM imidazole. We assessed via (a) SDS-PAGE (Bio-Rad 4-20%, 120V 65 min) and noted a band (black box) at ~38 kDa in fraction 20-26 via (b) WB analysis (100A, 30 min) for α-6xHis (1:5000). There was no observed 6xHis present across soluble protein or fractions from the elution.

2.2.4 Expression tests for pET28-b(+) 6xHisGST-Braf35

With a Braf35 containing pET28-b(+) vector in hand, we next moved to establish which conditions would result in expression of soluble protein. I decided to test across three induction temperatures (37 °C, 30 °C, 20 °C) at two different IPTG concentrations (0.75 mM and 0.25 mM).

I performed an electrochemical transformation to a BL21 (DE3) (Novagen) expression strain and inoculated an overnight starter flask (37 °C, 200 rpm). Upon reaching confluency, 100 mL flasks of LB (GeneseeScientific 11-125) media with kanamycin were inoculated to 1% v/v_t and incubated (37 °C, 200 rpm) to an OD_{600} of 0.4

and induced with IPTG (0.75 or 0.25 mM). Time points were taken at 0, 1, 2, and 4 hours for the 30 °C and 37 °C samples and overnight for samples at 20 °C.

The time point samples were via SDS-PAGE (Bio-Rad 4-20%, 120V, 65 min) with Coomassie staining (Figure 27). The expected molecular weight was 67 kDa, which we noted across all conditions (Figure 27a), particularly at the 2- and 4-hour time points for an IPTG concentration of 0.25 mM at 30 °C induction and both of the 20 °C overnight time points. I further assessed this utilizing WB (100V, 30 min transfer, 1:5000 α-6xHis) (Figure 27b). The band we noted via Coomassie stain directly below the 75 kDa ladder marker is consistent with the band which appears via WB, suggesting the presence of 6xHisGST-Braf35.

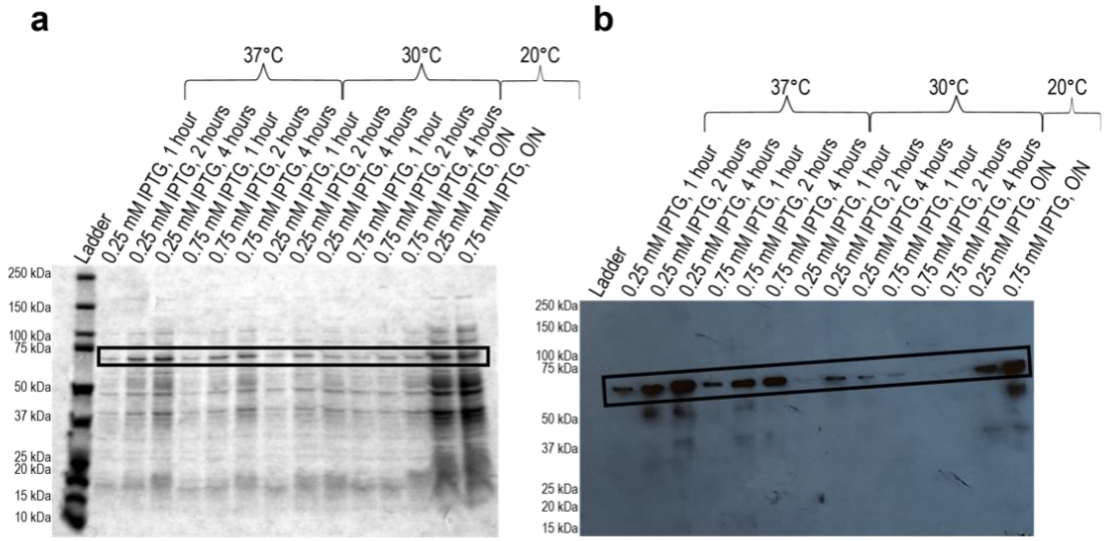


Figure 27: Expression test of 6xHisGST-Braf35. A BL21 (DE3) cell line containing the 6xHisGST-Braf construct was induced at 37 °C, 30 °C, 20 °C utilizing 2 different IPTG concentrations (0.75 mM, 0.25 mM). Time points were taken at 0, 1, 2, and 4 hours for the higher two temperatures and overnight for the lower temperature. and then subjected to SDS-PAGE (Bio-Rad 4-20%, 120V, 65 min) with subsequent (a) Coomassie staining and WB analysis (100A, 30 min). Visually, there was a consistent band across all conditions, WB analysis revealed the presence of a 6xHis tag at the appropriate MW, suggesting the presence of 6xHisGST-Braf35.

2.4 Expression and purification of HDAC1/2

Histone deacetylase 1 and 2 (HDAC1/2) are Class I metalloenzymes which likely utilize zinc to mediate hydrolysis of the acetyl-group from a lysine residue (Scheme 2). Similar to KDM1A, HDAC1/2 have been implicated in a series of pathophysiological functions, particularly throughout various cancer phenotypes, and there are currently several FDA approved HDAC inhibitors *(215,267–269)*.

HDAC1/2 are commonly found in complex with KDM1A and CoREST, interacting with CoREST through the ELM2/SANT domain, including in the 5-mer

complex we sought to express and purify in *E. coli (73)*. HDAC1/2 has been expressed and purified primarily via mammalian cell culture, though there are reports of utilization of Sf9 cells for obtaining a catalytically active protein *(262)*. To date there are no reports of a catalytically active HDAC1/2 expressed and purified in *E. coli*. The most likely reason for a catalytically inactive enzyme from *E. coli* is that there are no PTM mechanisms in *E. coli* and HDAC1/2 have been shown to require phosphorylation for activity *(218)* on an unstructured region on the C-terminus of the two proteins. Another member of the Class I HDAC family, HDAC8, has been shown to be oppositely affected by phosphorylation, and does not contain the unstructured tail region of HDAC1/2 (Figure 28). Interestingly, the Fierke lab has published a series of papers over the past decade detailing their efforts and success in expressing and purifying a catalytically active HDAC8 from *E. coli (220–222,270,271)*.

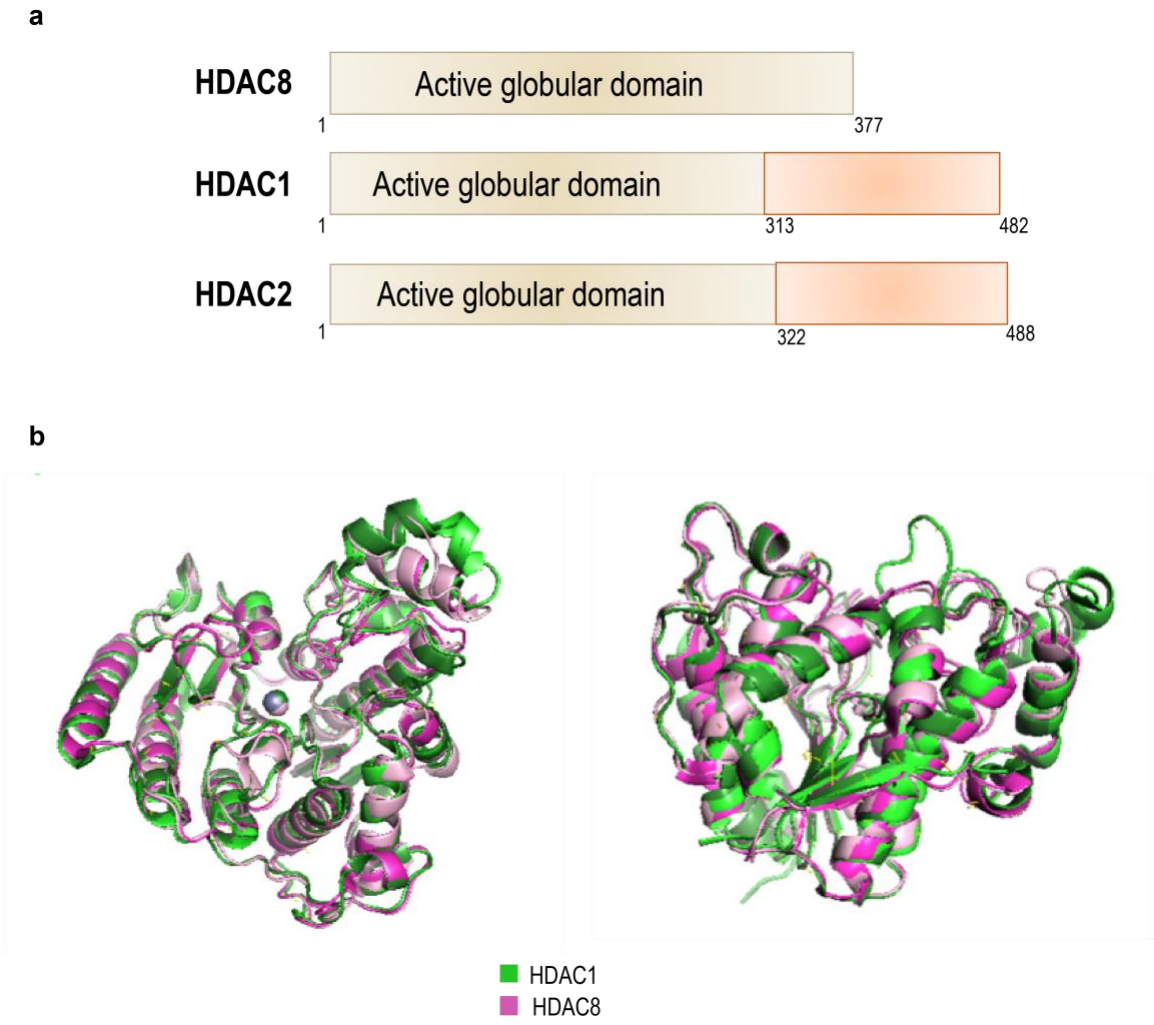


Figure 28: Overlay of the active globular domains of HDAC1 (green) and HDAC8 (pink). (a) both HDAC1/2 and HDAC8 share an active globular domain. While HDAC1/2 have an additional C-terminal tail, the (b) active-site containing globular domains very closely align between the proteins, suggesting a potential for soluble expression and purification utilizing a construct of HDAC1/2 without the C-terminal region

In order to access large quantities of active and pure HDAC1/2, I envisioned multiple routes towards gaining access to these proteins, each of which are detailed below. Firstly, I established a method of purification utilizing a 6xHis tagged protein construct. Secondly, I produced constructs which do not contain the phosphorylated C-

terminal tail of HDAC1/2 to mimic HDAC8. Finally, I produced both MBP and GST constructs and expression profiles to provide different handles for pull-down assays or purification methods should a metal-free method be necessary.

2.4.1 Cloning efforts towards HDAC1/2 constructs

Previous lab member Jonathan Burg had previously cloned HDAC1/2 into pET30-b(+) constructs with a thrombin cleavage site directly proceeding the protein sequence and preceding the C-terminal 6xHis tag. With these constructs in hand, we were able to readily utilize standard restriction digestion cloning protocol towards production of constructs with both MBP and GST tag systems. In each case, we utilized a previously utilized backbone of pET28-b(+) with either an N-terminal MBP or 6xHisGST tag followed by a TEV protease cleavage and then the protein sequence.

A restriction digest was performed on pET28-b(+) backbones (6xHis GST and MBP tag) and pET-30b constructs containing HDAC1/2 with restriction enzyme NdeI and XhoI (NEB, 37°C, 1 hr). The results of the digest were assessed with by gel (0.8% agarose, 80V, 30 min,1 kB ladder) and the backbones and inserts were subsequently excised from the gel and purified (Qiagen) (Figure 29) in preparation for ligation.

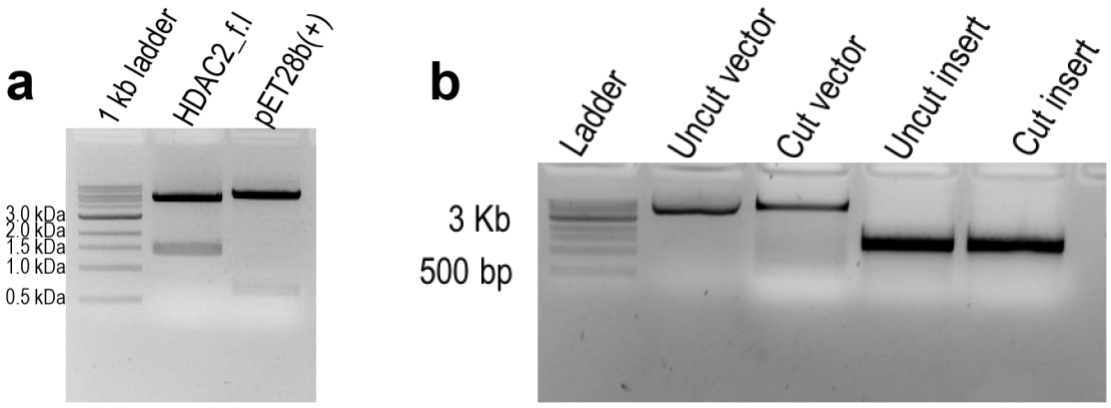


Figure 29: Agarose gel following restriction digest of two pET28-b(+) vectors containing a 6xHisGST tag (a) and an MBP tag (b) for HDAC2 (a) and HDAC1 (b). Digest was run over the course of 1 utilizing CutSmart buffer and 20 U of both XhoI and NdeI. A 1.2% agarose gel was run (80V, 30 min, 1kB ladder) of the cut pet28-b(+) vector and inserts from their original backbone vector (pET30-b).

The ligation reaction was performed at 3:1and 4:1 insert:vector molar ratios and incubated overnight (NEB T4K, 16 °C). The products were purified (EMD Millipore Pellet Paint) and resuspended in 16 μL nuclease-free water. We performed an electrochemical transformation with 5 μL of the purified ligated DNA and mixed them with an aliquot of DH10α cells. Directly following the electro-shock, we diluted the cells to 1 mL in super optimal broth with catabolite repression (SOC) and incubated them for 1 hour at 37 °C in a rotating incubator, allowing the freshly transformed cells recovery time before plating them on kanamycin plate overnight at 37 °C.

Single colonies were isolated and the plasmid DNA isolated (Zymo mini-prep) and sequence confirmed by EtonBiosciences for: pET28-b(MBP-HDAC1), pET28-b(MBP-HDAC2), pET28-b(GST-HDAC1), and pET-28b(GST-HDAC2).

2.4.1 Expression of an HDAC-6xHis pET30-b construct

While Dr. Burg had previously reported expression conditions, we sought to confirm these conditions in our hands as well as establish if in the incorporation of exogenous zinc would increase the expression of the protein.

Having confirmed the construct sequence, we next moved to expression of the protein in *E. coli* towards purification. The published protocol notes a 20 °C induction temperature, but not an IPTG concentration, so I attempted an initial expression test with a final IPTG concentration of 0.5 mM IPTG.

I performed an electrochemical transformation with both the pET30-b HDAC1 plasmid as well as a constitutively active GroESL plasmid to a BL21Star (Novagen) expression strain and inoculated an overnight starter flask (37 °C, 200 rpm). Upon reaching confluency, flasks of 2X-YT media with 50 μM kanamycin and 25 μM chloramphenicol were inoculated to 1% v/v_t and incubated (37 °C, 220 rpm) to an OD_{600} of 0.8 at which time the flasks were moved to 19 °C for 10 minutes and then induced to a final IPTG concentration of either 0.1 mM or 1.0 mM and grown overnight. Cells were subsequently pelleted (SLA-3000, 5000rpm, 4 °C, 15 min) and stored at -20 °C.

To test the solubility, we utilized BugBuster reagent per manufacturer's protocol and analyzed via SDS-PAGE (homemade 12% agarose gel, 140V, 80 min) with Coomassie staining (Figure 30). An overexpression band at roughly 57 kDa, the expected MW, was visualized in the induced and soluble protein lanes for the 1.0 mM IPTG

concentration. Interestingly, at 0.1 mM IPTG, the protein was still expressed, but appears insoluble.

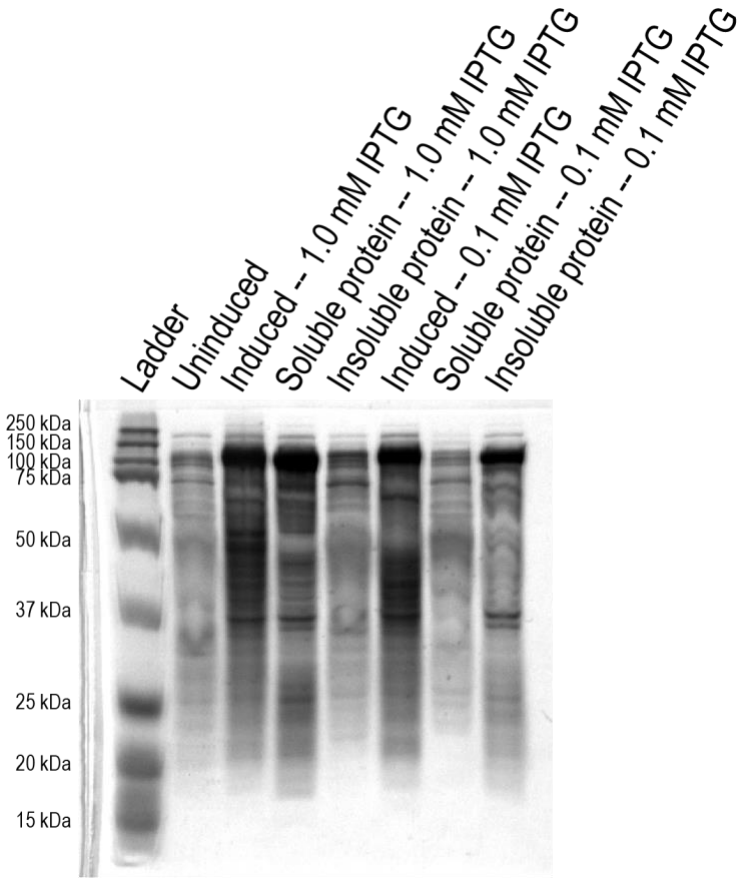


Figure 30: Test of solubility of HDAC1 at 0.1 mM and 1.0 mM IPTG in 2X-YT media. Cells were grown at 37 °C to an OD_{600} of 0.8 upon which they were induced at 19 °C and grown overnight. We noted over expression at the expected MW for 1.0 mM IPTG.

I next sought to assess the potential benefit of exogenous zinc in HDAC1 expression. As HDAC1/2 are metalloenzymes, incorporating additional amounts of zinc into the growth conditions will allow for less stress on the *E. coli* cells, enabling a potential higher level of expression to be obtained.

The expression test was performed as above at 1.0 mM IPTG only and with and without the incorporation of 100 μM $Zn(SO_4)$ at the inoculation of the induction cultures. There appeared to be no large difference in the plus and minus exogenous zinc samples suggesting that additional zinc was not required for expression; however, we opted to retain the usage of exogenous zinc in the event that it would aid solubility upon purification (Figure 31).

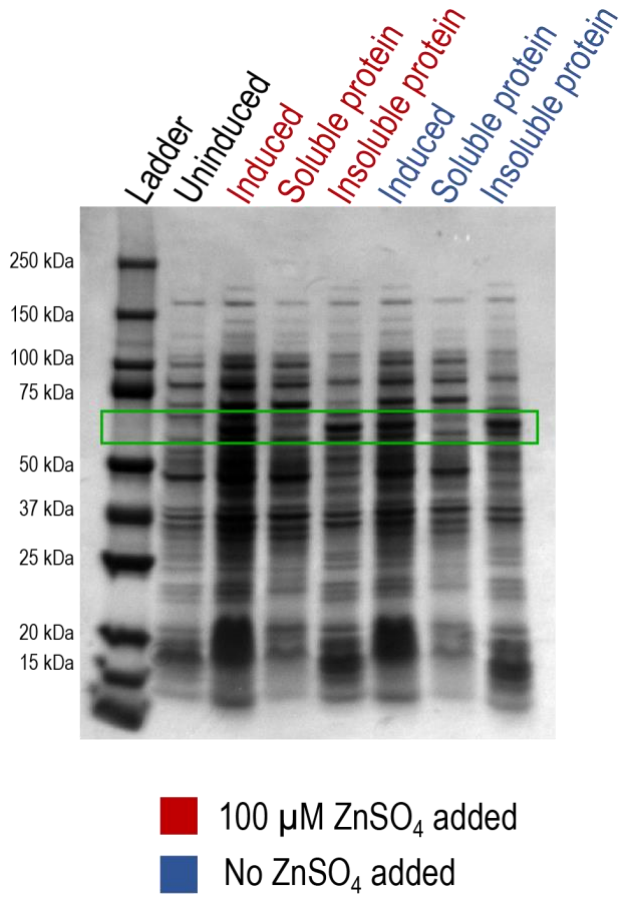


Figure 31: Assessing the importance of exogenous zinc on the expression of HDAC1. We noted no significant changes in protein expression with the introduction of exogenous zinc into the expression of HDAC1.

2.4.2 Purification of an 6xHisGST-HDAC pET28-b(+) construct

A thawed pellet of HDAC2 was resuspended in a 25 mM HEPES buffer system at pH 8.0 containing 300 mM NaCl, 10 mM imidazole, 10% glycerol, 1X protease inhibitor cocktail (Biotool 14001), lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626) and then mechanically lysed in a high-pressure cell disruption system. The resulting cell lysate was subjected to ultracentrifugation to remove insoluble material (42,000 *x g*, 4 °C, 45 min). The supernatant was then loaded onto a 25 mL IDA

chelating sepharose column previously charged with nickel at a flow rate of 2.0 mL/min and column was washed extensively with a 25 mM Tris-HCl buffer system at pH 8.0 containing 300 mM NaCl, 10 mM imidazole, and 10% glycerol. I performed a step 60 mM imidazole hold, a subsequent wash with 0.01% triton X-100, and then a gradient up to 300 mM imidazole. Fractions collected were analyzed by SDS-PAGE (Bio-Rad 4-20%, 120V, 65 min) with Coomassie staining (Figure 32). We noted a band across all of the fractions that corresponded to the correct MW of 57 kDa, suggesting a slow elution of the protein over the course of the imidazole gradient.

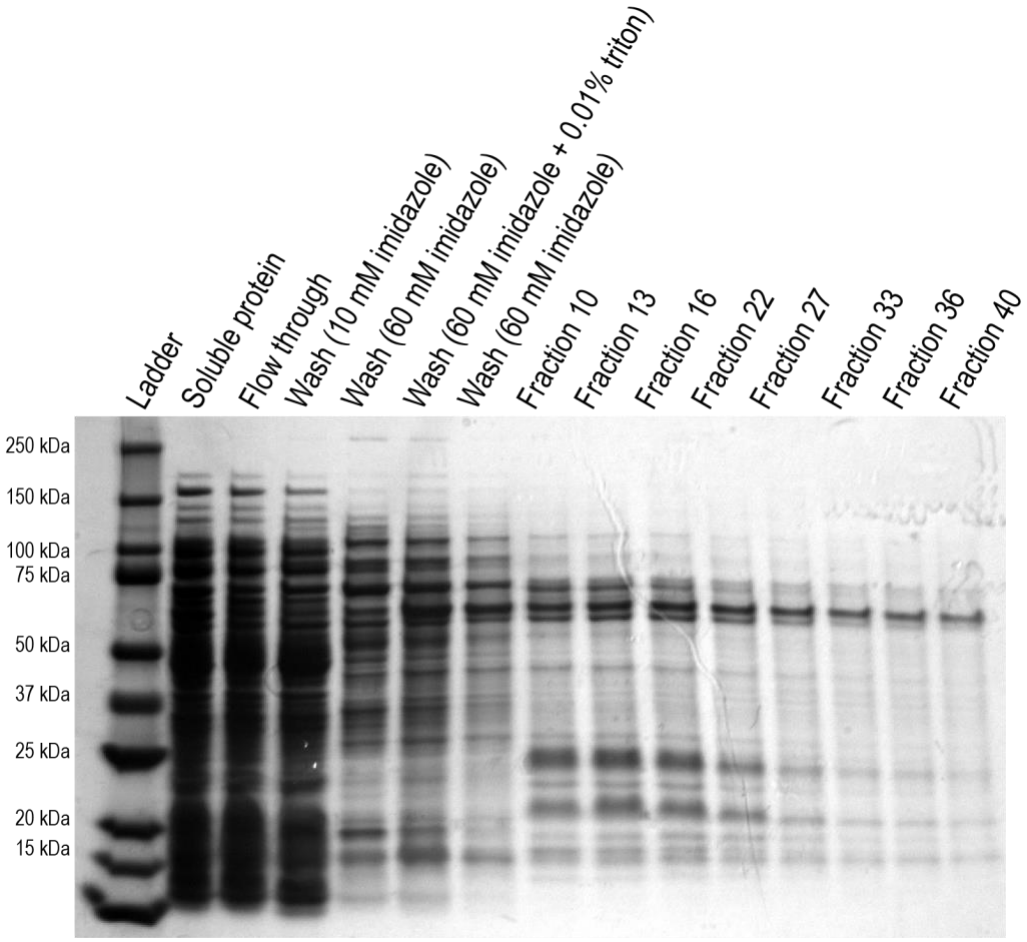


Figure 32: Initial purification attempt of HDAC2 utilizing nickel-immobilized affinity chromatography

Since HDAC2 has a predicted pI of 5.8, we purified HDAC2 additionally using anion Q sepharose IEC as a secondary purification step. The protein was dialyzed overnight into a 25 mM HEPES buffer at pH 7.0 containing 50 mM NaCl and 2 mM BME, loaded onto a 20 mL Q-sepharose column at a rate of 2 mL/min and then eluted over a gradient up to 1M NaCl. SDS-PAGE (Bio-Rad 4-20%, 120 V, 65 min) with Coomassie staining (Figure 33) was utilized to visualize this column purification step. We noted a band at the expected MW across fractions 20 to 24. They were combined and

concentrated via centricon concentrator (10 MWCO, 2000 *x g*, 4°C) to 1.3 mg/mL and was stored at -20°C in 40% glycerol.

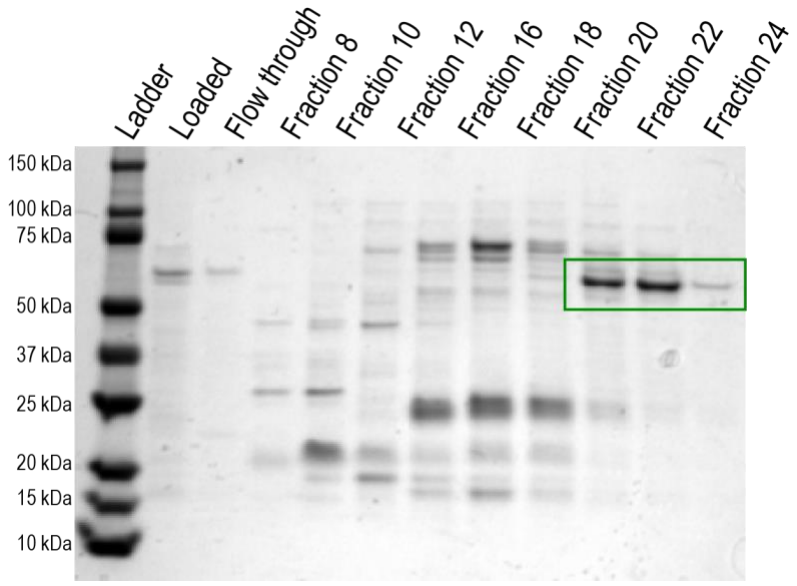


Figure 33: Utilization of IEC to further purify 6xHis-HDAC2. Utilizing a Q sepharose column, we were able to significantly separate the desired 57 kDa MW band from the contaminating proteins.

Additional optimization of the protocol was pursued. As zinc is likely the metal utilized by HDAC1/2, we wanted to assess utilization zinc-immobilized affinity chromatography step to avoid introduction of a non-native metal into the protein. We performed a cell lysis and initial purification step as above and analyzed the purification step via SDS-PAGE (Bio-Rad 4-20% 120 V, 65 min) with Coomassie staining (Figure 34).

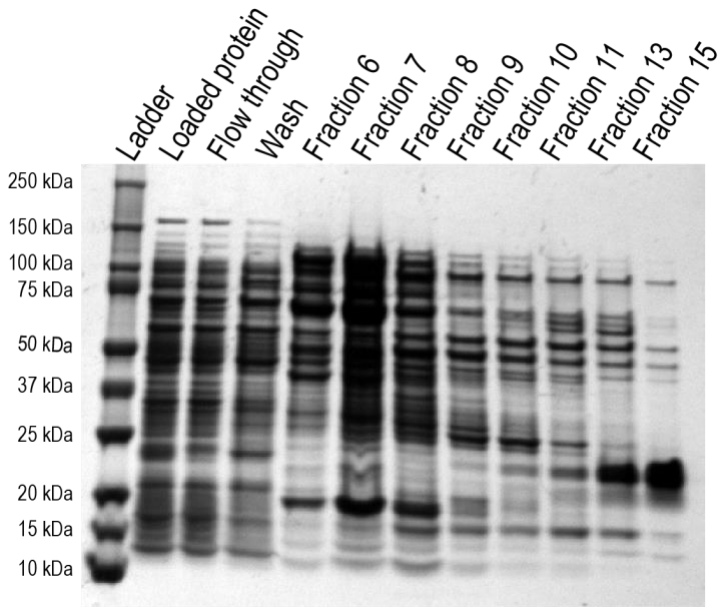


Figure 34: Purification profile utilizing zinc-immobilized affinity chromatography of 6xHis-HDAC.

This resulted in a significantly less purified protein after the initial purification step. I anticipate that a Q-sepharose column would be a successful second purification step, though this was not assessed. However, I have demonstrated that if it is found that nickel is being utilized by HDAC1/2 and it is not the native metal, it could be that this is a potentially viable initial purification step utilizing a HDAC with a 6xHis tag.

2.4.3 Purification of an MBP-HDAC protein

Having produced a construct for HDAC1/2 with an MBP tag, we wished to assess the feasibility of purification of HDAC1 utilizing this tag. As discussed previously, we have established an in-house method to purify MBP-tagged proteins in a cost-effective manner utilizing potato starch. I utilized this method in a batch elution to result in relatively pure HDAC1; however, there was significant loss of protein over

wash steps, suggesting that another method may be more amenable to purification to gain access to higher levels of protein.

A thawed pellet of *E. coli* BL21 (DE3) cells transformed with the pET-28b(+) MBP-HDAC1 plasmid was resuspended in a 25 mM HEPES buffer at pH 8.0 containing 200 mM NaCl, 1 mM EDTA, 1X protease inhibitor cocktail (Biotool 14001), lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626). The cells were mechanically lysed in a high-pressure Avestin cell disruption system and clarified via ultracentrifugation (40,000 *x g*, 4 °C, 45 min). The resulting supernatant was incubated with potato starch and then washed 3x via swinging bucket rotor (500 rpm, 15 min, 4 °C) and then eluted with the addition of 10 mM maltose. We assessed this purification method via SDS-PAGE (Bio-Rad 4-20%, 120 V, 65 min) with Coomassie staining (Figure 35). While the expected band (100 kDa) was obtained with relatively high purity, we noted that there was an apparent significant loss of protein across the initial binding and wash steps. This loss of protein suggests that we reached the binding capacity of the potato starch and a significant amount more would be required. It may also be that this protein construct is not able to bind as tightly to potato starch as we noted with KDM1A:CoREST, and so may not be the most optimal method for purifying HDAC1/2.

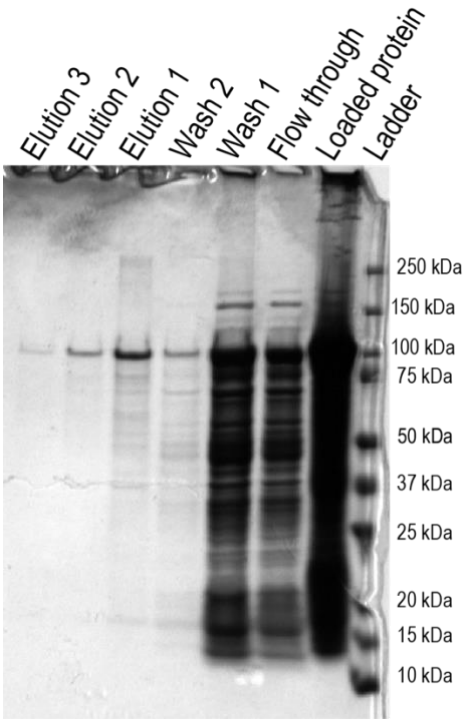


Figure 35: Purification of MBP-HDAC1 utilizing potato starch in a batch elution manner. While we obtained relatively pure protein at the correct MW (100 kDa), we noted significant loss of protein in the wash steps.

2.5 Expression and purification of MTA1$_{162\text{-}335}$

The majority of my efforts in protein expression and purification were towards the goal of forming the 5-mer complex KDM1A was initially reported in *(113)*; however, as described in Chapter 1, epigenetic enzymes can be coopted by multiple complexes within cellular environments to target different gene subsets. A core demethylase-scaffolding protein-deacetylase containing both KDM1A and HDAC1/2 is found within several complexes. In many of these cases, the scaffold protein is CoREST; however, in the nucleosome remodeling and histone deacetylase (NuRD) complex, the scaffold protein is metastasis tumor associated protein 1 (MTA1). MTA1 contains a bromo-

adjacent-homology (BAH) domain, a reader domain involved in nucleosomal binding, ELM2 and SANT domains in a similar layout to CoREST, as well as a zinc finger domain which has been shown to be involved in protein:protein interactions *(72,217,272–277)* (Figure 36a). Through the ELM2 and SANT domains, MTA1 has been demonstrated via X-ray crystallography to bind HDACs and has additionally been reported to bind KDM1A through an unestablished mechanism (Figure 36b) *(217)*.

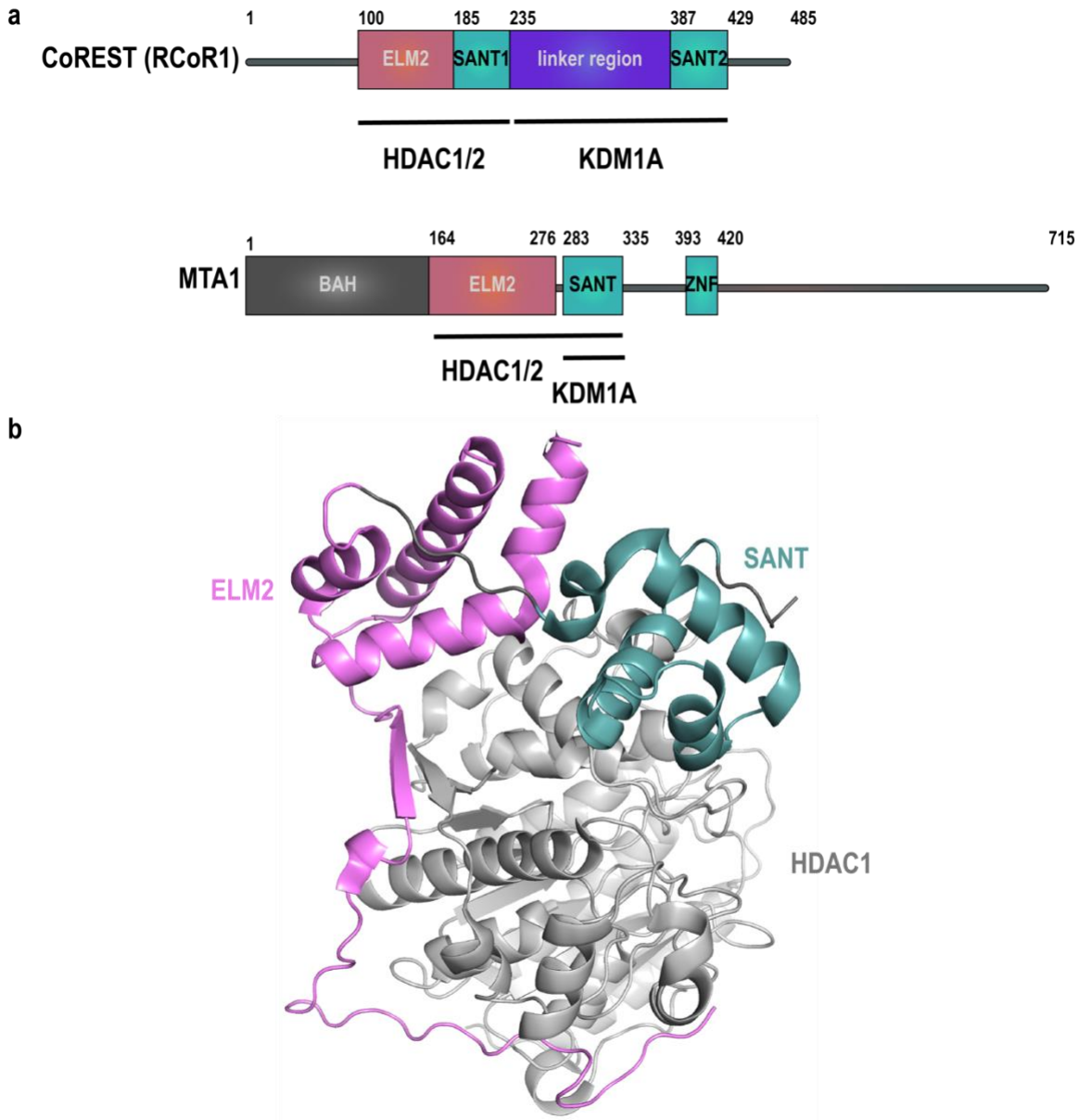


Figure 36: Domain map of MTA1 with reported binding sites of KDM1A and HDAC1/2 in comparison to CoREST. (a) MTA1 contains ELM2 (pink) and SANT (teal) domains similar to CoREST and additionally zinc finger and bromo-adjacent homology (BAH) domains. (b) A crystal structure exists of MTA1$_{165\text{-}340}$:HDAC1$_{8\text{-}376}$ depicting the mode of binding between the two proteins.

As both MTA1 and CoREST function as scaffold proteins for KDM1A and HDAC1/2, I was interested to explore the interaction between KDM1A and MTA1

further. In the interaction between CoREST:KDM1A, previous lab member Sunhee Hwang utilized ITC to demonstrate that the majority of the binding affinity originated from the linker region of CoREST to interact with KDM1A. The evidence that KDM1A binds MTA solely through the SANT domain suggests that MTA1 may employ a different binding mode than CoREST *(72)*. This is interesting because it suggests that we could potentially produce chemical biology tools to either stabilize or obstruct the interaction of one scaffold protein over another, directing KDM1A towards complexes *in vivo* to enable us to establish the role of KDM1A in specific complexes.

Based on previous reports to express and purify $MTA1_{162-335}$, I initially produced an N-terminal 6xHis-tagged $MTA1_{162-335}$ with a TEV protease cleavage site directly preceding the desired protein. In our hands, we were unable to obtain soluble protein based on the conditions reported and have now been able to express and purify an N-terminal 6xHis-GST $MTA1_{162-335}$ protein with a TEV protease cleavage site directly preceding the desired protein.

2.5.1 Cloning $MTA_{162-335}$ into a pET15-b and pET28-b(+) backbone

In order to produce a construct containing $MTA1_{162-335}$, I first obtained a pUC-57 plasmid containing the full-length MTA1 gene codon optimized for *E. coli.* The obtained a gBlock from IDT containing a codon optimized sequence for *E. coli* flanked by N-terminal NdeI and C-terminal XhoI cleavage sequences. Based on literature preference, I chose first to access the pET15-b construct utilizing a PCR amplification of the DNA

sequence corresponding to residues 162-335 followed by a standard restriction digest, ligation, and transformation. We chose PfuTurbo DNA polymerase (NEB) for our PCR, amplifying the DNA sequence corresponding to MTA1 residues 162-335 with NdeI and XhoI restriction digest sequences at each end of the sequence, respectively (Figure 37).

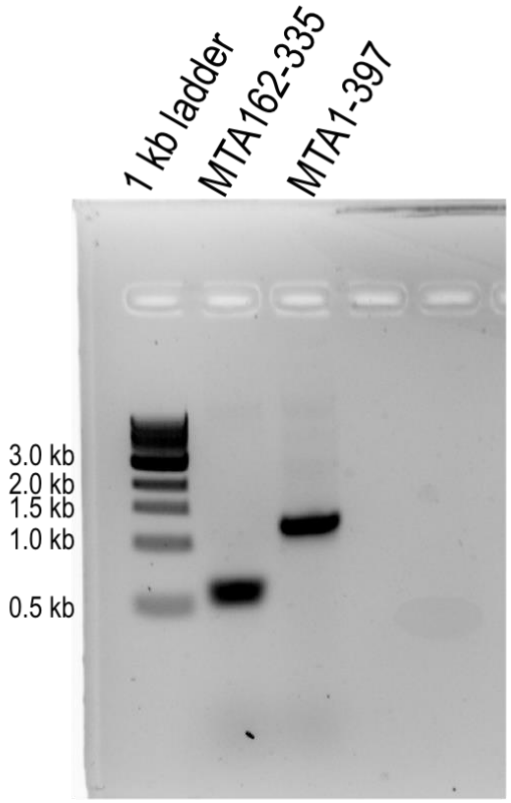


Figure 37: PCR amplification of a DNA fragment corresponding to MTA1 residues 162-335 flanked with restriction enzymes NdeI and Xho1. We utilized manufacturer's protocol for PfuTurbo DNA Polymerase PCR protocol at an annealing temperature of 62 °C

Following a PCR reaction clean up (Zymo), we incubated 1 μg of both pET15-b backbone and insert, separately, with CutSmart buffer (NEB), and 20 U of NdeI and

XhoI (NEB), dosing in an additional 20 U of each enzyme halfway through the digest (37 °C, 3 hrs.). Following completion of the digest, I ran an 0.8% agarose gel (80V, 30 min,1 kB ladder) to compare the cut backbone to the original vector and allow for separation between the backbone, and the insert that had been digested out of the pET15-b vector (Figure 38).

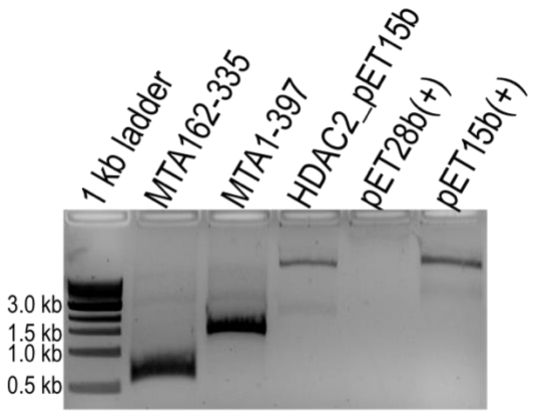


Figure 38: Agarose gel following restriction digest of a pET15-b vector and MTA1 PCR amplification. Digest was run over the course of 3 hours utilizing CutSmart buffer and 40 U of both XhoI and NdeI. A 1.2% agarose gel was run (100V, 45 min, 1kB ladder) of the cut pET15-b vector. In this lane, we noted a lower molecular weight band corresponding to the original sequence between the two restriction enzymes – KDM1A.

We noted two distinct bands in the lane where the vector had been cut, indicating the backbone and the sequence that had been digested from the original vector, corresponding to a downward shift of the cut vector from the uncut vector

control, both indicating a successful restriction digest. We excised the digested backbone and PCR fragment from the gel and the and performed a Qiagen gel extraction kit on both the backbone and the insert, as per protocol instructions.

For the ligation reaction, we measured the concentrations of the purified pieces and ran two ligation reactions (NEB) at 3:1 insert:vector molar ratio and incubated overnight at 16 °C. In order to remove excess enzyme and salts that could inhibit an electrochemical transformation, we performed a DNA purification utilizing EMD Millipore PelletPaint Co-Precipitant, as per manufacturer's protocol. We resuspended the pellet in 16 μL nuclease-free water. We performed an electrochemical transformation with 5 μL of the purified ligated DNA and mixed them with an aliquot of DH10β cells. Directly following the electro-shock, we diluted the cells to 1 mL in super optimal broth with catabolite repression (SOC) and incubated them for 1 hour at 37 °C in a rotating incubator, allowing the freshly transformed cells recovery time before plating them on an ampicillin plate overnight at 37 °C.

With the pET15-b vector in hand, we additionally cloned $MTA1_{162\text{-}335}$ into a pET28-b(+) backbone. The pET28-b(+) vector we utilized has a 6xHis-GST followed by a TEV protease cleavage sequence directly preceding the protein of interest and required the same restriction enzymes as pET15-b. Performing an additional restriction digest as above, we ran a 1.2% agarose gel (100V, 45 min,1 kB ladder) to separate the backbone and insert pet28-b(+) fragment (Figure 39).

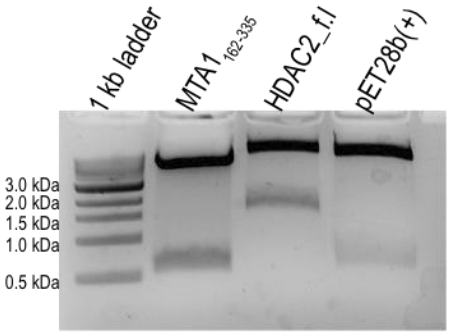


Figure 39: Agarose gel following restriction digest of a pET28-b(+) vector and pET15-b 6xHis-MTA1$_{162-335}$. Digest was run over the course of 3 hours utilizing CutSmart buffer and 40 U of both XhoI and NdeI. A 1.2% agarose gel was run (100V, 45 min, 1kB ladder) of the cut pET15-b vector. In this lane, we noted a lower molecular weight band corresponding to the original sequence between the two restriction enzymes – CoREST$_{286-493}$.

We noted two distinct bands in the lane where the vector had been cut, indicating the backbone and the sequence that had been digested from the original vector, corresponding to a downward shift of the cut vector from the uncut vector control, both indicating a successful restriction digest. We excised the digested backbone from the gel and performed a Qiagen gel extraction kit on both the backbone and the insert, as per protocol instructions.

For the ligation reaction, we measured the concentrations of the purified pieces and ran two ligation reactions (NEB) at 3:1and 6:1 insert:vector molar ratios and incubated overnight at 16 °C. In order to remove excess enzyme and salts that could inhibit an electrochemical transformation, we performed a DNA purification utilizing EMD Millipore PelletPaint Co-Precipitant, as per manufacturer's protocol. We

resuspended the pellet in 16 μL nuclease-free water. We performed an electrochemical transformation with 5 μL of the purified ligated DNA and mixed them with an aliquot of DH10β cells. Directly following the electro-shock, we diluted the cells to 1 mL in super optimal broth with catabolite repression (SOC) and incubated them for 1 hour at 37 °C in a rotating incubator, allowing the freshly transformed cells recovery time before plating them on a kanamycin plate overnight at 37 °C.

Upon obtaining both vectors with confirmed sequencing from Eton Biosciences, we proceeded to expression tests for both vectors to determine the best method to access a purified $MTA1_{162-335}$ from *E. coli*.

2.5.2 Expression of pET15-b 6xHis-$MTA1_{162-335}$

Having confirmed the construct sequence, we next moved to expression of the protein in *E. coli* towards purification. The published protocol notes a 20 °C induction temperature, but not an IPTG concentration, so I attempted an initial expression test with a final IPTG concentration of 0.5 mM IPTG.

I performed an electrochemical transformation to a BL21Star (DE3) (Novagen) expression strain and inoculated an overnight starter flask (37 °C, 200 rpm). Upon reaching confluency, 100 mL flasks of LB (GeneseeScientific 11-125) media with ampicillin and 1 mM $MgCl_2$ were inoculated to 1% v/v_t and incubated (37 °C, 220 rpm) to an OD_{600} of 0.4 at which time the flasks were moved to 21 °C for 10 minutes and then induced to a final IPTG concentration of 0.5 mM with an overnight timepoint taken,

pelleted in a tabletop centrifuge (8000 rpm, 4 °C, 2 min), supernatant removed, and the pellet stored at -20 °C until use to test expression and solubility.

To test solubility, we utilized BugBuster reagent per manufacturer's protocol and analyzed via SDS-PAGE (Bio-Rad 4-20%, 120 V, 65 min) with subsequent Coomassie staining (Figure 40). An overexpression band at roughly 21 kDa, the expected MW, was visualized in the induced and insoluble protein lanes. The BugBuster solubility test is harsh and may not be indicative of solubility, so I next opted to expand my expression conditions in an attempt to obtain soluble protein via BugBuster.

In order to assess whether I could obtain soluble protein, I chose three additional IPTG concentrations at a 21 °C induction temperature (0.1 mM, 0.25 mM, and 1.0 mM) as well as four IPTG concentrations at a lower 15 °C induction temperature (0.1 mM, 0.25 mM, 0.5 mM, and 1.0 mM). I inoculated an overnight flask from a BL21Star (DE3) (Novagen) glycerol stock and grew the cells overnight (37 °C, 200 rpm).). Upon reaching confluency, 100 mL flasks of TB media with ampicillin were inoculated to 1% v/v_t and incubated (37 °C, 200 rpm) to an OD_{600} of 0.4 at which time the flasks were moved to either 21 °C or 15 °C for 10 minutes and then induced to the appropriate final IPTG. An overnight timepoint was taken, pelleted in a tabletop centrifuge (8000 rpm, 4 °C, 2 min), supernatant removed, and the pellet stored at -20 °C until use to test expression and solubility.

To test solubility, we utilized BugBuster reagent per manufacturer's protocol and analyzed via SDS-PAGE (Bio-Rad 4-20%, 120 V, 65 min) with subsequent Coomassie staining (Figure 40). Again, we noted overexpression at the correct MW, however, a majority of the protein at the correct MW across all conditions appeared, widely insoluble.

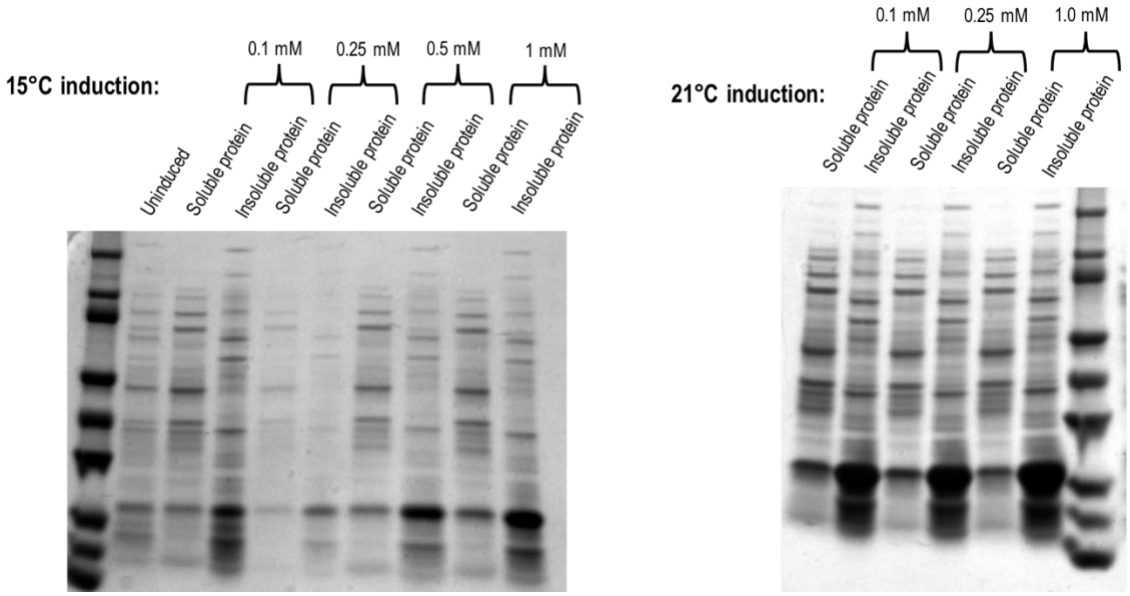


Figure 40: Expanded expression test of 6xHis-$_{MTA1162-335}$ from a 15 °C and 21 °C induction with a final IPTG concentration from 0.1 mM to 1.0 mMOD$_{600}$ of 0.4. Test expressions were subjected to a BugBuster solubility test. Expected MW of the protein was 21 kDa, which was visualized. We noted some apparent solubility of protein at the expected MW.

In parallel to the expanded expression test, I also sought to assess the utilization of two different expression cell lines. Historically, we have utilized BL21(DE3) strains, however, I wished to assess if a slightly different cell line, Rosetta, would produce more soluble protein. The published protocol utilized a Rosetta cell line, which incorporates a

pRARE plasmid which contains the tRNA that are highly utilized in *H. sapiens* but not *E. coli* to supplement the difference between codon usage. The two strains should be extremely similar otherwise, and as I had codon-optimized the MTA1 sequence for *E. coli*, I hypothesized that this would enable usage of a BL21 cell line. I performed an expression test as above for a final IPTG concentration of 0.5 mM at an OD_{600} of 0.4 and at 15 °C and 21 °C with BL21(DE3) and Rosetta(DE3) PlysS cell lines (Novagen). Time points were taken at 0, 2, 4 hours in addition to overnight. Samples were pelleted via tabletop centrifuge (8000 rpm, 4 °C, 2 min), the supernatant removed, and samples stored at -20 °C until utilization for BugBuster test. The results of the test were visualized via SDS-PAGE (Bio-Rad 4-20%, 120 V, 65 min) with Coomassie stain (Figure 41). Again, we noted an overexpression band, but no appreciable quantity of soluble protein for either strain.

The BugBuster test can be harsh and provide a false negative result for the solubility of proteins. Thus, even with the negative results in hand I was encouraged by the published report of 6xHis-$MTA1_{162-335}$ *E. coli* expression and purification and chose to proceed to attempt purification utilizing this construct (21 °C induction, 0.5 mM IPTG, OD_{600} 0.4, overnight).

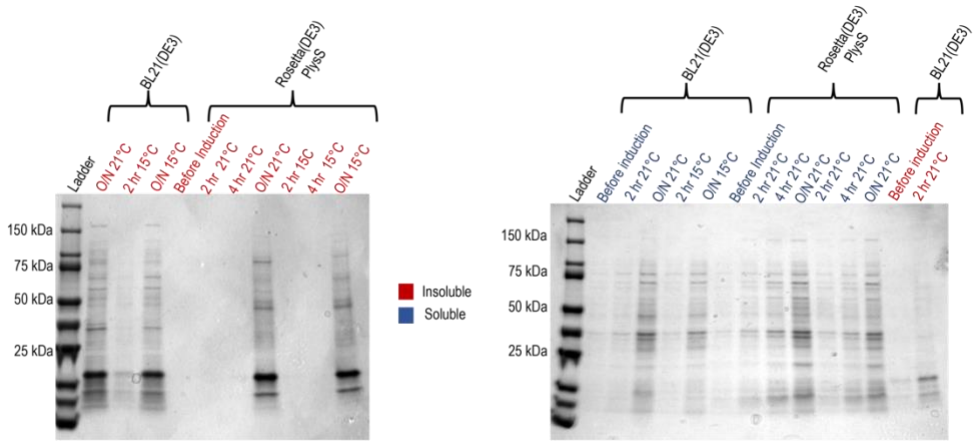


Figure 41: Expression test of 6xHis-MTA1$_{162-335}$ with two different cell lines (BL21(DE3) and Rosetta(DE3)PlysS. We noted similar expression patterns between both cell lines, with no noticeable soluble protein in either strain.

2.5.3 Purification of pET15-b 6xHis-MTA1$_{162-335}$

I wanted to utilize the purification strategy previously reported to assess if 6xHis-MTA1$_{162-335}$ would be soluble during a protein purification. The proposed strategy utilized nickel-immobilized affinity chromatography and then size exclusion chromatography to gain access to purified protein.

A thawed pellet was resuspended in a 50 mM Tris-HCl buffer system at pH 8.0 containing 350 mM NaCl, 10 mM imidazole, 5% glycerol, 1X protease inhibitor cocktail (Biotool 14001), lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626) and then mechanically lysed in a high-pressure cell disruption system. The resulting cell lysate was subjected to ultracentrifugation to remove insoluble material (42,000 *x g*, 4 °C, 45 min). The supernatant was then loaded onto a 25 mL IDA chelating sepharose column previously charged with nickel at a flow rate of 2.0 mL/min via benchtop

peristaltic pump. After completion of loading, the column was washed extensively with a 50 mM Tris-HCl buffer system at pH 8.0 containing 350 mM NaCl, 10 mM imidazole, and 5% glycerol utilizing a GE Healthcare fast protein liquid chromatography (FPLC) ÄKTA purification system (ÄKTA). The reported procedure method contained a step to 60 mM imidazole hold and then a gradient up to 200 mM imidazole. I performed a gradient and then hold at 60 mM imidazole and then a gradient up to 200 mM imidazole. At this point, I wished to assess the eluted fractions from this column by SDS-PAGE (Bio-Rad 4-20%, 120V, 65 min) with Coomassie staining (Figure 42). While there was a fairly pure band at fractions 20 and 23, corresponding to the correct MW of $MTA1_{162-335}$, there appeared to be very little protein present in the fractions, consistent with our previous tests demonstrating that 6xHis-$MTA1_{162-335}$ is not soluble in our hands.

For many of the purposes we envisioned utilizing purified $MTA1_{162-335}$ for we would require a significant level of protein. We had a construct of $MTA1_{162-335}$ with a GST tag, utilized historically for increasing solubility, so we opted to attempt a purification of the GST-containing $MTA1_{162-335}$ construct *(278)*.

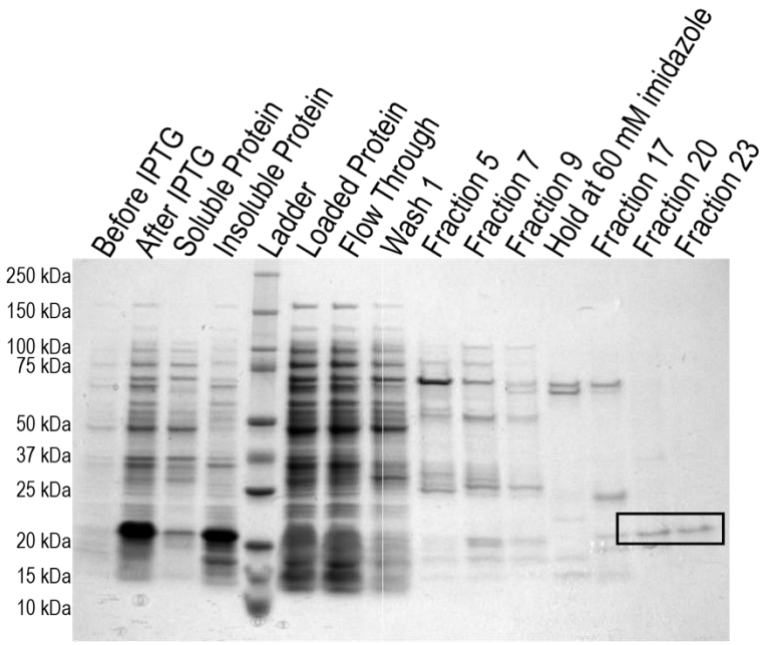


Figure 42: Testing solubility of MTA1$_{162-335}$ with reported buffer conditions and nickel-immobilized affinity chromatography. We utilized an imidazole gradient with a hold at 60 mM imidazole and then gradient to 200 mM imidazole. While there was a band (black box) consistent with the overexpression band we note in expression tests (left side of gel), the level of protein was relatively low for our purposes

2.5.4 Expression of pET28-b(+) 6xHisGST-MTA1$_{162-335}$

With a Braf35 containing pET15-b vector in hand, we next moved to establish which conditions would result in expression of soluble protein. I decided to test across three induction temperatures (37 °C, 30 °C and 21 °C) at two different IPTG concentrations (1.0 mM and 0.25 mM).

I performed an electrochemical transformation to a BL21 (DE3) (Novagen) expression strain and inoculated an overnight starter flask (37 °C, 200 rpm). Upon reaching confluency, 100 mL flasks of LB (GeneseeScientific 11-125) media with ampicillin were inoculated to 1% v/v$_t$ and incubated (37 °C, 200 rpm) to an OD$_{600}$ of 0.45

at which time the flasks were moved to their induction temperature and induced with IPTG to a final concentration of either 0.25 or 1.0 mM. Time points were taken at 0, 1, 2, and 4 hours for the 30 °C and 37 °C samples whereas I only took a time point for the overnight growth samples at 21 °C due to the slowed growth of the two lower temperature samples.

The time point samples were spun down in a tabletop centrifuge (8000 rpm, 4 °C, 2 min), supernatant removed, and the pellet stored at -20 °C until use. I subsequently ran SDS-PAGE (Bio-Rad 4-20%, 120 V, 65 min) with Coomassie staining (Figure 43a,c) in addition to utilizing WB analysis for the presence of GST (Figure 43b,d). By WB analysis, there was a strong expression across all conditions at the expected MW (51 kDa); however, there also were apparent truncation products at most of the conditions tested with the most protein present 4 hours post-30 °C induction temperature at a final IPTG concentration of 0.25 mM.

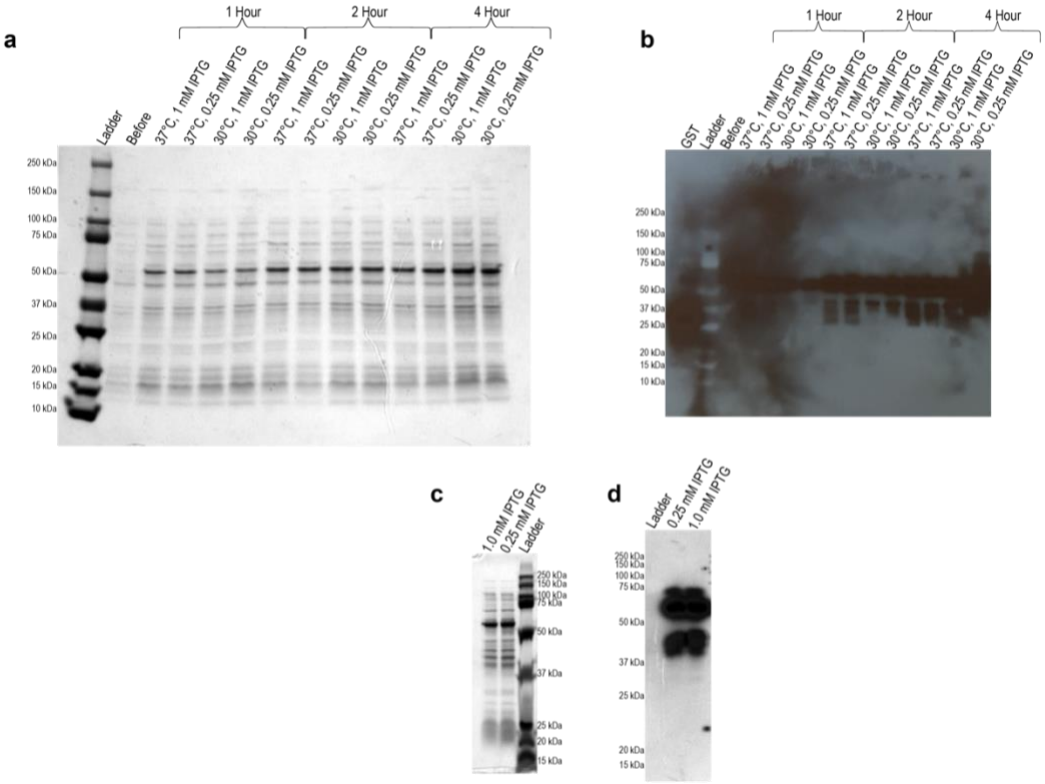


Figure 43: Expression test of 6xHisGST-MTA1$_{162-335}$. A BL21 (DE3) cell line containing the 6xHisGST-MTA1$_{162-335}$ construct was induced at 37 °C, 30 °C, and 21 °C utilizing 2 different IPTG concentrations (1.0 mM, 0.25 mM). Time points were taken at 0, 1, 2, and 4 hours for the higher two temperatures and overnight for the lower temperature. and then subjected to SDS-PAGE (Bio-Rad 4-20%, 120V, 65 min) with subsequent (a,c) Coomassie staining and (b,d) WB analysis (100A, 30 min). Visually, there was a consistent band across all conditions, WB analysis revealed the presence of a 6xHis tag at the appropriate MW (51 kDa), suggesting the presence of 6xHisGST-MTA1$_{162-335}$.

2.5.5 Purification of pET28-b(+) 6xHisGST-MTA1$_{162-335}$

Having established expression conditions for this construct, I moved forward with purification of this construct. I envisioned being able to purify this protein both with and without the 6xHisGST tag system. While obtaining cleaved protein is important for structural and biophysical techniques, retaining a handle is beneficial for

initial pull-down binding assays with MTA1$_{162-335}$. To obtain clean uncleaved protein, I hypothesized an initial glutathione agarose column as an initial stepped followed by IEC or SEC as a fine-tuning purification step

A thawed pellet was resuspended in PBS with 1X protease inhibitor cocktail (Biotool 14001), lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626) and then mechanically lysed in a high-pressure cell disruption system. The resulting cell lysate was subjected to ultracentrifugation to remove insoluble material (43,000 *x g*, 4 °C, 45 min). The supernatant was then loaded onto a 10 mL glutathione agarose column (Genesee Scientific) at a flow rate of 1.0 mL/min via benchtop peristaltic pump. After completion of loading, the column was washed extensively with a 50 mM Tris-HCl buffer system at pH 8.0 containing 350 mM NaCl, 10 mM imidazole, and 5% glycerol utilizing a GE Healthcare fast protein liquid chromatography (FPLC) ÄKTA purification system (ÄKTA) at 1 mL/min. The protein was eluted with 10 mM reduced glutathione (Sigma) in 50 mM Tris HCl (pH 8.0) with 50 mM NaCl. Eluted fractions from this column were assessed by SDS-PAGE (Bio-Rad 4-20%, 120V, 65 min) with Coomassie staining (Figure 44). The bands we had previously noted to likely be truncation products were prominent in the elution, indicating they are indeed truncation products as they readily bound to the glutathione column. However, we did not note a large soluble fraction at the expected 51 kDa. This suggested to us that either this construct was still

not soluble even with a GST-tag, or that a different buffer with more stabilizing components, such as glycerol would be required.

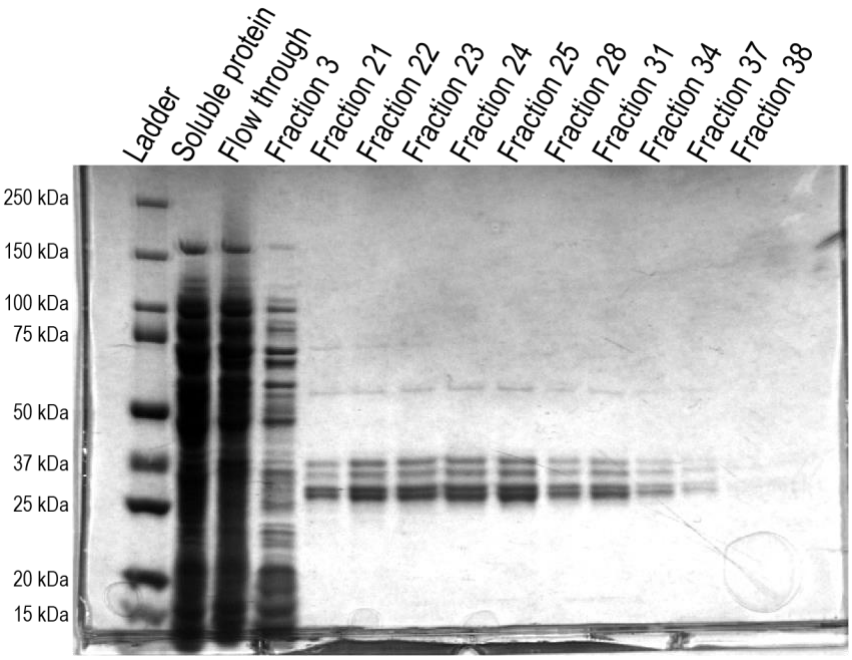


Figure 44: Initial purification attempt of 6xHisGST-MTA1$_{162-335}$ utilizing a glutathione agarose column. Protein was lysed in a PBS buffer and attempted to elute with 10 mM reduced glutathione in 50 mM Tris-HCl (pH 8.0), 50 mM NaCl, 1 mM BME. While the truncation products previously noted in the expression tests were present and bound the column, we noted little solubility of the band at 51 kDa.

I hypothesized that utilizing a buffer containing a higher ionic strength as well as another stabilizing component, glycerol, would help in the potential solubility of the protein. A thawed pellet was resuspended in a 50 mM Tris-HCl buffer system at pH 8.0 containing 350 mM NaCl, 10 mM imidazole, 5% glycerol, 1X protease inhibitor cocktail (Biotool 14001), lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626) and then mechanically lysed in a high-pressure cell disruption system. The resulting cell lysate was subjected to ultracentrifugation to remove insoluble material (45,000 *x g*, 4 °C,

45 min). The supernatant was then loaded onto a 25 mL IDA chelating sepharose column previously charged with nickel at a flow rate of 2.0 mL/min via benchtop peristaltic pump. After completion of loading, the column was washed extensively with a 50 mM Tris-HCl buffer system at pH 8.0 containing 350 mM NaCl, 10 mM imidazole, and 5% glycerol utilizing a GE Healthcare fast protein liquid chromatography (FPLC) ÄKTA purification system (ÄKTA). An imidazole gradient was then implemented up to 300 mM imidazole over a 15 CV to elute bound protein from the column.

At this point, I wished to assess the eluted fractions from this column not only by SDS-PAGE (homemade 12% acrylamide gel, 140 V, 80 min) with Coomassie staining (Figure 45) and noted a strong band at 51 kDa, corresponding to 6xHisGST-MTA1$_{162-335}$. The truncation products were still present and directly aligned with elution of the band at 51 kDa. I next sought to separate the truncation products. We utilized several cell pellets, each with the same first step of purification, to focus on optimization of the second step of this purification, which are outlined below. In summary, we utilized IEC, glutathione agarose, and SEC in attempt to separate the truncation products.

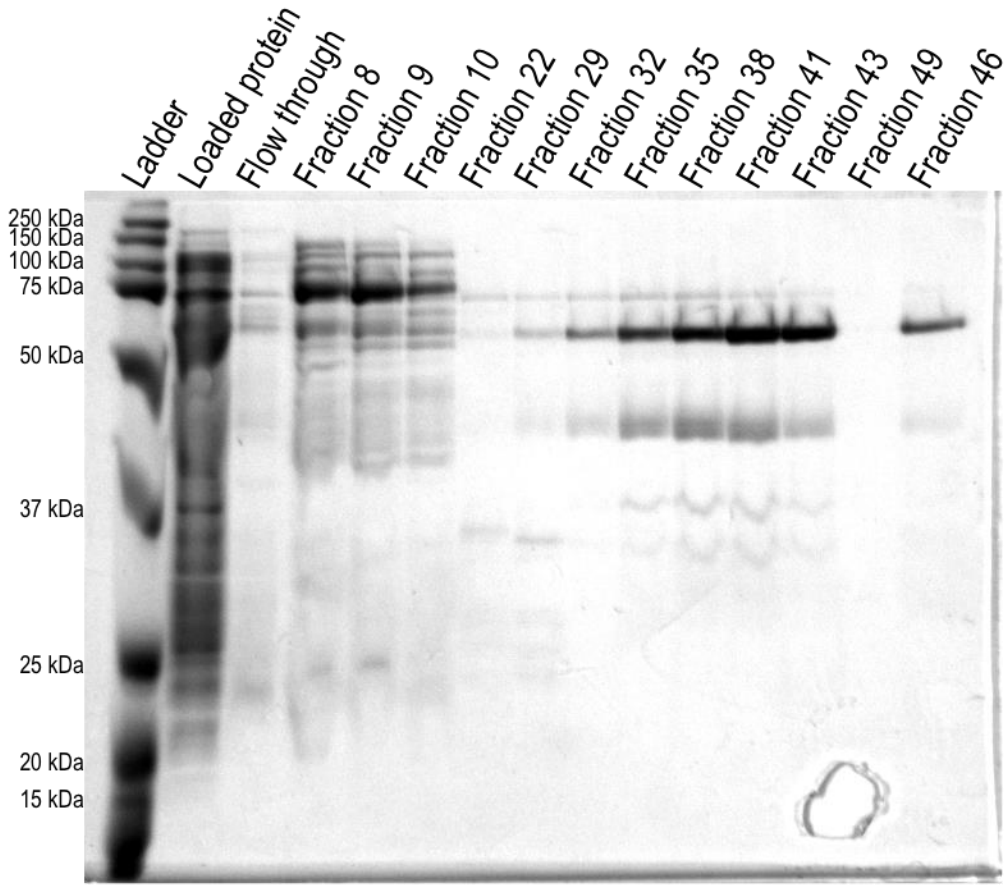


Figure 45: Purification of MTA1$_{162-335}$ via an initial nickel-immobilized affinity chromatography

Following the first purification step, we next wanted to assess binding of the 51 kDa band to glutathione agarose. We performed an overnight dialysis into a PBS buffer with 1 mM BME of pooled fractions containing 6xHisGST-MTA$_{162-335}$. The fractions were loaded onto an equilibrated 10 mL glutathione agarose column. Utilizing a GE Healthcare fast protein liquid chromatography (FPLC) ÄKTA purification system (ÄKTA), the column was washed and then protein eluted with 10 mM reduced glutathione in PBS with 1 mM BME. SDS-PAGE (12% Bio-Rad gel, 120 V, 80 min) with

Coomassie staining (Figure 46) was utilized to visualize this column purification step. Indicative of being the 6xHisGST-MTA1$_{162-335}$ protein at 51 kDa, complete binding of this band to the glutathione agarose column and elution upon introduction of reduced glutathione was visualized. While a higher MW band at roughly 75 kDa was able to be separated with this column step, as anticipated, we saw little separation between the full-length and truncation products.

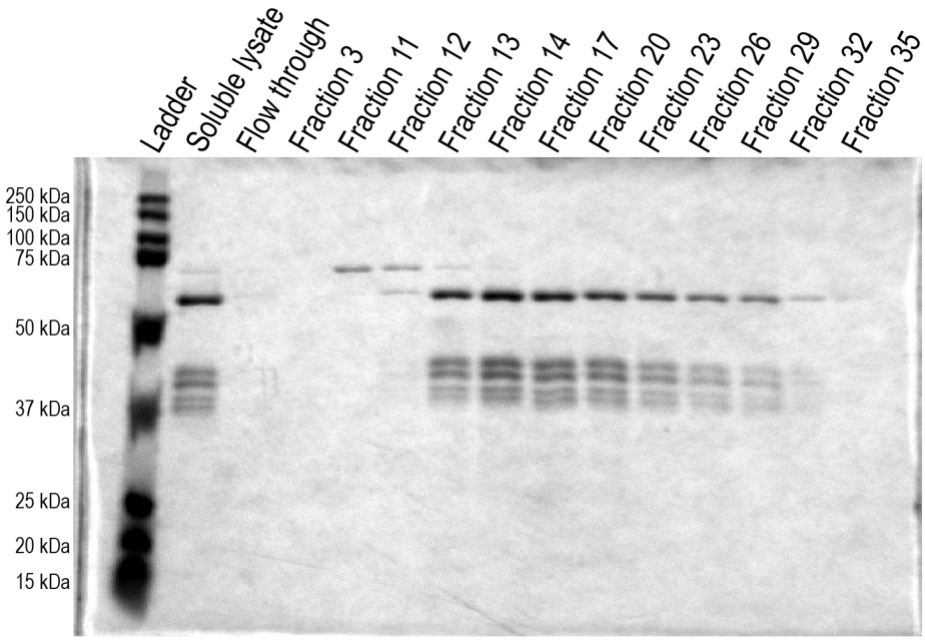


Figure 46: Utilization of a glutathione agarose column as a second step of purification for MTA1$_{162-335}$. While a higher MW band at roughly 75 kDa was able to be separated utilizing this column, there was little separation, as anticipated, between the full-length band at 51 kDa and the truncation products.

Following this, due to a predicted isoelectric point of 5.75, I next wanted to assess binding and separation utilizing a strong IEC resin, Q-sepharose. Protein was dialyzed overnight into a 25 mM HEPES buffer at pH 7.0 containing 50 mM NaCl and 10 mM

BME, loaded onto a 20 mL Q-sepharose column at a rate of 2 mL/min. Utilizing a GE Healthcare fast protein liquid chromatography (FPLC) ÄKTA purification system (ÄKTA), the column was washed and then eluted over a gradient up to 1M NaCl. SDS-PAGE (12% Bio-Rad gel, 120 V, 80 min) with Coomassie staining (Figure 47) was utilized to visualize this column purification step. Similar to the glutathione agarose column, we saw separation between the desired 51 kDa MW band and the upper 75 kDa MW band but no apparent separation with the truncation products.

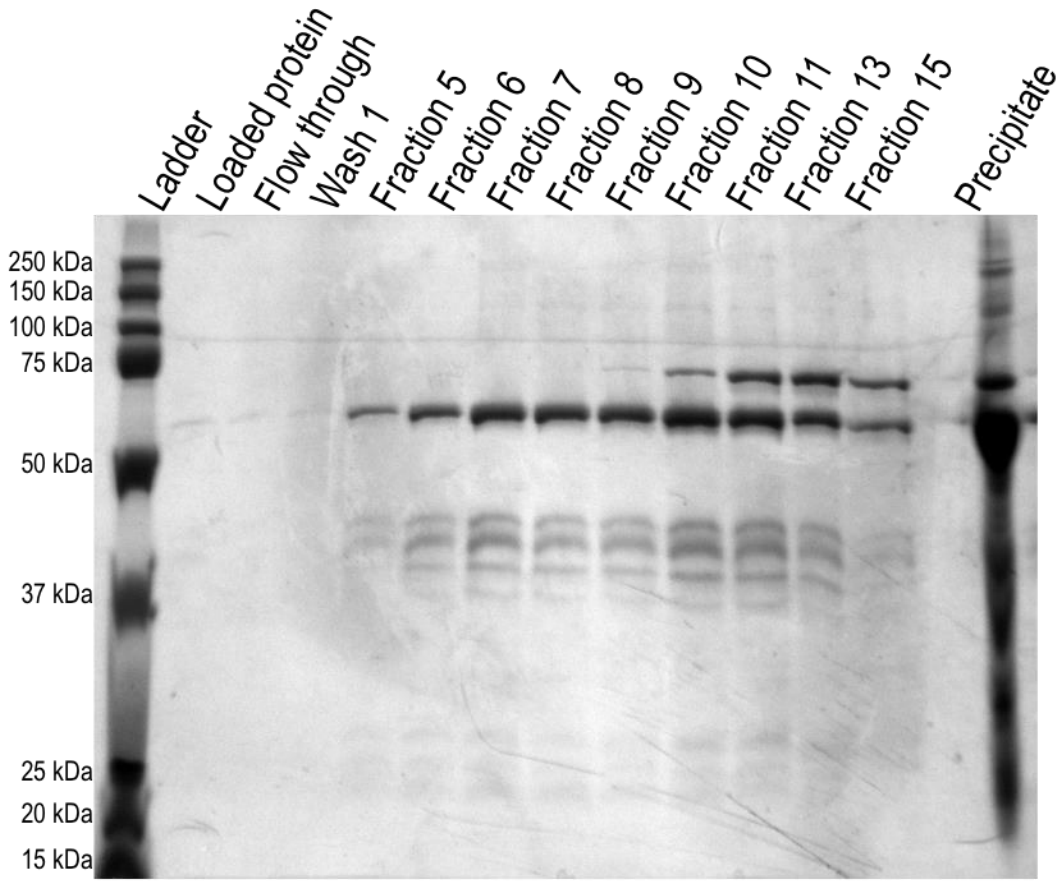


Figure 47: Utilization of IEC to further purify MTA1$_{162\text{-}335}$ from truncation products. While a higher MW band at roughly 75 kDa was able to be separated utilizing this column, there was little separation, as anticipated, between the full-length band at 51 kDa and the truncation products.

Finally, we utilized SEC to assess whether we could separate the truncation products from the desired MW band. The protein was concentrated down and loaded on to a GE Healthcare Superdex HiLoad 16/600 200pg column pre-equilibrated in PBS and run at 0.5 mL/min. SDS-PAGE (4-15% Bio-Rad gel, 120 V, 55 min) with Coomassie staining (Figure 48) was utilized to visualize this column purification step. Interestingly,

using SEC, we were able to slightly separate the truncation products from the desired band, however, the 75 kDa band was not separated utilizing this method.

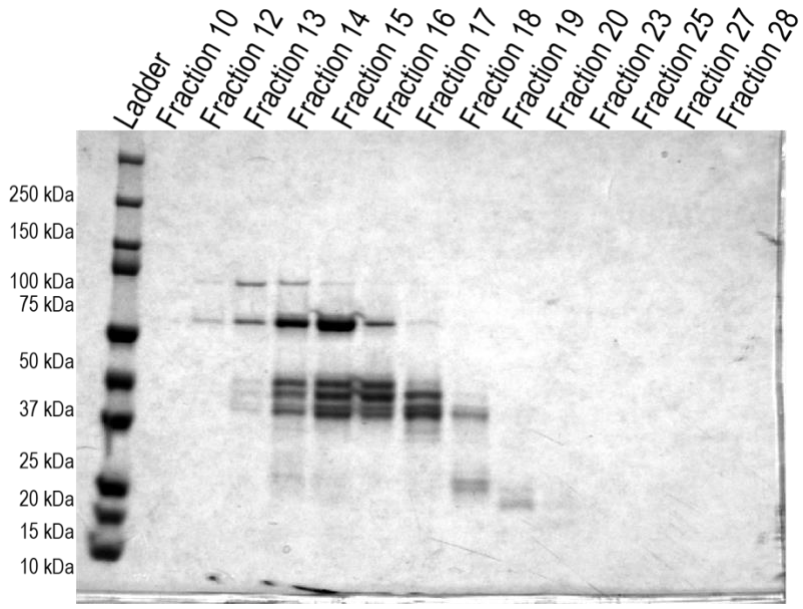


Figure 48: Utilization of SEC to further purify MTA1$_{162-335}$. A GE Healthcare Superdex HiLoad 16/600 200pg column in PBS was utilized (0.5 mL/min) to test separation between the 51 kDa desired band and the truncation products. Utilizing this method there was a slight shift in the truncation product elution from the 51 kDa MW band, but no separation with the upper 75 kDa band.

2.5.6 Proposed future purification strategy

While we were unable to completely separate the truncation product from the desired 51 kDa 6xHisGST-MTA1$_{162-335}$ MW band, we are able to consistently obtain 3-4 mg of protein per purification strategy. Based on SDS-PAGE analysis, we believe that, should the tagged protein be necessary, utilization of a two-step nickel-immobilized affinity chromatography followed by IEC will allow for protein with purity to allow for initial binding studies. For a tag-less purification, I performed a test TEV protease

cleavage post-nickel-immobilized affinity chromatography (Figure 49) to demonstrate complete cleavage of 6xHisGST-MTA1$_{162-335}$. The TEV protease utilized is tagged with a 6xHis tag which will allow for a second nickel-immobilized affinity chromatography step to remove the protease as well as the cleaved tag while purifying MTA1$_{162-335}$ in high yield.

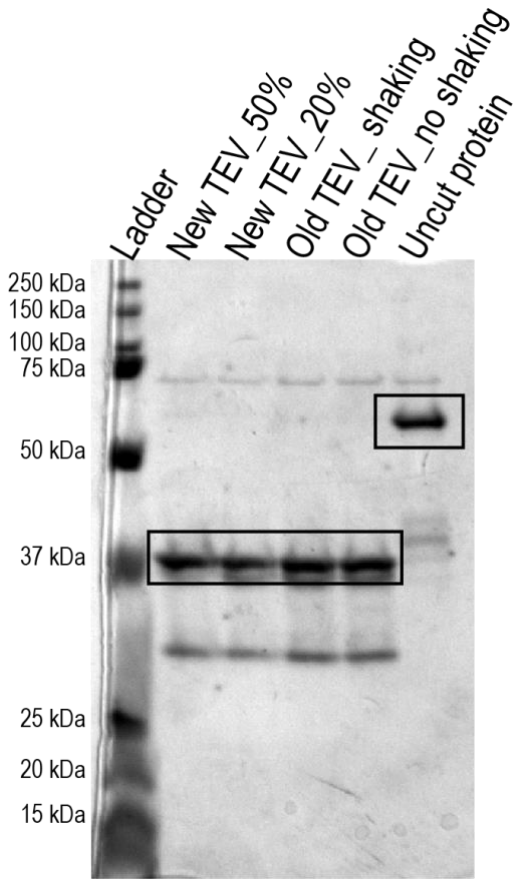


Figure 49: TEV protease cleavage test of 6xHisGST-MTA1$_{162-335}$. We utilized several different TEV protease conditions, including percent glycerol storage of the protease and shaking the overnight cleavage reaction versus not. In all conditions we saw complete cleavage of the protein.

2.6 Expression and purification of a codon optimized CoREST$_{286-493}$

As discussed above, the interface between KDM1A and CoREST has been widely defined via biochemical and biophysical methods, including a series of crystal structures. As we looked towards utilizing the KDM1A:CoREST$_{286-493}$ interface as a model system to assess our tools of biophysical characterization and protein:protein interface disruption, we noted that in the hands of previous lab members, CoREST$_{286-493}$ purified still contained truncation products. I produced a new construct of CoREST$_{286-493}$ that was codon optimized and allowed for expression and purification of CoREST$_{286-493}$ without previously noted truncation products.

2.6.1 Cloning

In order to obtain a codon optimized construct of CoREST$_{286-493}$, I first obtained a pUC-57 plasmid containing the full-length CoREST1 gene codon optimized for *E. coli* flanked by N-terminal NdeI and C-terminal XhoI cleavage sequences. With an expression protocol already established for pET28-b(+), we envisioned a PCR amplification of CoREST$_{286-493}$, subsequent restriction digest of both the PCR product and the pET28-b(+) vector containing the CoREST$_{286-493}$ sequence not codon optimized for *E. coli* expression, followed by subsequent ligation. We chose PfuTurbo DNA polymerase (NEB) for our PCR, amplifying the DNA sequence corresponding to CoREST residues 286-493 with NdeI and XhoI restriction digest sequences at each end of the sequence, respectively (Figure 50).

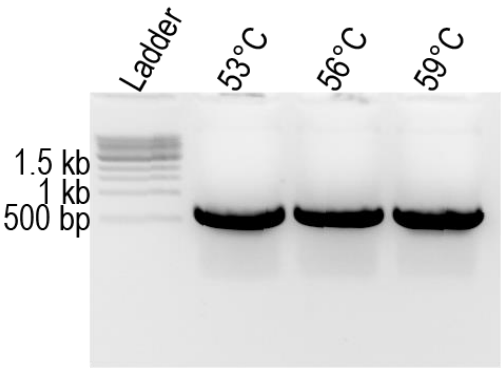


Figure 50: PCR amplification of codon optimized CoREST$_{286-493}$ from a pUC57 vector containing the full-length CoREST sequence. We utilized PfuTurbo DNA polymerase and standard manufacturer protocol for this. The annealing temps were based on calculated T_m for the forward and reverse primers we designed.

Following a PCR reaction clean up (Zymo), we incubated 1 µg of both pET15-b backbone and insert, separately, with CutSmart buffer (NEB), and 20 U of NdeI and XhoI (NEB), dosing in an additional 20 U of each enzyme halfway through the digest (37 °C, 3 hrs.). Following completion of the digest, I ran an 0.8% agarose gel (100V, 30 min,1 kB ladder) to compare the cut backbone to the original vector and allow for separation between the backbone, and the insert that had been digested out of the pET28-b(+) vector (Figure 51). We excised the digested backbone and PCR fragment from the gel and the and performed a Qiagen gel extraction kit on both the backbone and the insert, as per protocol instructions.

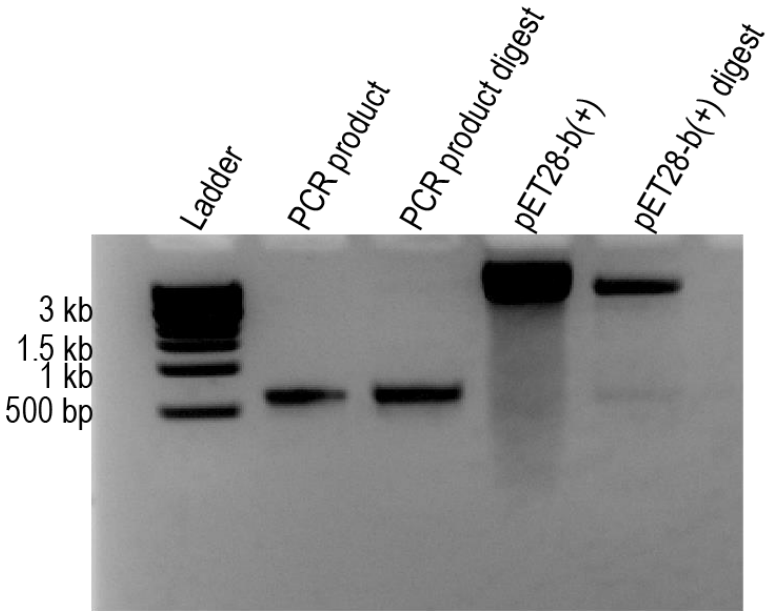


Figure 51: Agarose gel following restriction digest of a pET28-b(+) vector and CoREST$_{286-493}$ PCR amplification. Digest was run over the course of 3 hours utilizing CutSmart buffer and 40 U of both XhoI and NdeI. A 0.8% agarose gel was run (100V, 30 min, 1kB ladder) of the pre- and post-digest PCR product and vector.

For the ligation reaction, we measured the concentrations of the purified pieces and ran two ligation reactions (NEB) at 3:1 and 4:1 insert:vector molar ratio and incubated overnight at 16 °C. In order to remove excess enzyme and salts that could inhibit an electrochemical transformation, we performed a DNA purification utilizing EMD Millipore PelletPaint Co-Precipitant, as per manufacturer's protocol. We resuspended the pellet in 16 μL nuclease-free water. We performed an electrochemical transformation with 5 μL of the purified ligated DNA and mixed them with an aliquot of DH10β cells. Directly following the electro-shock, we diluted the cells to 1 mL in super optimal broth with catabolite repression (SOC) and incubated them for 1 hour at

37 °C in a rotating incubator, allowing the freshly transformed cells recovery time before plating them on an ampicillin plate overnight at 37 °C.

With the codon optimized pET28-b(+) vector in hand, we proceeded to the purification of the codon-optimized sequence from BL21(DE3) cells. I envisioned a purification strategy wherein we utilized an initial nickel-immobilized affinity chromatography purification step followed by an overnight TEV protease cleavage, a second nickel-immobilized affinity chromatography step, and a final SEC purification step as required.

2.7. Discussion and future directions

Epigenetic enzymes play a critical role in the regulation of gene expression. While there has been a significant effort towards elucidating active site mechanisms towards potential therapeutics, there is another layer of complexity that has largely been disregarded until recently. Indeed, epigenetic enzymes function within the context of multimeric complexes wherein coregulatory proteins contribute additional specificity and selectivity to the enzyme. Each coregulatory partner imparts either a change to the selectivity or specificity of the enzyme towards its substrate or defines the gene targets of the epigenetic complex as a whole. In order to assess how KDM1A is affected by its coregulatory partners, we have made significant progress towards the expression and purification of each coregulatory partner and enzyme within a stable KDM1A-containing 5-mer complex.

While purification of full-length CoREST has not proven successful in the past, we have accessed full-length KDM1A:CoREST via a pET-DUET1 vector allowing for KDM1A to act as a solubility handle for CoREST to bind. We have accessed these two proteins via two different solubility tags and purification methods, depending on which is more accessible with reagents on-hand. With full-length CoREST in hand, there are several key future directions we envision. We hypothesize that CoREST acts as the scaffold protein for the formation of the 5-mer complex, so achieving expression and purification of full-length CoREST will allow for pull-down assays to compare which proteins bind to KDM1A:CoREST$_{286\text{-}493}$ to KDM1A:CoREST. Furthermore, while the main binding affinity between KDM1A:CoREST is contributed via the linker region of CoREST, it may be that other domains of CoREST contact KDM1A and provide additional stability to the interaction with KDM1A that have been previously undocumented.

With full-length CoREST in hand, we are now also able to access a catalytic demethylase-deacetylase catalytic core through the binding of KDM1A and HDAC1/2 to CoREST. We have produced as well as demonstrated expression and several purification methods for HDAC1/2, depending on the need for an affinity handle. We have demonstrated that HDACs can be accessed with purity at a mg scale, which will allow for the series of experiments we envision to establish activity as well as binding. As discussed above, phosphorylation by the kinase CKII has been reported to be critical

for activity, we anticipate that an *in vitro* phosphorylation will result in active protein. Inositol phosphate has additionally been reported as a potential co-factor for activity, so addition of this either during expression or during purification may be required *(143)*. In either case, the Fluor-de-lys assay is a well-established fluorescence assay that can be utilized to test the deacetylation via our HDAC1/2 proteins. If both of these methods fail, we hypothesize that utilization of the truncated HDAC1/2 constructs may result in accessing active HDAC1/2 due to the precedence of HDAC8 activity.

Lacking activity, it may be that HDAC1/2 are still conformationally folded so that binding and the binding of HDAC1/2 to CoREST can be determined. We hypothesize that it is very similar to MTA1:HDAC1/2 and exploring this interaction interface will allow us to understand how different scaffold proteins can utilize similar binding methods in these epigenetic complexes. Interestingly, a 2018 study reported initial activity studies of KDM1A and HDAC1/2 from a purified 3-mer catalytic core complex, which we would like to explore further and also access the activity of HDAC1/2 within the 3-mer catalytic core complex *(261,279)*.

We have also reported the expression and initial purification attempts of both Braf35 and BHC80. We anticipate that, upon obtaining both proteins, we will not only be able to show the formation of the 5-mer complex, but also be able to perform critical pull-down assays. Each of these assays will allow us to gain a better understanding of which proteins are required for binding and formation of the 5-mer complex. While we

anticipate that CoREST is the core protein for the 5-mer complex to form, it is unclear what is required for BHC80 and Braf35 to be associated. It may be that BHC80 and Braf35 provide substrate specificity to position KDM1A. They may also function to target specific genes through 'reader' functions of these two coregulatory proteins.

With the 5-mer complex in hand we will be able to utilize CryoEM to gain a more complete understanding of how the proteins interact. We envision utilizing a technique similar to what was used in the recent 2018 study to establish the PCR2 complex via CryoEM wherein each subunit was tagged with GFP to track the position of the protein within the complex *(193)*.

Establishing the interaction interfaces will enable us to utilize techniques such as hydrogen-deuterium exchange with mass spectrometry (HDX-MS) coupled with SPR or ITC to establish the binding affinities of the interfaces. We will also be able to establish key interactions for complex formation. This will allow us to produce targeted interaction inhibitors to disrupt formation of this specific 5-mer complex. This will allow us to elucidate the role of this 5-mer complex, and KDM1A within the context of this complex, *in vivo*.

Overall, the efforts we have made in cloning, expression, and purification of a stable KDM1A-containing 5-mer complex will enable us to understand the role of KDM1A within the context of the biological complexes it is found within. Long-term we

hope that this information will allow us to understand the role of KDM1A within the context of cancers and lay the foundation for producing targeted interaction inhibitors.

2.8 Experimental Methods

Primers

Table 3: Primers were purchased from Eton Biosciences and resuspended to a final concentration of 100 mM

Construct	Forward (5′-3′)	Reverse (5′-3′)
Codon optimized CoREST$_{286-493}$	CCCCACGAATTC CATATGCCGACG GAAACC	GAGCCTGAATTC CTCGAGTTAGTG ATGGTGATGG
pET-28b(+) MTA1$_{162-335}$	AGCTGAATTCCA TATGGCGGATAA AGGTG	CAAAGAATTCCT CGAGTTATTAAC GATCGGTGGTT
pET-15b MTA1$_{162-335}$	AGCTGAATTCCA TATGGCGGATAA AGGTG	CAAAGAATTCCT CGAGTTATTAAC GATCGGTGGTT

KDM1A:CoREST expression and purification

6xHis-KDM1A:MBP-CoREST$_{N\Delta 60}$

A pETDuet-1 vector containing N-terminal MBP tagged CoREST (residues 60-485) and 6xHis tagged KDM1A (pET15b, residues 151-852) in multiple cloning sites 1 and 2, respectively was used for expression and purification. The plasmid was transformed into electrocompetent BL21 Star (DE3) *E. coli* cells (ThermoFisher). Cells were grown in TB medium with 100 μg/μL ampicillin at 30 °C with shaking (200 rpm) for 40 hours due to leaky expression. Cells were collected using an SLA-3000 rotor (4225 *x g*, 4 °C, 15 minutes) and lysed with an Emulsiflex C-5 high-pressure homogenizer

(Avestin, Inc.) in Ni-IMAC wash buffer containing 50 mM Tris-HCl buffer at pH 7.4 containing 350 mM NaCl, 10 mM imidazole, 5% glycerol, 1X Halt protease inhibitor cocktail, lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626). The resulting crude extract was spun for 40 min at 4 °C and 44000 *x g*, using a Ti-45 rotor to pellet insoluble material. The supernatant was further purified via 15 CV imidazole gradient to 300 mM after extensive washing via Ni-IMAC (GE healthcare 17057501). After buffer exchange via Amicon Stirred cell into a buffer containing 20 mM Tris-HCl (pH 7.4), 200 mM NaCl, 1 mM EDTA, and 1 mM DTT, the protein was loaded onto a 10 mL amylose resin (NEB) at <1 mL/min, washed extensively and eluted with 10 mM maltose. A second Ni-IMAC purification following overnight TEV protease cleavage with the same buffer system as before following which the fractions containing a 1:1 ratio of the proteins was concentrated via EMD Millipore centrifugal filter unit (10 kDa MWCO, 2000 *x g*, 4 °C, 20 min) to a volume less than 1 mL and then loaded onto a 16/600 S200 Sephacryl column which was eluted for 1.5 CV (20 mM HEPES (pH 7.8), 200 mM NaCl, 1 mM DTT) at 0.7 mL/min. The purified protein was concentrated and stored (40% glycerol, -20 °C) until future use.

6xHis-KDM1A:GST-CoREST$_{N\Delta 60}$

A pETDuet-1 vector containing N-terminal MBP tagged CoREST (residues 60-485) and 6xHis tagged KDM1A (pET15b, residues 151-852) in multiple cloning sites 1 and 2, respectively was used for expression and purification. The plasmid was

transformed into electrocompetent BL21 Star (DE3) *E. coli* cells (ThermoFisher). Cells were grown in TB medium with 50 μg/μL kanamycin at 30 °C with shaking (200 rpm) to an OD_{600} of 0.8 and then induced (0.25 mM IPTG, 4 hr, 30 °C). Cells were collected using an SLA-3000 rotor (4225 *x g*, 4 °C, 15 minutes) and lysed with an Emulsiflex C-5 high-pressure homogenizer (Avestin, Inc.) in Ni-IMAC wash buffer containing 50 mM Tris-HCl buffer system at pH 7.4 containing 350 mM NaCl, 10 mM imidazole, 5% glycerol, 1X Halt protease inhibitor cocktail, lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626). The resulting crude extract was spun for 40 min at 4 °C and 43000 *x g*, using a Ti-45 rotor to pellet insoluble material. The supernatant was further purified via 15 CV imidazole gradient to 300 mM after extensive washing via Ni-IMAC(GE healthcare 17057501). After overnight dialysis into PBS buffer containing 1 mM BME, the solution was loaded onto a 10 mL glutathione agarose column (Genesee Scientific) at 1 mL/min, washed extensively and eluted with 10 mM reduced glutathione. The fractions containing a 1:1 ratio of the proteins was concentrated via EMD Millipore centrifugal filter unit (10 kDa MWCO, 2000x g, 4 °C, 20 min) to a volume less than 1 mL and then loaded onto a 16/600 S200 Sephacryl column which was run for 1.5 CV (20 mM HEPES (pH 7.8), 200 mM NaCl, 1 mM DTT) at 0.7 mL/min. The purified protein was concentrated and stored (40% glycerol, -20 °C) until future use.

HDAC1 expression and purification

pET-28b(+) MBP-HDAC1

A pET-28b(+) vector containing N-terminal MBP tagged HDAC21was used for expression and purification. The plasmid was transformed into electrocompetent BL21 (DE3) *E. coli* cells (ThermoFisher). Cells were grown in TB medium with 50 μg/μL kanamycin and 1 mM $MgCl_2$ at 30 °C with shaking (220 rpm) until induced with IPTG (1 mM, OD_{600} 0.45, 15 °C) and grown overnight. Cells were collected using an SLA-3000 rotor (4225 *x g*, 4 °C, 15 minutes) and lysed with an Emulsiflex C-5 high-pressure homogenizer (Avestin, Inc.) in a wash buffer a 25 mM HEPES buffer system at pH 8.0 containing 200 mM NaCl, 1 mM EDTA, 1X protease inhibitor cocktail (Biotool 14001), lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626). The resulting crude extract was spun for 40 min at 4 °C and 44000 *x g*, using a Ti-45 rotor to pellet insoluble material. The supernatant was incubated with 10 g/L of Bob's Red Mill potato starch that had been washed 3x with nuclease-free water and 3x with wash buffer in a swinging-bucket rotor centrifuge (500 rpm, 4 °C, 5 minutes). The lysate was incubated for 4 hours at 4 °C with rotation to promote binding. The protein lysate was washed 3x-15-minute incubation with wash buffer before pelleting the starch and removing supernatant via swinging-bucket rotor centrifuge (500 rpm, 4 °C, 15 min) before a 2-hour incubation with wash buffer containing 10 mM maltose. The eluted protein was separated from the potato starch via swinging-bucket rotor centrifuge (500 rpm, 4 °C, 15 min) for a total of 3 elution fractions, the second and third with only a 15-minute incubation period before centrifugation.

pET-30b HDAC2-6xHis

A pET-30b HDAC2 was used for expression and purification. The plasmid was transformed into electrocompetent BL21 Star (DE3) *E. coli* cells (ThermoFisher). Cells were grown in 2X-YT medium with 50 μg/μL kanamycin, 1 mM $MgCl_2$, and 1 mM $Zn(SO_4)$ at 30 °C with shaking (220 rpm) to an OD_{600} of 0.5 and then induced (1.0 mM IPTG, 15 °C, overnight). Cells were collected using an SLA-3000 rotor (4225 *x g*, 4 °C, 15 minutes) and lysed with an Emulsiflex C-5 high-pressure homogenizer (Avestin, Inc.) in Ni-IMAC wash buffer containing 25 mM HEPES buffer system at pH 8.0 containing 300 mM NaCl, 10 mM imidazole, 10% glycerol, 1X protease inhibitor cocktail (Biotool 14001), lysozyme (1.0 mg/L Sigma L6876), and 0.4 mM PMSF (Sigma P7626). The resulting crude extract was spun for 40 min at 4 °C and 43000 *x g*, using a Ti-45 rotor to pellet insoluble material. The supernatant was further purified via step to 60 mM imidazole, 60 mM imidazole with 0.01% Triton X-100, 60 mM imidazole, and then subsequent gradient to 300 mM imidazole. After overnight dialysis into 25 mM HEPES buffer at pH 7.0 containing 50 mM NaCl and 2 mM BME, the solution was loaded onto a 20 mL Q-sepharose for Fast Flow (GE Healthcare) 2 mL/min, washed extensively and eluted over a gradient up to 1 M NaCl. The fractions containing the protein were combined and concentrated via EMD Millipore centrifugal filter unit (10 kDa MWCO, 2000x g, 4 °C, 20 min).

CoREST$_{286-493}$ expression and purification

An N-terminal 6xHisGST-CoREST$_{286-485}$ with a TEV cleavage site immediately preceding CoREST$_{286-493}$ was utilized for expression and purification. The plasmid was transformed into electrocompetent BL21(DE3) *E. coli* cells (Novagen). Cells were grown in TB medium with 50 μg/μL kanamycin at 37 °C shaking (200 rpm) to an OD_{600} of 1.0, upon which IPTG was added to a final concentration of 1.0 mM and the induced cells were grown overnight at 19 °C. Cells were collected using an SLA-3000 rotor (4225 *x g*, 4 °C, 15 minutes) and lysed with an Emulsiflex C-5 high-pressure homogenizer (Avestin, Inc.) in Ni-IMAC wash buffer containing 50 mM Tris-HCl (pH 8.0), 350 mM NaCl, 5% glycerol, supplemented with lysozyme (1.0 mg/L Sigma L6876), and 1x Protease Inhibitor Cocktail (Biotool 14001). The resulting crude extract was spun for 40 min at 4 °C and 40000 *x g*, using a Ti-45 rotor to pellet insoluble material. The supernatant was further purified via Ni-IMAC (GE Healthcare 17057501), overnight TEV protease cleavage, a secondary Ni-IMAC (GE Healthcare 17057501) column and stored at -20 °C (25 mM HEPES (pH 7.4), 138 mM NaCl, and 40% glycerol).

3. Utilization of SELEX towards production of a second-generation KDM1A:CoREST inhibitor

3.1 Introduction

The protein:protein interfaces between coregulatory proteins are still poorly understood in the case of many epigenetic complexes. The subunits of each complex have widely been established with co-precipitation or pull-down assays. In order to elucidate how coregulatory proteins affect the catalytic efficiency or substrate specificity of an enzyme, it is critical to map the interactions. Furthermore, obtaining biochemical and biophysical information about the distribution of affinity as well as conformational changes due to binding between two proteins will provide a basis through which to target critical interfaces.

In order to approach exploring these interactions in a systematic method, I first chose to establish the utility of these methods with the well-established model system of KDM1A:CoREST$_{286\text{-}482}$. In addition to several available X-ray crystallography structures of KDM1A:CoREST$_{286\text{-}482}$, Dr. Sunhee Hwang in our lab has previously extensively analyzed the interface through epitope mapping, peptide competition assays, and isothermal calorimetry (ITC) *(56,60,61,117,280)*. This foundation has enabled me to utilize hydrogen-deuterium exchange with mass spectrometry to further assess the binding interface between KDM1A and CoREST$_{286\text{-}493}$ and demonstrate that the binding affinity is spread across the linker region between the two proteins. However, I will show that the apical end of KDM1A appears to be locked into place upon binding

$CoREST_{286-493}$ suggesting an importance for CoREST to stabilize this unique protein domain.

The ongoing goal of our research has been not only to probe the interface but produce targeted probes of the KDM1A:CoREST interface. Previous lab members utilized a peptide inhibitor which was not able to disrupt pre-formed KDM1A:$CoREST_{286-493}$ complex *in vitro* and suffered from stability issues *in vivo*. Dr. Jennifer Link Schwabe and I turned to RNA aptamer to derive selective disruptors of the KDM1A:CoREST interaction. I will discuss our progress and future directions herein.

3.2 Hydrogen-deuterium Exchange with Mass Spectrometry

To assess the dynamic nature of the KDM1A:CoREST interaction, we compared relative hydrogen-deuterium exchange (HDX) rates between KDM1A in isolation versus bound to $CoREST_{286-482}$ with Dr. Eric Soderblom at the Duke GCB Proteomics and Metabolomics Core Facility. Briefly, both proteins were expressed via *E. coli* and purified as described. The purified proteins were subjected to a 10-fold D_2O dilution for 0, 1, and 5 minutes and subsequent quenching in chilled 0.3% formic acid and on-column pepsin digestion (Figure 52). Resulting peptides were then resolved using a Waters nanoAcuqity UPLC system coupled to a Waters Synapt G2 high resolution accurate mass tandem mass spectrometer via a electrospray ionization source. Peptides identified using Protein Lynx Global Server (v2.5.2) with a score of 7.0 were included in the final analysis for a final peptide FDR < 1.0%. The overall coverage for KDM1A was 99.2%

(Figure 52e) with a redundancy of 10.6, and 95.5% for $CoREST_{286-482}$ with a redundancy of 10.2. In addition to manual inspection, peptides carried forward for HDX comparison adhered to the subsequent criteria: 1) presence across all time points in both (-) and (+) $CoREST_{286-482}$ conditions, 2) percent difference of uptake between (+) and (-) $CoREST_{286-482}$ at the 1- and 5-minute time points to be 30% and 3) a p-value < 0.05 as assessed by a two-tailed students t-test. Our 30% cutoff was chosen by visualization of a box and whisker graph: the 10% and 90% tails lay at 29.7 and 35.9%.

After manual and statistical evaluation of each peptide, 43 peptides remained. With Jennifer Link Schwabe, we observed several regions with a decrease in relative deuterium uptake as a function of $CoREST_{286-482}$; we did not observe any significant peptides experiencing an increase in relative deuterium uptake. Each of the 43 peptides were mapped to an existing KDM1A:$CoREST_{286-482}$ structure (pdb:2iw5, Fig. 7a,b,c). The largest change was observed near the top of the Tower domain, where KDM1A TαA residues 440-451 (Figure 52a,b, red) experienced a >20% decrease in deuterium uptake and the two regions bordering the turn between TαA and TαB (residues 452-467 and 478-490; orange) experienced a 10-20% decrease in deuterium uptake (Figure 52b). We noted that the change in deuterium exchange in this region is strikingly differentiated from that of the lower region of TαA, despite both forming direct interactions with CoREST. This difference did not correlate with a change in solvent-accessible surface area compared to the lower region of TαA (data not shown). We therefore hypothesized

that the alterations we observed were instead due to a shift in flexibility of the Tower upon CoREST binding.

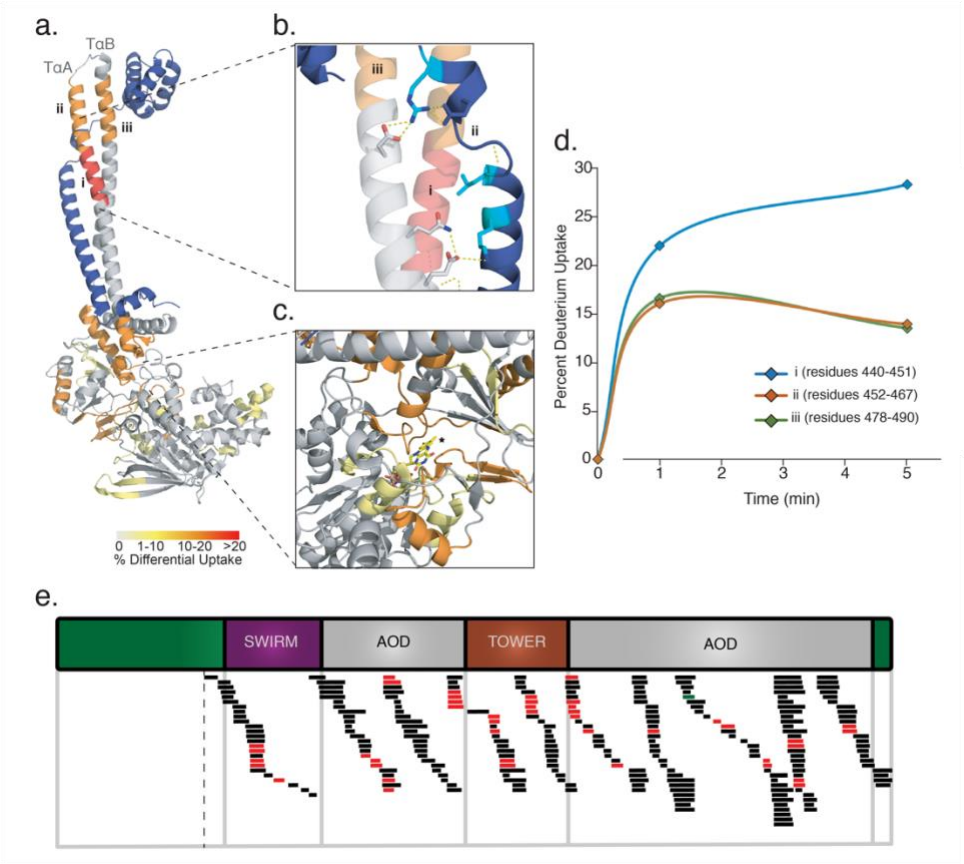


Figure 52: HDX-MS of KDM1A with respect to the presence of CoREST$_{286-492}$ (a) Differential deuterium uptake between corresponding peptides in the ± CoREST samples were mapped onto an available LSD1/CoREST co-crystal structure (PDB 2IW5) as a heat map. Red represents the greatest decrease in deuterium uptake (20%), while gray areas experienced no change. No peptides experienced an increase in deuterium uptake as a function of CoREST binding. (b) Regions of protection in the tower region correspond to CoREST mutations that weakened affinity for LSD1 (cyan). (c) Several regions adjacent to the FAD (black star) binding pocket and active site experienced a decrease deuterium exchange in the presence of CoREST. (e) Tryptic LSD1 peptide fragments identified by MS represented 99.2% of the complete protein sequence, giving excellent coverage. Note that the construct of LSD1 used was NΔ150, which is noted by the black dashed vertical line.

To determine if any conformational differences in the Tower domain were detectable due to the presence of CoREST$_{286-482}$, all available KDM1A-containing structures in the PDB database were assembled in the Bio3D webserver, with exclusion

only of structures containing KDM1A primary sequence mutations or binding partners other than CoREST *(281)*. Importantly, only the KDM1A chain of each structure was included in inquiry; hence structure analysis was unbiased by the presence of CoREST, accounting only for structural deviations of KDM1A. All KDM1A chains were structurally aligned by superimposition to an invariant core. Deviations in positioning between structures were visualized as a heatmap of root mean square fluctuation (RMSF) value per residue (Figure 53a). Within all structures, the AOD and SWIRM domains display very little variation. Similarly, RMSF values per residue were highest in Tower domain residues (Figure 53b) increasing near the apical end. Interestingly, maximum RMSF values were much lower when comparing only structures of KDM1A in isolation or only co-crystal structures of KDM1A/CoREST, suggesting that the Tower domain exists in two distinct conformational states depending upon the presence or absence of CoREST

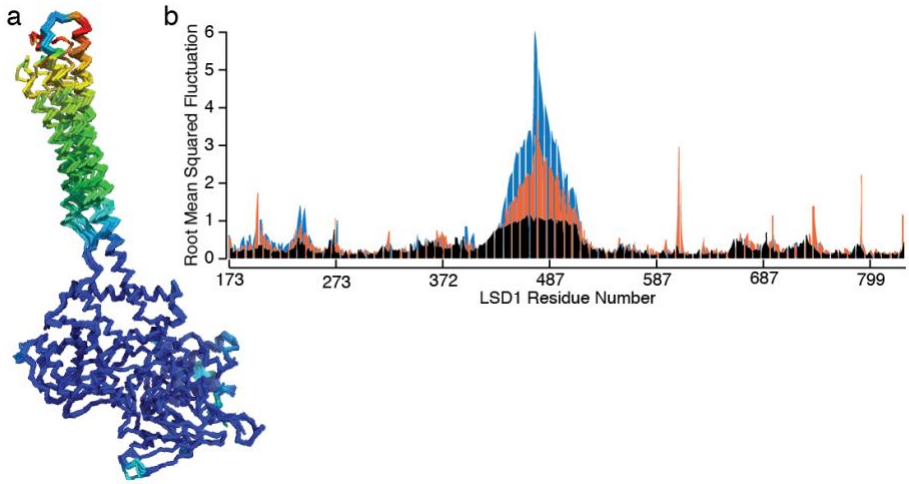


Figure 53: Fluctuations between all KDM1A and KDM1A:CoREST crystal structures focused around tower region. (a) KDM1A chains from all crystal structures were structurally aligned and RMSF values between corresponding residues overlaid as a heatmap onto the alignment (dark blue to red). We noted the highest RMSF variance in the apical end of the tower domain. (b) RMSF values per residue displayed the greatest fluctuation in the tower domain region (residues 415-515). Interesting only looking at the KDM1A-only structures (orange) orKDM1A/CoREST co-crystal structures (black) demonstrated a lower magnitude of RMSF, suggesting two states exist between KDM1A-only and when bound to CoREST.

To test if these deviations were sufficient to discriminate CoREST-bound KDM1A structures from isolated KDM1A structures, all were clustered by RMSD and visualized as a dendrogram (Figure 54a). Most structures were partitioned into two major clusters, which were distinguished by either the presence or absence of CoREST. Similarly, principle component analysis (PCA) grouped structures into the same clusters (Figure 54b). In both cases, one structure lacking CoREST (PDB 2H94) was placed in its own cluster.*(63)* The reason for this distinction is unclear. However, the clustering of these structures is clearly discernible by Tower domain positioning (Figure 54c), with approximately 13 Å separating closely-grouped KDM1A and KDM1A/CoREST residues,

respectively, at the apical end of the Tower. The lone structure constituting its own cluster falls in between these two extremes. Interestingly, both clusters include structures of not only unoccupied KDM1A active sites, but also KDM1A:inhibitor, and KDM1A:H3$_{1-21}$.

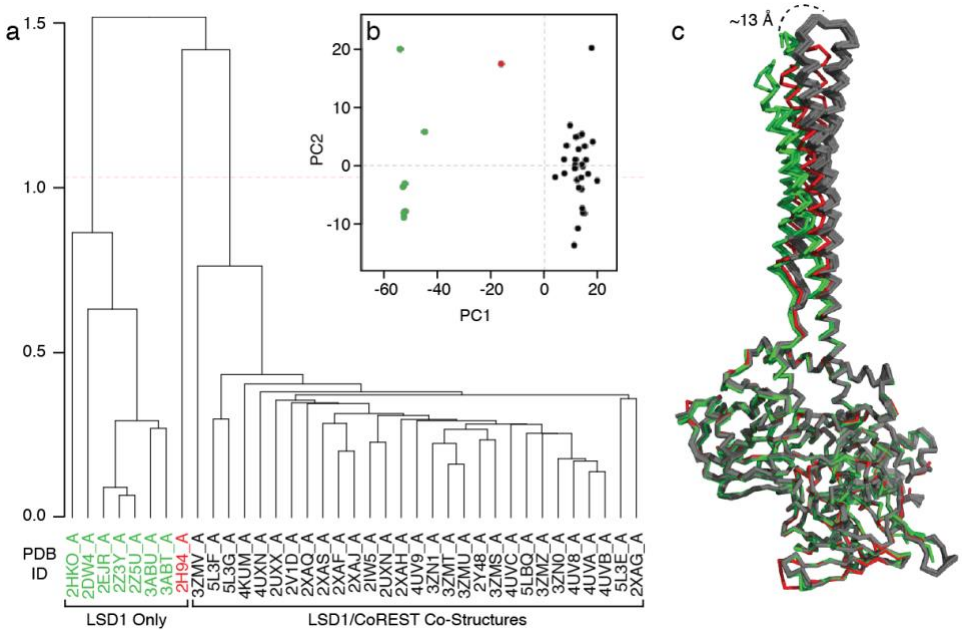


Figure 54: Clustering and PCA of KDM1A and KDM1A:CoREST crystal structures. (a) Clustering by RMSD successfully grouped KDM1A-only structures (green) separately from KDM1A:CoREST$_{286-482}$ crystal structures (black) and were similarly grouped in a(b) Principle component analysis (PCA). Interesting, in both cases the isolated LSD1 structure 2H94 was placed in its own cluster for no discernable reason. (c) Visual representation of the clustering demonstrates an easily identifiable shift with the apical end separating the (+/-) CoREST structures separated by approximately 13 Å with the 2H94 structures taking an intermediary form.

3.3 Efforts towards production of a second-generation inhibitor

The catalytic activity of KDM1A has been well-established as explored in Chapter 1. However, KDM1A has been demonstrated to be involved in many epigenetic

complexes which each affect separate pathways. Inhibition of the active site of KDM1A allows for a culminative analysis of the role of KDM1A in cells. However, in order to understand the role of KDM1A within a specific complex, we would need to disrupt the recruitment of KDM1A to the complex.

Our aim is to produce a high-affinity targeted inhibitor of the KDM1A:CoREST interface. Towards this end, previous lab members Sunhee Hwang and Jennifer Link Schwabe produced and tested an inhibitor of the KDM1A:CoREST interaction interface (iLCC) both *in vitro* and *in vivo* which is a peptide inhibitor of the linker region of CoREST. What they were able to establish was that while this peptide inhibitor was able to disrupt CoREST from binding to KDM1A, it was not able to disrupt pre-formed KDM1A:CoREST complexes; nor was it stable due to intracellular proteolytic degradation *in vivo*.

One method that has been utilized for the production of protein:protein interaction inhibitors is through the systematic evolution of ligands by exponential enrichment (SELEX) wherein an initial library is subjected to iterative rounds of binding selection and amplification to select towards a high-affinity binding motif. The Sullenger lab at Duke has utilized this method extensively towards disruption of protein:protein binding with respect to the blood coagulation cascade with well-established results and pM binding RNA aptamers *(282,283)*. They graciously provided us with their protocol

which we have been able to adapt towards establishing this as a method in our lab to produce RNA aptamers to disrupt epigenetic enzyme complexes.

3.3.1 SELEX efforts towards a KDM1A-specific RNA Aptamer

A general scheme for the SELEX protocol is provided in Figure 55. Briefly, a DNA library is generated based on an initial ssDNA template which in our case consisted of a 5′ and 3′ known sequence capping a 40 N-mer sequence: 5′-TCTCGGATCCTCAGCGAGTCGTCTG-N40-CCGCATCGTCCTCCCTA-3′ Utilizing a DNA Polymerase 1 fragment that has lost its exonuclease function but can still function as a polymerase, the NEB Klenow fragment produces the dsDNA library.

Upon production of the initial DNA library, the RNA library can be produced via T7 RNA Polymerase and subjected to a selection round wherein the library is incubated with protein (37 °C, 15 min) and then vacuum filtered over a nitrocellulose membrane so that protein and the RNA sequences which bind the protein will be retained and unbound RNA will flow through. The membrane is subjected to a phenol:chloroform:isoamyl alcochol (PCA) extraction to separate protein from nucleic acid.

The extracted RNA is then reverse transcribed to cDNA utilizing Roche AMV RT enzyme (10109118001), and then amplified via ThermoFisher Platinum Taq DNA Polymerase and the resulting amplified dsDNA is transcribed to RNA via T7 RNA Polymerase, completing the cycle. This cycle is completed for iterative rounds,

increasing the stringency via both the ratio between the RNA and protein and increasing the ionic strength. In both cases, the goal is to disrupt low-affinity interactions to allow for enrichment of high-affinity binding sequences.

The increase in affinity for the protein is generally tested every 1-2 rounds. In this case, the RNA library at the end of a selection round is dephosphorylated, and subsequently phosphorylated with ^{32}P, γ-ATP. Upon radio-labeling the RNA, a 96-well filtration binding assay is utilized to assess binding of the RNA pool across a range of protein concentrations to produce a binding curve. Ideally, as the stringency each of round increases, the overall demonstrated affinity for the protein will increase due to the enhancement and level of high-affinity RNA sequences in the pool. Upon reaching an apparent high affinity of binding, the cDNA of a selection round can be prepared for a next generation sequencing (NGS) run by PCR amplifying the pool with NGS adapter sequences and subjected to the massively paralleled sequencing. These results can then be utilized to pool for sequence motifs are present across the pool. Specific aptamers containing these motifs are then produced and individually tested for their binding affinity to the protein of interest.

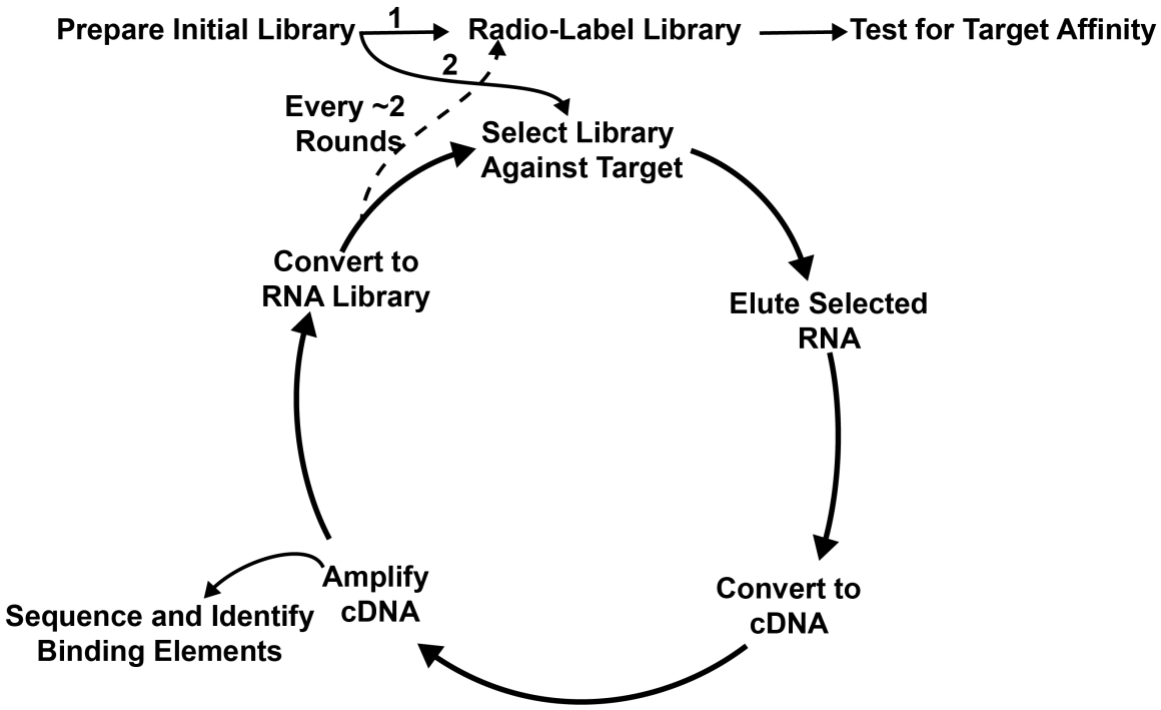


Figure 55: Protocol overview of SELEX. An initial library containing up to 10^{15} sequences is prepared and transcribed to RNA to be incubated with the target protein. The bound RNA is eluted, converted to cDNA, amplified and transcribed back to RNA. Every 1-2 rounds, the post-selection library is radiolabeled and tested for target affinity to assess the progression of high-affinity ligand enhancement

Dr. Jennifer Link Schwabe and I envisioned producing an RNA aptamer that was specific for KDM1A. Ideally, we would produce an RNA aptamer that would be specific towards the KDM1A:CoREST binding interface and allow us to disrupt this interface in not only newly translated KDM1A, but also preformed KDM1A:CoREST both *in vitro* and *in vivo*. With Jennifer Link Schwabe's initial efforts, we were able to produce the initial library and subject KDM1A to four rounds of selection in a HEPES buffer (pH 7.4) with 100 mM NaCl ionic strength. Between each round, we focused on increasing the ratio of available protein to RNA in hopes to select for RNA which were able to

selectively bind KDM1A (Table 4). At this point, we would have anticipated to see a shift upward in the adjusted fraction bound of the RNA in Round 4 over the RNA library in Round 0. However, upon utilizing the filter-binding assay to establish the binding curves, we noted no discernable increase in affinity (Figure 56). This suggests a lack of increased levels of any RNA able to bind tightly to KDM1A.

Table 4: Conditions and quantities of RNA and protein in each selection round

Round	*uM LSD1*	*uM RNA*	*RNA:Protein*	*Input RNA*	*Buffer*
1	5	20	04:01	2.0 nmol	E
2	2.5	10	04:01	1.0 nmol	E
3	1.25	10	08:01	1.0 nmol	E
4	1.25	15	12:01	1.5 nmol	E

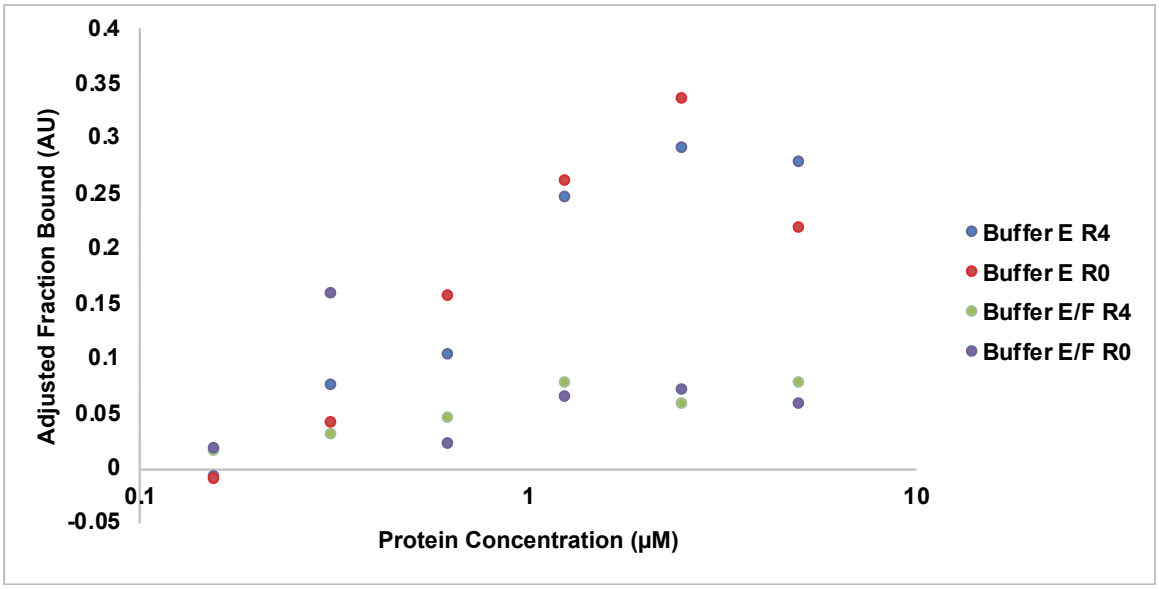


Figure 56: Results from initial trial with SELEX. We compared the binding affinity of the RNA library in Round 0 to the RNA remaining in Round 4 with binding buffers E and E/F which contain 50 and 100 mM NaCl, respectively. We saw no discernable change between the rounds in either buffer, suggesting a lack of increased affinity or selection for binding RNA

We observed no consistent trend between round 0 and round 4 for the binding buffer we had been utilizing (E). We additionally noted an overall decrease in the adjusted fraction bound in binding buffer (E/F) which has a 100 mM ionic strength over 50 mM in binding buffer (E). This suggests KDM1A on its own may not have a high affinity for RNA and that a significant number of iterative rounds would be required to produce an RNA aptamer that would be able to selectively bind KDM1A at high-affinity.

3.4 Discussion

Historically, the active sites of enzymes have been the center of drug target development; however, it is becoming increasing clear that the binding partners of many epigenetic enzymes dictate not only substrate selectivity, but also could impact substrate binding kinetics or spatially-related PTM sites. As protein-protein interaction (PPI) inhibitors are becoming more commonplace, we sought to utilize the KDM1A:CoREST$_{286\text{-}482}$ interface to establish potential regions of targeting along this interface. This interface was attractive to us as it has not only been predefined by crystal structures, but also has been deemed a critical scaffolding protein for directing KDM1A's demethylase activity. Additionally, there is evidence that replacing CoREST for another scaffolding protein, MTA, may redirect KDM1A's selectivity (H3K9 instead of H3K4) as well as which genes are targeted by the enzyme *(53,72)*. Finally, as we confirmed in our evolutionary sequence conservation studies, the Tower domain

architecture is unique to KDM1A; thus, targeting this interaction *in vivo* is unlikely to disrupt interactions of other related enzymes. For these reasons, we believe disruption of the Tower domain interactions may prove a novel strategy for selective KDM1A inhibition.

In exploring the CoREST linker domain with ITC, Dr. Sunhee Hwang initially identified several residues which weaken binding affinity upon mutation, L360A and K363A. These two residues exhibited a statistically significant change in K_d, suggesting that while the binding energy is distributed along the KDM1A:CoREST interface, it may be concentrated towards the apical end of the Tower domain. HDX-MS further supports these findings with residues 440-451 the apical end of the Tower domain experiencing the greatest protection from solvent-exposure in the presence of CoREST. This region of residues aligns with residues K360 and L363 that displayed the sharpest increase in K_d in the mutational ITC analysis. Interesting, there was also a 10-20% decrease in deuterium uptake at the base of the tower domain. Combined, these data suggest a model wherein the interaction between KDM1A and CoREST is spread along the entire surface but potentially utilizes a locking mechanism at either end of the interface to hold the proteins in place.

Upon further examination of the interaction, it became clear that not only could they be grouped by principle PCA by whether or not CoREST was present with the largest RMSD between the two groups centered around the Tower domain. Indeed, as

we follow the Tower domain towards the apex, the disparity between the crystal structures plus and minus CoREST only grows, which is supported by the HDX-MS data displaying the highest solvent-protection towards the apical end of the Tower domain. We hypothesize that CoREST induces a local shift and rigidification in the upper region of the Tower domain, leading to a higher binding energy. In alignment with our findings it has previously been shown that the mutation M448V of KDM1A prevented interaction with CoREST *in vivo*. The accidental mutation (M448V) is not only in the upper Tower region which displayed the highest level of HDX-MS protection, but also lies near CoREST residue L363 which significantly altered K_d in our mutational ITC analysis.

Together, these studies suggest that while there are no specific hot spot resides, CoREST does indeed bind most tightly to the apical end of the Tower, inducing a local shift and rigidification of KDM1A. This improved understanding of the KDM1A:CoREST interface demonstrates the dynamics of this interaction. We envision that the rigidification of KDM1A by CoREST may lay the groundwork for substrate binding and catalysis. While the timing events of the formation of KDM1A:CoREST containing complexes is unclear, we speculate that formation of this rigidified interface between the two proteins may act as a scaffold for binding of additional accessory proteins, lncRNA, and even the nucleosomal substrate.

Moving forward, we hope to generate probes capable of disrupting this interaction for the purpose of exploring the role of KDM1A within specific complexes *in vivo*. The utilization of a SELEX technique allows for the directed evolution of a high-affinity binding molecule. Unfortunately, our initial attempts did not suggest a strong affinity for LSD1 towards RNA; however, there are several future directions that we believe would allow access towards an RNA aptamer probe through SELEX.

First, the initial library can have upwards of 10^{15} sequences present and molecular modeling simulations have demonstrated high-affinity (at least 10^{-7} M K_d) sequences may be lost in the initial rounds due simply to random binding effects *(284)*. One mechanism to allow for the enrichment of high-affinity sequences is to alter the total concentration of protein, which has been shown to skew towards a higher-affinity binding fraction at a much more rapid pace than higher target protein concentrations. While there is some debate as to what level is appropriate, our utilization of a 1.25-5 μM range of KDM1A is at least 10-fold higher than the highest recommended range *(284,285)*. We also anticipate that NaCl concentration could play a role in allowing for an initial enhancement of high-affinity binders, and thus believe that a lower starting ionic strength, while retaining more non-specific binders initially, will allow for high-affinity binders to not be lost in interative selection rounds.

Further, there are several recent reports of one-step SELEX methods which result in high-affinity, specific binders *(286–288)*. In each method, a capillary column is utilized

wherein the target protein is pre-bound and the library of RNA is loaded onto the column to allow for binding to occur. In MonoLEX, the column is then fragmented width-wise and the RNA within each layer is subjected to further testing for binding target protein which is not resin-bound *(286)*. In the case of MonoLEX, a side-by-side comparison with SELEX was performed and the same binding motif was elucidated by both methods. While optimization of buffer system and resin-binding would need to be optimized, we hypothesize that MonoLEX would allow for a more rapid test of the feasibility in utilization of an aptamer to disrupt the KDM1A:CoREST interface.

We anticipate a tight-binding aptamer or peptide may act to outcompete the KDM1A:CoREST interaction both *in vitro* and *in vivo*. Disruption of KDM1A binding CoREST over other binding partners will inform the effect of KDM1A in complex-specific targets and presents a new method for exploring epigenetic enzymes.

3.5 Methods

Hydrogen-deuterium exchange with mass spectrometry (HDX-MS)

Proteins were expressed and purified from E. coli as previously described and ultimately prepared for use in 25 mM HEPES, 200 mM NaCl, 1 mM βME (pH 7.8). Concentrations were measured by Bradford assay and samples diluted to 10 μM in HEPES buffer. Using an automated HDX labeling LEAP robot, proteins were diluted 10x into either HEPES buffer in water (for t = 0) or in D_2O (for t = 1, 5 min). After the designated amount of time, exchange reactions were quenched with 45 μL 0.3% formic

acid, 1 mM TCEP at 4°C nuclease-free water. Proteins were then digested on an immobilized pepsin column (Waters Enzymate BEH pepsin column) at 200 μL/min at 20 °C for 2 min. 50 μL of each sample were analyzed by UPLC-MS (Waters nanoAcquity, Synapt G2). Samples were first trapped on a VanGuard C18 column (40 μL/min in 5% v/v water/acetonitrile, 0.1% formic acid), followed by separation on a 1.7 μm Acquity BEH130 C18 column (Waters) with a linear gradient from 5-35% acetonitrile, 0.1% formic acid over 7 min at 40 μL/min. All separation steps were performed at 1 °C. Peptide masses were analyzed by a tandem mass spectrometer via electrospray ionization source, with alternating high- and low-energy scans from 50-2000 m/z. Peptides were identified by comparison to the LSD1 and CoREST sequences in the Protein Lynx Server (v2.5.2). Each sample was analyzed in duplicate. 648 unique peptides were identified and filtered for analysis by sequences that could be identified in all samples. Deuterium uptake values were then subjected to a t-test for each – CoREST condition compared to the corresponding + CoREST peptide. Sites of differentiation were identified by percent deviation (increase or decrease) being consistent across all + CoREST time points, a pvalue of < 0.05 between ± CoREST samples, and manually inspection of peak picking. 43 peptides met all of these criteria. Regions of altered deuterium uptake were mapped onto a co-crystal structure of LSD1/CoREST (PDB 2iW5).

Analysis of Flexibility in LSD1 Crystal Structures

All analysis was performed using the Bio3D-web server. By manual comparison to the RCSB PDB, all available LSD1 structures were selected as an input (see list below). Only structures containing point mutations or binding partners other than CoREST or a short histone peptide were excluded. Additionally, only the LSD1 chain from each structure was included in analysis. All chains were structurally aligned using the invariant core mode and visualized as a heatmap of RMSF value per residue. RMSF values were also plotted as a function of residue position. Additional analyses of (i) only isolated LSD1 structures or (ii) only LSD1/CoREST crystal structures were also performed and these RMSF plots overlaid to illustrate the decrease in variability within these two states. All structures were clustered by RMSD using a single-linkage method and represented as a dendrogram. PCA analysis presented the same groups, with the majority of separation attained from PC1, for which the largest contribution was derived from tower resiudes (data not shown). Structures included in analysis include: isolated LSD1 structures 2H94, 2HKO, 2DW4, 2Z3Y, 2EJR, 2Z5U, 3ABU, 3ABT and LSD1/CoREST co-crystal structures 2IW5, 2UXN, 2UXX, 4KUM, 2V1D, 2Y48, 5L3E, 5L3F, 5L3G, 5IBQ, 2XAF, 2XAJ, 2XAG, 2XAH, 2XAQ, 2XAS, 4UXN, 3ZMS, 3ZMT, 3ZMZ, 3ZN0, 3ZN1, 4UV8, 4UV9, 4UVA, 4UVB, 4UVC, 3ZMU, 3ZMV.

Production of a SELEX Library

Initial ssDNA template ordered as follows, where in N40 represents 40 sequential randomized bases:

Template:
5′-TCTCGGATCCTCAGCGAGTCGTCTG-N40-CCGCATCGTCCTCCCTA-3′

5′ Primer: 5′-GGGGGAATTCTAATACGACTCACTATAGGGAGGACGATGCGG-3′

To produce the library, the 5′ primer is initial annealed to the DNA template (95 °C, 5 min) then snap-cooled (4 °C, 5 min).

Table 5: Conditions for annealing the 5' primer to the library template

Reagent	Stock Conc.	Final
Annealing Buffer	10X	1X
DNA Template	200 μM	1 nmole
5' Primer	300 μM	1.5 nmole
Nuclease-free Water		
Vt		

Table 6: dsDNA library preparation conditions with Klenow fragment

Reagent	Stock Conc.	Final
Annealing Reaction		
NEB2 (Klenow) Buffer	10X	1X
dNTPs	10 mM	0.5 mM
NEB Exo$^-$ Klenow	5 U/μL	30 U
Nuclease-free Water		
Vt		500 μL

After annealing, the dsDNA library reaction is prepared and incubated to allow for the reaction to occur (37 °C, 1.5 hours). After halting the reaction with EDTA (2

mM), the DNA is extracted with a PCA extraction wherein 800 μL of a phenol:chloroform:isoamyl (25:24:1) is added to the reaction and shaken vigorously. The mixture is spun in a table-top centrifuge (14,000 rpm, RT, 10 min), and the aqueous layer is extracted. The aqueous layer is further extracted in a similar fashion but with 800 μL chloroform:isoamyl alcohol (24:1). The resulting aqueous extraction is washed 3x with TE buffer in an Amicon-4 (10 kDa MWCO) and the concentrated library is measured via UV-visible spectrophotometer (SpectraMax i3x micro-volume microplate).

To transcribe the library to RNA, the reaction was assembled as described in Table 7 and 8 and allowed to proceed at 37 °C overnight. To remove template DNA from the reaction, 20 U of DNase was added to the reaction (NEB M0303) which incubated further (37 °C, 30 min) upon which time the reaction was halted with 2 mM EDTA and a PCA extraction performed as previously described with chloroform:isoamyl alcohol (24:1). The RNA was further gel purified with a homemade 12% denaturing acrylamide-urea gel in 0.5X TBE buffer (220 V, 5 hrs). The RNA was excised and transferred to a 15 mL conical tube and subsequently broken into pieces. TE was added to the tube, sample frozen on dry ice and then incubated with rocking overnight at 4 °C to allow for RNA elution. The RNA solution was subsequently filtered through a 0.45 μm cellulose acetate membrane and washed 3x with TE in swinging bucket rotor (3700 *x g*, RT, 15 min). The concentrated RNA library concentration was measured via UV-visible spectrophotometer (SpectraMax i3x micro-volume microplate).

Table 7: T7 RNA Polymerase reaction to produce initial RNA library

Reagent	Stock Conc.	Final
T7 RNAP Buffer	5X	1X
2'F NTP Mix	10X	1X
DNA Library	16.23 μM	250 pmole
IPPI		2.5 U
Y639F T7 Mutant		4 μL
Nuclease-free Water		
Vt		500

Table 8: 5x T7 RNA Polymerase reaction buffer components

Reagent	Final	Stock
Tris-HCl	200 mM	0.5 M
MgCl2	60 mM	1 M
PEG 8000	20% (w/v)	8000 g/mol
Spermidine HCl	5 mM	254.63 g/mol
Triton X-100	0.01% (w/v)	10%
Nuclease-free Water		
pH	8.0	
Vt	2 mL	

SELEX-selection round

A reaction containing 2 nmol RNA library and 10x binding buffer (Table 9) were incubated with ¼ of a nitrocellulose filter membrane (Sigma WHA10402506) to pre-clear and remove membrane-binding RNA (37 °C, 15 min). After briefly spinning, the reaction mixture was passed through a pre-wet 0.45 μm nitrocellulose GVS Centrex centrifuge filter (1500 rpm, RT, 5 min) and the reaction volume was adjusted to 100 μL with 1x wash buffer (Table 10). Protein and RNA were combined and incubated (37 °C,

15 min) and applied to a pre-equilibrated nitrocellulose membrane in a large vacuum manifold (200 torr) and washed with 5 mL 1x wash buffer. The protein-bound RNA was eluted from the membrane via PCA extraction as described above. Following extraction, the RNA was precipitated with 10 μM linear acrylamide followed by 3 M sodium acetate (20 μL) and 1 volume 100% cold ethanol. The precipitated RNA was pelleted (14,000 rpm, 4 °C, 30 min), and washed with 70% cold ethanol. The resulting pellet was dried on benchtop and resuspended in TE buffer before concentration was established via UV-visible spectrophotometer (SpectraMax i3x micro-volume microplate).

Table 9: 10x SELEX Binding Buffers

	Stock	Buffer D	Buffer E	Buffer E/F	Buffer F
HEPES (pH 7.4)	0.5 M	200 mM	200 mM	200 mM	200 mM
NaCl	5 M	250 mM	500 mM	1000 mM	1500 mM
$CaCl_2$	1 M	20 mM	20 mM	20 mM	20 mM
BSA (Fraction V)		0.10%	0.10%	0.10%	0.10%

Table 10: 1x SELEX Wash Buffers

	Stock	Buffer D	Buffer E	Buffer E/F	Buffer F
HEPES (pH 7.4)	0.5 M	20 mM	20 mM	20 mM	20 mM
NaCl	5 M	25 mM	50 mM	100 mM	150 mM
$CaCl_2$	1 M	2 mM	2 mM	2 mM	2 mM

To produce RNA for the next round of selection, the purified RNA is first reverse transcribed to cDNA, PCR amplified, then transcribed to RNA using T7 RNA Polymerase. The reverse transcription proceeds via mixing the components in Table 11

together and incubating at 80 °C for 5 minutes before the addition of 10 mM dNTPs (2.5 μL) and 1 μL Roche AMV RT enzyme (37 °C, 45 min) and the enzyme subsequently inactivated (95 °C, 5 min).

Table 11: Reaction components for reverse transcription of RNA post-selection

Component	Stock	Final
Roche AMV RT Buffer	5X	1X
3' Primer	10 μM	4 μM
RNA library		1/4 rxn
Nuclease-free Water		
Total Volume		46.5

Following production of cDNA, 10 PCR reactions are generated utilizing the components in Table 12 with the addition of 5 μL of the resulting cDNA reaction to initiate the reaction which proceeds following the cycle sequence in Table 13.

Table 12: Reaction conditions for PCR amplification of cDNA

Component	Stock	Final
PCR Buffer	10X	1X
$MgCl_2$	50 mM	1.5 mM
dNTPs	10 mM	0.2 mM
5' Primer	10 μM	1 μM
3' Primer	10 μM	1 μM
Nuclease-free Water		
Platinum taq		2 U
Total Volume		

Table 13: Cycle parameters for PCR amplification of cDNA

Step	Temp. (°C)	Time (min)	# Cycles
Initial Denaturation	94	5 min	1
Amplification	94	0.5	22
Annealing	55	0.5	
Elongation	72	0.5	
Final Elongation	72	1	1
Hold	4	infinite	1
Total Time		*39*	

The resulting PCR products are buffer exchanged and concentrated in TE buffer utilizing an Amicon-4 (10 kDa MWCO) and the concentration assessed on a UV-visible spectrophotometer (SpectraMax i3x micro-volume microplate). At this point, the T7 RNA Polymerase reaction can proceed as outlined above to begin the next selection round.

SELEX-affinity binding assessment

32P labeling of RNA transcripts

Freshly transcribed RNA transcripts were dephosphorylated (100 pmol, 50 °C, 1 hour) in a buffer containing 50 mM Tris (pH 8.0) and 60 U of BAP (ThermoFisher, 18011-015). The resulting mixture was purified via SPRI RNA beads as outlined in appendix 2 and the concentration was quantified via UV-visible spectrophotometer (SpectraMax i3x micro-volume microplate).

Table 14: RNA 3' dephosphorylation conditions

100 μL Reaction	Stock	Final
Tris-Cl (pH 8.0)	1 M	50 mM
BAP	15 U/μL	60 U
RNA	-	100 pmol
Nuclease-free Water		v/vt

Dephosphorylated RNA (3 pmol, 37 °C, 30 minutes) was ^{32}P labeled utilizing γ-32 ATP (BLU502Z250UC, Perkin Elmer) in a reaction as outlined in Table 15. The resulting mixture was purified via Bio-Rad Micro Bio-Spin P-30 Gel column (Tris buffer, RNase-free, 7326250) and radioactivity levels were assessed using a Beckman LS 6000SC scintillation counter.

Table 15: T4 Kinase ATP, γ-^{32}P labeling reaction buffer

20 μL Reaction	Stock	Final
PNK Buffer	10x	1x
γ-^{32}P ATP	10 mCi/mL	60 U
NEB T4 Kinase	10 U/μL	20 U
RNA		3 pmole
Nuclease-free Water		v/vt

96-well filter-binding assay

Protein was diluted into a binding buffer containing 50 mM HEPES (pH 7.4), 200 mM KCl, 25 mM $MgCl_2$, 0.1 mM EDTA, and 0.1% BSA to a final concentration of 6.67 μM. The protein was diluted 2-fold to 156 nM and incubated at 37 °C for 5 minutes. The ^{32}P-radiolabeled RNA was diluted into 5x binding buffer and incubated at 65 °C for 5

minutes and then at 37 °C for 5 minutes. To each protein sample was added RNA (2000 CPM), and the protein:RNA mixtures were incubated at 37 °C for 20 minutes. During this time, the dot blot manifold was assembled from bottom to top with filter paper, nylon membrane (FisherScientific 50-905-0176) and nitrocellulose membrane (VWR 28151-760), each pre-equilibrated in a 1x wash buffer containing 50 mM HEPES (pH 7.4), 200 mM KCl, 25 mM $MgCl_2$, and 0.1 mM EDTA. The wells were rinsed with 100 μL of 1x wash buffer. The protein:RNA mixtures were applied to each well, aspirated at 400 PSI, and subsequently washed to ensure separation of bound and unbound RNA. Membranes were wrapped 1x in saran wrap and exposed to the phosphor storage screen (GE Healthcare, 2828956475) for 1.5 hours. The screen was scanned by a Typhoon imaging machine and analyzed via ImageQuant v 5.2 software.

4. Defining a binding interaction between KDM1A and the long non-coding RNA HOTAIR

Epigenetic complexes are primarily composed of proteins with domains to increase specificity and selectivity of the catalytic machinery. The reader domains of proteins aid in targeting genes with specific DNA or histone PTMs, allow for regulatory control over which sites the epigenetic complex interacts with. However, there is still much about the molecular mechanisms of these events that remains unexplored. As the field of long non–coding RNAs (ncRNA) has expanded rapidly in the years since the ENCODE project began in the early 2000s and first published ncRNA in 2012, we have begun to understand that these ncRNA can play an additional role in epigenetic regulatory events including targeting epigenetic complexes to specific gene sites *(289–291)*.

One lncRNA, HOTAIR, directs the EZH2-containing PRC2 complex as well as KDM1A-containing complexes to HOXD gene loci. There has been significant work to establish the molecular interface between PRC2 and HOTAIR and evidence that disruption of this interface could decrease cancer cell viability *(153,164,165,251,252,292)*. However, there has been comparatively no work performed to establish the KDM1A:HOTAIR interface *(154,293)*. I have been able to make significant advances in elucidating the KDM1A:HOTAIR interaction.

In this chapter I will introduce the importance of non-coding RNA, and their roles within cells. I will discuss the importance of HOTAIR within the cell and the

current efforts to establish its molecular role and disruption of its protein:protein interactions. Finally, I will report my utilization of an electromobility shift assay (EMSA) to establish the minimal protein unit required for HOTAIR to bind. I found that CoREST is required for HOTAIR to bind. I also confirmed that only the first 300 bp of the KDM1A-binding domain of HOTAIR (D4) is required for protein-binding. I propose a region of protein-binding based on computational analysis and discuss the implications herein.

4.1 Introduction to non-coding RNA

While over 75% of the human genome is transcribed into mRNA, it is well-established that only 1-2% of those transcripts are translated into proteins. While an estimated 90-95% of the remaining non-coding mRNA is designated for ribosomal and transfer RNA (rRNA and tRNA, respectively), there was still a significant percentage of RNA that was, until recently, considered 'junk' non-coding RNA (ncRNA) *(294–296)*. However, through a massive paralleled sequencing endeavor in 2012, the ENCODE project set the stage for establishing that much of the remaining ncRNA had the potential to function in cellular functions *(289,297)*. Due to the large size variation of ncRNA, the distinction between long non-coding RNA (lncRNA) and the abundance of short non-coding RNA is an arbitrary cutoff at 200 nt. Both short and long non-coding RNAs have been demonstrated to have important biological functions as discussed below.

4.1.1 The roles of short non-coding RNAs in cells

Micro RNA (miRNA) and small interfering RNA (siRNA) are two classes of short non-coding RNAs which play a role in the epigenetic regulation of protein production, primarily via antisense recognition of mRNA. Each of these classes have been thoroughly reviewed elsewhere *(298–301)*; however, we will briefly discuss their product, structure, and function with respect herein.

4.1.1.1 micro RNA

Initial discovery of miRNA through their role in regulation of mRNA translation

Micro RNA (miRNA) are a class of 21-25 nt single stranded RNAs (ssRNA) which regulate the translation of mRNA. While the role of these genes had been shown via gene knockdown in the 1980s, it was not fully appreciated that these were non-coding genes until the characterization of lin-4 in *Caenorhabditis elegans* by the Ambros lab at Harvard University *(302,303)*. Previous studies had made clear that deletion of a 693 nt sequence, lin-4, produced a gain-of-function of lin-14 – a protein involved in *C. elegans* developmental growth *(302)*. The 1993 study established that not only was lin-4 within an intronic region, it was unlikely that the sequence was coded. Further, they established the production of 2 RNA transcripts, lin-4S and lin-4L, 22 and 61 nt in length, respectively *(302)*. While the exact mechanism of how lin-4S and lin-4L were able to negatively regulate lin-14 protein expression was unknown, the authors proposed an RNA:RNA interaction due to a partial antisense complementary between the lin-4 RNA

transcripts and lin-14 RNA. They noted that the decreased lin-14 protein expression occurred between L1 and L2 of *C. elegans* developmental stages, which correlated with an increase in the levels of lin-4 *(302,304)*. More recent studies have confirmed these original findings, that the miRNA lin-4, originally named lin-4S, negatively regulates lin-14 through an antisense interaction in the lin-14 3'UTR *(305,306)*.

Role of miRNA in biological and pathobiological functions

Since the mechanism of miRNA interference was first noted in lin-4, there have been a significant number of not only miRNA that have been established, but also proteins and pathways that they affect. One of the hallmarks of an miRNA:mRNA interaction is an imperfect binding which allows for promiscuity within miRNA targeting *(300,307)*. Thus, as miRNA can have many cellular targets, this convolutes establishing specific roles or how specific interactions with mRNA affect the cell. We will briefly discuss two miRNA interactions that have been shown to involve the mRNA of KDM1A are miRNA-329 and miRNA-137.

miRNA-329 has been implicated in regulating angiogenesis and recently reported to potentially play a role in gastric cancer. One study rigorously studied the potential interplay between miRNA-329 and KDM1A in gastric cancer *(308)*. Utilization of five separate gastric cancer cell lines revealed that in each of these, the level of miRNA-329 was significantly decreased compared to the control cell line whereas the levels of KDM1A mRNA was significantly increased *(308)*. Furthermore, a luciferase

assay suggested targeting of WT miRNA-329 to a KDM1A–containing plasmid over mutant miRNA-329 when the expression of miRNA-329 was forced in these cell lines via transfection, and short-interfering RNA of KDM1A were utilized, not only was colony formation mitigated and apoptosis enhanced, but tumor formation in mice was inhibited *(308)*. Interestingly, apoptosis was higher in cells with knockdown of KDM1A via siRNA over those with both KDM1A knockdown and miRNA-329 overexpression. This speaks to the many potential roles of miRNA within a cell upon expression *(308)*.

Another miRNA, miRNA-137, has been implicated in due to epidural anesthetic treatment of bupivacaine *(309)*. Upregulation of miRNA-137 was demonstrated upon treatment of the anesthetic both in a clinical setting and on dorsal root ganglion (DRGN) cells *in vivo (310)*. The levels of KDM1A were shown to be downregulated upon increased levels of miRNA-137 *(309)*. A knockdown of miRNA-137 via an antisense oligo approach with a parallel transfection of KDM1A to increase KDM1A levels revealed a disruption in normal growth of the DRGN cells while a decrease miRNA-137 alone upon introduction of bupivacaine demonstrated a neuroprotective effect. Similarly, overexpression of KDM1A alone demonstrated a significant decrease in the rates of apoptosis *(309)*. It should be noted that another study showed that miRNA-137 was increased upon introduction of ketamine into a system, however, they noted a hippocampal neuroprotective affect. These conflicting results warrant further studies,

however, it is not unlikely that the same miRNA can have opposing affects in different cell types based on the number of targets of each miRNA *(310)*.

4.1.1.2. Short interfering RNA

Short interfering RNA (siRNA) are mostly studied in their context of knockdown studies in *in vivo* studies, and while they are also produced endogenously, there is significantly less known about specific siRNA *(311,312)*. However, they do have several distinct differences from miRNA that do make them a distinct class and worth noting. First, pre-miRNA is composed of an asymmetrical stem, which results in a bulge, with an apical loop. The asymmetrical stem allows for distinction between the guide strand, with the bulge, and the passenger strand, without the bulge. siRNA are dsRNA which are formed without an apical loop but are an exact match between the two sides of the RNA *(313)*. Both additionally form distinct RISC complexes, however, while miRNA are able promiscuous with their targets, siRNA only have one target mRNA with which their sequence is antisense to *(307)*. This complete complementary between siRNA and target mRNA creates a much longer interface upon binding and allows for the Argonaut protein 2 to cleave the mRNA, resulting in downregulation *(307,314,315)*.

4.1.2. The roles of long non-coding RNAs in cells

The cut-off for long non-coding RNA (lncRNA) has been arbitrarily established to be greater than 200 nt. It has been established that many lncRNA are spliced and polyadenylated as a coding RNA would be. A lack of a Kozak translation initiation

sequence appears to preclude these transcripts from translation *(294,316)*. However, there is evidence that some lncRNA are translated, though this is still a highly debated area *(298,317,318)*.

The field of lncRNA and their roles and functions within cells is still a rapidly expanding field and there is still much we do not understand about the molecular basis for many of these functions in cells. It is likely that lncRNA, like miRNA, may have multiple targets or functions depending on cellular context. The roles of lncRNA have been closely linked to epigenetic transcriptional expression and repression via: target mimicry, mRNA processing, mRNA disruption, and epigenetic complex scaffolding. Below we will explore how lncRNA have been implicated in each of these processes and new areas of these functions that still need to be explored and established.

Target mimicry

In the case of target mimicry or target 'decoy,' lncRNA base pair with to titrate the level of effector proteins such as transcription factors (TF) or RNA–binding proteins thus inhibiting the role of the effector proteins in the cellular context. TF can target a wide number of genes and a shift in the level of protein available can result in genome-wide changes in expression. One example of this is between the lncRNA P21-associated NcRNA DNA damage-activated (*upst:CDKN1A:-4845*, PANDA) and the transcription factor NF-YA *(319)*. Nuclear transcription factor Y subunit (NF-Y) is a bidirectional CCAAT sequence binding effector protein composed of three subunits to form a

heterotrimeric complex with subunit a (NF-YA) providing the sequence specificity *(320–324)*. NF-YA has been linked to regulation of genes involved in cell cycle (*topo IIα*, *cdc2*, cyclins and *cdc25C*), regulation of embryonic basement membranes (*Sox7*, *Sox17*), hematopoiesis (gamma (γ)-globin genes (*HBG1* and *HBG2*)), as well as apoptotic pathways (*Sox2*, transforming growth factor β type II receptor, *FAS*) *(325–330)*.

In exploring the role of *CDKN1A* in DNA damage response, Hung et al showed that a 1.5 kb RNA was transcribed by p53 transcriptional activation, but they could find no subsequent evidence of translation of the RNA. Upon further investigation, they established that this was a non-coding RNA, PANDA, able to sequester NF-Y via interaction with NF-YA, which in turn repressed transcription of pro-apoptotic genes presumably to allow for the DNA damage response to ensue *(331)*. While this link has been established, there are additional studies which have cited an increase of PANDA leading to changes in transcriptional activation that were NF-Y independent. Thus, it will be interesting to establish how the NF-YA functionality of PANDA is limited to specific cellular contexts as well as what the other functions of PANDA are *(319,332,333)*. Furthermore, there is still much work that needs to be completed to establish not only the molecular mechanism of this interaction. NF-YA contains the CCAAT recognition sequence, but it is unclear whether there is a previously uncharacterized RNA-binding domain of NF-YA.

Insight into the potential mechanism of the PANDA:NF-YA binding may be found by looking at another lncRNA, growth arrest-specific 5 (Gas5). Gas5 has been reported as a sensitizer for apoptosis, and shown to be downregulated in breast cancer *(334)*. Part of the mechanism for the pro-apoptotic function of Gas5 is likely due to its binding to the DNA binding domain (DBD) of the glucocorticoid nuclear hormone receptor (GR) *(335)*. The interaction between GR and Gas5 was suggested to stimulate a sensitization to pro-apoptotic factors by mitigating the anti-apoptotic potential of glucocorticoids. GR bind to the glucocorticoid receptor element (GRE), and computational studies with Gas5 suggest a dsRNA region that mimics the GRE. While there is only a crystal structure of the GRE:GR interaction, modeling predicts that a critical G:C base pairing in the canonical interaction is conserved in the RNA mimic. This is particularly interesting because dsRNA generally forms an A-form helix that is 20% wider than that of A-form DNA, thus making it unlikely that an RNA would be able to mimic the secondary structure of DNA *(336)*. Similar to PANDA, Gas5 has been implicated in transcriptional regulatory functions independent of its interaction with the GR *(337)*.

These are likely not the only instances of mimicry by lncRNA mechanism of mimicry by lncRNA; however, what is clear is that the tools to assess what functions lncRNA have *in vivo* are extremely context specific and will require significant

validation and mechanistic studies to truly understand how they are targeted or targeting specific proteins within a cell.

mRNA processing

Before transport to the cytoplasm, mRNA go through several processing steps to increase their stability, 5'capping and 3' polyadenylation, as well as splicing to remove the introns. However, many proteins have several different isoforms which stems from differential splicing of introns referred to as alternative splicing (AS) which occurs in a cellular context-specific manner in many cases *(338–340)*.

Nuclear paraspeckle assembly transcript 1 (NEAT1) is a 4kb lncRNA which has been implicated in the formation and stability of nuclear paraspeckles *(341–343)*. The function of paraspeckles is still being established, though it is likely the function to localize proteins within nuclear compartments, and there is evidence that nuclear paraspeckles play a role in pre-mRNA splicing *(342)*. Regardless of the role of NEAT1 in paraspeckles, the lncRNA has a function in adipogenesis *(344)*. Abolishment of NEAT1 knockdown with siRNA showed a significant decrease in the production of PPARγ2, which is a key protein in the establishment of adipogenesis in response to glucose metabolism along with its isoform PPARγ1. Interestingly, in response to insulin, the levels of NEAT1 showed a correlation with the phosphorylation levels of an insulin-associated splicing protein, Serine and Arginine Rich Splicing Factor 40 (SRp40) *(344)*. Cooper et al noted that knockdown of SRp40 lowered levels of both PPARγ2 but

overexpression of SRp40 only affected levels of PPARγ2. While they performed no binding experiments, they proposed that an interaction with NEAT1 primes SRp40 for phosphorylation, impacting which in turn activates the protein and alters the balance between the production of the two PPAR isoforms *(344)*. Future studies will be critical to truly establish if SRp40 and NEAT1 are interacting, or if there is an additional interacting factor in this mechanism.

Other proteins such as MALAT1 (NEAT2) , LINC01133, GOMAFU have been shown to have similar functions to NEAT1 where an interaction with a splicing factor affects the outcome of isoform formation. Each of these four lncRNA have also been shown to be dysregulated in disease states such as hepatocellular cancer (MALAT1/NEAT2), schizophrenia (GOMAFU), and colorectal cancer (LINC01133) *(345–347)*. MALAT1 and LINC01133 have also been implicated in binding to binding epigenetic enzymes such as EZH2, and KDM1A in separate cancer-types, again bringing to light the diverse functions that each lncRNA can retain depending on context within the cell *(348,349)*.

miRNA disruption

In addition to target mimicry in the nucleus via targeting effector proteins, lncRNA have been shown to act as mRNA target mimics in the cytoplasm. As miRNA do not require full complementary to affect the processing of an mRNA, it is not surprising that lncRNA are able to interact with miRISC complexes to affect the

availability of mRNA for translation. These lncRNA are specifically referred to as competing endogenous lncRNA (ceRNA) *(350–352)*.

One example of this is within differentiating and self-renewal cells. Key protein markers in undifferentiated cells including the transcription factors Oct4, Sox2, and Nanog. The levels of these proteins, and thus the differentiation potential of a cell, is mediated in part by miRNA-145 which targets the mRNA of these proteins for degradation *(353,354)*. The lncRNA lincRNA-RoR had been previously suggested to perpetuate the self-renewal potential of cells in an undifferentiated state. lincRNA-RoR contains two separate sites with a high binding potential to miRNA-145 and mutation of these sites was shown to increase the levels of Oct4, Sox2, and Nanog and decrease the number of differentiated cells *(354)*. Interestingly, while not statistically significant, there was an increase between mutating only one of the miRNA-145 interaction sites and mutating both at the same time. This suggests that lincRNA-RoR may utilize both of these binding sites for targeting miRNA-145 *(355)*.

Epigenetic complex scaffolding

While each PTM may be associated with either transcriptional activation or repression for a gene, it is in fact, a concert of multiple PTMs on a gene's enhancer and promoter regions which sum together to either promote or repress transcription. This 'histone code hypothesis' relies on rapid and precision addition and removal of each PTM to match the ever-changing landscape of the cellular environment requiring

multiple epigenetic enzymes to function together. One method the cells have adopted to increase the efficient turnover of multiple PTMs is through multimeric complexes which contain multiple enzymes. Thus, a singular complex can interact with a histone and lead to changes at multiple histone residues, potentially even spanning multiple nucleosomes *(250)*.

Another technique that is utilized by cells to increase the synchrony of PTM changes and transcriptional regulation is through lncRNA scaffolding whereby multiple epigenetic complexes with RNA binding regions interact with different regions of the same lncRNA which is capable of targeting a specific subset of genes.

One of the best studied examples of this is also one of the first examples of not only a trans-acting lncRNA was of HOX transcript antisense RNA (HOTAIR) which is transcribed from the HOXC locus, but functions along HOXD loci. The homeobox (Hox) genes, which are critical for embryonic development in mammals, are evolutionarily conserved from plants and fungi, although their genes are not clustered as they are now in animals *(356,357)*. The first protohox cluster was established upon gene duplication in *Cnidaria (356)*. The number of Hox gene clusters has expanded from *Caenorhabditis elegans* with one cluster with only five genes to *Homo sapiens* with four clusters of 39 genes *(356,358)*. Presently, there are 10 Hox genes in humans which have been linked to a specific disease due to gene mutations *(359)*. While the mechanism is unclear, there is also evidence that changes in Hox gene expression can play a fundamental role in cancer *(356,360,361)*.

It had been long established that H3K27me3 marks were important for specifically Hox gene regulation, and a large-scale screen for lncRNA along the HOX-coding regions of human chromosomes by the Chang lab in 2010 revealed that a lncRNA localized to the HOXD locus and could potentially be affecting gene transcription in at least the HOXD loci *(154)*. This work implicated a relationship between HOTAIR, KDM1A, and SUZ12. siRNA silencing of HOTAIR and subsequent ChIP-seq assays in a parallel study suggested that not only were KDM1A and SUZ12 colocalizing to the same genes in a time-dependent manner, but that their localization was in response to the presence of HOTAIR *(163)*.

4.1.3 The HOX antisense transcript RNA was the first discovered trans-acting lncRNA

The discovery that HOX antisense transcript RNA (HOTAIR) was a lncRNA transcribed in the HOXC loci but localized to the HOXD loci was important as it was the first time a lncRNA was reported to function in trans. The Chang lab established that the changes in H3K27me3 and H3K4me2 levels due to the knockdown of HOTAIR was precipitated from an interaction with EZH2 and LSD1. Indeed, the 2148-nt long lncRNA was able pull-down EZH2 and LSD1 *(362)*. Further, *in vitro* production of smaller HOTAIR fragments identified that nucleotides 1-600 were able to interact with both EZH2 and nucleotides 1500-2148 were required for LSD1 binding. Interestingly, while this study showed that a 1821-2148 fragment was unable to bind to LSD1, it was not shown whether 1500-1820 alone was able to bind to LSD1 *(163)*.

HOTAIR is an antisense transcript on chromosome 12 band 12q13 partially overlapping HOXC11. Comprised of six exons in humans, regions of HOTAIR appear to be at least partially evolutionarily conserved in mammals *(251,363)*. The Pyle lab noted in 2015 that there was a significant level of covariation at the highly structured regions of HOTAIR, particularly in regions reported to be involved in protein binding. For example, 45% of nucleotides in a 30 bp helix (Helix 7) within HOTAIR D1, and two other helices retain 52 and 57% covariance across 33 mammalian species retaining HOTAIR. Helix 7 also appeared conserved in mouse HOTAIR (mHOTAIR), which is conflicting with other reports regarding the function of HOTAIR *(251,364)*. While there was reported sequence covariance by the Pyle lab, the Duboule lab reported in 2011 that mHOTAIR had not retained exon 1 of HOTAIR and lost significant PRC2 binding functionality *(365)*. Further, they reported that upon knockdown of the HOXC loci in cells, they did not note changes H3K27me3 in the HOXD loci. This study was replicated in 2016 with mice where more significant changes in gene transcription in HOXC11 and HOXC12 where the *HOTAIR* gene is positioned were noted than, again, in the HOXD loci where HOTAIR has been reported to be functionally relevant for humans *(366)*.

Even though the evolutionary conservation of lncRNA overall is much more complex, this has brought some debate as to the functional importance of HOTAIR in mice, and thus, in humans. In the 2011 study, they proposed that the knockdown of the entire HOXC loci could have altered another transcript that compensated for the lack of

HOTAIR; however, the knockdown in the mice was a much smaller region so this is a less likely explanation *(365)*. The Duboule lab hypothesized that the lack of notable change in gene expression may be due to a significant change in HOTAIR between mice and humans resulting in a change in functionality of the enzyme *(364)*.

It should also be noted that the level of H3K4me2 was not reported in either of these studies. It would be interesting to see if these levels at HOXD loci, presumably due to the HOTAIR:KDM1A interaction, were also unaffected when HOTAIR was knocked-out. Overall, it is clear that there is more work that needs to be done to understand how HOTAIR has evolved and its functional importance in other mammals. Clearly, establishing the relationship between sequence or structure conservation and covariance of non-coding RNA to function and evolution has proven more challenging than with proteins. This is, in part, due to the flexible nature of RNA and their inherently more flexible structures than proteins. Particularly for lncRNA it is still unclear how much a change in sequence informs a change in structure. If proteins are recognizing a functional element such as a stem or hairpin on a lncRNA, as long as the overall structure is retained, it may be that the function is conserved.

While the functional importance of HOTAIR in mice has been somewhat debatable, the Pyle lab has reported a structural analysis of HOTAIR *(251)*. They performed three footprinting techniques in parallel, SHAPE, DMS, and Terbium cleavage to establish the secondary structure of the lncRNA. They demonstrated further that there

are four distinct domains of HOTAIR wherein Domain 1 and Domain 4 span the nucleotides suggested to interact with PRC2 and LSD1, respectively, by the Chang lab. The analysis reveals that HOTAIR is highly structured with over 50 helices, 72 loops (terminal and internal) and 19 junction points *(251)*. The high level of structure to HOTAIR is contrary to other well-established lncRNA such as Xist that is proposed to bind proteins via a sequence repeat. The Pyle lab proposed that the highly structured HOTAIR may, in fact, interact with proteins through a more complex interplay between both sequential and structural elements *(251)*.

4.1.4 HOTAIR is dysregulated in several pathobiological and cancer phenotypes

Similar to many epigenetic enzymes and complexes, since it's discovery in 2010, HOTAIR has been implicated in a series of cancer phenotypes (Table 16). In most cases, upregulation of HOTAIR has been noted, and it's potential for a biomarker in several cancer types including digestive cancers and esophageal squamous cell carcinoma has been explored and proposed. Generally speaking, we see that an upregulation of HOTAIR leads to a poorer patient outcome. While the mechanism for HOTAIR within each cancer phenotype is beginning to be established, most current evidence only suggests a link between HOTAIR and the cancer.

Table 16: Cancer phenotypes HOTAIR has been implicated in

Cancer Type	*Role of HOTAIR*	*Reference*
Renal cell carcinoma	Sponges miRNA-124	*(367)*
Oesophageal squamous cell carcinoma	Sponges miRNA-125 and miRNA-143	*(157)*
Hepatocellular carcinoma	SNP intronic region Role in EMT Sponges miRNA-23b-3p	*(158,368–370)*
Glioma	Sponges miRNA-126-5p	*(371,372)*
Oral carcinoma stem cells	Role in EMT	*(89)*
Prostate cancer	Sponges miRNA-34-a Interaction with EZH2 and DNMT1	*(374,375)*
Breast cancer	Role in HMGA2 axis Sponges: miRNA-20a-5p and miR-206	*(152,154,253,376)*
Colorectal cancer	Role in Wnt/β-catenin pathway Role in Bcl-w signaling pathway Sponges miRNA-203a-3p	*(155,377)*
Ovarian cancer	Estradiol induced up-regulation	*(159,165)*
Pancreatic cancer	Gemcitabine induced up-regulation	*(228,378,379)*
Cervical cancer	Interaction with STAT3	*(380–384)*
Gallbladder cancer	Sponges miRNA-130a	*(162)*
Gastric cancer	Role in NF-κB pathway	*(160,161,379,385,386)*
Retinoblastoma	Role in EMT miRNA-613/c-met axis	*(387)*

As is evident from Table 16, while HOTAIR binds KDM1A and EZH2, HOTAIR is capable of acting as a molecular sponge and miRNA disruption. However, the molecular basis for many of these interactions, and how these interactions lead to a

cancerous phenotype is less than clear. Interestingly, it may be possible that some of the miRNA implicated overlap with protein-binding regions of HOTAIR, disrupting a structural element or precluding a sequence-specific interaction to occur between HOTAIR and a scaffolding protein.

Single Nucleotide Polymorphism of HOTAIR

One interesting case in hepatocellular carcinoma (HCC) and the role of HOTAIR involves an intronic region that has been established to be a splicing enhancer of the initial transcript. In Chinese populations, it was shown that a C to T single nucleotide polymorphism (SNP) , genotype rs920778, increased the risk of HCC, and increasingly so in those that had ever drank versus those who had never consumed alcohol *(381)*. This genotype and associated likelihood of poor survival outcome has also been implicated in Chinese populations for breast cancer, cervical cancer, esophageal squamous-cell carcinoma (ESCC), and gastric cancer *(159,161,368,381,386,388)*. In the case of gastric cancer and ESCC, it was additionally shown that the T allele was associated with an increased level of HOTAIR *(386,388)*. However, a study on a Turkish population showed no statistical significance with the T allele in gastric cancer, and a study on a Portuguese population with glioma's observed a statistically significant lower survival rate in both rs920778 TT and CC patients than CT patients but not between TT and CC patients *(371,385)*.

While the T allele seems correlated with a poor prognosis in Chinese populations, it is unclear why these results do not hold constant across different populations. It could be due to sample size or population differences. While note of the SNP is of potential value towards understanding the functional role HOTAIR plays within disease, there is still a significant amount that remains unknown about SNPs and their overall role in disease susceptibility and variation.

4.1.5 The HOTAIR:PRC2 interface suggests a benefit for disrupting lncRNA:epigenetic complex interactions

To understand more completely the role of the HOTAIR:PRC2 interaction in a physiological as well as a pathophysiological cell type, disruption of the HOTAIR:PRC2 complex has been extensively explored. A 2013 study established a minimal binding unit of both protein and HOTAIR RNA that was required for the HOTAIR:PRC2 interaction *(252)*. Utilizing the 300-nucleotide unit of HOTAIR previously reported to bind PCR2 (nucleotides 1-300), the Balasubramanian lab demonstrated that while the trimethylase, EZH2, alone can bind to $HOTAIR_{1\text{-}300}$ at a K_d of 755 ± 45 nM this was lower than the minimal active 3-mer PRC2 unit of EZH2:EED:SUZ12 at 165 ± 16 nM *(252)*. EZH2:SUZ12 demonstrated a binding constant similar to EZH2 alone, suggesting that SUZ12 plays no role in HOTAIR binding whereas EZH2:EED interacted at HOTAIR at a similar K_d to the 3-mer unit *(252)*. This is interesting because EED alone had a K_d >5μM.

The Lee lab at Harvard Medical School further showed that EZH2 is capable of binding non-specific RNA, but selectivity is increased when EZH2 is bound with EED or

SUZ12 *(389,390)*. There have been no further structural studies to assess which residues or regions are involved in HOTAIR binding, but it should be noted that there are no canonical RNA-binding domains present in either EZH2 or EED. Additionally, separate from EZH2 and EED, coregulatory protein JARID was demonstrated via PAR-CliP and RIP-qPCR to bind to lncRNA Meg3 *(391)*. It is unclear whether PRC2 can interact with both lncRNA at the same time, or if binding is mutually exclusive.

Exploring the region of HOTAIR that specifically interacts with PRC2 has additionally been explored. After establishing that EZH2:EED was the minimal binding unit of HOTAIR D1, the Balasubramanian lab at the University of Cambridge established via radio-labeled EMSA assays that the minimal binding unit of PRC2 was an 89-mer (nucleotides 212-300) *(252)*. While the secondary structure of the 89-mer was established via complementary RNase I and RNase 4 which cleave single stranded RNA (ssRNA) and double stranded RNA (dsRNA), respectively, this is not representative of the structure that the Pyle lab established in 2015 *(251,252)*. Specifically, between the two structures there is a low conservation of paired and unpaired nucleotides, thus it is unclear whether PRC2 is binding a structural element is lowly conserved the 89-mer fragment or is dependent on sequence.

In a follow-up study, peptide-nucleic acids (PNA) complementary to regions of 89-mer fragment were produced *(165)*. PNAs have the benefit of being intractable to both nucleases and peptidases, making them more attractive for *in vivo* work and

potential therapeutics. In this study, one PNA was able to ablate the HOTAIR:PRC2 interaction *in vitro.* Further, in prostate and ovarian cancer cell lines, utilization of the PNA resulted in decreased proliferation of the cells; and a smaller tumor size in mice with xenograft ovarian tumors was additionally recorded *(165)*. Similarly, a small molecule disrupter of the 89-mer region of the HOTAIR:EZH2 interface was able to decrease the growth of a MDA-MB-231 tumor in a mouse xenograft model in a study published earlier this year *(164)*.

These studies show merit towards disruption of these epigenetic scaffolding lncRNA:protein interactions with respect to disease. Since HOTAIR targets both EZH2 and KDM1A, we were interested in elucidating the KDM1A:HOTAIR interface in order to disrupt this interface. In conjunction with HOTAIR:EZH2 targeted disruption, we believe this could lead to, at least, an additive effect of targeting both epigenetic complexes.

4.1.6 Elucidation and disruption of the HOTAIR:KDM1A binding interface

With evidence that disruption of the HOTAIR:PRC2 interaction can mitigate the growth of cancerous cells and tumor size, we hypothesize that disruption of the HOTAIR:KDM1A interaction could produce a similar affect. Further, we are interested if disruption of both complexes simultaneously could display an additive, or even synergistic, effect in the cancerous cell types in which both complexes, as well as HOTAIR, have been implicated.

As there is little information regarding the molecular basis for the HOTAIR:KDM1A interaction, we sought to first establish this. Our goals for this project were to (a) establish which residues of KDM1A were responsible for interacting with HOTAIR, (b) which nucleotides of HOTAIR contributed to binding KDM1A, and (c) establish methods which could be utilized to disrupt the HOTAIR:KDM1A interface.

The remainder of this chapter will address our efforts to establish which residues of KDM1A are critical for HOTAIR binding via radiolabeled binding assays and initial cross-linking proteomics attempts. We have been able to successfully establish techniques to yield high levels of *in vitro* RNA in our lab. I have also established an EMSA binding assay to show the binding of HOTAIR D4 to KDM1A:CoREST$_{286\text{-}482}$ and establish a binding affinity of this interaction. I have shown that the first 300 bps of HOTAIR D4 are critical for binding KDM1A:CoREST$_{286\text{-}482}$, and that the SANT2 domain is not required for binding to occur.

4.2 Results

In order to establish the molecular basis for the KDM1A:HOTAIR interaction, we first hoped to utilize a binding assay to establish which proteins were required for HOTAIR binding. We utilized a radiolabeled electromobility shift assay (EMSA) to establish binding curves and demonstrate the requirement for KDM1A and CoREST to be bound for HOTAIR to bind.

4.2.1 Experimental Design

The Chang lab at Harvard University reported that Domain 4 was responsible for HOTAIR binding KDM1A through an *in vitro* pull-down assay with HOTAIR fragments as bait. We first sought to replicate these findings in our lab. While there are several viable techniques to utilize for general binding studies, we decided that an electromobility shift assay (EMSA) with ^{32}P–RNA would best suit our needs. The utilization of radioactivity provides a higher specificity and resolution than other tags such as fluorescence or biotin. Further, the utilization of EMSA over a filter binding assay as an initial study would allow us to ascertain if there were non-specific or multiple binding events occurring through the presence of multiple upward band shifts.

The general setup we followed for *in vitro* RNA radiolabeling and subsequent EMSA is displayed in Figure 57. Briefly, we chose to PCR amplification from a pcDNA3 plasmid that was gifted to us from the Chang lab at Harvard University. Within the PCR amplification of the HOTAIR fragment of interest, the primers we designed incorporated a T7 promoter sequence on the 3′ end for *in vitro* transcription on the and additional nucleotides with a restriction digest cut site on the 5′ end to promote primer binding and PCR amplification. Proceeding PCR amplification, the unpurified PCR fragments were subjected to consecutive DPNI and BAMHI (1 hr, 37 °C, NEB) digests to remove the template DNA as well as the extra 5′ nucleotides. BAMHI was chosen as the restriction enzyme as it's recognition site and cleavage specificity would leave only one

additional nucleobase, guanine, at the end of the RNA transcript. Following each digest, each fragment was examined on a 3% agarose gel to confirm production of the DNA fragment and subsequent removal of the additional 5′ nucleotides, several examples of this are displayed in Figure 58.

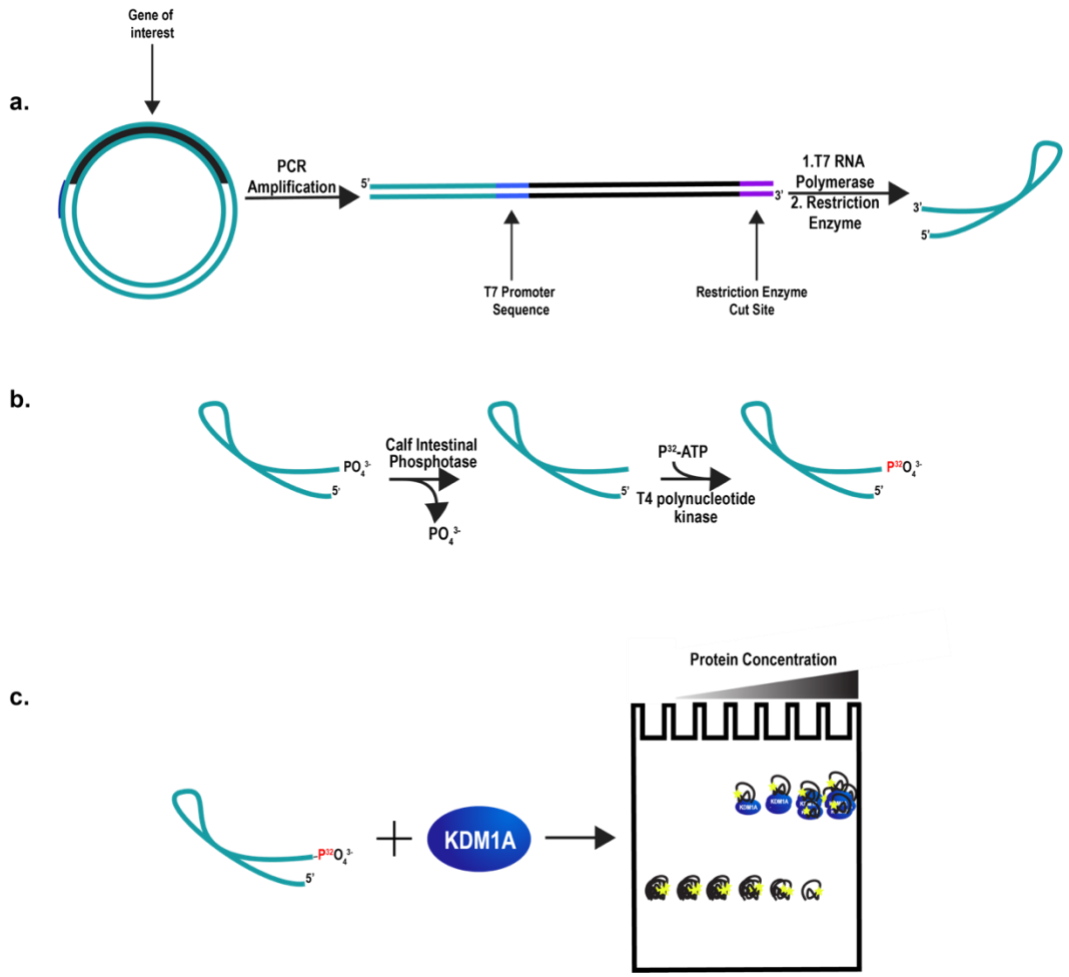


Figure 57: General scheme for *in vitro* RNA EMSA assay wherein (a) PCR amplification from a plasmid containing the gene of interest (HOTAIR) is performed to produce the desired fragments. Subsequent DPN1 and BAMH1 digestions are performed prior to an *in vitro* T7 RNA polymerase reaction. (b) Post–T7 RNA polymerase reaction and SPRI bead cleanup, a dephosphorylation of 100 pmol of RNA and subsequent T4 kinase reaction with ATP, γ-^{32}P, on 3 pmol of RNA were

performed. (c) following validation of radiolabeled RNA, an electromobility shift assay (EMSA) assay was performed utilizing a 6% non-denaturing acrylamide gel.

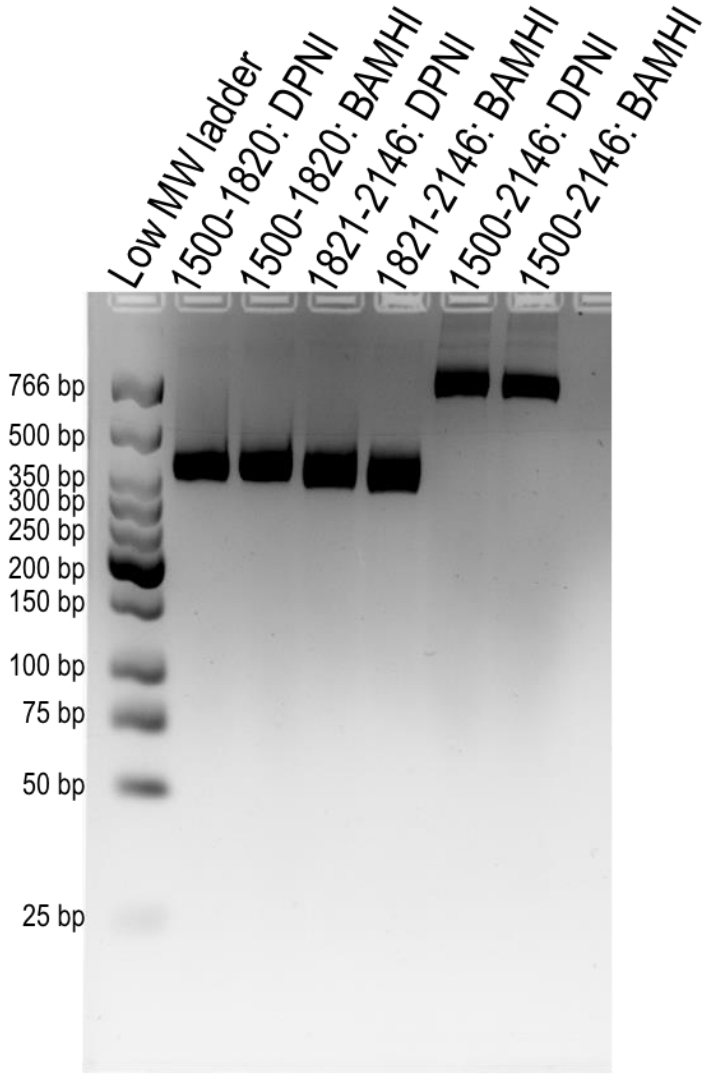


Figure 58: Results of PCR amplification of HOTAIR$_{1500-1820}$, HOTAIR$_{1821-2146}$, and HOTAIR D4, pre– and post–BAMHI digest on a 3% agarose gel (120V, 90 min), utilizing the NEB Low MW ladder (N3233)

Production of *in vitro* HOTAIR fragments

Upon DPNI and restriction digests, we next purified the PCR products with solid phase reverse immobilized beads (SPRI). While beads in the buffers specific for DNA and RNA purification are available for purchase, we purchased GE Sera-mag magnetic beads with a carboxylate modified surface (cat number 65152105050250). Described in Appendix a, we were able to utilize these beads as a homemade and extremely cost-effective version of the beads available for purchase. Incubation of the DNA product with 1.8x beads which contains PEG desolvates the RNA towards an interaction with the carboxylated surface of the beads. The beads can then be pelleted using a magnet, washed with 70% ethanol, and eluted at low volumes of water or TE buffer. Upon elution, concentration of the purified DNA product was established with the Micro-volume Microplate of the UV-visible spectrophotometer (SpectraMax i3x micro-volume microplate).

We initially utilized a store-bought T7 RNA polymerase (NEB M0251) with subsequent gel purification; however, the yield of these reactions was low, requiring a large number of reactions to be run. The Weeks lab generously gifted us with a T7 RNA polymerase plasmid along with a protocol for protein expression and purification that they utilized in their lab. The purification as well as initial tests and comparisons to the

store-bought enzyme are outlined in appendix a, but overall, upon utilization of the in-house T7 RNA polymerase protein and reaction buffer, we saw a 5.5–fold increase in RNA quantity utilizing a gel extraction for purification and 29.9–fold increase in RNA quantity utilizing RNA SPRI bead purification procedure (Appendix a, Table 34).

However, while our yield was greatly increased, we noted that now there were, what we hypothesized to be, truncation products (Figure 59). It is possible that these truncations products were always present with the store bought T7 RNA polymerase but just at such a low level that we had not detected them previously. However, as we wanted to utilize the SPRI bead procedure for clean-up along with the in-house T7 RNA polymerase, RNA with minimal truncation products was required.

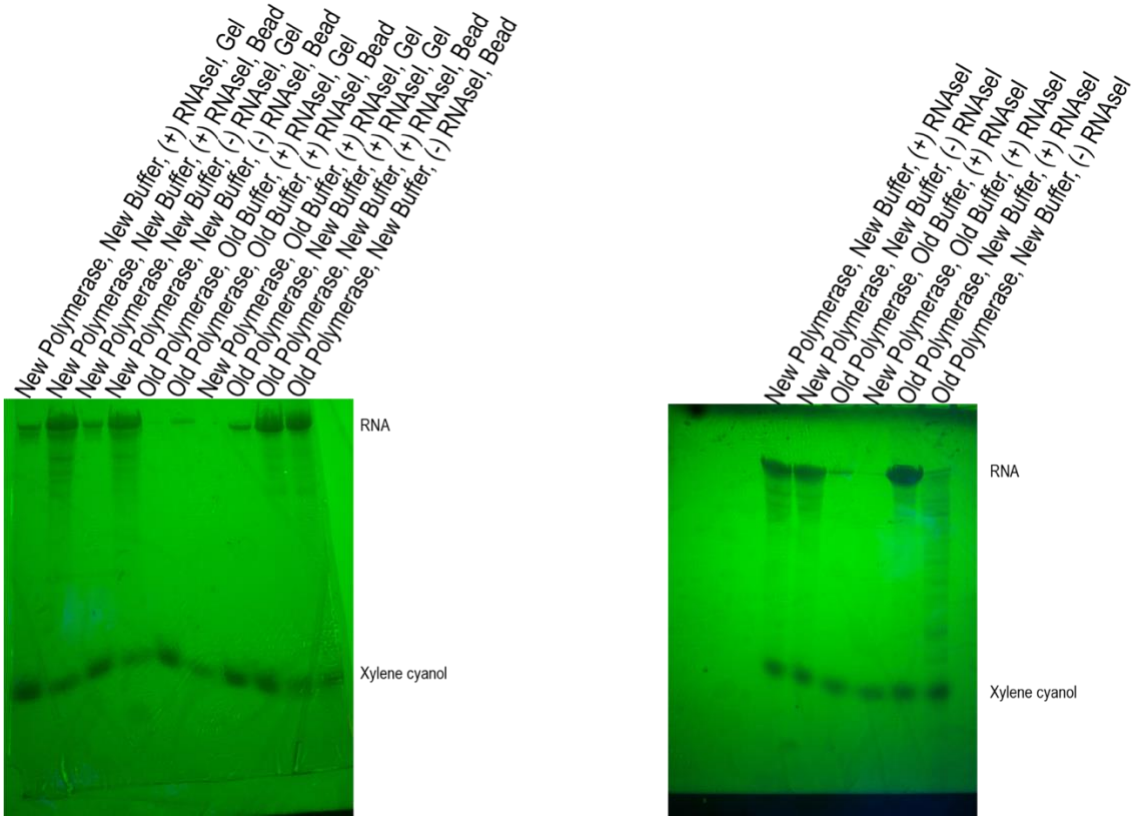


Figure 59: Comparison of in-house (New) T7 RNA Polymerase to store-bought (old) T7 RNA polymerase under different conditions with buffers and RNaseI

We initially attempted several different reaction buffer conditions with no visual decrease in the amount of truncation products (Figure 60a), and it was only upon varying the time of the reaction that we were able to demonstrate a decrease in the level of truncation products across RNA products with similar concentrations (Figure 60b, Table 17).

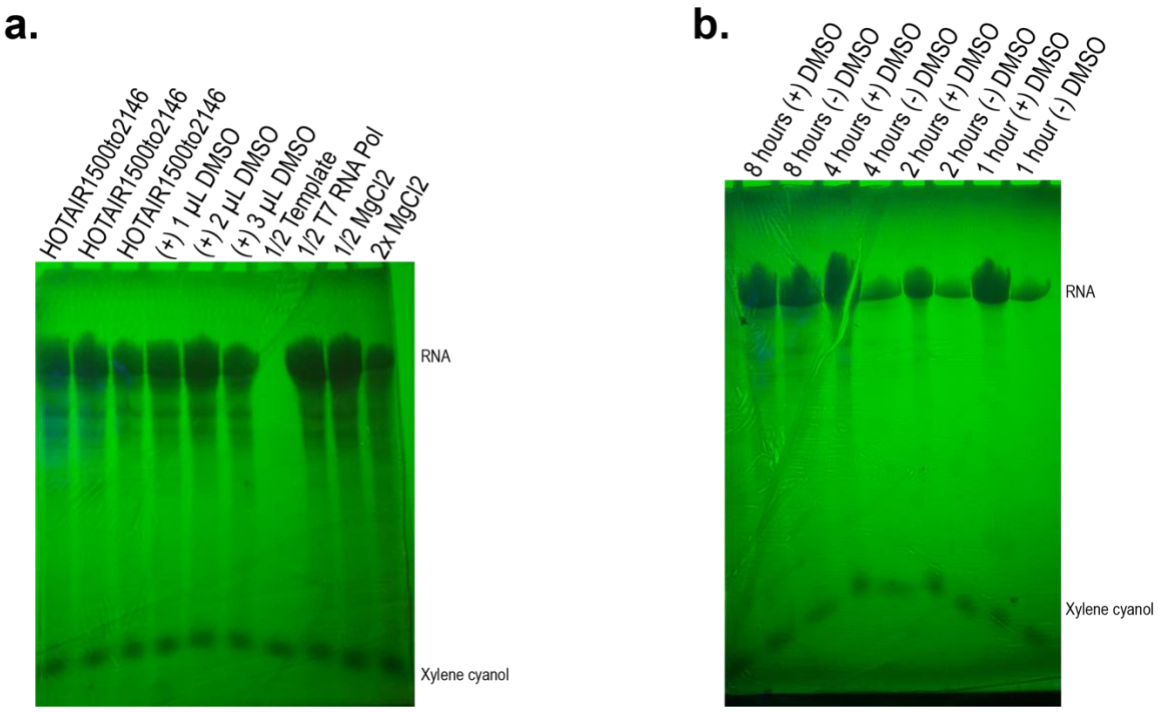


Figure 60: Optimization efforts to remove truncation products from HOTAIR D4 *in vitro* production utilizing in-house T7 RNA Polymerase and buffer system. (a) varying concentrations of DMSO, enzyme, and other buffer components revealed little change to truncation products; however, (b) upon varying the reaction time, we were able to decrease the level of truncation products while maintaining a high level of *in vitro* RNA production.

Table 17: Concentration (ng/uL) of HOTAIR D4 produced utilizing in-house T7 RNA polymerase and buffer system at differing time points plus and minus DMSO

Sample	Concentration (ng/µL)
8hour (plus)	1652.9
8hour (minus)	2864.3
4 hour (minus)	1624.1
4 hour (plus)	2921.6
2 hour (plus)	3190.4
2 hour (minus)	1969.4
1 hour (plus)	3565.2
1 hour (minus)	1930.9

With an optimized method to produce *in vitro* fragments of HOTAIR RNA in hand, and KDM1A and $CoREST_{286-482}$ which had been previously purified (Chapter 2), we turned towards utilization of a radiolabeled electromobility shift assay (EMSA) to assess binding of HOTAIR to KDM1A.

End-labeled ^{32}P-HOTAIR RNA

While there are several different methods to radiolabel RNA, we chose to label the 3′ phosphate via a dephosphorylation (30 min 50 °C, Thermo 18011051) reaction with 100 pmol of RNA followed by an SPRI bead clean-up and T4 kinase reaction with γ-^{32}P ATP (30 min, 37 °C, NEB M0201). Following a spin column clean up (Bio-Rad 7326250), the level of radiation was ascertained via scintillation counter (Beckman LS 60000SC).

Each EMSA was run utilizing a 6–point 2–fold protein dilution curve from 5 μM to 156 nM with a no protein control, keeping the level of RNA constant at 50,000 cpm/μL. The RNA was heated to 65 °C and then equilibrated at 37 °C for 5 minutes in the binding buffer (50 mM HEPES (7.4), 200 mM KCl, 25 mM $MgCl_2$, 0.1 mM EDTA, 0.1% BSA), and then mixed with the protein dilutions and incubated to ensure binding (20 min, 37 °C). Samples were mixed with 4x loading dye and loaded onto a homemade 6% TBE non-denaturing gel and run in cold buffer with ice packs to mitigate heat disruption of the complex and RNA degradation (100V, 120 min). Upon the dye front reaching the bottom of the gel, the run was halted, and the gel was saran wrapped to begin transfer to phosphor storage screen (1.5 hour). A typical resulting image of the

binding of HOTAIR$_{1500-2146}$ to KDM1A and KDM1A:CoREST$_{286-482}$ is depicted in Figure 61a. The resulting image was scanned using a Typhoon 9200 machine and ImageQuant v5.2 software was utilized to quantify the level of bound and unbound RNA.

4.2.2 Results

4.2.2.1. HOTAIR D4 requires pre-bound KDM1A:CoREST$_{286-482}$ for binding

We initially attempted to show binding between HOTAIR Domain 4 (HOTAIR D4) and KDM1A; however, our assay revealed little to no binding across the concentration range (Figure 61a,b).

Interestingly, HOTAIR Domain 1 requires both EED and EZH2 to bind, so we hypothesized that, in fact, more than just KDM1A was required for HOTAIR D4 to interact with KDM1A. With KDM1A:CoREST$_{286-482}$ already in hand from previous studies in the lab, we performed an EMSA with a pre-formed KDM1A:CoREST$_{286-482}$ complex to HOTAIR D4 which revealed a significant shift in the ^{32}P signal, suggesting binding (Figure 62a,c).

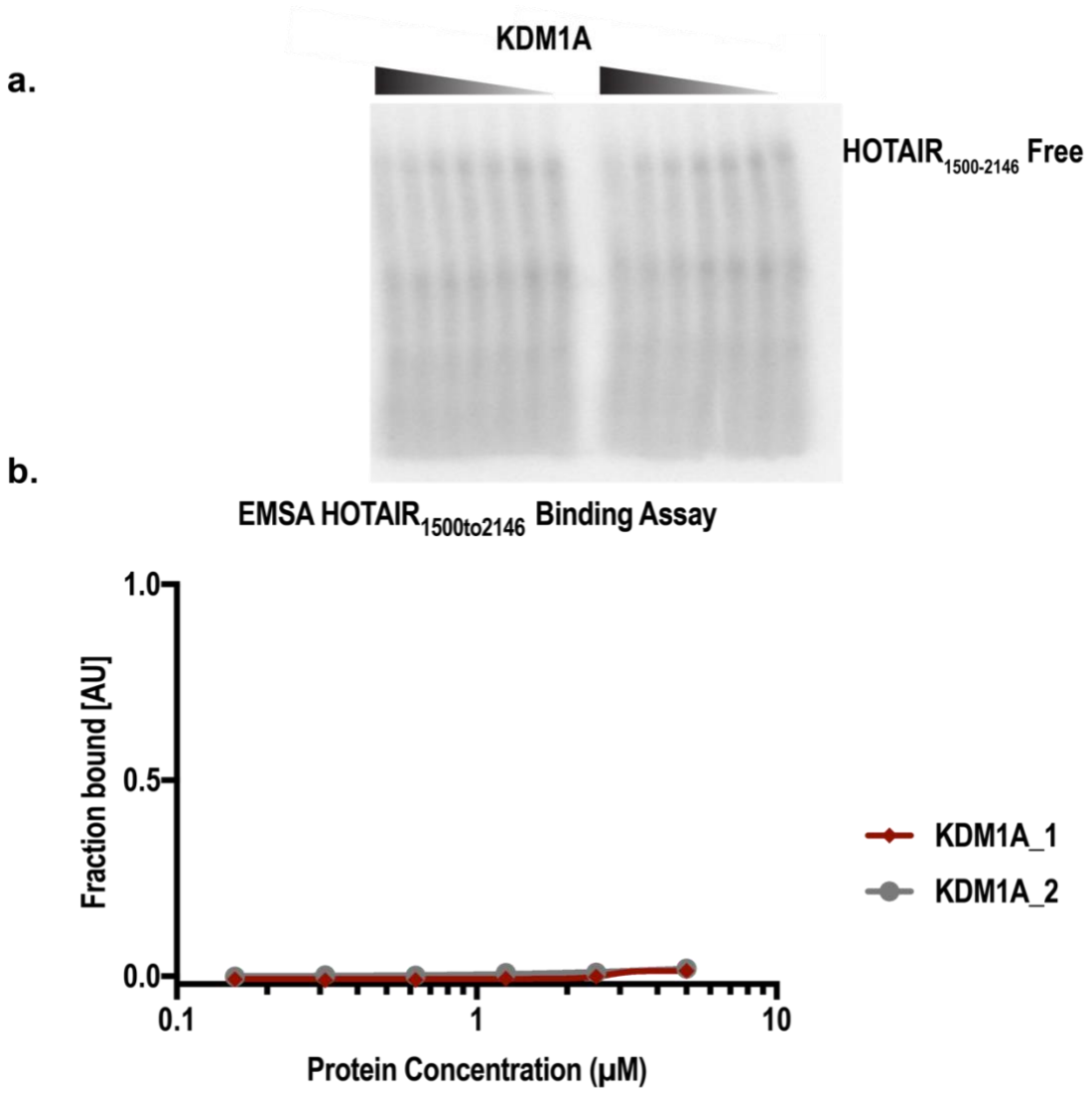


Figure 61: EMSA assay between KDM1A and HOTAIR D4. We utilized ^{32}P labeled end-labeled RNA to perform the assay with a 20 min incubation period at 37 °C between protein and RNA. (a) Is the visualization of the gel via phosphorimager and (b) the quantitation of the fraction bound vs the protein concentration.

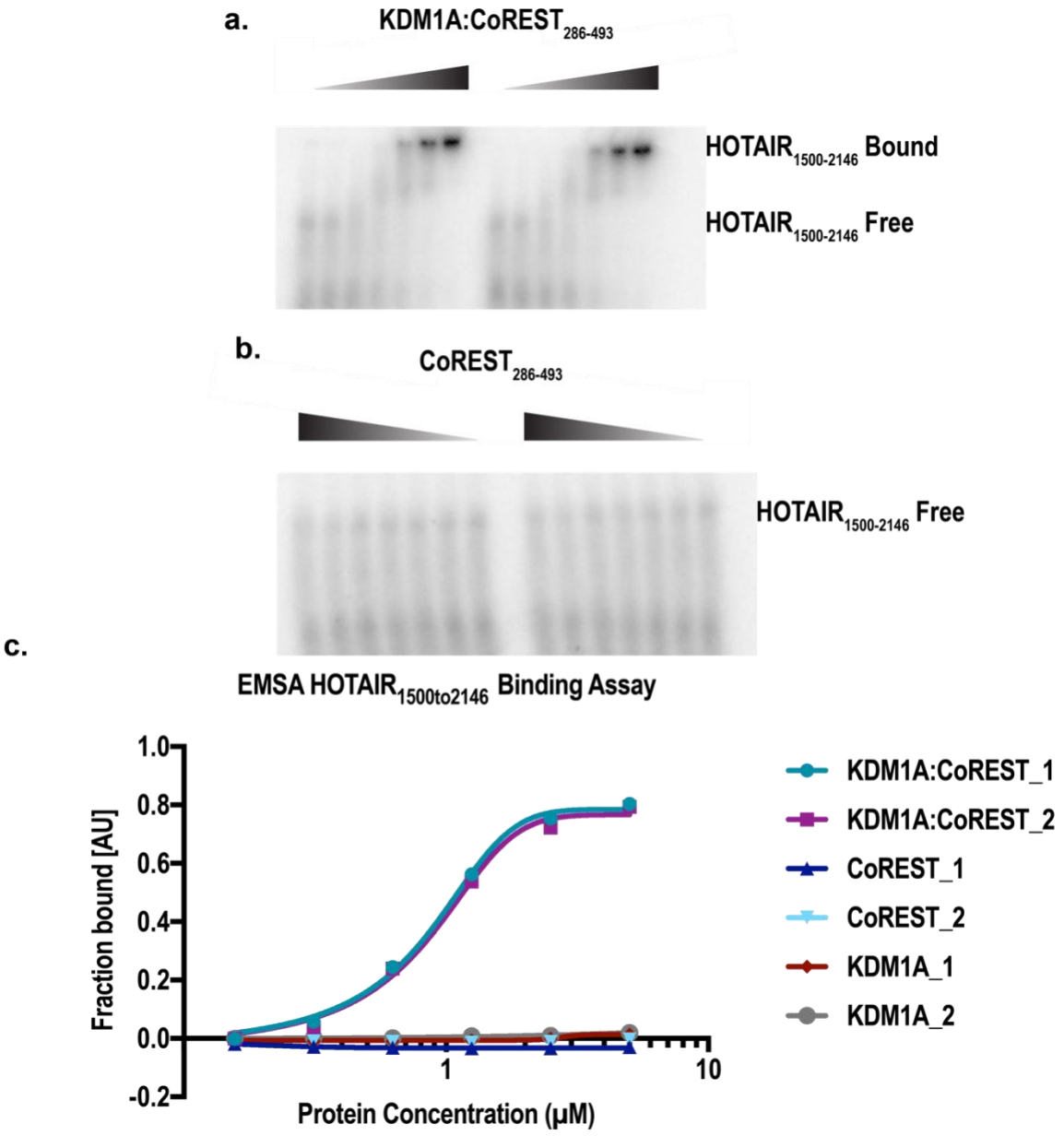


Figure 62: Binding assay utilizing HOTAIR D4 to assess binding to KDM1A and CoREST revealed that (a) while pre-bound KDM1A:CoREST$_{286\text{-}482}$ is able to effectively bind to HOTAIR D4, in a manner that produces a (c) binding curve suggesting a low μM affinity binding, while (b) there is no apparent binding to either KDM1A or CoREST$_{286\text{-}482}$ individually.

As the SANT 2 domain (residues 38-432) was reported to bind nucleosomal DNA by the Mattevi lab, we wanted to assess if CoREST alone was able to bind HOTAIR D4 *(118)* and performed an EMSA utilizing HOTAIR D4 binding to only $CoREST_{286-482}$. However, we observed no binding HOTAIR D4 and $CoREST_{286-482}$, alone even at 5 mM protein concentrations (Figure 62 b,c).

With a binding curve between KDM1A:$CoREST_{286-482}$ and HOTAIR D4 in hand, we sought to establish an apparent K_d for the equation. In a standard K_d calculation, we would utilize a saturating ligand, however, due to the nature of radioactivity, the receptor/protein in this case is in great excess to the RNA; thus, the EMSA measures the fraction of bound RNA rather than free protein concentration. To relate the bound fraction of RNA to the K_d the following equation has been proposed where in $[P_t]$ represents the free protein concentration at equilibrium. A non-linear least square fit can be utilized over several f values for multiple $[P_t]$ to establish the dissociation constant for the interaction. I performed the EMSA assay with KDM1A:$CoREST_{286-482}$ in biological triplicate to ascertain the apparent K_d to HOTAIR D4 to be 1.71±0.38 μM, which is roughly half of the previously reported K_d of HOTAIR to EZH2:EED as discussed above (Figure 63) *(252)*.

$$f = \frac{[RP]}{[R] + [RP]} = \frac{[P_t]}{[P_t] + K_d} = \frac{1}{1 + (\frac{K_d}{[P_t]})}$$

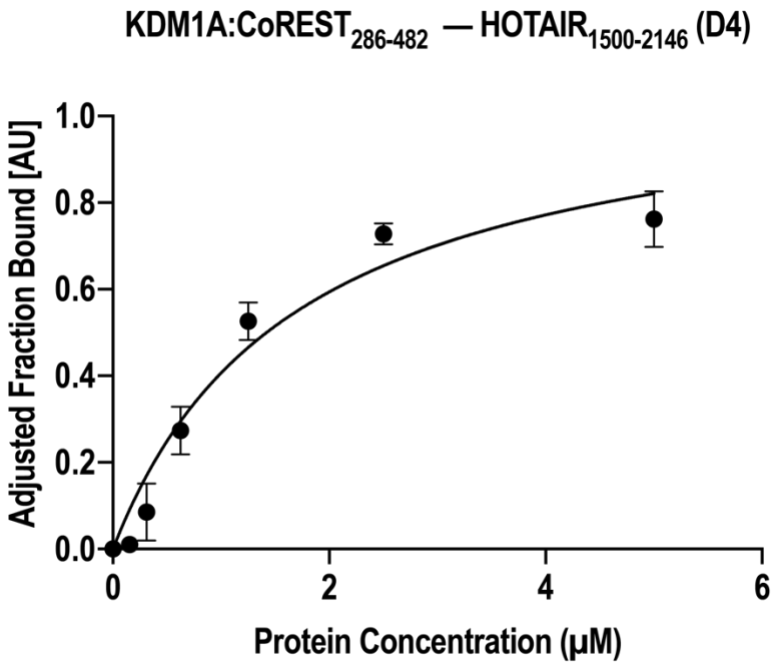


Figure 63: Binding curve of the KDM1A:Linker to HOTAIR D4 EMSA in triplicate. Produced in Prism utilizing the non-linear regression model.

4.2.2.2 Computational program aaRNA predicts tower domain is critical for HOTAIR:KDM1A-CoREST$_{286-482}$ interface

To create a targeted approach towards disrupting the RNA:protein interface, we wanted to more fully elucidate the molecular basis for the binding. Due to the lack of binding of CoREST alone to HOATIR D4, it is unlikely that the SANT2 domain was unlikely to be major contributor of the binding affinity. Furthermore, the HDX-MS data of the KDM1A:CoREST$_{286-482}$ complex discussed in Chapter 3 suggested to us that the catalytic core was also likely not the main arbiter of this interaction as the change in deuterium uptake in this domain was relatively limited upon CoREST$_{286-482}$ binding. First considering the residues of the protein which were involved in the interface, we elected to look first at a computational model to inform our future experiments. We

hypothesized that HOTAIR D4 interacted with the pre-bound KDM1A-CoREST$_{286-482}$ along an interface with contributions from both proteins. To assess the probability of this, we chose to next utilize a computational RNA-binding potential prediction program.

There are several different computational programs available which predict RNA-binding potential of proteins *(392–394)*. Most of these programs either consider only the sequence or the overall structure of the proteins. In the former case, while each residue may have an individual RNA-binding potential and affinity, there is no information for whether these residues are solvent accessible, or if the adjacent residues from folding would preclude RNA-binding ability. Conversely, the structures that only take overall fold into account, base RNA-binding capacity on structural similarity to structures that have been previously published. This is limiting based on the difficulty of obtaining RNA:protein crystal structures and thus a narrow range of RNA-binding protein structures compared to the number of proteins that likely make non-canonical interactions with RNA.

aaRNA was designed to bridge both sequence and structural data by assessing not only the overall structure and RNA-binding potential of the protein, but by also assessing the local landscape of each residue with respect to its RNA-binding potential. Their theoretical results closely matched experimental data and we felt most confident in

the reported results and statistics of this program and moved forward with this program over the other available programs.

Utilizing the pdb file 2iw5 for this study, we were able to gain several key pieces of information. First, looking at the overlay of the residues with the highest propensity for binding to RNA as predicted by the aaRNA algorithm, we were able to establish that there are residues (Figure 64, red) across both structures that have an RNA-binding potential. However, there is one region that was of interest to us. This region spans along the tower domain of KDM1A (residues 429-442, **E**H**W**KKIV**K**T**Q**E**E**L**KE**) and the linker region of CoREST (residues 349-357, **KRQ**I**QN**I**KQ**). While this region does contain many residues with a higher propensity for RNA-binding, two things of note were that there were a significant number of acidic residues across both proteins as well as a lack of an apparent pocket or region for the RNA to bind across in what might be considered a canonical binding manner.

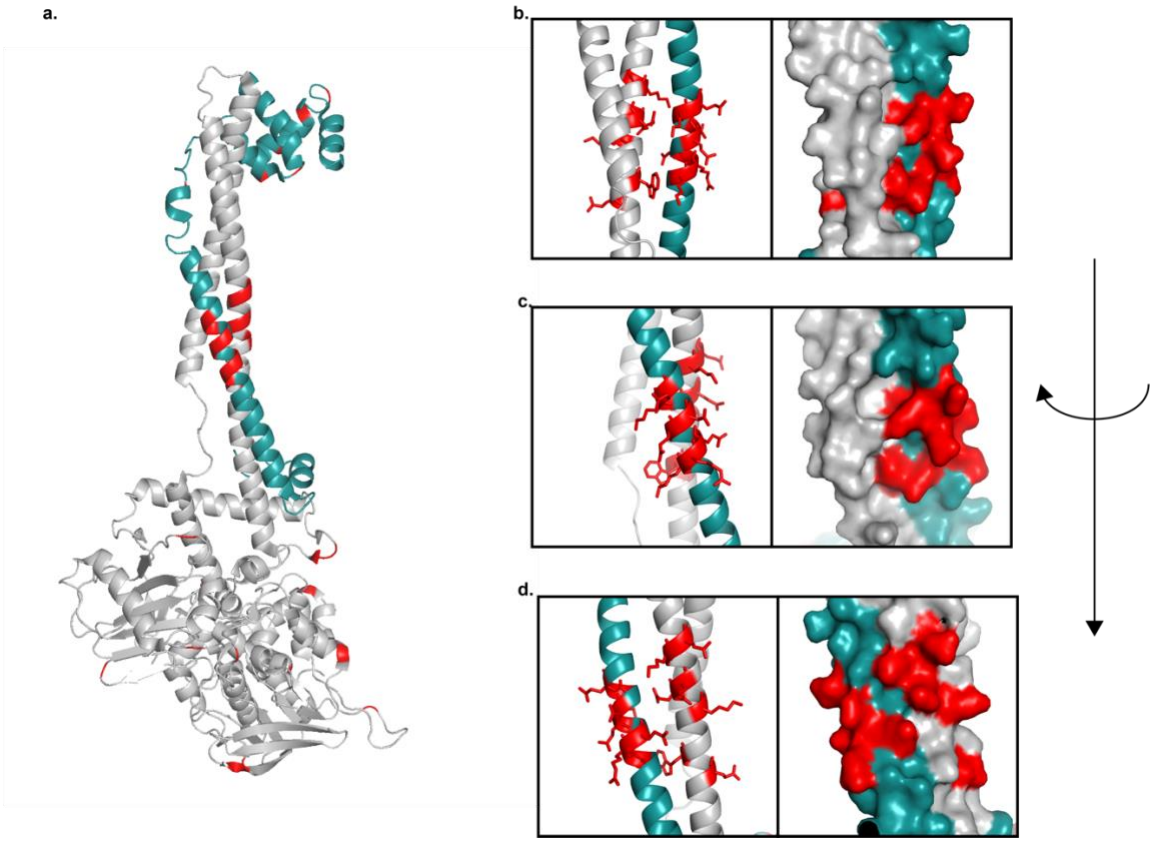


Figure 64: aaRNA predicted residues from the structure 2iw5 with the highest propensity (red) to interact with RNA nucleotides. (a) depicts the full structure with residues throughout with a potential for RNA binding. (b),(c) A region of potential RNA binding appears to span two α-helices on KDM1A and CoREST along the KDM1A tower domain.

The aaRNA software is also able to compute the propensity of binding to each potential dinucleotide pair. Looking at only the residues which have over a 50% propensity for RNA-binding capacity, the relative heatmap depicts the likelihood of each residue to bind to each dinucleotide pairing (Figure 65). Interestingly, for both KDM1A (Figure 65b) and CoREST (Figure 65a), there is an apparent preference for dinucleotides containing a G (AG, CG, GA, GC, GG) over any of the other potential

pairings for almost all of the residues.

a.

Residue	Propensity	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
312	0.51059	0.104	0.091	0.118	0.056	0.084	0.11	0.136	0.075	0.136	0.13	0.134	0.096	0.056	0.076	0.101	0.05
346	0.56	0.133	0.1	0.141	0.059	0.1	0.14	0.164	0.084	0.167	0.17	0.161	0.108	0.057	0.08	0.113	0.052
347	0.58758	0.137	0.104	0.151	0.063	0.104	0.139	0.166	0.091	0.169	0.176	0.161	0.112	0.062	0.088	0.118	0.052
348	0.6126	0.13	0.101	0.143	0.061	0.103	0.14	0.163	0.087	0.163	0.167	0.159	0.111	0.061	0.085	0.115	0.052
350	0.68415	0.146	0.118	0.18	0.074	0.111	0.151	0.185	0.099	0.189	0.194	0.18	0.128	0.069	0.101	0.132	0.063
351	0.5298	0.12	0.098	0.139	0.059	0.098	0.122	0.156	0.082	0.157	0.163	0.159	0.103	0.057	0.078	0.114	0.05
353	0.71551	0.152	0.122	0.182	0.073	0.112	0.16	0.209	0.1	0.209	0.211	0.192	0.133	0.066	0.099	0.139	0.057
354	0.69991	0.148	0.116	0.18	0.075	0.112	0.156	0.198	0.103	0.191	0.206	0.182	0.136	0.07	0.101	0.139	0.064
371	0.53161	0.124	0.096	0.111	0.052	0.09	0.12	0.141	0.077	0.136	0.136	0.126	0.091	0.056	0.071	0.097	0.05
400	0.52762	0.116	0.091	0.121	0.059	0.09	0.118	0.143	0.077	0.136	0.147	0.144	0.096	0.057	0.077	0.11	0.05
403	0.65182	0.129	0.108	0.165	0.072	0.109	0.14	0.17	0.095	0.171	0.175	0.166	0.124	0.067	0.099	0.128	0.063
418	0.51293	0.101	0.087	0.134	0.06	0.09	0.117	0.145	0.079	0.142	0.147	0.15	0.103	0.058	0.08	0.113	0.052
423	0.66301	0.105	0.096	0.143	0.059	0.09	0.11	0.156	0.077	0.163	0.161	0.158	0.102	0.054	0.074	0.118	0.05
426	0.69392	0.151	0.13	0.199	0.085	0.121	0.161	0.207	0.116	0.211	0.208	0.192	0.146	0.082	0.112	0.154	0.072

b.

Residue	Propensity	AA	AC	AG	AU	CA	CC	CG	CU	GA	GC	GG	GU	UA	UC	UG	UU
172	0.70515	0.151	0.126	0.196	0.085	0.119	0.16	0.201	0.113	0.204	0.195	0.194	0.161	0.079	0.113	0.142	0.08
204	0.62132	0.11	0.091	0.126	0.063	0.094	0.12	0.141	0.083	0.135	0.14	0.146	0.104	0.061	0.085	0.111	0.052
247	0.50742	0.119	0.091	0.117	0.054	0.09	0.118	0.136	0.076	0.133	0.135	0.132	0.094	0.057	0.076	0.095	0.05
275	0.63598	0.148	0.121	0.18	0.071	0.112	0.148	0.181	0.1	0.19	0.189	0.174	0.127	0.067	0.096	0.131	0.057
276	0.74282	0.174	0.143	0.21	0.084	0.126	0.172	0.225	0.12	0.242	0.232	0.205	0.157	0.082	0.111	0.16	0.07
324	0.56442	0.147	0.12	0.153	0.072	0.111	0.148	0.165	0.096	0.174	0.176	0.163	0.118	0.071	0.096	0.129	0.063
403	0.59774	0.159	0.129	0.174	0.072	0.116	0.156	0.194	0.108	0.196	0.201	0.175	0.131	0.073	0.097	0.139	0.065
404	0.58599	0.147	0.119	0.156	0.068	0.11	0.151	0.176	0.096	0.178	0.182	0.166	0.118	0.066	0.088	0.129	0.053
429	0.53189	0.1	0.073	0.103	0.046	0.074	0.102	0.118	0.062	0.114	0.12	0.125	0.078	0.044	0.059	0.083	0.041
431	0.5726	0.125	0.095	0.119	0.056	0.094	0.13	0.146	0.079	0.135	0.149	0.143	0.097	0.058	0.077	0.103	0.05
436	0.54014	0.13	0.098	0.126	0.059	0.099	0.134	0.152	0.082	0.145	0.155	0.152	0.103	0.058	0.079	0.105	0.052
438	0.52493	0.11	0.079	0.107	0.052	0.084	0.113	0.124	0.068	0.119	0.124	0.126	0.086	0.053	0.068	0.086	0.046
439	0.51973	0.093	0.067	0.094	0.043	0.07	0.097	0.107	0.059	0.104	0.109	0.111	0.072	0.043	0.056	0.075	0.041
442	0.52851	0.096	0.072	0.091	0.046	0.074	0.099	0.109	0.06	0.104	0.109	0.113	0.077	0.045	0.058	0.078	0.044
443	0.53378	0.084	0.057	0.075	0.036	0.061	0.084	0.087	0.05	0.084	0.089	0.091	0.059	0.037	0.048	0.063	0.037
610	0.53103	0.119	0.104	0.125	0.054	0.09	0.118	0.144	0.075	0.155	0.144	0.143	0.091	0.053	0.071	0.103	0.046
611	0.53441	0.146	0.121	0.179	0.071	0.111	0.15	0.176	0.098	0.19	0.187	0.181	0.124	0.066	0.091	0.131	0.055
645	0.59597	0.103	0.081	0.115	0.057	0.083	0.107	0.136	0.071	0.127	0.133	0.132	0.096	0.056	0.07	0.101	0.051
684	0.57522	0.126	0.104	0.123	0.059	0.102	0.128	0.141	0.083	0.142	0.136	0.142	0.103	0.062	0.079	0.104	0.053
750	0.5734	0.137	0.111	0.139	0.059	0.103	0.14	0.156	0.087	0.167	0.162	0.155	0.102	0.058	0.08	0.113	0.052
758	0.56164	0.156	0.128	0.188	0.074	0.116	0.157	0.199	0.108	0.208	0.201	0.184	0.136	0.072	0.1	0.139	0.065
791	0.64588	0.15	0.123	0.169	0.069	0.111	0.145	0.179	0.098	0.186	0.187	0.171	0.121	0.066	0.093	0.128	0.056

Figure 65: Heatmap depicting the propensity of each dinucleotide to interact with the specified residue where (a) is the results for CoREST and (b) is the results for KDM1A. The bold/underlined residues are those within the interface defined by Figure 64.

4.2.2.3 Utilization of EMSA assays to narrow down the minimal protein binding unit of the HOTAIR:KDM1A-CoREST$_{286\text{-}482}$ interface

With the goal of establishing which protein residues were directly interacting with HOTAIR in mind, we next performed several EMSA assays to narrow down the minimal protein complex required for binding.

We first wanted to confirm that SANT2 was not required for the binding of HOTAIR to KDM1A:CoREST$_{286\text{-}482}$. CoREST$_{289\text{-}383}$ (hereafter referred to as Linker)has been previously expressed and purified by Dr. Jennifer Link Schwabe in our lab. We

were able to optimize and shorten the length of purification to an initial Nickel-IMAC column followed by a TEV overnight cleavage and dialysis, subsequent Nickel-IMAC and concentration. Upon concentration, the sample was subjected to SEC on a Hi-Load Sephacryl S-200 (GE Healthcare) to high purity as assessed by SDS-PAGE and western blot (Figure 66a,b) and at a yield of 0.291 mg/mL.

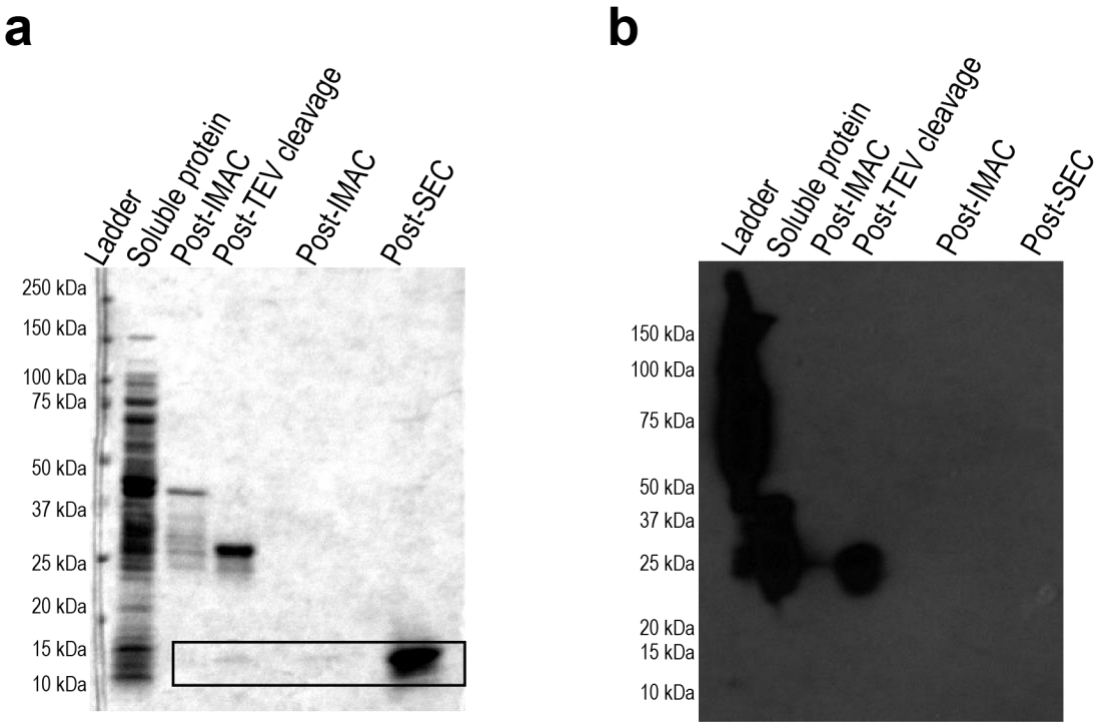


Figure 66: Overall purification of 6xHisGST-CoREST$_{293-380}$. We followed a purification strategy utilizing an initial nickel IMAC purification followed by an overnight TEV protease cleavage, a second nickel IMAC purification (chelating sepharose for fast flow, GE Healthcare), and final purification with SEC. Protein was concentrated and stored at 0.291 mg/mL in 40% glycerol with a buffer containing 25 mM HEPES (pH 7.4) with 138 mM NaCl and 1 mM BME

Upon obtaining purified Linker, as both proteins were in the same buffer, we directly moved to a 1:1 incubation with KDM1A with Linker at 37 °C for 1 hour, and then an EMSA assay with HOTAIR D4. We noted that we still saw binding of HOTAIR D4 to KDM1A:Linker (Figure 67). Interestingly, the overall percent bound appeared lower for KDM1A:Linker compared to KDM1A:CoREST, which could suggest some contribution or stability added by the SANT2 domain of CoREST.

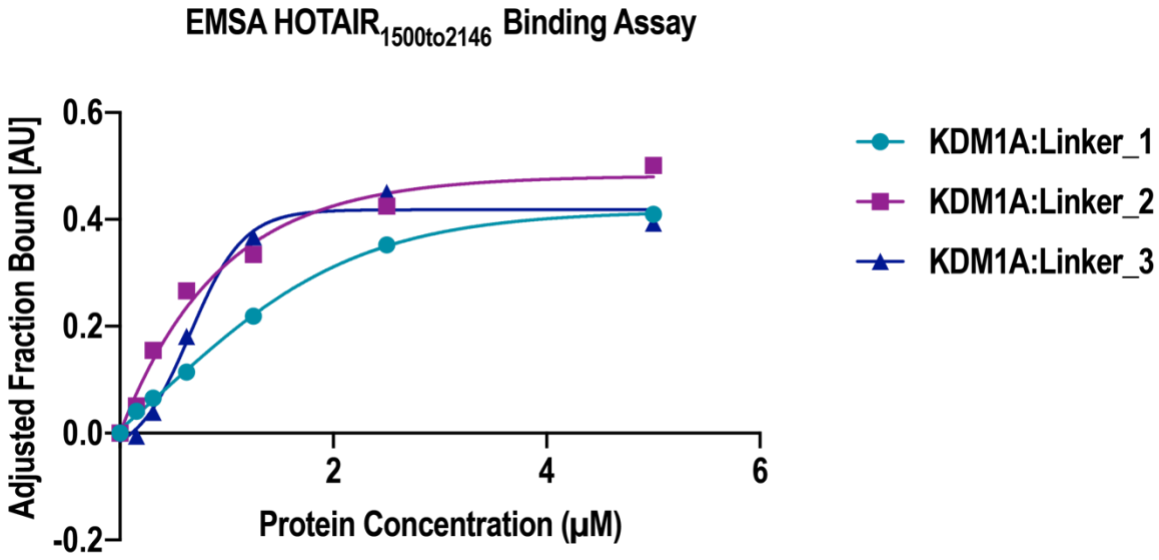


Figure 67: Initial EMSA assay between KDM1A:Linker and HOTAIR D4. We utilized ^{32}P labeled end-labeled RNA to perform the assay with a 20 min incubation period at 37 °C between protein and RNA.

In a similar fashion to KDM1A:CoREST$_{286\text{-}482}$, we established the apparent K_d of HOTAIR D4 to KDM1A:Linker to be 1.29±0.34 μM (Figure 68). This is within the error of the K_d of HOTAIR D4 to KDM1A:CoREST$_{286\text{-}482}$ suggesting there is a loss of binding affinity without the presence of the SANT2 domain, and that the binding interface lies along the tower domain and linker region of KDM1A and CoREST.

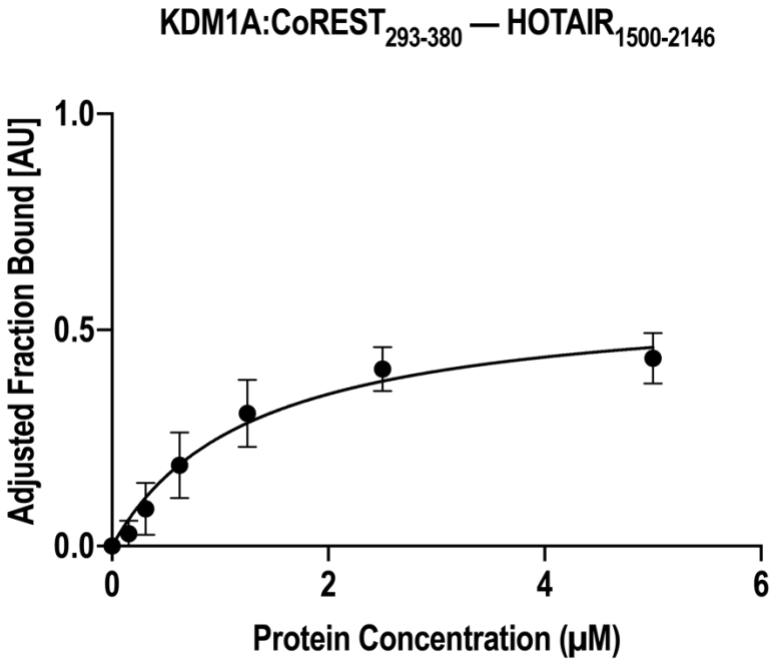


Figure 68: Binding curve of the KDM1A:Linker to HOTAIR D4 EMSA in triplicate. Curve and K_d were produced in Prism utilizing the non-linear regression model for One Site – specific binding.

4.2.2.4 Utilization of EMSA assays to narrow down the minimal RNA nucleotide region of the HOTAIR:KDM1A-CoREST$_{286-482}$ interface

Upon demonstrating an apparent requirement of both KDM1A and CoREST$_{286-482}$, we next sought to establish if we could narrow down the region of HOTAIR D4 that the two proteins were interacting with. We divided the 648–base domain in half to nucleotides 1500-1820 and 1821-2148 and *in vitro* transcribed, post-PCR as described, with no apparent truncation products visualized by denaturing acrylamide urea gel (400 V, 3.5 hr). We performed EMSA assays as above with KDM1A, CoREST$_{286-482}$, and KDM1A:CoREST$_{286-482}$ for both HOTAIR$_{1500-1820}$ and HOTAIR$_{1821-2148}$ (Figure 69). For

HOTAIR$_{1500-1820}$, as expected, binding was demonstrated only with both proteins present (Figure 69a). However, only at saturating concentration of KDM1A:CoREST$_{286-482}$ was binding demonstrated for HOTAIR$_{1821-2148}$ (Figure 69b).

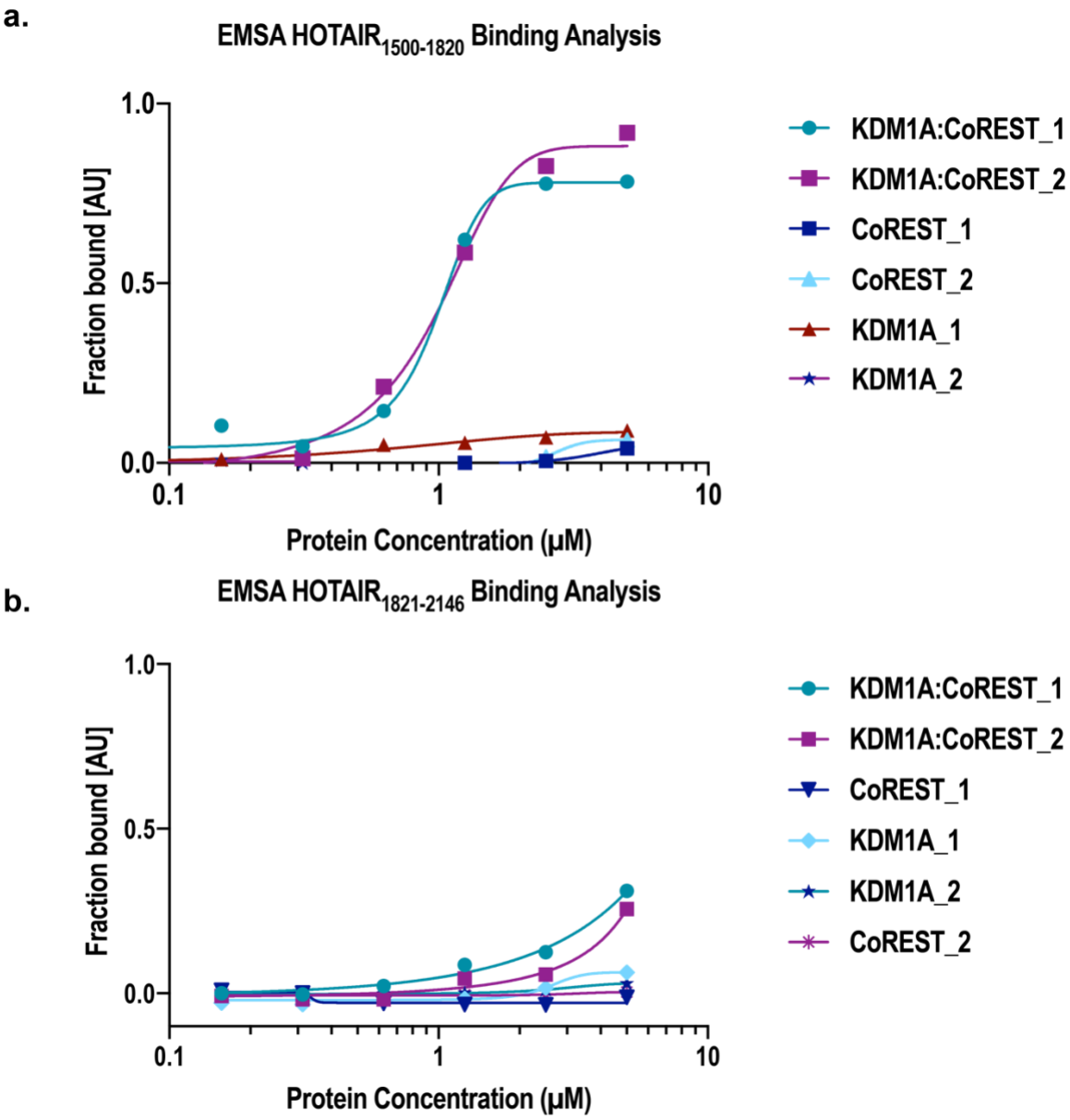


Figure 69: Binding assay utilizing HOTAIR D4 fragments to assess binding to KDM1A and CoREST revealed that (a) while HOTAIR$_{1500-1820}$ is able to bind KDM1A:CoREST$_{286-482}$, (b) HOTAIR$_{1821-2146}$ is unable to interact with KDM1A:CoREST$_{286-482}$ at relevant concentrations.

As previously, we were able to obtain binding curves for HOTAIR$_{1500-1820}$ to KDM1A:CoREST$_{286-482}$ in triplicate and establish the apparent K_d to be 2.97±0.96 μM

(Figure 70). The lower dissociation constant coupled with a higher variability as demonstrated in standard error could suggest that this interaction is less stable than the with full D4, and that the second half of HOTAIR D4 may contribute stability to the interaction.

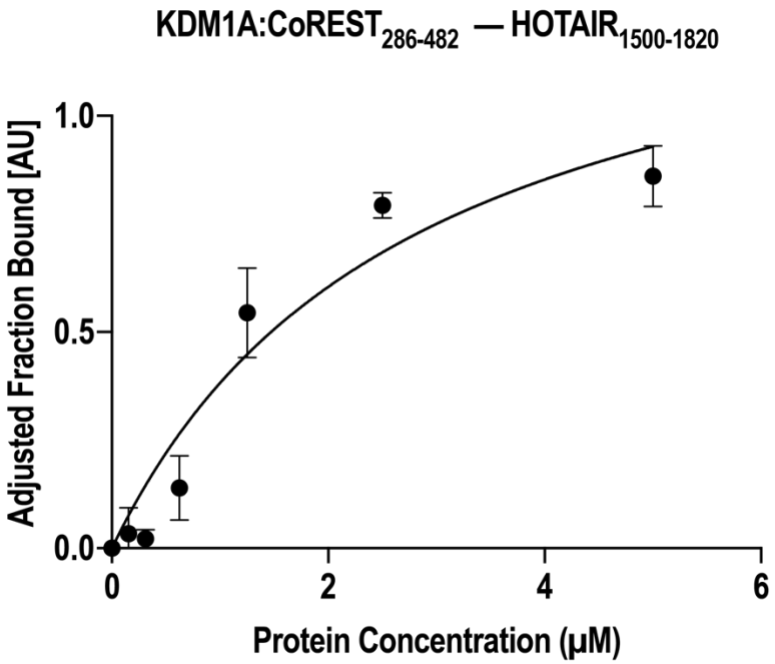


Figure 70: Binding curve of the KDM1A:CoREST$_{286-482}$ to HOTAIR$_{1500-1820}$ EMSA in triplicate. Curve and K_d were produced in Prism utilizing the non-linear regression model for One Site – specific binding.

4.3 Discussion

As the field of lncRNA continues to rapidly grow, we are beginning to gain a much better appreciation towards the extend of the interplay between non-coding RNA and epigenetic gene regulation. Not only can a ncRNA affect the transcriptional activation or repression of a gene, each ncRNA can have very separate targets and functions under different cellular contexts. This added level of complexity to the system

can lead to convolution of any cellular-based assays and data regarding interactions and mechanisms. We believe that beginning in an *in vitro* system, while not a complete biological mimic, can provide a basis to inform what may be occurring *in vivo*.

With the goal of not only disrupting the HOTAIR:KDM1A interaction, but also to establish which nucleotides and residues are involved in the interface, we first utilized EMSA assays to establish which proteins and regions of HOTAIR are critical for binding to occur. We were surprised to find that CoREST and KDM1A alone were unable to facilitate any appreciable binding to HOTAIR D4. The apparent requirement for both proteins is interesting as KDM1A is involved in complexes where CoREST is not present, such as when in complex with the nuclear hormone androgen and estrogen receptors, Herpes Simplex Viral Complex, or SNAI1 complex *(48)*. In the case of nuclear hormone receptor recruitment, KDM1A has been reported to be H3K9 specific, not H3K4 *(58,111,149,395)*.

It is unclear how this change in enzyme specificity occurs, but a requirement for CoREST may be a mechanism through which HOTAIR is able to selectively target complexes with CoREST-specific regulatory proteins, and enzyme specificity. It may also be that additional complex components impart further binding affinity with HOTAIR, driving the interaction with a specific complex. Increased affinity with addition of binding partners is a trend that we see with PRC2 and HOTAIR. The K_d that we established was roughly 2-fold higher than the previously established K_d for

$HOTAIR_{1-300}$ with EZH2. Upon addition of EED or EED and SUZ12 for binding $HOTAIR_{1-300}$ the affinity increased almost 4-fold to 147 and 165 nM, respectively, with the full PRC2 complex displaying an apparent 93 nM affinity for $HOTAIR_{1-300}$ *(252,389)*.

With the additional coregulatory proteins I gained access to in Chapter 2, we will be able to assess whether any of these coregulatory proteins contribute additional binding energy. It will be important to utilize techniques such as ChIRP and new *in vivo* crosslinking methods to establish which complexes interact with HOTAIR. Establishing the full KDM1A-containing complexes which interact with HOTAIR would allow us to establish the biologically relevant interactions to explore.

While neither KDM1A nor CoREST contain a canonical RNA-binding domain, CoREST does contain a SANT2 domain that was reported to bind DNA. While unlikely that the domain is able to bind both DNA and RNA, there are cases of targeted mimicry as discussed above wherein an RNA was able to mimic a specific fold and orientation to allow for binding at a DNA-binding domain of a protein. We thus utilized a pre-bound KDM1A:Linker to assess binding without the SANT2 domain. We demonstrated that there was still appreciable binding to HOTAIR D4 by this complex at a similar dissociation constant to the protein complex containing the SANT2 domain. This suggests that the majority of the binding interface resides along the tower domain and linker region of KDM1A and CoREST, respectively. We do not anticipate that the AOD or SWIRM domain contribute to the binding to HOTAIR based on the HDX-MS results

we obtained in Chapter 3. Upon binding CoREST, these regions were relatively unaltered, suggesting a lack of conformational change due to the presence of CoREST, which is born out in available crystal structures of KDM1A versus KDM1A:CoREST. Since CoREST does not induce a conformational change to the AOD or SWIRM domains of KDM1A we hypothesize that the binding lies along the coiled-coil interactions between KDM1A and CoREST tower domain and linker region.

The computational aaRNA study we performed lends further credence to the importance of the parallel α-helices along the Linker and Tower domains of KDM1A and CoREST. The region that was suggested to retain the highest RNA-binding propensity is interesting for several regions. While there are a high proportion of lysine residues present, which would be expected due to the negatively charged phosphate backbone, the high number of acidic glutamines was interesting to us. A comprehensive analysis of RNA:protein interacting structures in the pdb in 2012 revealed several features of RNA-binding proteins *(396)*. Firstly, they divided the landscape of the RNA-binding regions into "dented," "intermediate," and "protruded." It was demonstrated that dented, concave, regions were more likely to interact with ssRNA nucleotides. Further, the intermediate and convex "protruded" protein regions did not significantly bind to ssRNA over dsRNA, or vice versa *(396)*. They finally demonstrated that the frequency of each amino acid to be in the RNA-binding region of the protein, and

further, whether that amino acid was interacting with the phosphate, the ribose, or the base *(396)*.

Based on the classifications utilized in this study, we would classify the proposed RNA-binding region to be "intermediate" and "protruded." Looking at the residues present in our proposed binding region, lysine residues are highly enriched in the "protruded-paired-phosphate" binding group, glutamate, glutamine, and aspartate are intermediately enriched in the protruded-paired-ribose, intermediate-unpaired-phosphate/ribose, and intermediate-paired-ribose groups *(396)*. Coupled with the knowledge that HOTAIR is highly structured, we can hypothesize that a dsRNA region of HOTAIR lies across the interface formed between KDM1A and CoREST. It is unclear whether this RNA:protein interaction is sequence specific or structurally specific and will require more biophysical techniques such as x-ray crystallography.

It is interesting to note that the α-helices of KDM1A and CoREST are reminiscent of dsRNA binding domains. These canonical binding domains consist of two parallel α-helices which provide the majority of the interactions and a three-stranded β-sheet which provides additional non-specific contacts to stabilize the interaction. It may be that the interface is reminiscent of this form of an RNA-binding domain *(397–400)*.

We also began to look at regions within HOTAIR D4 that are contributing to the interaction. We found that the first 300 nt of HOTAIR retains the binding with KDM1A:CoREST$_{286\text{-}482}$, which agrees with the Chang lab's report that only the second 300

nt of HOTAIR D4 was unable to pull-down KDM1A. While we considered moving forward with assessing smaller binding fragments of HOTAIR, we were concerned with retaining critical structural elements within the lncRNA. We thus decided to move towards more probative RNA footprinting techniques that would allow us to assess which nucleotides were involved in the RNA:protein interface.

Overall, we have established that the minimal binding unit between HOTAIR and KDM1A requires the linker region of CoREST. I utilized radiolabeled EMSA assays to establish binding curves and the dissociation constant between KDM1A:CoREST$_{286\text{-}482}$ and KDM1A:Linker to both be 1.71±0.38 μM and 1.29±0.34 μM, respectively. Coupled with computational prediction software, we hypothesize that the interaction interface lies along an interface formed between the tower domain of KDM1A and the linker region of CoREST. I further affirmed that HOTAIR$_{1500\text{-}1820}$ is sufficient for binding to KDM1A:CoREST$_{286\text{-}482}$. In the future we hope to establish which specific residues are involved in binding to HOTAIR D4.

4.4 Experimental methods

4.4.1 Recombinant Protein Expression and Purification

Lysine-specific demethylase 1 (KDM1A/LSD1)

An N-terminal 6xHis tagged KDM1A (pET15b, residues 151-852) was used for expression and purification. The plasmid was transformed into electrocompetent BL21(DE3) *E. coli* cells (Novagen). Cells were grown in TB medium with 50 μg/μL

kanamycin at 37 °C with shaking (200 rpm) to an OD_{600} of 0.6, upon which IPTG was added to a final concentration of 0.5 mM and the induced cells were grown overnight at 21 °C. Cells were collected using an SLA-3000 rotor (4225 *x g*, 4 °C, 15 minutes) and lysed with an Emulsiflex C-5 high-pressure homogenizer (Avestin, Inc.) in Ni-IMAC wash buffer containing 50 mM sodium phosphate (pH 7.8), 300 mM NaCl, 5% glycerol supplemented with 0.4 mM PMSF (Sigma P7626), lysozyme (1.0 mg/L Sigma L6876) and 1X Protease Inhibitor Cocktail (Biotool 14001). The resulting crude extract was spun for 40 min at 4 °C and 40000 *x g*, using a Ti-45 rotor to pellet insoluble material. The supernatant was further via Ni-IMAC (GE Healthcare 17057501) overnight TEV protease cleavage, a secondary Ni-IMAC (GE Healthcare 17057501), and SEC(GE Healthcare 28989335). The final concentration of KDM1A was determined by absorption spectroscopy at 458 nm, and the protein was stored at -20 °C a buffer containing 25 mM HEPES (pH 7.4), 138 mM NaCl, and 50% glycerol.

REST Corepressor 1 (CoREST)$_{286-482}$

An N-terminal 6xHisGST-CoREST$_{286-482}$ with a TEV cleavage site immediately preceding CoREST$_{286-482}$ was utilized for expression and purification. The plasmid was transformed into electrocompetent BL21(DE3) *E. coli* cells (Novagen). Cells were grown in TB medium with 50 μg/μL kanamycin at 37 °C shaking (200 rpm) to an OD_{600} of 1.0, upon which IPTG was added to a final concentration of 1.0 mM and the induced cells were grown overnight at 19 °C. Cells were collected using an SLA-3000 rotor (4225 *x g*, 4

°C, 15 minutes) and lysed with an Emulsiflex C-5 high-pressure homogenizer (Avestin, Inc.) in Ni-IMAC wash buffer containing 50 mM Tris-HCl (pH 8.0), 350 mM NaCl, 5% glycerol, supplemented with lysozyme (1.0 mg/L Sigma L6876), and 1x Protease Inhibitor Cocktail (Biotool 14001). The resulting crude extract was spun for 40 min at 4 °C and 40000 *x g*, using a Ti-45 rotor to pellet insoluble material. The supernatant was further purified via Ni-IMAC (GE Healthcare 17057501), overnight TEV protease cleavage, a secondary Ni-IMAC (GE Healthcare 17057501) column, and stored at -20 °C (25 mM HEPES (pH 7.4), 138 mM NaCl, 40% glycerol)

KDM1A:CoREST$_{286\text{-}482}$

The two proteins were purified separately as described above. They were incubated for 60 minutes and stored at -20 °C in a buffer containing 25 mM HEPES (pH 7.4), 138 mM NaCl, and 50% glycerol

KDM1A:CoREST$_{289\text{-}383}$

An N-terminal 6xHisGST-CoREST$_{289\text{-}383}$ with a TEV cleavage site immediately preceding CoREST$_{289\text{-}383}$ was utilized for expression and purification. The plasmid was transformed into electrocompetent BL21(DE3) *E. coli* cells (Novagen). Cells were grown in TB medium with 50 μg/μL kanamycin at 37 °C shaking (200 rpm) to an OD_{600} of 1.0, upon which IPTG was added to a final concentration of 1.0 mM and the induced cells were grown overnight at 19 °C. Cells were collected using an SLA-3000 rotor (4225 *x g*, 4 °C, 15 minutes) and lysed with an Emulsiflex C-5 high-pressure homogenizer (Avestin,

Inc.) in PBS buffer supplemented with lysozyme (1.0 mg/L Sigma L6876), and 1x Protease Inhibitor Cocktail (Biotool 14001), and 2 mM BME (FisherScientific BP176). The resulting crude extract was spun for 40 min at 4 °C and 40000 *x g,* using a Ti-45 rotor to pellet insoluble material. The supernatant was further purified via glutathione agarose (Prometheus), overnight TEV protease cleavage, and a secondary Ni-IMAC (GE Healthcare 17057501) column. Upon concentration, previously purified KDM1A was added at a 1:1 ratio and after two hours, the protein complex was further purified over SEC (GE Healthcare 28989335) and stored at -20 °C in a buffer containing 25 mM HEPES (pH 7.4), 138 mM NaCl, and 40% glycerol.

X. laevis Histone H3$_{1\text{-}135}$

Histone protein constructs from X. laevis, as described by Luger et al., were transformed into BL21(DE3) E. coli (Novagen). Cells containing the expression plasmid were used to inoculate LB (GeneseeScientific 11-125) medium supplemented with ampicillin to 1% inoculum and were grown at 37 °C with 200 rpm shaking. Protein expression was induced with the addition of IPTG to a final concentration of 200 μM at an OD_{600} of 0.6 and incubated for an additional 2 hr at 37 °C. Cells were harvested by centrifugation as above (SLA-3000 rotor, 4225 *x g*, 4 °C) and stored at -20 °C until use. Pellet was thawed and resuspended in wash buffer containing 50 mM Tris-HCl (pH 7.5), 100 mM NaCl, 1 mM Na-EDTA, 1 mM BME (FisherScientific BP176) , lysozyme (1.0 mg/L Sigma L6876), and 1x Protease Inhibitor Cocktail (Biotool 14001).

The mixture was gently stirred on ice for 10 min and viscosity of the solution was reduced and remaining cells were lysed with an EmulsiFlex-C5 high-pressure homogenizer (Avestin, Inc.). The resulting crude extract was spun for 20 min at 4 °C and 26000 *x g*, using an SS-34 rotor to pellet insoluble material. Supernatant was discarded, and the resulting pellet was washed by resuspension and centrifugation three times in 50 mL of wash buffer containing 1% (v/v) Triton X-100 (Sigma). To remove the detergent, the pellet containing the inclusion bodies was additionally washed two times with wash buffer. The pellet, which contained the inclusion bodies was soaked in 1 mL of dimethyl sulfoxide (DMSO) for 30 minutes at room temperature. A 50 mL volume of 6M guanidinium hydrochloride, containing 20 mM sodium acetate (pH 5.2) and 10 mM DTT was slowly added and unfolding was allowed to proceed overnight at room temp. Unsolubilized material was removed via centrifugation at 26000 *x g*. Sample was then dialyzed extensively at 4 °C against water that contained 1 mM in 3.5 kDa MWCO (Thermo Scientific) dialysis tubing to remove DMSO, After dialysis, 4L 3x at 4 °C, spin down the pellet on swinging bucket rotor at 2000 *x g* pour off supernatant, resuspend in 10 mL ddH2O with 1 mM BME then flash freeze in liquid N2 and proceed to lyophilization. Lyophilized powder was stored at -20 °C.

T7 RNA Polymerase

A pT7-911 construct containing T7 RNA Polymerase was generously gifted by the Weeks lab at UNC Chapel Hill. After transformation into BL21(DE3) *E. coli*

(Novagen) cells, colonies containing the expression plasmid were used to inoculate LB (GeneseeScientific 11-125) medium supplemented with 100 μg/μL ampicillin to 1% inoculum and were grown at 37 °C with 200 rpm shaking. Protein expression was induced with the addition of IPTG to a final concentration of 0.1 mM at an OD_{600} of 0.6 and incubated for an additional 3 hr at 37 °C. Cells were harvested by centrifugation as above (SLA-3000 rotor, 4225 *x g*, 4 °C) and stored at -20 °C until use. Pellet was thawed and resuspended in wash buffer containing 50 mM Tris-Cl (pH 7.8), 300 mM NaCl, 5% glycerol, 1x Protease inhibitor cocktail (Biotool 14001), and 5 mM BME (Fisher Scientific BP176). Cells were lysed with an EmulsiFlex-C5 high-pressure homogenizer (Avestin, Inc.). The resulting crude extract was pelleted using an SS-34 rotor (26000 *x g*, 4 °C, 20 minutes), and the supernatant was applied to a pre-equilibrated Ni-IMAC resin (GE Healthcare 17057501). Batch purification was utilized with iterative rounds of washing with wash buffer supplemented with 10 mM and then 20 mM imidazole in triplicate (700 *x g*, 4 °C, 2 minutes). A final elution step was performed with 400 mM imidazole, and the resin was pelleted after 30 minutes of incubation at 4 °C (700 *x g*, 4 °C, 2 minutes), in duplicate. The protein was stored at 1 mg/mL in a buffer containing 20 mM potassium phosphate (pH 7.5), 0.1 mM EDTA, 50% glycerol, 2 mM DTT.

4.4.2 *in vitro* Transcription of RNA

A pcDNA3 plasmid containing the full-length sequence of HOTAIR was generously gifted by the Chang lab at Harvard University. Primers outlined in Table 18

were utilized for the PCR amplification of specific fragments of HOTAIR utilizing Platinum Taq polymerase (ThermoFisher 10966018) as outlined in Table 18 and 19.

Table 18: Primers utilized for PCR amplification of HOTAIR fragments

HOTAIR NTS	FORWARD	REVERSE
1500-1820	Gatagatcctaatacgactcactataggg ACCAGCCCTAGCCTTTGGAAG	GGTCAGGCGG GATCCACTTATTTAAGTGTTCTCCT
1821-2146	Gcatggatcctaatacgactcactataggg ATGCATGTAGACACAGAAG	Gctaggcggatcc GAAAATGCATCCAGATATTA
1500-1697	Gatagatcctaatacgactcactataggg ACCAGCCCTAGCCTTTGGAAG	Ggtgcgggatcc GGAACAAGCATTATATAATT
1597-1820	Gattcatcctaatacgactcactataggg TCTAACTGGCAGCACAGAGC	GGTCAGGCGgatcc ACTTATTTAAGTGTTCTCCT
1697-2146	Gattcatcctaatacgactcactataggg CATACAGGAGTGATTATGCAG	Gatgccggatcc GAAAATGCATCCAGATATTA
1500-2146	Gatagatcctaatacgactcactataggg ACCAGCCCTAGCCTTTGGAAG	Gctaggcggatcc GAAAATGCATCCAGATATTA
1500-1820 Antisense	Aaactcagtccgcggatcc ACCAGCCCTAGCCTTTGGAAG	Ctatctaggattatgctgagtgatatcc CACTTATTTAAGTGTTCTCCTAT
Full length	Aaatctatctaatacgactcactatag GGACTCGCCTGTGCTCTGGAG	Gctaggcggatcc GAAAATGCATCCAGATATTA

Table 19: PCR buffer conditions for HOTAIR$_{1\text{-}2146}$ and HOTAIR$_{AS1500\text{-}1820}$

100 µL rxn	Stock	Final
PCR Buffer	10X	1X
MgCl2	100 mM	3.0 mM
dNTPs	10 mM	0.2 mM
5' Primer	10 µM	0 µM
3' Primer	10 µM	0.2 µM
Nuclease-free water	v/vt	v/vt
Template	50 ng/µL	25 ng
Platinum taq		2 U

Table 20: PCR buffer for additional HOTAIR fragments

100 µL rxn	Stock	Final
PCR Buffer	10X	1X
MgCl2	50 mM	1.5 mM
dNTPs	10 mM	0.2 mM
5' Primer	10 µM	0 µM
3' Primer	10 µM	0.2 µM
Nuclease-free Water	v/vt	v/vt
Template	50 ng/µL	50 ng
Platinum taq		2 U

Table 21: PCR amplification conditions for HOTAIR fragments

Step	Temp (°C)	Time (min)	# of cycles
Initial denaturation	94	5	1
	94	0.5	35
Amplification	64.0	0.5 min/kb	35
	72	1	35
Final elongation	72	10	1
Hold	4	Infinite	1

After PCR amplification, the reactions were subjected to consecutive rounds of DPN1 (NEB M0176, 20 U) and BAMH1 (NEB R3136, 20 U) digestion at 37 °C for 1 hour. Subsequently, the reactions were purified using homemade SPRI DNA beads as outlined in appendix 2 and the concentration was quantified with a UV-visible spectrophotometer (SpectraMax i3x micro-volume microplate).

The T7 RNA polymerase reaction was performed utilizing a homemade buffer outlined in Table 22 and allowed to proceed in a 37 °C heat block for 1-8 hours as determined via visualization on a denaturing urea gel. The reaction was halted by addition of DNase I (NEB M0303, 2U, 37 °C, 30 minutes) and subsequently 20 mM

EDTA. The reaction was purified using homemade SPRI RNA beads as outlined in appendix 2 and the concentration was quantified with a UV-visible spectrophotometer (SpectraMax i3x micro-volume microplate).

Table 22: In-house T7 RNA Polymerase buffer conditions

Reagent	Stock	Final
rNTPs	25 mM	5 mM
RNaseI	40 U/μL	1 U/μL
Tris (pH 8.0)	1 M	40 mM
MgCl2	1 M	25 mM
Spermidine	100 mM	2.5 mM
Triton X-100	1%	0.01%
DTT	1 M	10 mM
New T7RNAP	0.8mg/mL	--
H2O	v/vt	v/vt
Template	20-100 ng/μL	150-750 ng
pyrophosphatase	0.2 U/μL	0.05 U

4.4.3 Electromobility Shift Assay

RNA was ^{32}P-labeled as described in Filter Binding Assay section. Protein was diluted into a binding buffer containing 50 mM HEPES (pH 7.4), 200 mM KCl, 25 mM $MgCl_2$, 0.1 mM EDTA, and BSA to a final concentration of 6.67 μM. The protein was diluted 2-fold to 156 nM and incubated at 37 °C for 5 minutes. The ^{32}P-radiolabeled RNA was diluted into 5x binding buffer and incubated at 65 °C for 5 minutes and then 37 °C for 5 minutes. To each protein sample was added RNA (50000 CPM), and the RNA:protein mixtures were incubated at 37 °C for 20 minutes. The samples were mixed with 2x loading dye and loaded and run on a homemade 6% TBE gel (100V, 120

minutes, 4 °C). The gel was removed from cassette, wrapped in saran wrap and exposed to phosphor storage screen (GE Healthcare, 2828956475) for 1.5 hours. The screen was scanned by a Typhoon Imaging machine and the bound versus free bands analyzed via Image Quant v5.2 software.

5. Utilization of RNA footprinting techniques to characterize the HOTAIR:KDM1A-CoREST$_{286-493}$ interface

The results of the binding assays provided us with the insight that CoREST$_{286-493}$ is critical for HOTAIR:KDM1A binding. We also confirmed that the first 320 nucleotides of domain 4 of HOTAIR (1500-1820) contribute the majority of the binding affinity, as previously suggested by the Howard Chang lab at Harvard *(163)*.

We are interested in understanding the molecular basis of this interaction, and thus we wanted to establish which nucleotides of HOTAIR specifically interacted with KDM1A:CoREST$_{286-493}$. Towards this end, we utilized RNA footprinting technique selective 2'-hydroxyl acylation analyzed by primer extension with mutational analysis profiling (SHAPE-MaP) as established by the Weeks lab at UNC *(401,402)*. This technique enabled us to ascertain that the highest level of impact in both reactivity and concentration of nucleotides affected by the presence of KDM1A:CoREST$_{286-493}$ was localized to a specific arm within the first 300 nt of HOTAIR D4.

We next utilized a photochemical crosslinking technique also developed by the Weeks lab to define which HOTAIR D4 nucleotides were likely directly interacting with KDM1A:CoREST$_{286-493}$. The arm that was significantly altered in SHAPE-Map again demonstrated the highest number of crosslinked nucleotides. Unexpectedly, a dominant number of guanine and cytosine bases are likely interacting with protein and there are six G:C paired nucleotides of which five are within the first 300 nt of HOTAIR D4. Each

of these G:C paired nucleotides is flanked by either a G or C nucleotide, suggesting a 4-mer binding motif.

Based on our results and current knowledge of RNA binding proteins, we propose that the arm is a significant binding feature through which KDM1A:CoREST$_{286-493}$ bind. It may be that the G:C paired nucleotide 4-mer sequences each provide a low-affinity binding motif that contribute to the overall binding affinity between HOTAIR D4 and KDM1A:CoREST$_{286-493}$.

5.1. Background

While RNA and DNA are similar, there are fundamental differences between the two nucleic acids which results in major structural and functional differences. Furthermore, the techniques utilized to explore RNA, particularly lncRNA is severely limited by the size and stability of the nucleic acid. Thus, we will first explore the basic structure and chemistry underlying RNA, then several of the common RNA characterization tools and finally overview the RNA footprinting techniques available for lncRNA.

5.1.1. RNA structure and chemistry

Primary structure

Looking first at the basic molecular structure of RNA, similar to DNA the primary structure of an RNA nucleic acid is comprised of nucleotides linked together through a 5′ to 3′ phosphodiester linkage. As the name would suggest, the sugar of RNA

does contain the 2′ hydroxyl group, unlike DNA. The four most common RNA bases: adenine, guanine, cytosine, and uracil, follow a similar base pairing pattern to DNA with uracil replacing adenine in the A:T pairing (Figure 71). The free energy calculations of the Watson-Crick hydrogen bonding are ΔG -5.53 and -4.42 kcal/mol for G:C and A:U, respectively *(403)*. NMR studies utilizing by the LiWang lab utilized ^{13}C splitting patterns on C2 of adenosine to demonstrate that the isotope splittings were larger on RNA than DNA *(404)*. They correlated this to a stronger hydrogen bonding in RNA over DNA. While additional indirect experimental and theoretical studies have agreed with these findings, they argue that the increased splitting pattern is merely due to the absence of a methyl group on uracil *(405,406)*

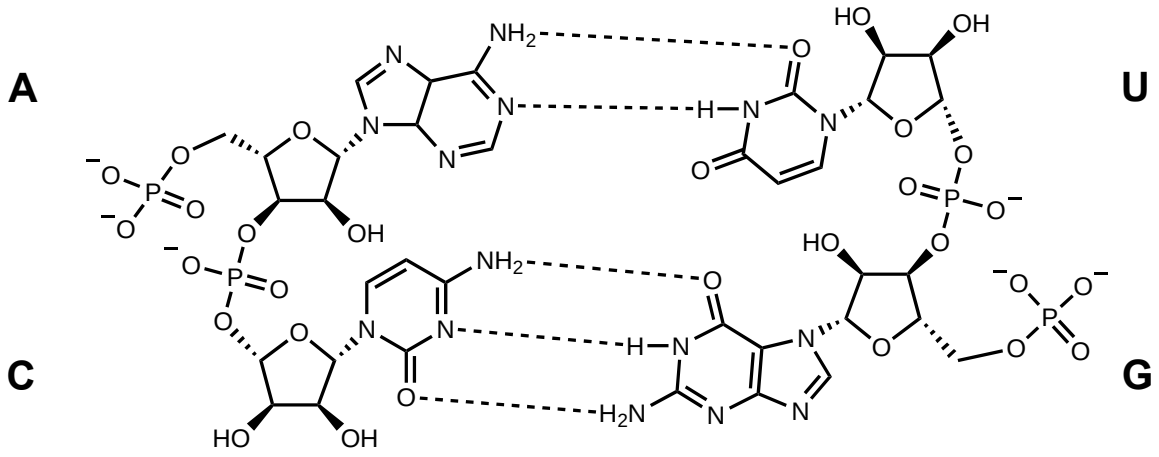
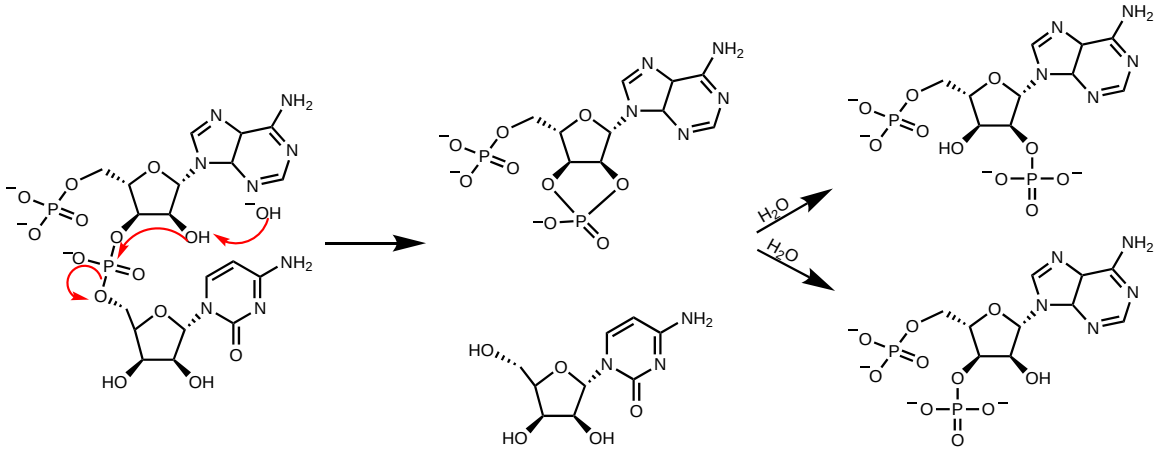


Figure 71: Watson-Crick hydrogen bond pairing of RNA nucleotides

The utilization of the ribose sugar in RNA results in a less stable nucleic acid, particularly at a higher pH due to a higher potential for hydrolysis, as outlined in Scheme 3 *(407,408)*.

Scheme 2: 2'-Hydroxyl mediated RNA hydrolysis reaction in alkaline conditions.



A series of studies have demonstrated that *in vitro*, this can be catalyzed by the presence of Mg^{2+} ions in an alkaline environment. Generally, Mg^{2+} is considered stabilizing to RNA secondary and even tertiary structure through both direct inner-sphere and water mediated outer-sphere coordination events of, primarily, the phosphodiester backbone *(271,409,410)*. The structure of RNA is generally mediated by the presence of Mg^{2+}; for example, the Pyle lab demonstrated that HOTAIR molecules were the most structurally similar under 25 mM $MgCl_2$ conditions *(251)*.

Under alkaline conditions, the Mg^{2+} ions form magnesium-aqua ions which are able to non-specifically cleave RNA *(411)*. Facilitated via in-line geometry, the hydrated Mg^{2+} functions as a Bronsted base to abstracts a proton from the 2′ hydroxyl to generate

a 2′ oxyanion which, in turn, attacks the phosphorus group *(412,413)*. This results in a 2′,3′-cyclic phosphate via the oxyanion and scission between the two nucleotides *(414)*. RNase A utilizes a similar mechanism via a catalytic triad of two His residues (12 and 119) and lysine 41 to catalyze the reaction *(415)*.

While Mg^{2+} is the most relevant for *in vitro* work, many different metals including Zn^{2+}, Pb^{2+}, and Tb^{3+}, are able to facilitate the hydrolysis reaction through not only nucleophilic activation, as expected, but additionally (Figure 72) leaving group stabilization, non-bridging oxygens coordination, nucleophilic activation, and induced intramolecularity *(410,413,416–420)*. Additionally, while divalent metals generally function in stability and catalysis via the phosphate backbone, there is evidence that monovalent ions such as potassium and sodium play a role interactions with nucleobases *(421)*.

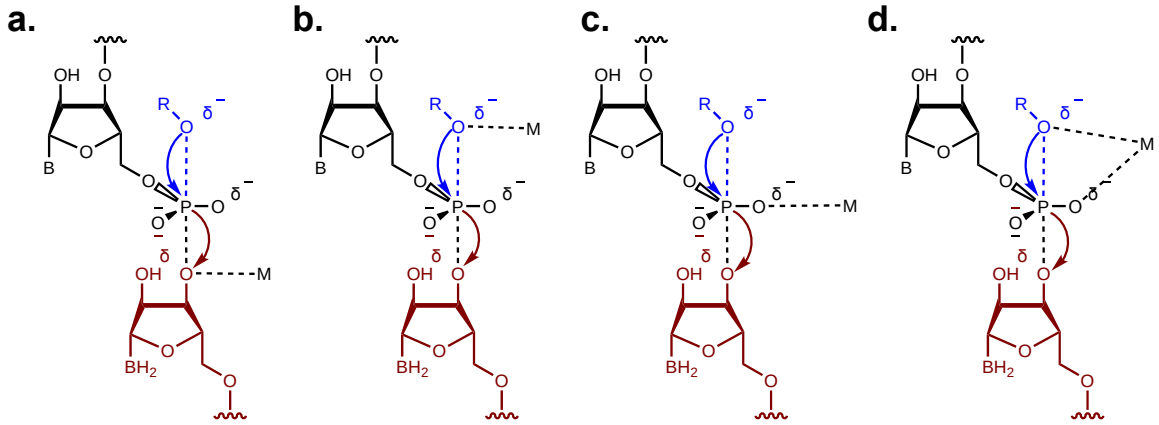


Figure 72: Metal ions can act through several different mechanisms to catalyze RNA-hydrolysis (a) Lewis acid stabilization, (b) activation of the nucleophile, (c) coordination of phosphate oxygen atom, (d) induced intermolecularity (Adapted from *(413)*)

Another key structural difference between DNA and RNA is in the conformation of the sugars. Unlike DNA, RNA does not have a complementary strand to interact with, and thus does not form the canonical "B-form" left-handed helix of DNA. However, many RNAs are intricately structured via intramolecular interactions to facilitate stability, and likely, function of the RNA. The most common secondary structural formation of RNA is a right-handed "A-form" helical structure which is due, in part, to the predominantly C3′-endo conformation of the RNA ribose of C2′-endo conformation of the DNA deoxyribose due to the presence of the 2′-hydroxyl group on the ribose *(422,423)*. Mono-ribonucleotides of bases adenine, guanine and uracil exist predominantly in the C2′-endo conformation with cytosine split between C2′- and C3′-endo conformations and it may be that upon formation of a nucleic acid that the shift to C3′-endo is initiated *(424)*.

Generally, this is the less favored conformation as it shortens the length of the backbone, bringing the negatively charged phosphate groups together from 7.0 to 5.9 Å, however, computational models have suggested that the C3′-endo conformation is energetically favored for ribose nucleotides by 2.0 kcal/mol with a 4.0 kcal/mol energy barrier to transition to a C2′-endo conformation *(425,426)*.

While the C2′-endo conformation may not be energetically preferred, there is evidence that these conformations may be favored in functional RNA elements. A molecular dynamics study suggested that the 2′-hydroxyl group is able to hydrogen

bond with basic residues, resulting in a torsion shift in the RNA which may induce a conformational fit between the RNA and protein *(427,428)*. This has been seen in an RNA:protein crystal structure (2XFM) wherein an RNA nucleotide is stabilized into an C2′-endo conformation by a lysine residue, but not when that residue is mutated to alanine stabilizes an RNA nucleotide *(427)*. This is interesting as C2′-endo conformations have been shown to be overrepresented in functional RNA motifs. In one instance, the ribozyme RNase P contains a conserved C2′-endo nucleotide which functions as a 'molecular timer' by modulating the timing of the rate-limiting step for the ribozyme folding *(429)*.

Secondary Structure of RNA

The secondary structure, when complete base-pairing occurs is, overall, similar to A-form helical structure of DNA; however, A-form is distinct from the canonical B-form of DNA (Table 23). The major differences of note are the twist of A-form RNA and B-form DNA, right-handed and left-handed, respectively, as well as the small minor groove and larger major groove of B-form DNA compared to the larger minor groove and smaller major groove of A-form RNA. Additionally, these changes in spacing and grooves results in a change in hydration patterns between the two nucleic acids. In B-form DNA, every single phosphate is hydrated, and there is a so-called "spine-of-hydration" down A:T rich regions of the major groove of the helix *(425)*. In A-form RNA, due to the increased proximity of each nucleotide, one water molecule is able to

bridge two phosphates, decreasing the overall number of water molecules. This has been demonstrated to be a less stable conformation, and may contribute, in part, to the flexible nature of RNA over B-form DNA *(425)*.

Table 23: Comparison of A form DNA and RNA characteristics *(336,430–434)*

	A-form RNA	*A-form DNA*	*B-form DNA*
Helical Twist (°)	32.7	32.7	34.3
Rise per bp (Å)	2.81	2.56	3.4
Minor groove depth (Å)	11.1	11.0	5.8
Major grove depth (Å)	4.2	2.3	12.3
Base-pair tilt (°)	16.7	22.0	-6
Radius (nm)	1.2	1.15	1.0

While the main conformation of pairing RNA are large looping structures to form A-form RNA helices, there are also smaller motifs that form due to mismatches in bases upon folding. Several of the most common and basic secondary structure motifs are outlined in Figure 73 which are common throughout all RNA, regardless of size.

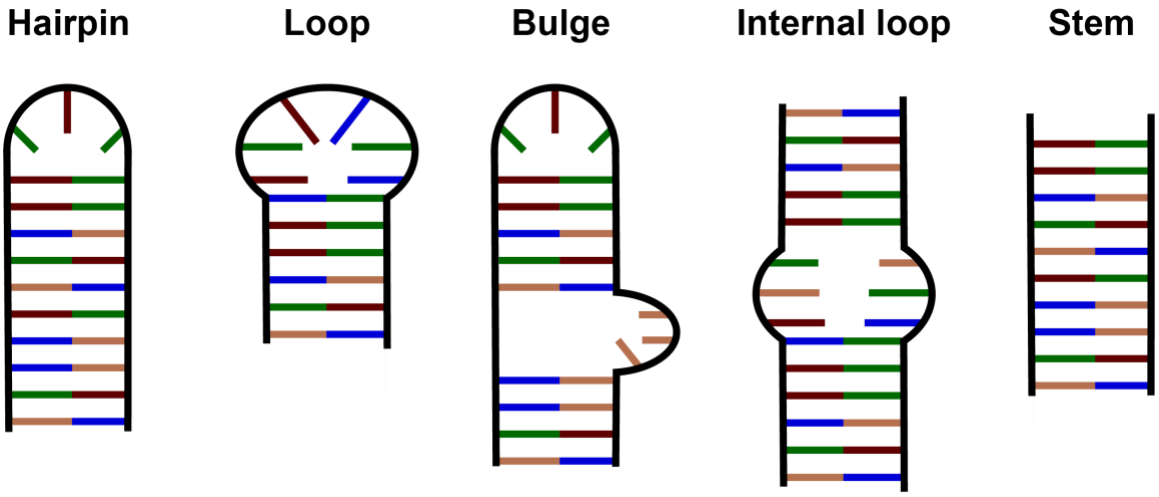


Figure 73: RNA secondary motifs which are common throughout RNA structures

Several of these motifs have been reported to potentially be important for binding proteins *(435)* . For example, a tetraloop is a 4 nt loop at the end of a stem. While hairpin loops are the most common structural motif throughout RNA, tetraloops are well reported throughout RNA and several thermodynamic studies have been performed to establish that the most stable display up to a 20 °C increase in melting temperature compared to other hairpin loops and even other tetraloops *(435)*. This increased thermodynamic stability is due to G:C base pairing at the first and last nucleotides in the tetraloop, base stacking along the stem directly preceding the loop, 2' hydroxyl hydrogen bonds within the loop, and potentially even non-canonical hydrogen bonding patters such as Hoogsteen or non-Watson-Crick within the loop as well *(436–440)*.

In addition to these secondary structural motifs interacting with RNA-binding proteins, the A-form of RNA also plays a role in RNA-protein binding. In fact, there are several crystal structures in the literature of dsRNA binding domain proteins bound to RNA. Indeed, dsRNA binding domains consist of a α-β-β-β-α wherein the first α-helix contacts the minor groove in a sequence specific manner, there are additional minor contacts between second and third β-sheets, and additional major groove phosphodiester backbone contacts are made with the second α-helix *(398,400)*.

5.1.2. Limitations of traditional structural biology approaches with long non-coding RNA

Traditional approaches to assessing the interaction interface of a protein utilizing biochemical and structural biology techniques generally include a combination of methods including NMR, x-ray crystallography, mutational scanning, limited proteolysis, pull-down assays, and HDX-MS. With proteins, the variation of individual amino acids including hydrophobicity, potential for steric hinderance, as well as non-covalent interaction capabilities can greatly limit the number of possible conformations proteins may form. The general stability of proteins allows them to be crystalized with, generally, minimal intervention.

Conversely, when approaching the structure of RNA, even small RNA such as a tRNA, present a much more difficult challenge. The chemical similarity in the four bases, and increased flexibility due to the backbone structure enables the RNA to adopt many more conformations, calculated to be 1.8^N possibilities where N is the length of the RNA *(441)*. The non-covalent interactions that are formed between nucleobases are, overall, weaker than the myriad of interactions that proteins can make including salt bridges and disulfide bonds. Thus, due to the high conformational flexibility of RNA at a tertiary level structural elucidation of RNA is more difficult. Free energy models of RNA secondary structures appear to be generally consistent with the experimentally determined secondary structures in short RNA, suggesting that thermodynamics may

play a major role . This is likely the case for lncRNA as well, however, prediction models are currently less accurate towards RNA over 400 nt, precluding many lncRNA.

Nuclear magnetic resonance (NMR has been critical for small RNA such as the HIV TAR RNA and other small RNA which fall below the 50kDa size limitation of this instrumentation *(442,443)*. It should be noted that a 50 nucleotides length is more ideal for obtaining high resolution data from NMR, even though medium resolution data can be obtained up to a 100-nucleotide length which correlates to roughly 50 kDa *(438)*. NMR can provide a significant amount of information regarding RNA including the base-pairing patterns as well as the dynamics and conformational equilibria of the structure*(444–446)*. Furthermore, this instrument can be utilized to assess ion binding along the RNA structure. Beyond the size limitation, evaluating the data can prove tricky as the similarity in the four nucleotides limits the chemical shifts of imino proteins in Watson-Crick pairing to 12-15 ppm. Fortunately, non-canonical pairings display an upfield shift, and there is additionally shift dispersion in secondary motifs such as hairpins, bulges, and internal loops *(444,447,448)*.

However, even the molecular weight of the shortest lncRNA is going to exceed the size restriction for NMR, and X-ray crystallography has, thus far, not provided additional help. The first X-ray crystal structure was of yeast tRNA for phenylalanine in 1974 at a 3Å; however, the flexibility of RNA remains the major limitation to production of crystals *(449)*.

The secondary structure motifs of RNA are relatively stable, but the tertiary structures are much less rigid and pose an issue for the production of crystals. Even with RNA:protein crystals, the regions of the RNA that are not directly interacting with the protein are generally much more difficult to crystalize due to a high B factor which is an indicator of disorder or motion. Techniques utilized to combat this include truncating the RNA, inducing a rigidifying tetraloop, however, these methods risk losing biological structure and thus relevance. Even if crystals are produced, they generally are much more likely to decay upon exposure to the diffraction ray, thus optimization with protectants is generally extensive. Heavy atom incorporation is a standard tool utilized by crystallographers towards clarifying phases and solving the structure for proteins. In RNA, this has also proven successful with lanthanides and osmium(III) hexamine, which both are believed to replace magnesium ions, though osmium (III) hexamine to a higher extent due to its ability to form outer-sphere interactions similar to magnesium. However, even with the advances available, there are no published structures of a lncRNA, likely due to the overall flexibility of the structures.

5.1.3 Techniques to study the structure of lncRNA *in vitro*

While traditional biophysical techniques remain intractable for lncRNA, there are several footprinting techniques which allow us to gain useful information into the structure of the lncRNA. erbium, dimethyl sulfate (DMS), and SHAPE are each common techniques that have been utilized to inform the structure of an RNA. Each provides

slightly differing information that can be coupled with computationally predicted free-energy folding patterns towards a highly-predictive secondary, and potentially tertiary lncRNA structure.

Terbium

Since the 1970s, trivalent lanthanum terbium has been utilized in various techniques to assess RNA. Initially noted to fluoresce upon complexation with the phosphate of DNA nucleobases, a similar occurrence was noted in 1978 for RNA *(450)*. As discussed above, terbium has also been shown to catalyze RNA cleavage reactions, particularly at unpaired RNA, the underlying principle being that 2'-hydroxyl groups of ssRNA will be more flexible and accessible than those that are involved in pairing *(413,451)*.

The level of terbium in a reaction, the reaction time, as well as the concentration of magnesium present in the sample have all been shown to affect the level of cleavage produced by terbium, thus a significant amount of optimization and validation is required for this technique. Interestingly, due to the tunability of terbium cleavage and its bio-orthogonal role to magnesium in stabilizing RNA structures, there have been reports of utilizing terbium towards establishing transition states of tertiary structure of tRNA *(452)*.

Dimethyl sulfate

Another method of RNA footprinting is utilization of a nucleobase modifying reagent such as dimethyl sulfate (DMS). In this technique, nucleobases are methylated, with those not involved in base pairing with another nucleobase much more likely to be methylated than those actively hydrogen bonding with another nucleobase *(453–457)*. Traditionally, the resulting cDNA products are then run on an agarose gel and compared against sequencing standards, in a standard Sanger sequencing method, to establish which nucleotides were modified.

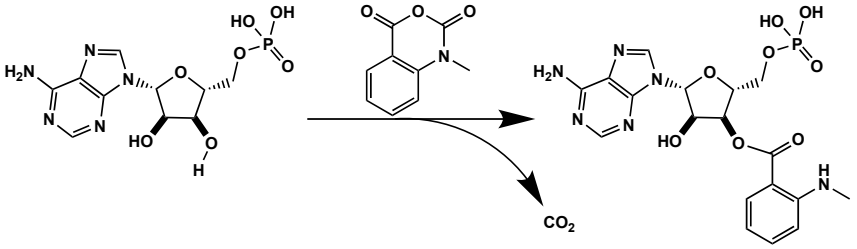
While capable of methylating all four bases, methylation of N1-A and N3-C are the only two modifications that stall reverse transcription *(458)*. Thus, DMS treatment is limited to information at adenine and cytosine, though there are other reagents available that can perform analogous chemistry at guanine and uracil bases *(454)*. One major drawback to DMS treatment that should be noted is that if a paired nucleobase is modified, the chances that the nucleobase it was paired to also being modified are significantly increased, adding to the inaccuracy of the overall data collected *(459)*.

Selective 2′-hydroxyl acylation analyzed by primer extension

Established by the Weeks lab at UNC Chapel Hill, SHAPE utilizes nucleobase modifying chemicals such as 5-nitroisotoic anhydride (5-NIA) or 1-methyl-7-nitroisatoic anhydride (1-M7) as nucleobase modifying reagents. When incubated with RNA, each compound rapidly acylates the 2′-hydroxyl of nucleotides with flexible sugar moieties (Scheme 4) *(401)*. The adducts that are formed on the RNA can then be used to establish

which nucleotides are more likely to be disordered, or not base pairing with other nucleotides. Unlike HDX-MS, SHAPE reactivities are not related or dependent on solvent accessibility. The 2′-hydroxyl group of flexible, or disordered, nucleotides, will have a higher rate of sampling conformations *(460)*. This will increase the nucleophilicity of the 2′-hydroxyl group *(460)*. Increased nucleophilicity of the 2′-hydroxyl will, upon addition of the modifying reagent, increase the amount that the nucleotide will be modified whereas RNA nucleotides which are interacting with another nucleotide will be less likely to sample these conformations, lowering their rate of modification *(460,461)*. The SHAPE reactivity has been demonstrate to be inversely correlated to the NMR generalized order parameter which depicts the nucleotide spatial range of motion, which demonstrates that, indeed, a high SHAPE reactivity is related to a high local nucleotide flexibility and local disorder *(462)*. These reactivities of modification can be utilized to establish the potential of single-stranded versus double-stranded at a nucleotide resolution and utilized in a secondary structural computational modeling program.

Scheme 3: SHAPE reaction with 5-NIA modifying an adenosine nucleotide



The RNA was extended via adaptive primers in the initial SHAPE protocol *(401)*. Each modified nt caused the reverse transcriptase to stall, terminating the extension process at the adduct. The resulting DNA fragments were run on a gel to establish which nucleotides had been modified. However, this only allowed for one modification to be captured per RNA strand, and also did not give much information regarding how common each of the modifications were.

Mutational analysis profiling technology

Upon any of the above footprinting methods being employed, originally, cDNA was formed, via reverse transcriptase that stalled upon reaching one of the modified nucleotides. The resulting fragments were run on a gel and compared against standards to establish which nucleotides along the RNA were modified. The production of cDNA via reverse transcriptase limited the level of depth of the reads to one per transcript, thus making it more difficult to establish frequency of a mutation. Further, it is likely that the further down the RNA strand the mutation was, the less likely it would be to be transcribed, again, producing a disparity between the reality of the mutation rate and what was observed *(463)*.

The advent of mutational analysis profiling, MaP, allowed for extension of the entire RNA to occur via production of conditions which allowed for the reverse transcriptase to extend past the modification *(464–467)*. In this updated protocol, the modifications are incorporated as a mutation in cDNA which can be sequenced in massive parallel by next generation sequencing (NGS). This allows for the direct comparison of mutations rates at a nucleotide level of a SHAPE modified to DMSO controlled samples to establish reactivity of each nucleotide. Processing via SHAPEmapper software developed by the Weeks lab allows for an overall higher resolution of the reactivities and coverage of the RNA at a nucleotide level by the MaP technology *(468)*.

5.2 SHAPE-MaP

In order to establish which nucleotides of HOTAIR domain 4 were involved in the interaction with KDM1A:CoREST$_{286\text{-}493}$, we utilized the SHAPE-MaP in collaboration with the Weeks lab at University of North Carolina at Chapel Hill. This technique establishes the flexibility of the 2′ hydroxyl group based on acylation frequency. This can, in turn, be correlated to a likelihood of the nucleotide being single stranded (ssRNA) or double stranded (dsRNA). The collected SHAPE reactivities can also be utilized to inform computational free energy secondary structure modeling programs.

We sought to compare the SHAPE reactivities of HOTAIR D4 with and without the presence of KDM1A:CoREST$_{286\text{-}493}$. The majority of the experiments the Weeks lab

utilizes SHAPE-MaP for are *in vivo* which required an initial protocol optimization to allow for RNA recovery from an *in vitro* system. Post-optimization we were able to obtain reproducible results wherein 83 nt displayed a 1.5-fold change in SHAPE reactivity due to the presence of KDM1A:CoREST$_{286\text{-}493}$, and 22 further demonstrated a 2.0-fold change in SHAPE reactivity. All but two of these nt were decreasing in reactivity due to the presence of protein, suggesting an overall stabilization of RNA. When mapped to a computationally modeled secondary structure, the highest concentration of SHAPE reactive nt falls along an arm of HOTAIR D4. The potential reasons and implications of these data are discussed herein.

5.2.1 Protocol and Optimization

5.2.1.1 Protocol

The standard SHAPE-MaP protocol utilized is as follows (Figure 74). RNA is equilibrated in folding buffer at 37 °C. Post-equilibration, a concentrated solution of SHAPE-reagent in DMSO is added to each sample and incubated for 10 minutes to allow for modification to occur. Post-modification, the samples are cleaned via GE healthcare G-50 spin column which allows for removal of unreacted SHAPE-reagent. The production of cDNA from the modified RNA utilizes NEB super script II (SSII) reverse transcriptase coupled with an in-house buffer and thermal cycler protocol as reported to allow for the reverse transcriptase to continue passed the modified nt *(401,402)*. Upon reaching a modified nt, instead of stalling and producing a truncated cDNA transcript,

the current protocol enables the reverse transcriptase to accommodate the SHAPE-modified nt, with the high likelihood of incorporating a mismatch base pair at the site of modification. These modifications lay the basis for comparison of samples treated with SHAPE-reagent versus a DMSO control. Post-production of cDNA, DNA is produced using a standard NEB second strand synthesis protocol.

To prepare the samples for next generation sequencing, adaptors must be incorporated to the ends of the sequences through utilization of a standard Nextera preparation protocol wherein the DNA are subjected to a tagmentation protocol. This produces smaller DNA fragments to enable better coverage during the NGS run. A subsequent PCR reaction allows for the incorporation of unique index primers at each end of the DNA fragments for each sample to enable separation between sample sequences in the analysis post-NGS.

After NGS, the samples are subjected to SHAPEmapper which separates the sequences by index primers, assesses the quality of each sample, and reports SHAPE-reactivities at a nt level. Superfold is additionally utilized due to its ability to provide information on overall RNA stability and incorporate SHAPE-reactivities into a free energy secondary structure computational RNA model.

To understand the changes in SHAPE-reactivities and structure due to the presence of KDM1A:CoREST$_{286-493}$, I envisioned utilizing SHAPE-MaP. I chose four samples, (+/-) SHAPE-reagent and (+/-) protein to subject to the SHAPE-MaP protocol. I

hypothesized that we would be able to ascertain which nucleotides were affected by the presence of protein which would, in turn, enable us to narrow down regions of HOTAIR D4 that were potentially forming a binding interface with KDM1A:CoREST$_{286-493}$.

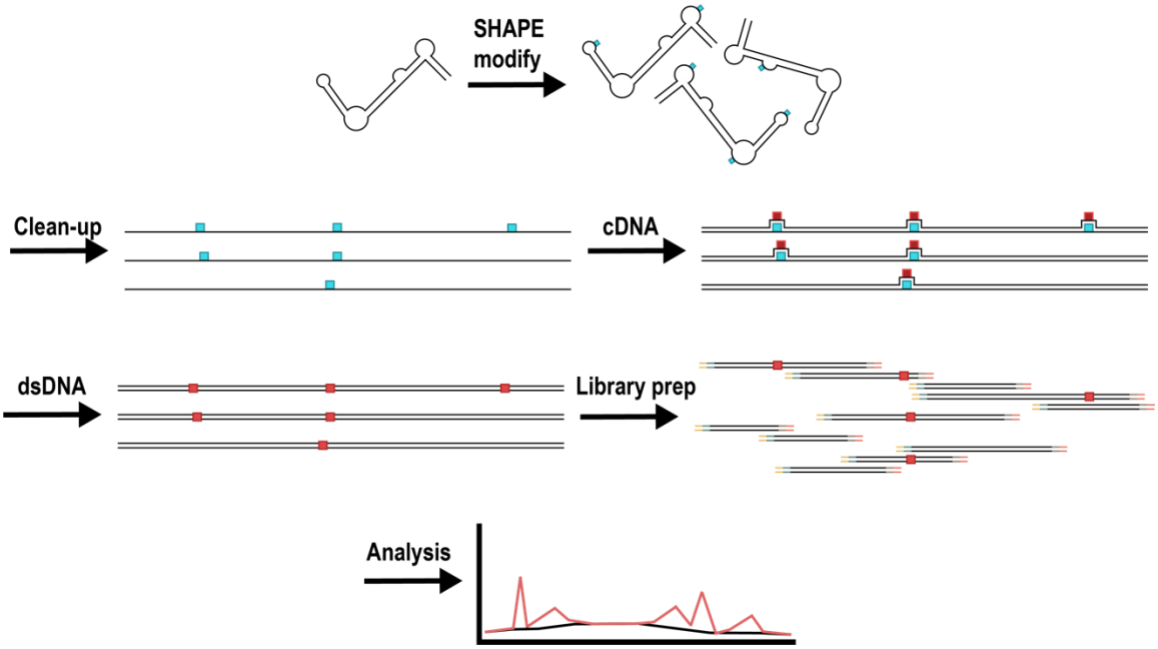


Figure 74: SHAPE-MaP workflow. An RNA is incubated with a SHAPE-modifying reagent such as 5-NIA. Excess reagent is removed via GE Healthcare G-50 spin column, SPRI bead clean up, and subjected to reverse transcriptase (NEB SSII). Following production of cDNA, dsDNA is produced via a second strand synthesis procedure before Nextera library preparation, MiSeq NGS run, and subsequent computational analysis with SHAPEmapper and Superfold

5.2.1.2 SHAPE-MaP efforts and optimization

I began with utilization of the initial SHAPE-MaP protocol provided by the Weeks lab (Figure 74). Briefly, 2 pmol of RNA was snap cooled and then incubated for 20 minutes at 37 °C in a folding buffer (100 mM HEPES (pH 7.4) 100 mM NaCl, and 10 mM $MgCl_2$). KDM1A:CoREST$_{286-493}$ was added at a 50 μM concentration and the

reactions were incubated for an additional 20 minutes at 37 °C. A solution of 5-NIA (250 mM or DMSO control) was added and the reactions run at 37 °C for 10 minutes.

The original protocol at this point called for a clean-up of the reaction simply via a GE Healthcare G-50 spin column proceeded by an Agencourt RNAclean XP bead clean-up to separate the protein from the RNA before proceeding to reverse transcription. However, I was, consistently, unable to recover appreciable quantities of the Protein(+) control (5-NIA(-)) as quantified by Qubit.

I hypothesized that the interaction between HOTAIR and KDM1A:CoREST$_{286\text{-}493}$ was disrupted by the 5-NIA in a way which affected the binding affinity enough to allow for separation via SPRI beads. However, due to either a high binding affinity, long interface, or both, the unmodified reaction was unable to be separated by SPRI beads alone. Through several test reactions, we were able to establish that a 1-hour proteinase K incubation at 37 °C post-SHAPE-reagent treatment and a subsequent Trizol extraction allowed for recovery of RNA (Figure 75).

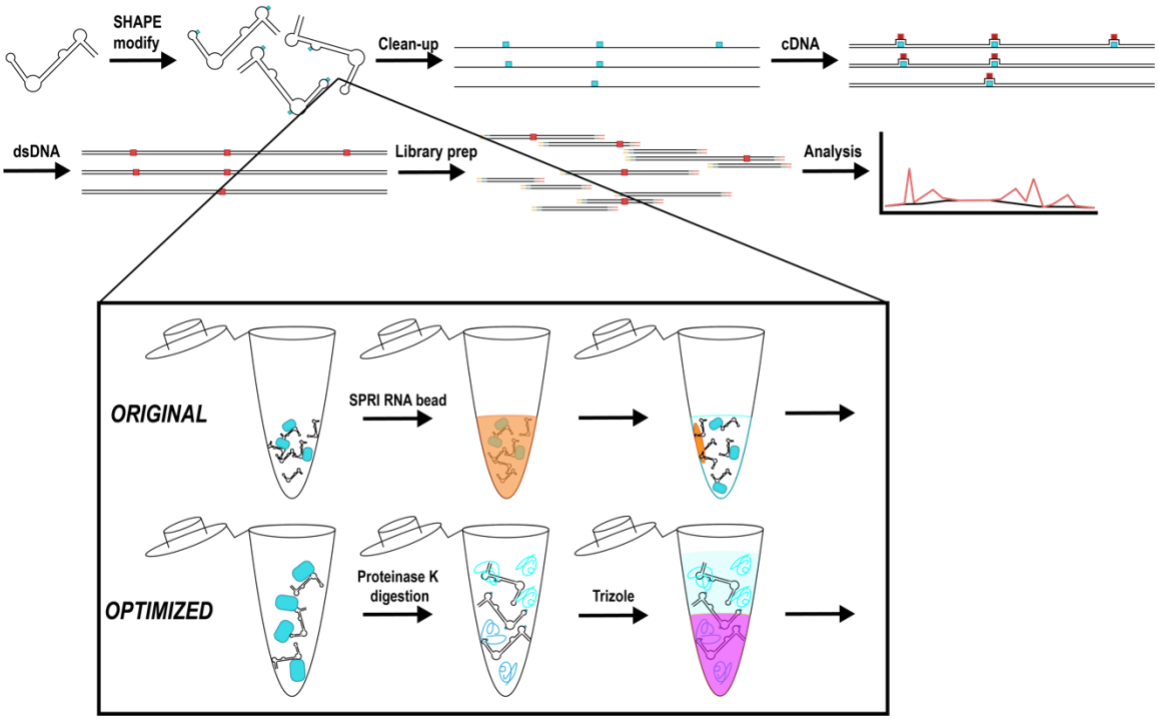


Figure 75: Optimization of SHAPE-MaP for *in vitro* system. Due to the nature of the RNA:protein interaction, we had to implement a more stringent clean-up protocol including a proteinase K digestion and Trizol cleanup as opposed to the original SPRI RNA bead cleanup

The recovered RNA was then subjected to reverse transcription and subsequent second strand synthesis as described in the methods. Due to the length of HOTAIR D4, we chose to utilize a Nextera preparation prior to NGS. The tagmentation step in this preparation procedure ensures more complete sequencing coverage. The prepared library was run utilizing a MiSeq v2 kit and the resulting data was analyzed as described below.

5.2.2 SHAPE-MaP Results

5.2.2.1 Establishing nucleotides which demonstrate a significant change in SHAPE-reactivity

Following optimization of the *in vitro* SHAPE-MaP protocol utilizing 5-NIA as the SHAPE nucleobase modifying reagent, we were able to obtain raw reactivities HOTAIR D4 through the SHAPEmapper2 program developed by the Weeks lab *(468)*, in duplicate. For each sample, we established the SHAPE reactivity of 5-NIA modification compared to a DMSO control through the natural log fold change of the two raw values for the RNAonly and RNA:protein samples separately.

To compare the RNAonly and the RNA:protein SHAPE reactivities and account for biological sample variation, we normalized the SHAPE reactivities between samples via minimizing the median value between the two sets of SHAPE reactivities. Utilizing two tiers of stringency to look at nucleotides of interest, 1.5– and 2.0–fold cutoff, we looked at the absolute fold change of RNAonly to RNA:protein SHAPE reactivities for each replicate (Figures 76 and 77).

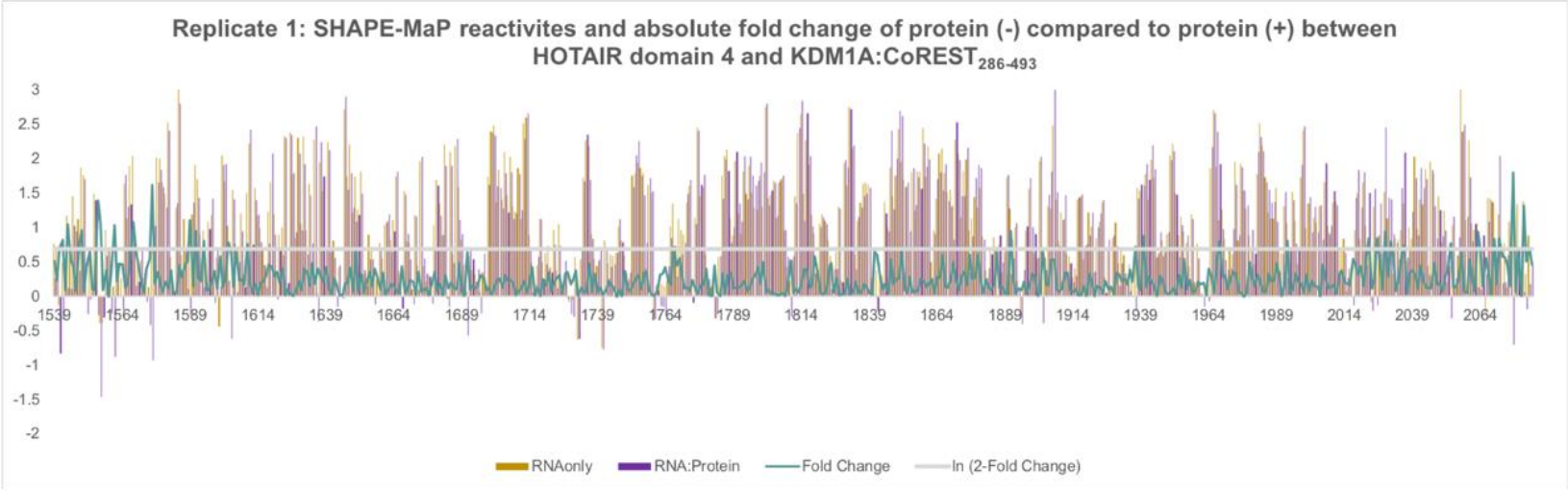


Figure 76: SHAPE-MaP reactivities of HOTAIR D4 for the first replicate. The samples were incubated 250 mM 5-NIA incubation in a folding buffer (50 mM HEPES (pH 7.4), 100 mM NaCl, 10 mM $MgCl_2$, 37°C, 10 min). We utilized an optimized clean-up procedure prior to NGS sequencing. We utilized SHAPEmapper to analyze the data, shown above. The bars represent the reactivities of the RNAonly and RNA:protein samples and the green line represents the ΔSHAPE reactivity due to the presence of protein.

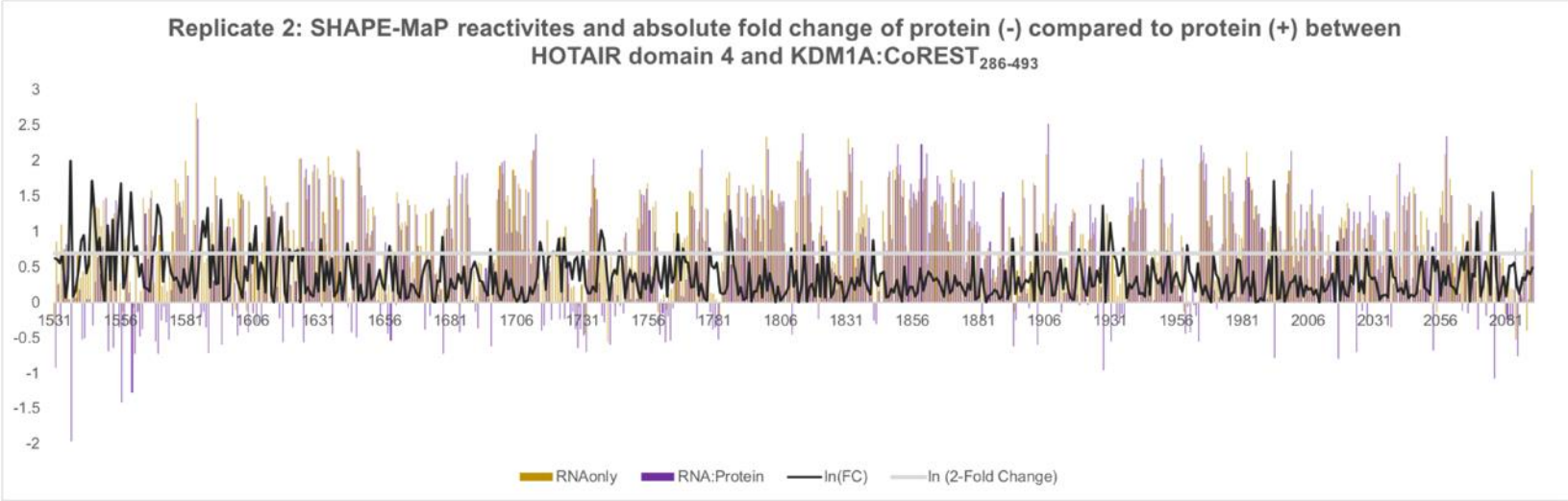


Figure 77: SHAPE-MaP reactivities of HOTAIR D4 for the second replicate. The samples were incubated 250 mM 5-NIA incubation in a folding buffer (50 mM HEPES (pH 7.4), 100 mM NaCl, 10 mM $MgCl_2$, 37°C, 10 min). We utilized an optimized clean-up procedure prior to NGS sequencing. We utilized SHAPEmapper to analyze the data, shown above. The bars represent the reactivities of the RNAonly and RNA:protein samples and the green line represents the ΔSHAPE reactivity due to the presence of protein.

To establish which nucleotides had a significant fold change due to the presence of KDM1A:CoREST$_{286-493}$, I next determined the nucleotides that demonstrated an absolute fold change of at least 1.5 across both replicates (Table 24), and which additionally displayed a 2.0–fold change across both replicates (Figure 78, Table 25). I found 83 nt displayed a 1.5–fold change, and 22 further displayed a 2.0–fold change.

I also established whether the fold change was due to an increase or decrease in SHAPE reactivity for each of these nucleotides removing those which had opposing fold change directions between the two replicates. Interestingly, only two nucleotides at the 1.5–fold change, and none at the 2.0–fold change, cutoffs displayed an increase in SHAPE reactivity due to the presence of KDM1A:CoREST$_{286-493}$.

A decrease in SHAPE-reactivity is most likely due to a decrease in flexibility of the 2′-hydroxyl group via stabilizing interactions. In this case, the decreased reactivity is likely due to stabilization of the 83 nt due to new interactions either, directly, with the protein or, indirectly, through induced conformational shifts from ssRNA to dsRNA in the presence of protein.

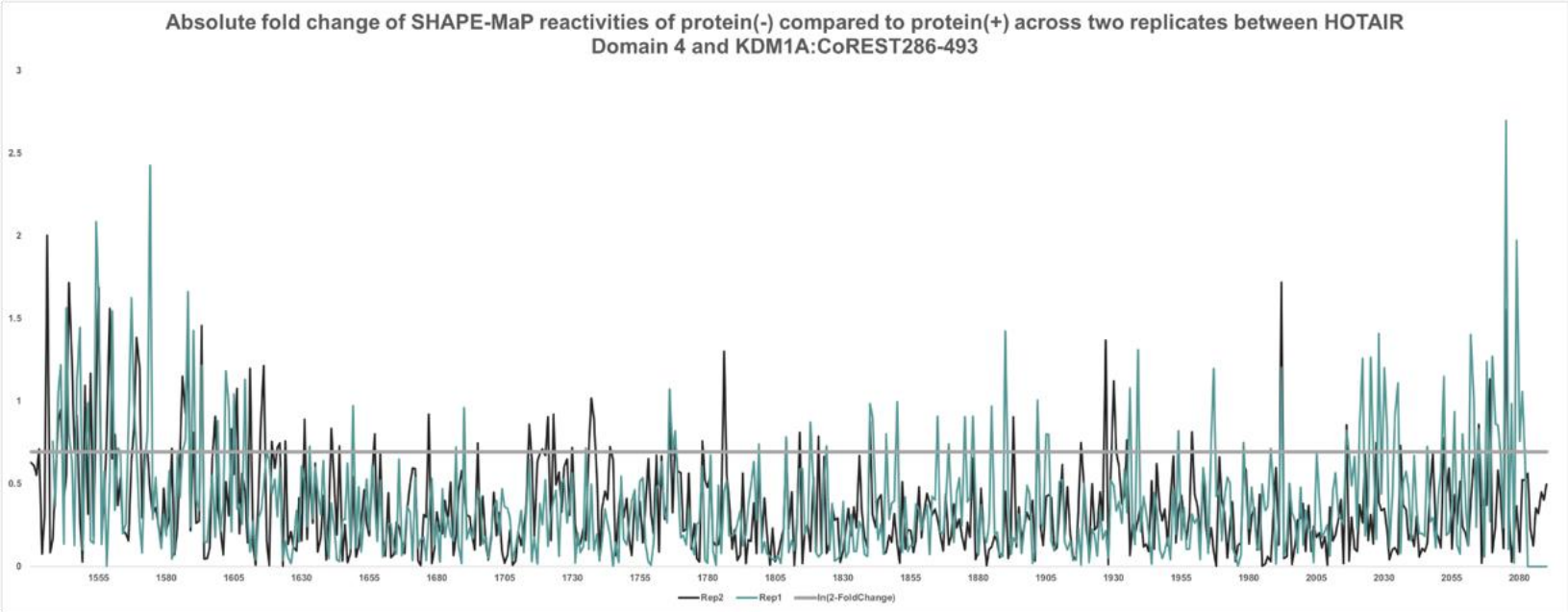


Figure 78: Absolute fold change of SHAPE-MaP reactivities across two replicates, normalized for technical variation between samples. A gray line denotes the cut-off of a 2-fold change between (+) and (-) protein samples

Table 24: HOTAIR D4 nucleotides and corresponding nucleobase demonstrating a 1.5-fold change across both replicates

HOTAIR D4 Nt	*Nucleobase*	*Change due to protein*
1541	C	Decrease
1542	A	Decrease
1544	C	Decrease
1545	C	Decrease
1546	A	Decrease
1548	G	Decrease
1551	U	Decrease
1555	C	Decrease
1556	C	Decrease
1560	C	Decrease
1561	U	Decrease
1564	U	Decrease
1568	C	Decrease
1569	G	Decrease
1570	C	Decrease
1573	C	Decrease
1574	U	Decrease
1575	C	Decrease
1586	U	Decrease
1587	G	Decrease
1588	G	Decrease
1589	C	Decrease
1591	C	Decrease
1594	C	Decrease
1598	U	Decrease
1603	C	Decrease
1606	G	Decrease
1610	C	Decrease
1617	C	Decrease
1620	C	Decrease

1621	U	Decrease
1623	U	Decrease
1632	C	Decrease
1636	U	Decrease
1650	G	Decrease
1657	U	Decrease
1660	U	Decrease
1673	U	Decrease
1691	C	Decrease
1696	U	Decrease
1715	G	Decrease
1721	G	Decrease
1725	C	Decrease
1728	G	Decrease
1731	G	Decrease
1738	G	Decrease
1749	U	Decrease
1754	A	Increase
1762	C	Decrease
1764	C	Decrease
1767	C	Decrease
1769	U	Decrease
1770	G	Decrease
1779	G	Decrease
1782	G	Decrease
1787	A	Decrease
1788	G	Decrease
1810	C	Decrease
1815	A	Decrease
1820	U	Decrease
1824	G	Decrease
1841	G	Decrease
1879	A	Increase
1891	U	Decrease
1903	G	Decrease

1912	U	Decrease
1920	C	Decrease
1930	C	Decrease
1931	U	Decrease
1935	G	Decrease
1953	G	Decrease
1964	U	Decrease
1970	G	Decrease
1979	C	Decrease
1993	A	Decrease
1996	U	Decrease
2017	C	Decrease
2053	U	Decrease
2057	U	Decrease
2064	G	Decrease
2073	G	Decrease
2076	C	Decrease
2082	G	Decrease

Table 25: HOTAIR D4 nucleotides and corresponding nucleobase demonstrating a 2.0-fold change in reactivity across both replicates

HOTAIR D4 Nt	*Nucleobase*	*Change due to protein*
1541	C	Decrease
1542	A	Decrease
1545	C	Decrease
1548	G	Decrease
1555	C	Decrease
1556	C	Decrease
1569	G	Decrease
1570	C	Decrease
1587	G	Decrease
1588	G	Decrease
1589	C	Decrease
1591	C	Decrease

1594	C	Decrease
1767	C	Decrease
1769	U	Decrease
1810	C	Decrease
1841	G	Decrease
1903	G	Decrease
1993	A	Decrease
2017	C	Decrease
2053	U	Decrease
2076	C	Decrease

5.2.2.2 Superfold demonstrates increases in structural stability due to the presence of KDM1A:CoREST$_{286-493}$

To assess the change in predicted stability of the nt due to the presence of protein, we next assessed the structural stability of HOTAIR D4 both alone and under the presence of KDM1A:CoREST$_{286-493}$ with the Superfold programming system established by the Weeks lab *(469)*. Superfold utilizes SHAPE-MaP reactivities to establish a partition function across each input RNA to assess regions that are well-defined over those that are not. It should be noted that the current method utilized in Superfold assumes that base pairs are local within an RNA and there are no pairings between nucleotides over 600 bases apart, based on evidence that looking at local folding in lncRNA is more accurate than utilizing traditional global minimal free energy *(470)*.

With a partition function in hand, Superfold next assesses the Shannon entropy at a nucleotide level. In this case, the Shannon entropy represents the likelihood of a base pair, and thus, how well established a particular structure is within a local region of 55 nucleotides within the overall RNA. A high Shannon entropy region paired with a high SHAPE reactivity suggests over that region of nucleotides there is a high flexibility and multiple conformations are likely. Conversely, a low Shannon entropy region coupled with low SHAPE reactivity suggests a well-defined conformation. Visualized via arc plots, these calculations can inform regions of stability or instability of structure within an RNA, and changes in stability of conformation upon the introduction of protein in our system (Figures 79 and 80). The overlaid red bars note the first nucleotide along the linear sequence in a predicted base pair which exhibit at least a 50% increase in stability upon introduction of protein. We noted significant overlap between the nucleotides which demonstrate a 50% increase in stability and the nucleotides which demonstrated a significant decrease in SHAPE-reactivity. This suggests that, indeed, these nucleotides are undergoing a stabilization due to the presence of protein, confirming the interaction between KDM1A:CoREST$_{286\text{-}493}$, and potentially indicating nucleotides of interest that may contribute to the binding affinity.

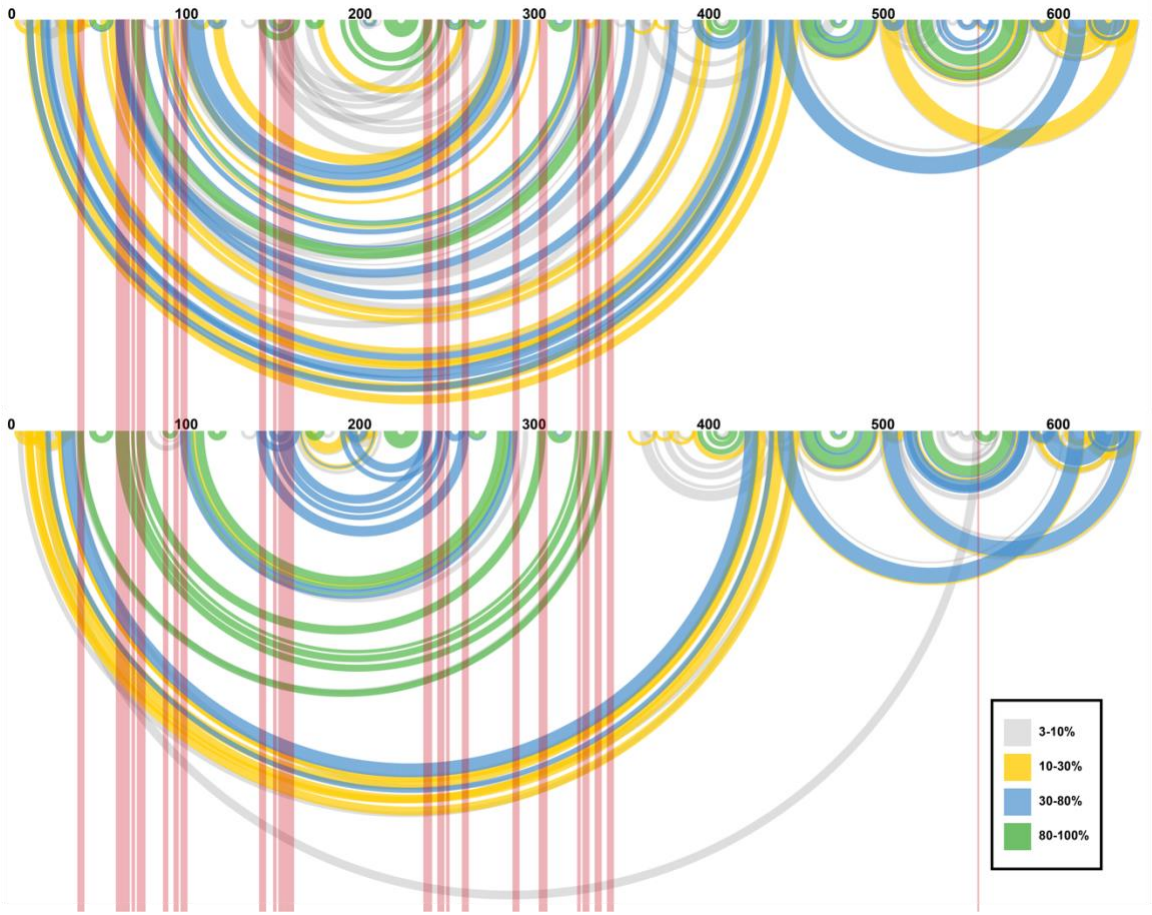


Figure 79: Superfold Arc plot comparing the stability of base paired nucleotides with (top) and without (bottom) KDM1A:CoREST$_{286\text{-}493}$ wherein the red bars represent increases instability by 50% for the first replicate.

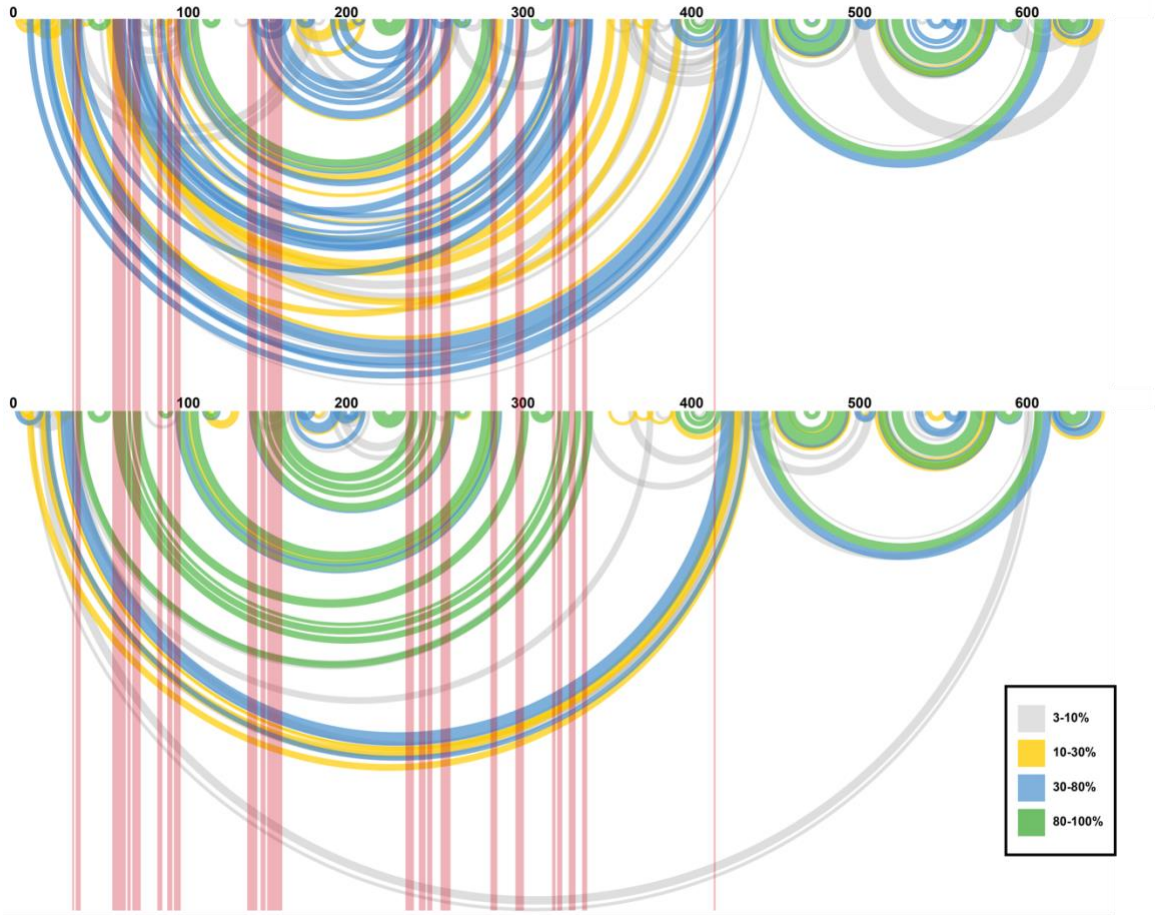


Figure 80: Superfold Arc plot comparing the stability of base paired nucleotides with (top) and without (bottom) KDM1A:CoREST$_{286-493}$ wherein the red bars represent increases instability by 50% for the second replicate.

5.2.2.3 HOTAIR D4 secondary structure demonstrates high level of secondary structure

These calculations, along with SHAPE-reactivities, were incorporated into a minimal free energy and structural analysis to output the most likely structure of the lncRNA based on the free energy coupled with SHAPE-reactivities *(464)*. We observed that between the two replicates, the structures were, as expected, extremely similar. The

NGS sequencing for the second replicate had a higher coverage, and we utilized this to display the nucleotides which displayed a change due to the presence of protein.

We first compared our structure to that of the previously reported structure by the Pyle lab *(251)*. To our satisfaction, the structures overall appeared closely aligned. The largest structural difference we noted was that two large loops in our structure, labeled as L1 and L5 in Figure 81, were split into a loop followed by a hairpin/stem motif in the Pyle lab structure. In addition to SHAPE-MaP, the Pyle lab additionally utilized DMS and terbium treatment to establish their SHAPE-MaP structure, so it is not surprising that our structures are not in perfect agreement. However, overall, the general bulge, stem, and hairpin motifs, remain structurally in agreement.

HOTAIR D4 is significantly structured, in our model we note 20 helices, 17 terminal loops (7 being tetraloops), 7 internal loops with more than 2 junctions (3 being 3-way junctions), and 7 bulges. Both tetraloops as well as 3-way junctions have both been demonstrated to have functions both in protein recognition RNA:protein binding, in the former, and protein binding and ribozymes, in the latter *(435,471–476)*. Many lncRNA, particularly those with biological functions, have been demonstrated to have a much higher level of secondary structure compared to mRNA *(477)*. Thus, verification of the structural complexity of HOTAIR D4 compared to previous reports as well as noting the several biologically-relevant motifs was promising.

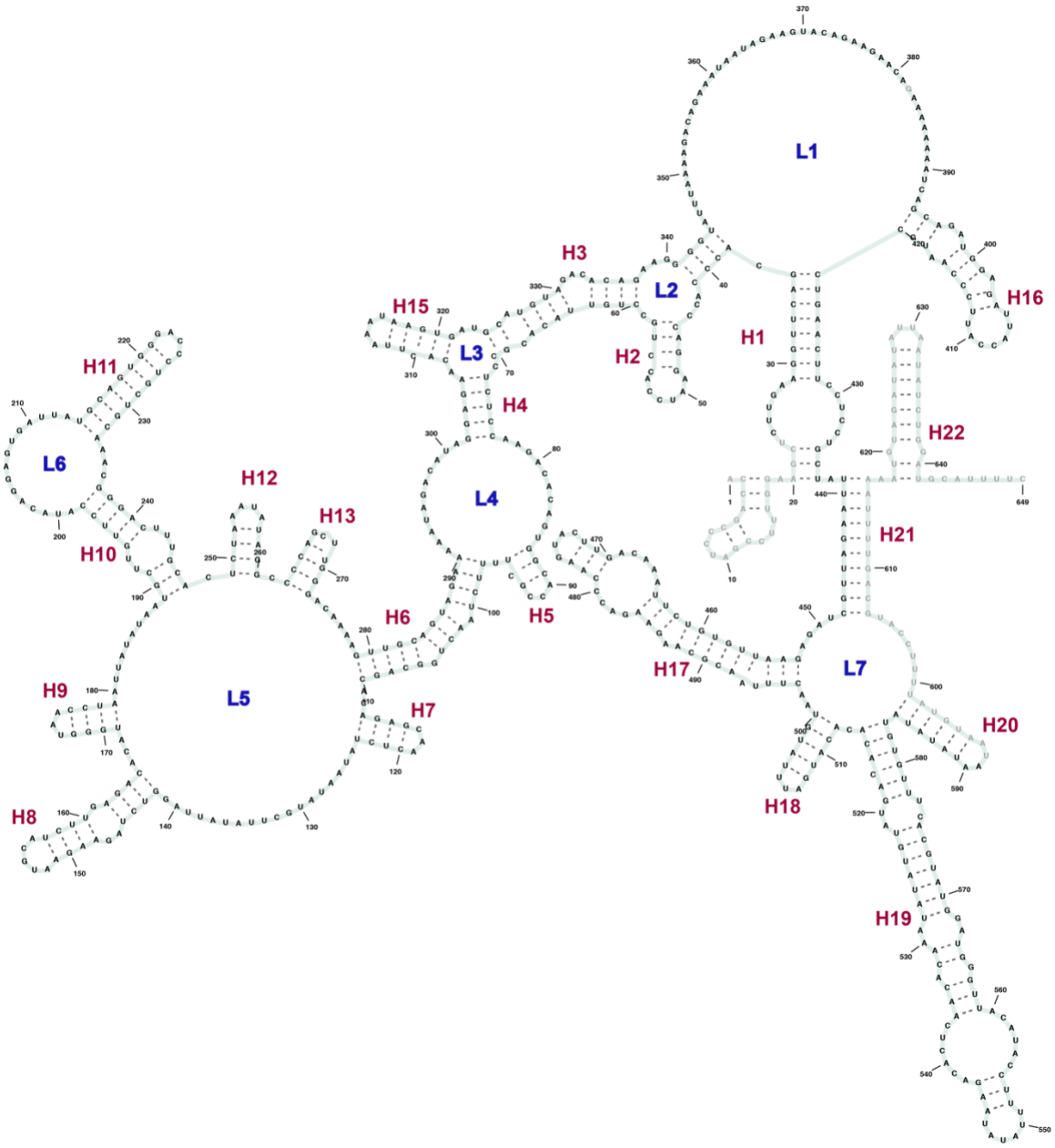


Figure 81: Computationally modeled secondary structure of HOTAIR D4. The computational program utilized, Superfold, incorporates the SHAPE-reactivities derived from 5-NIA modification of HOTAIR D4, and subsequent workup and NGS sequencing. Our structure is closely aligned with previously reported structures and

contains a significant level of secondary structure including multiple biologically relevant motifs including 3-way junctions and tetraloops.

5.2.2.4 A concentration of significant SHAPE-reactive nucleotides appears along a HOTAIR D4 arm

Finally, we overlaid the 83 nucleotides which demonstrated a significant fold change between RNAonly and RNA:protein samples, for both a 1.5–fold change (orange) and a 2.0–fold change (red) (Figure 82). Interestingly, we noted that, while the significant changes in SHAPE reactivities were spread throughout HOTAIR D4, there was a larger number of nucleotides with a significant SHAPE reactivity fold change, particularly at a 2–fold change, along an arm spanning nucleotides 60-110 and 257-379, which we highlighted in Figure 83.

While this arm contains the highest concentration of SHAPE-reactive nucleotides with a change in reactivity due to the presence of KDM1A:CoREST$_{286\text{-}493}$, it is interesting that the remaining significant nucleotides span the entire domain. This could suggest global structural changes to the RNA, or perhaps disruption of a tertiary structure upon protein binding.

We have also modeled the SHAPE-reactive nucleotides with a change in reactivity due to the presence of protein onto a second Superfold computational structure (Figure 84b). This structure represents the model which Superfold predicts the secondary structure of the protein based on the RNA:protein SHAPE-reactivities, as opposed to the RNAonly SHAPE-reactivities that generated the initial structure.

Superfold is unable to account for the presence of protein and is not indicative of the secondary structure when protein is present, but changes to the structure do, again, suggest that protein is binding (Figure 83b). We note that overall the structures remain similar with the exception of the closing of what was denoted as L5 and is highlighted in green in Figure 83.

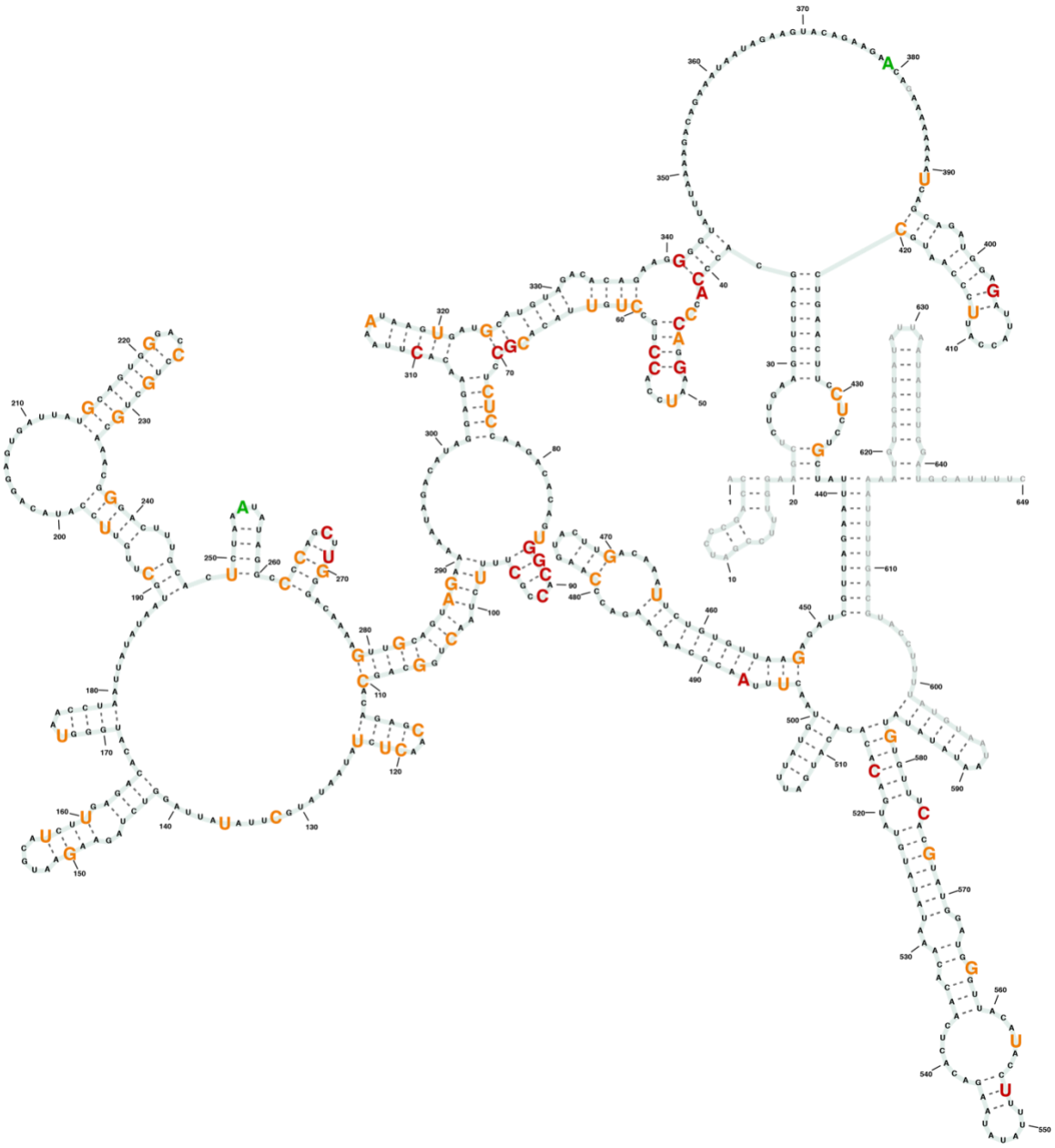


Figure 82: Overlay of the SHAPE-reactive nucleotides with a significant change in SHAPE-reactivity and the Superfold predicted secondary structure of HOTAIR D4 alone.

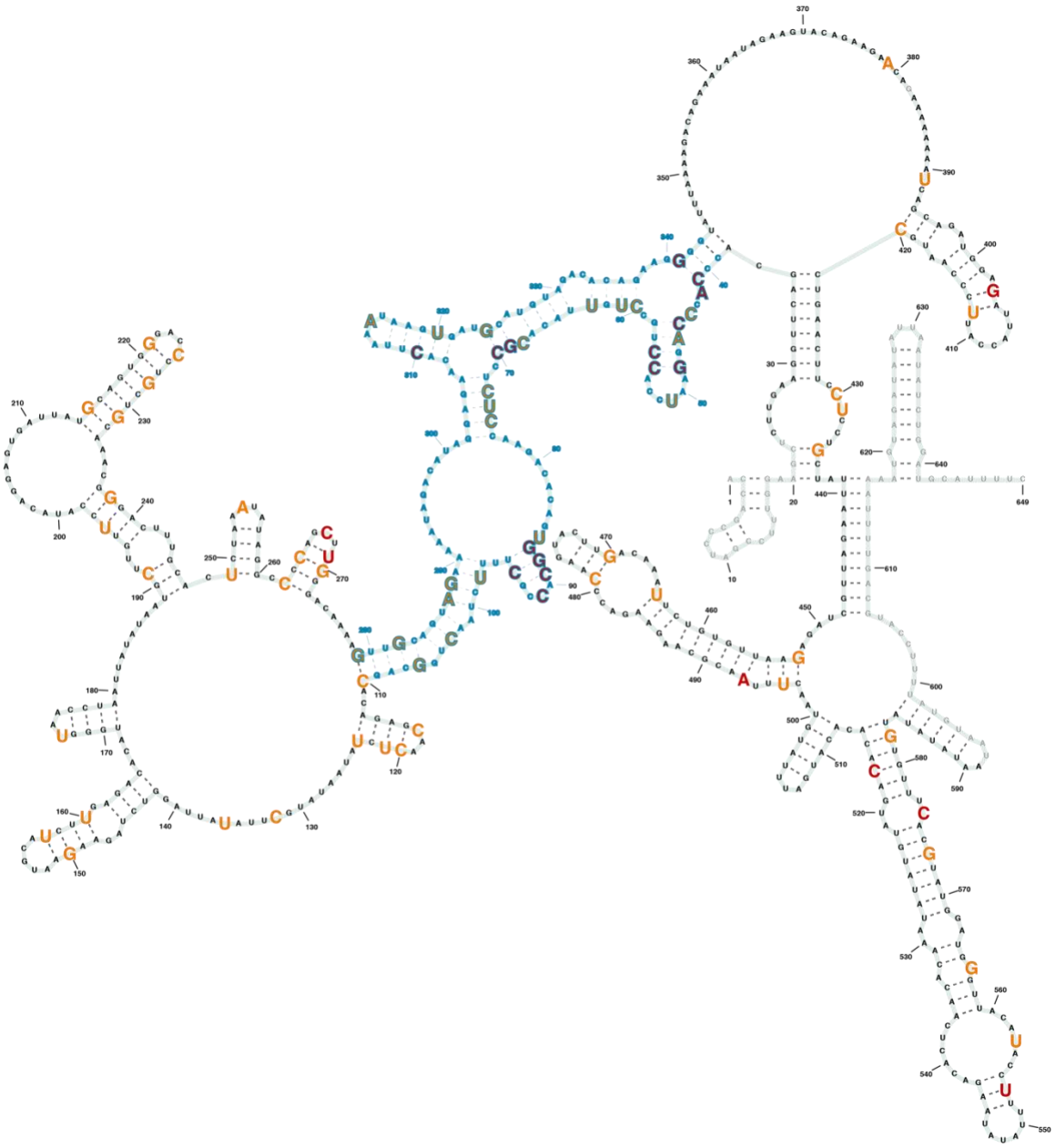


Figure 83: Arm within HOTAIR D4 displays the highest concentration of SHAPE-reactive nucleotides which decrease in reactivity due to the presence of KDM1A:CoREST$_{286\text{-}493}$. Of the 22 nt displaying 2.0–fold decrease in SHAPE-reactivity, 15 are found within this arm, and 44% of the 83 nt with at least a 1.5 –fold decrease in SHAPE-reactivity are concentrated to this arm.

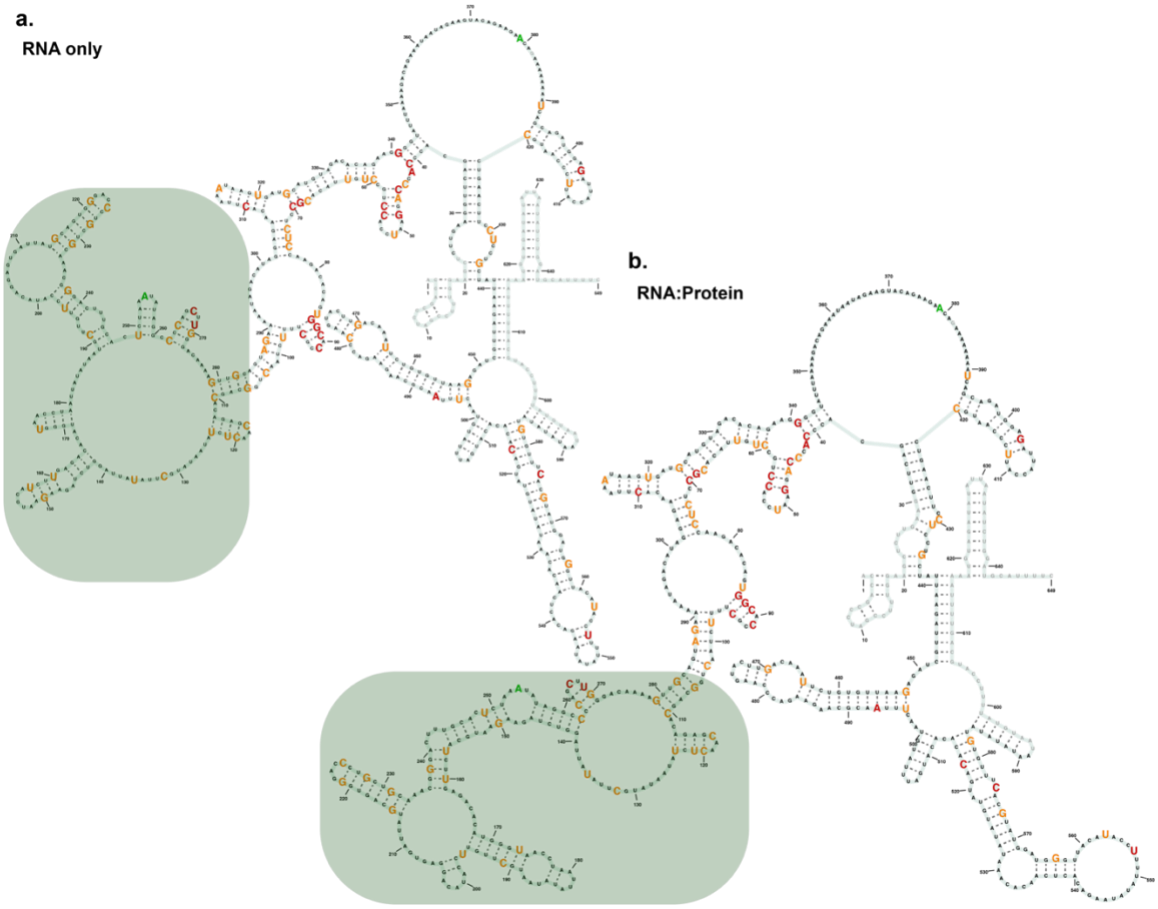


Figure 84: Side-by-side comparison of the computational models of HOTAIR D4 based on SHAPE-MaP reactivities with and without KDM1A:CoREST$_{286\text{-}493}$

Table 26: Percentage of each nucleobase for a 1.5– and 2.0–fold change

Nucleobase	*1.5 –fold change*	*2 –fold change*
A	8.43	8.70
U	27.71	13.04
G	28.92	26.09
C	34.94	52.17

5.2.3 SHAPE-MaP Discussion

We have utilized SHAPE-MaP as an RNA footprinting technique to assess changes in HOTAIR D4 due to the presence of KDM1A:CoREST$_{286\text{-}493}$. The optimization of the protocol to include a proteinase K digestion suggested that, while this interaction may not be tight binding based on the EMSA binding studies we performed in Chapter 4, there is an extensive network of hydrogen bonding which disallows the easy separation via carboxylate-coated magnetic beads. This is of interest to those utilizing SHAPE-MaP in an *in vivo* method, as there may be critical RNA:protein interactions which are not being elucidated due to loss of RNA in the clean-up process.

Upon optimization of the protocol, we first attempted SHAPE-MaP with HOTAIR$_{1500\text{-}1820}$ with and without the presence of KDM1A:CoREST$_{286\text{-}493}$ and found that while we observe binding in the filter binding assays, the structure we obtained did not correlate with previously established structures (data not shown). While this suggested a potential sequence-dependence, instead of a structural-dependence, of protein binding, we next sought to establish the structure of HOTAIR D4, which has been previously reported by the Pyle lab. Gratifyingly, the overall structure we obtained for HOTAIR D4 without protein closely resembled that of the previously reported Pyle lab HOTAIR D4 structure *(251)*.

We were able to obtain SHAPE-MaP reactivities for HOTAIR D4 with and without the presence of KDM1A:CoREST$_{286\text{-}493}$ in duplicate. Looking first at the nucleotides that demonstrated a 1.5– or 2.0–fold change as a function of protein, we noted that the significant nucleotides displayed a bias towards guanine and cytosine bases (Table 26). We have found no reported bias towards these nucleotides in other SHAPE studies, or any explanations based on the chemical reaction thus suggesting this bias towards regions of HOTAIR D4 is real.

Upon mapping the nucleotides with a significant fold change, we were able to take note of several interesting features. Firstly, we confirm that HOTAIR is highly structured. While most RNA likely form some level of secondary structure, it is generally believed that those that display a higher degree of secondary motifs and structural components possess a functional *(409,477)* component as opposed to mRNA which, overall, remain single stranded. Secondly, we noted changes to SHAPE reactivity across the entire HOTAIR D4 structure, which suggests a global structural shift due to the presence of protein. There is precedence for a large conformational shift of RNA induced by protein binding, so this was not overly surprising *(429,453,478,479)*. It was, however, interesting that only one nucleotide displayed a 1.5–fold increase in SHAPE reactivity. This suggests that there is an overall structural stabilization of HOTAIR D4 when in the presence of KDM1A:CoREST$_{286\text{-}493}$, presumably due to binding. As some

RNA are only able to be crystalized in the presence of protein, the stabilization offered by proteins is not unexpected.

Comparing the regions that displayed stabilization, we noted that this correlated with an arm (residues 1538-1610, 1779-1844). This arm displayed the most concentrated level of SHAPE reactivity decrease, and the highest number of nucleotides with a 2.0–fold decrease. This correlation between an increase in stability at the region of high SHAPE reactivity decrease suggests that this region is particularly impacted by the presence of protein, and likely plays a role in the binding interface.

The potential importance of this binding arm for an interaction with KDM1A:CoREST$_{286\text{-}493}$ is mirrored in our EMSA assays wherein only the first 300 nt of HOTAIR D4 were able to display binding. This binding arm is almost entirely comprised of this first 300 nt fragment of HOTAIR, and thus may retain enough of the elements that are required for binding KDM1A:CoREST$_{286\text{-}493}$.

There are several cases in which proteins interact along a larger 40-60 nt sequence which is primarily stem, likely A-form, RNA *(397,480)*. While SHAPE-MaP does not provide direct evidence of protein binding, the concentration along a specific arm suggests that this arm is critical towards HOTAIR D4 binding to KDM1A:CoREST$_{286\text{-}493}$.

SHAPE-MaP provided us insight into the changes in reactivity due to the presence of protein. We noted that there is an arm within HOTAIR D4 that has an increased concentration of significant fold-change of SHAPE reactivities, which we have named the binding arm. Additionally, this region correlates with paired nucleotides which display an increase instability due to the presence of protein. These data suggest to us that this arm is necessary for KDM1A:CoREST$_{286\text{-}493}$ binding. We hoped to utilize RNA:protein crosslinking to assess which nucleotides were, in fact, interacting with KDM1A:CoREST$_{286\text{-}493}$.

5.3 RnP-MaP

Having utilized SHAPE-MaP to assess the changes in RNA secondary structure due to the presence of KDM1A:CoREST$_{286\text{-}493}$, we sought to establish which specific nucleotides were directly interacting with the proteins. A post-doc in the Weeks lab, Dr. Chase Weidmann, has designed a photochemical cross-linking technique, RnP-MaP, which enables the crosslinking of RNA nucleotides in close proximity to their protein partners. Upon crosslinking, the protein is degraded, and treated with the same MaP protocol as described in section 5.2.1.2. We chose to utilize this crosslinking strategy due to its complementarity to the SHAPE-MaP data we had previously collected.

Similar to SHAPE-MaP, we needed to optimize this protocol for an *in vitro* purified system as opposed to an *in vivo* sample. Following optimization, we observed

32 nucleotides with a significant upregulation of mutation rate, suggesting that these nucleotides were directly crosslinking with KDM1A:CoREST$_{286\text{-}493}$. Of these nucleotides, 84.38% were either guanine or cytosine and 62.50% of those were base pairing. We also note the clustering along the same arm as previously shown to be significant in SHAPE-MaP, and a potential for a 4-mer sequence motif that occurs within the RnP-activated nucleotides. In the discussion herein, we discuss the potential binding implications these results have towards the nature of the RNA:protein binding interface.

5.3.1 Protocol and Optimization

The photochemical crosslinking RnP-MaP procedure utilizes a photochemical crosslinking reagent wherein the RNA is incubated with the reagent with and without protein. Upon quenching the chemical crosslinking reaction, the samples are subjected to 365 nm UV light to activate the photo-crosslinking reaction. In the original protocol, the samples are then treated similarly to SHAPE-MaP with the added step of a Proteinase K digestion after removal of excess reagent via GE Healthcare G-50 spin column (Figure 85). Proteolysis of the proteins within the samples allows for a small adduct to remain on the RNA to be incorporated into the cDNA sequence as a mutation, analogous to the SHAPE-reagent modification in SHAPE-MaP. Following NGS, the data are subjected to Shapemapper2, and then subsequently RNPMapper.

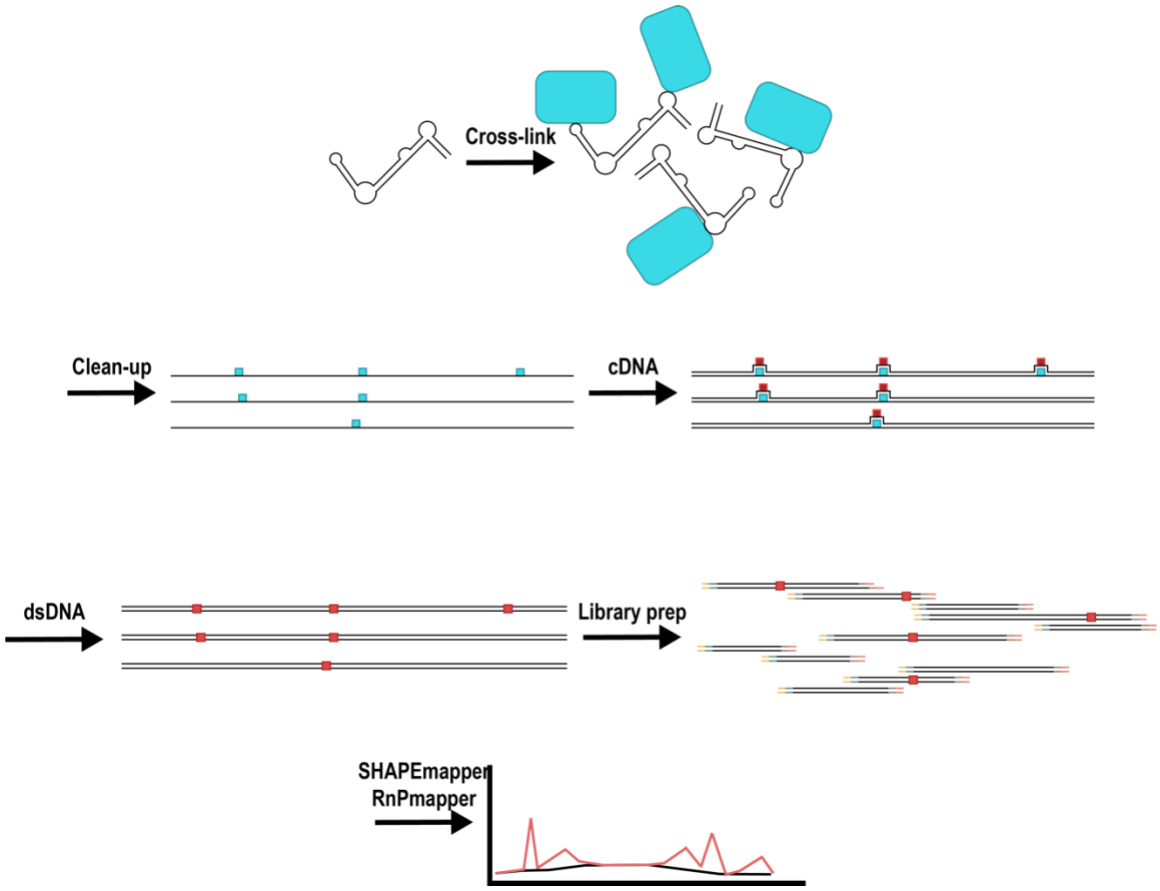


Figure 85: RNP-MaP workflow. An RNA is incubated with a SHAPE-modifying reagent such as 5-NIA. The resulting reaction mixture is cleaned up to remove excess reagent and subjected to reverse transcriptase. Following production of cDNA, dsDNA is produced before library preparation for NGS sequencing. The data are analyzed utilizing computational programs SHAPEmapper and RNPMapper

The technique that has been developed in the Weeks lab has solely been utilized for *in vivo* treatment of cells to recover whole cell RNA:protein interactions; however, we saw no reason why this technique could not be adapted for *in vitro* applications. We decided upon samples including RNA (+/- RnP-reagent) and (+/- KDM1A:CoREST$_{286\text{-}493}$).

Briefly, 2 pmol of RNA was heated to 95 °C and, following a snap cool, was incubated with a folding buffer containing 100 mM HEPES (pH 8.0), 100 mM NaCl, and 10 mM $MgCl_2$ for 20 minutes at 37 °C. Protein was then added to the RNA:protein samples to a final concentration of 50 mM and the resulting mixtures were incubated (37 °C, 20 min) to facilitate binding. NHS-diazirine (SDA), or DMSO control, was then added to a final concentration of 10 mM and incubated (37 °C, 10 min) before quenching with 50 mM Tris (pH 8.0). Each sample was then spread across a 12–well plate and the plate was crosslinked at 365 nm (3x, max energy). The crosslinked samples were then transferred to tubes with a 2x proteinase K buffer and incubated (37 °C, 1 hour). Following a Trizol extraction, the quantity of RNA was checked via Qubit wherein we discovered the RNA:protein DMSO control samples did not contain quantifiable levels of RNA.

We hypothesized that, similar to SHAPE-MaP, the size of the interface or the stability of the RNA:protein complex was inhibiting the separation between HOTAIR D4 and KDM1A-CoREST$_{286-493}$. We performed a Proteinase K time-course digestion, without RNP-MaP crosslinking treatment, to assess how rapidly Proteinase K was able to digest the proteins. I was able to recover over 90% of the RNA from the overnight sample, suggesting that the RNA would be stable for a longer proteinase K digestion.

After performing RNP-MaP treatment with an overnight proteinase K digestion

and utilizing a more stringent PCA extraction over a Trizol extraction, we were able to attain RNA from the RNA:protein DMSO control at quantities in line with quantities recovered from the other samples (Figure 86). The recovered RNA was then subjected to reverse transcription and subsequent second strand synthesis as described in the methods. Due to the length of Domain 4 of HOTAIR, we chose to utilize a Nextera preparation prior to NGS. The prepared library was run utilizing a MiSeq v2 kit and the resulting data was analyzed as described below.

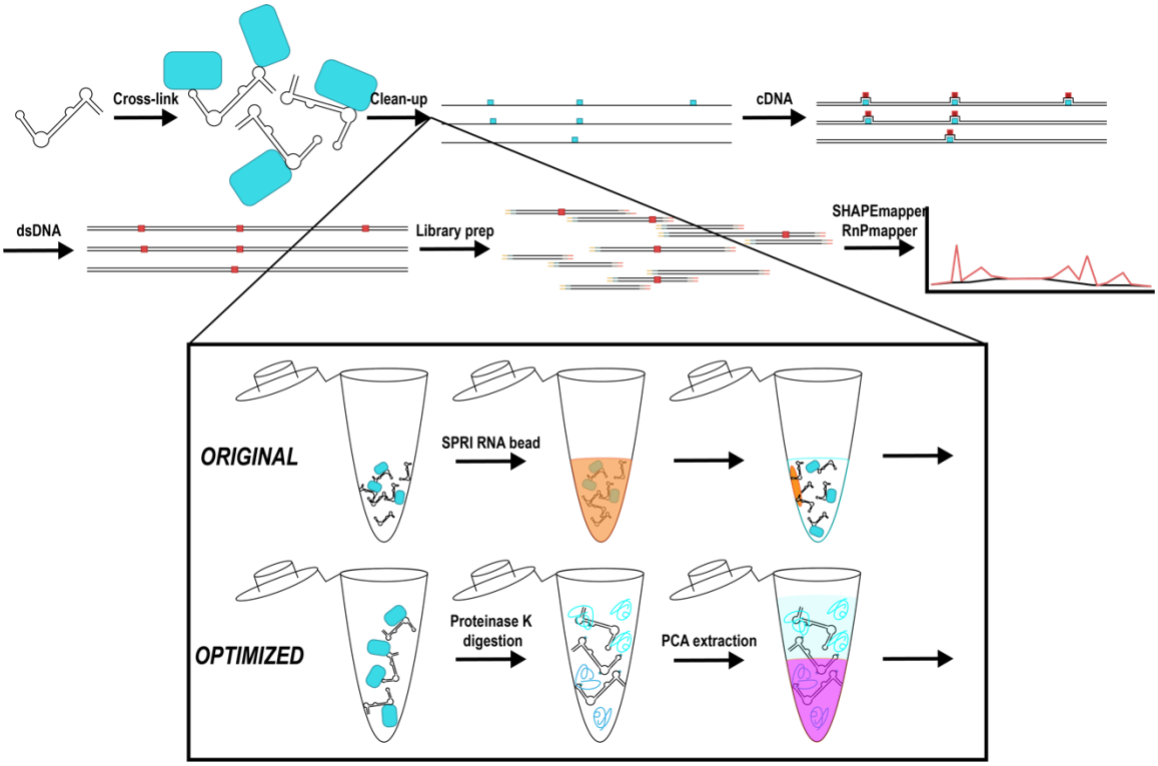


Figure 86: Optimization of RNP-MaP protocol for an *in vitro* system. Due to the nature of the RNA:protein interaction, we had to implement a more stringent clean-up protocol including a proteinase K digestion and more stringent PCA cleanup as opposed to the original SPRI RNA bead cleanup

5.3.2 RnP-MaP Data

Following NGS sequencing, the data were subjected to the ShapeMapper2 and the resulting profiles were submitted to RNP-MaPper 2 which calculates the raw difference in SHAPE reactivities between the DMSO control and the SDA treated samples. RNP reactivities are established through taking the raw difference for the RNAonly and the RNA:protein samples to the Z-score of the sample population. The resulting RNP reactivities for the RNAonly and RNA:protein samples were compared via natural log fold change. Based on pilot RNP-MaP studies performed by the Weeks lab, we utilized a 1.2–fold cutoff in addition to the nucleotides being in the top 10% percentile of changes in at least two of our three replicates to establish which nucleotides were likely interacting in protein by an increased RNP reactivity (Table 27), which resulted in 32 nucleotides.

Table 27: RNP-MaP nucleotides with a 1.2-fold cutoff in addition to being in the top 10% percentile of changes

Nucleotide	*Nucleobase*
1531	G
1535	A
1540	C
1541	C
1548	G
1555	C
1576	C
1579	G
1605	G
1634	U
1636	U
1655	C
1694	G
1732	C
1741	C
1760	G
1783	C
1796	G
1803	G
1815	A
1841	G
1857	G
1903	G
1930	C
1953	G
1960	G
1962	C
1983	G
1986	G
2024	U
2055	C
2065	G

It has been demonstrated that uracil and thymine have the highest rate of photochemical reactivity, so we expected a potential bias towards these nt *(481,482)*. As demonstrated in Table 28, I do not see any pyrimidine bias, in fact, 84.4% of the nucleotides likely making a direct interaction with protein are guanine and cytosine nucleobases, suggesting a binding preference by KDM1A:CoREST$_{286\text{-}493}$ for guanine and cytosine nucleobases over adenine and uracil.

Table 28: RNP-MaP percentage of nucleobases that displayed at least a 1.2-fold change due to the presence of protein

Nucleobase	*1.2 -fold change*
A	6.2
U	9.4
G	50.00
C	34.4

We then mapped where these nucleotides were structurally on the secondary structure representation we established with SHAPE-MaP. Upon placement of arrows to represent 1.2 – and 1.5 –fold change cutoffs (red, black, respectively), we noted several points of interest (Figure 87). Firstly, all but eight of the RnP-activated nucleotides were base paired. Furthermore, there were six instances where the RnP-activated nucleotides were base pairing with another RnP-activated nucleotide. All of these RnP-activated base pairs were between a guanine and a cytosine (Figure 89). Upon further examination, four of these base pairs, all within the binding arm, were flanked by either an additional C or G base pair and a subsequent adenine nucleotide. This suggests a potential sequence motif that either provides sequence-specific interactions for protein recognition or is mediating a structural element which is readily recognized by the proteins.

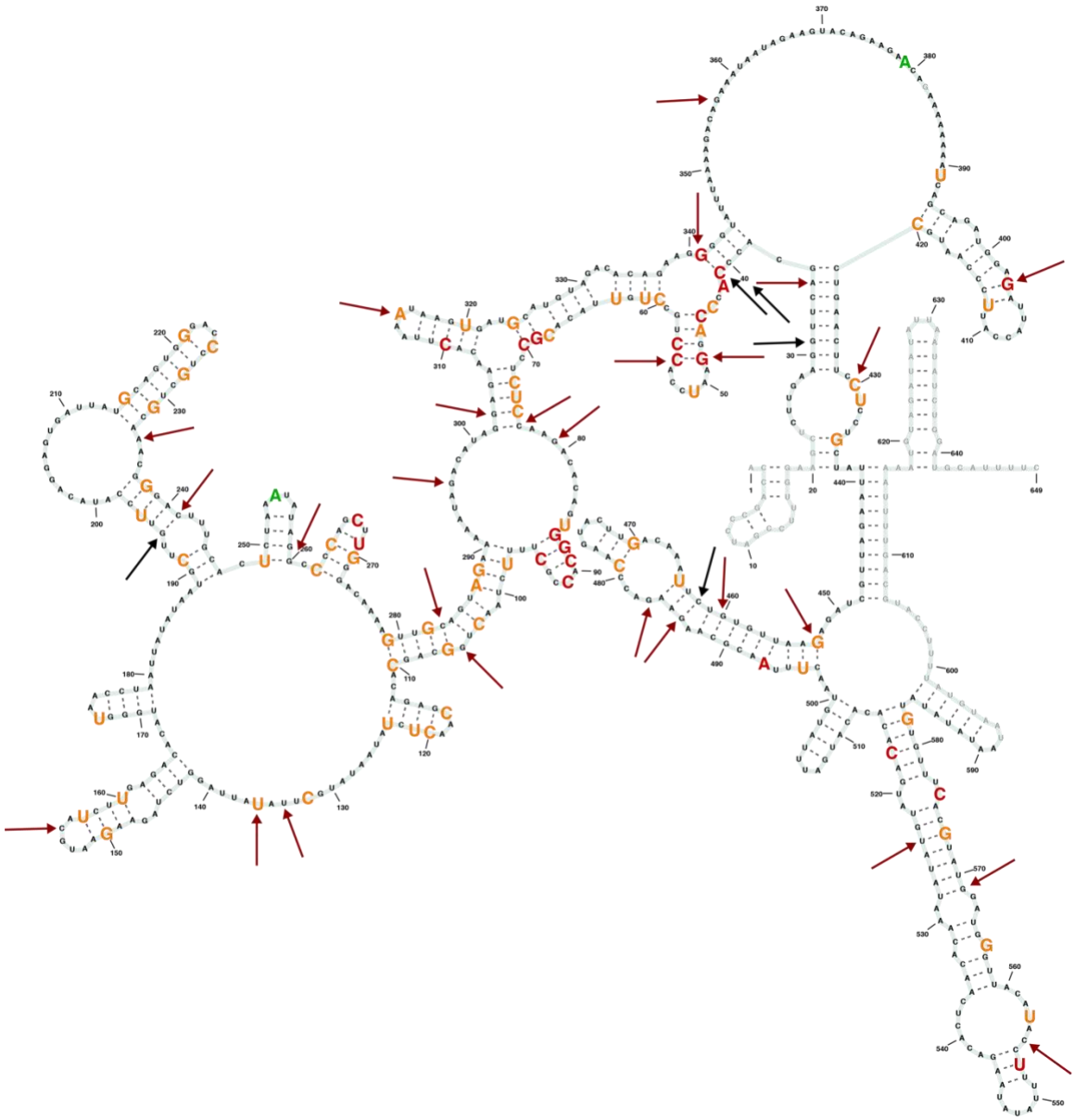


Figure 87: Representation of RNP-MaP data overlaid on a computational model of HOTAIR D4 (RNA only) with previously discussed SHAPE reactivities. Arrows represent nucleotides potentially directly interacting with protein due to a 1.2– or 1.5–fold change (red and black, respectively) due to the presence of KDM1A:CoREST$_{286\text{-}493}$.

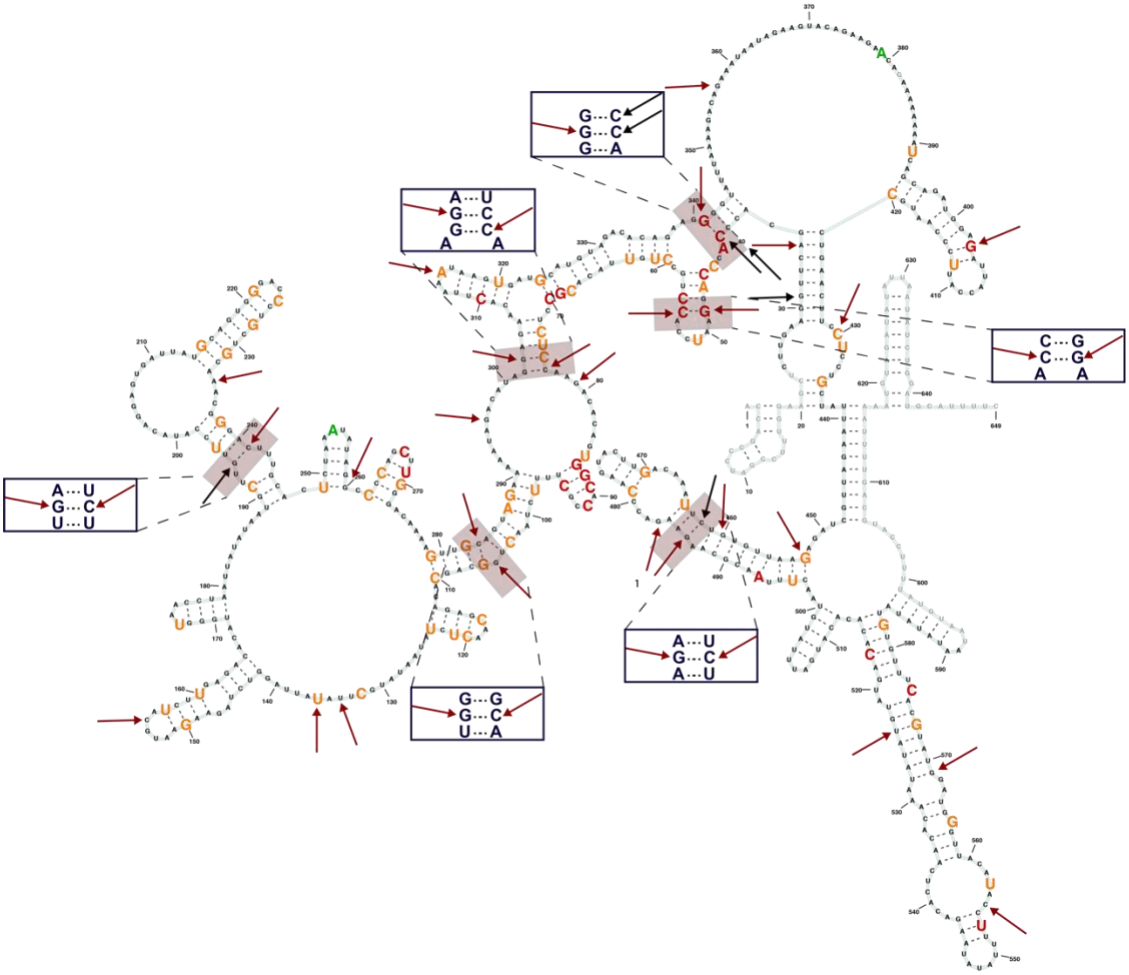


Figure 88: Of the nt which were crosslinked to KDM1A:CoREST$_{286\text{-}482}$, a significant number were G:C bp, which are highlighted above. We noted that the sequence surrounding these nt is also similar in purine-purine dinucleotides on either side with an adjacent adenine.

5.3.3. RNP-MaP Discussion

Following an optimization of the protocol for RNP-MaP to be amenable to for an *in vitro* system, we were able to assess which nucleotides are most likely directly

interacting with KDM1A:CoREST$_{286-493}$ to contribute to binding. Interestingly, again we note a large spread of nucleotides throughout the structure of the RNA, instead of a more localized region as might be expected. The binding arm we noted for SHAPE-MaP did, again, display the highest number of RNP reactive nucleotides, which was promising.

However, canonical RNA-binding protein domains tend to recognize a short strand of RNA, 4-12 nt, so this large of a potential binding interface was unexpected *(397–400,483,484)*. Upon further investigation, we noted that not only were a majority of the nucleotides either guanine or cytosine (84.4%), but 62.5% of the nucleotides that were indicated to potentially bind KDM1A:CoREST$_{286-493}$ were also predicted to be base paired. This was interesting as, considering the predicted binding region along KDM1A:CoREST$_{286-493}$ may bind paired nucleotides at a higher level due to its lack of binding pocket as discussed more extensively in the discussion of Chapter 4 *(396)*.

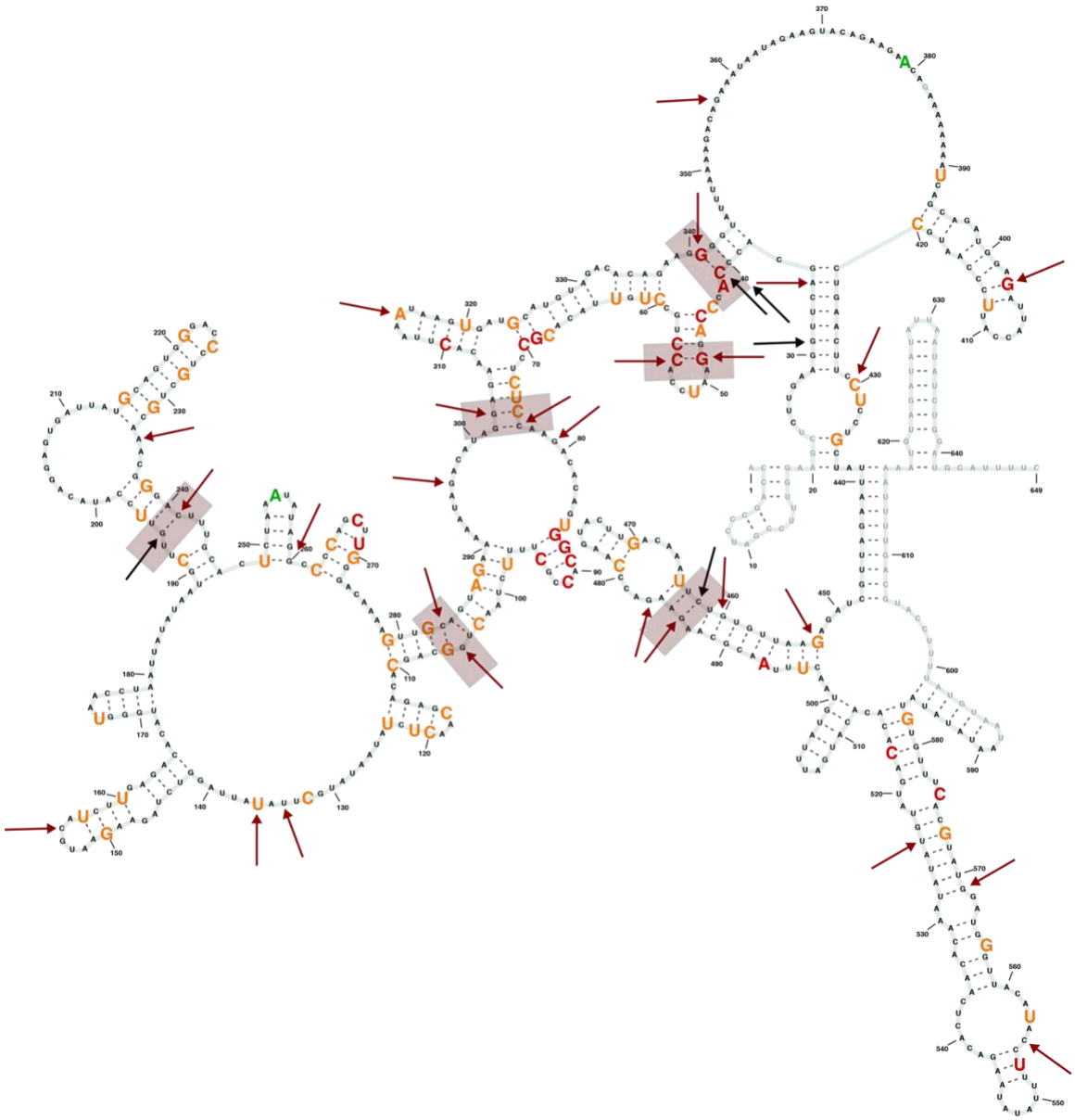


Figure 89: Several of the RNP-activated residues were base pairing. All were GC base pairs, highlighted here in red

We noted that several of the RNP-activated residues were base pairing, and that all were all were G:C base pairs (Figure 88). Further, all but two of these GC pairs

resided within the first 300 nt of HOTAIR D4 ($HOTAIR_{1500-1820}$), where binding was previously demonstrated with the EMSA assay. We assessed the similarity of the sequences 10 nucleotides upstream and downstream of each strand. If there were sequence similarities, this might suggest a sequential or structural motif that the proteins bind to. On both strands of the RNP-activated G:C base pairs, all were either GG of CC with on strand displaying CG, but there were no longer-range apparent sequence similarities. In chapter 4, aaRNA suggested that many of the proposed protein binding region along KDM1A:$CoREST_{286-482}$ had a high reactivity to dinucleotides containing a guanine. At the sites of the G:C base pairs, the is at least one guanine-containing dinucleotide (Figure 88).

The Wooten lab at UNC at Chapel Hill has hypothesized that some proteins may not bind specifically to one high-affinity motif of a lncRNA, but instead display binding when there is an abundance of lower-affinity 'k-mers' along a lncRNA *(355)*. Each of the short 5-8 nt k-mers contributes to the overall affinity between the RNA and protein, and only when the RNA contains a certain number of k-mers is the protein able to demonstrate binding. The Wooten lab demonstrated that the k-mers may not be the same sequence, so it may be that they form all form a similar structural motif that the protein is able to interact with *(355)*.

There has been long-standing evidence for this in the field with respect to trinucleotide repeat RNAs. An increased number of proteins are able to interact when the number of CAG trinucleotide repeats in exon one the Huntington's disease *HD* gene exceeds 37. The hypothesis is that there are a minimum number of low-affinity repeats required for certain proteins to bind, contributing in part to the disease. Additionally, dsRNA-binding protein staufen1 was demonstrated to interact more strongly with a structured dsRNA over a completely A-form dsRNA *(397)*. Staufen-1, along with TRBP both demonstrated a 'sliding' ability along a structured 55 bp stem that contained several of the binding sequences each has been shown to display affinity for via FRET *(397)*. Thus, we hypothesize there are several low-affinity binding sites along HOTAIR D4, likely around the RNP-activated GC base pairs we noted. The continued affinity to only $HOTAIR_{1500-1820}$ is likely due four of the six RNP-activated GC base pairs residing within this window.

5.4. Future directions: Towards elucidation of the KDM1A:CoREST:HOTAIR interaction

As our understanding of lncRNA expands, so does our understanding of their interactions with proteins. We have shown that HOTAIR D4 exhibits a structural stabilization due to the presence of $KDM1A:CoREST_{286-493}$. We have shown that these regions of stabilization match with regions that display a concentration of decreased SHAPE reactivity due to the presence of protein. Furthermore, we have demonstrated

nucleotides which may directly interact with KDM1A:CoREST$_{286-493}$ and suggested that the binding interaction with HOTAIR may be through a series of low-affinity motifs concentrated within the first 300 nt of HOTAIR D4.

Moving forward, it would be interesting to test the importance of the binding arm we noted in HOTAIR binding to KDM1A:CoREST$_{286-493}$. We envision the production of a linear DNA fragment which contains the sequence of both strands of the binding arm linked between nucleotides 110 and 279 via a tetraloop. We chose this end of the arm to link with a tetraloop because the predicted structural loop at this end of the arm is much smaller and thus, shrinking the loop to a 4 nt sequence is less likely to perturb overarching structural components of the arm. Tetraloops are a well-represented structural component in RNA as discussed above, and we envision usage of the UCCG sequence as it is reported to be one of the most thermodynamically stable *(436,472)*. Upon T7 RNA Polymerase II production and clean-up of the RNA, we expect that an initial SHAPE-MaP experiment will be performed on the RNA only to ensure that the overarching structure of the RNA has not been disrupted by the tetraloop. With the RNA fragment in hand, we will be able to perform EMSA assays as reported in chapter 4 as well as follow-up with SHAPE-MaP and RNP-MaP if binding is present.

If binding is not present in this binding arm and the structural integrity is retained, we propose a more systematic shortening of the 300 nt sequence. We were able

to obtain binding with this fragment, which suggests that a critical binding motif lies within this region.

We additionally propose the removal of the binding arm from HOTAIR D4 as a control to assess whether the binding arm is truly the critical component for binding. Removal of the binding arm sequence and linkage via a short stem with no sequence similarity to the loop should allow for the overall structure of the remaining HOTAIR D4 to be retained. Similar experiments as above can then be performed to assess the level of binding affinity that was concentrated within the binding arm.

A crystal structure would be the most helpful tool in understanding the interaction between KDM1A:CoREST$_{286\text{-}493}$. It is most likely that a crystal structure of the full HOTAIR D4 will not able to be obtained. If the binding arm is established to be the critical region for protein binding, it may be that utilization of just this arm in the crystal structure will result in useable crystals. Fortunately, KDM1A:CoREST$_{286\text{-}493}$ has been commonly crystallized and provides a starting point for conditions through which we can produce crystals with RNA.

5.5 Future Directions: Towards disruption of the KDM1A:CoREST:HOTAIR interface

In parallel to further elucidation of the interface between KDM1A:CoREST and HOTAIR, we return to the third aim of this project which was to disrupt the interface in a targeted manner. In addition to utilization of a SELEX aptamer, I have designed two

additional paths which can be tested in parallel to establish the most suitable method to disrupt this interaction. The first route utilizes an antisense oligonucleotide (ASO) approach while the second employs small molecule inhibitors both towards disruption of the RNA:protein interface. In each of these two cases, we envision these methods targeting the RNA in a highly specific manner that can be translated to *in vivo* work.

ASO knockdown is a common approach towards disruption of mRNA translation and is currently being explored in modern therapeutics such as exon skipping *(314,485,486)*. ASOs function by base pairing with a critical region of the RNA, disrupting either protein binding or producing a conformational shift which disrupts the function of the RNA. However, similar to peptide inhibitors being susceptible to peptidases, RNAs are susceptible the nucleases, potentially lowering their efficacy and lifetime within the cell.

The utilization of a peptide nucleic acid (PNA) has been utilized in response to this pitfall of ASO. Essentially, a peptide backbone is utilized with the R-groups as nucleotides. The different spatial arrangement precludes PNAs from nuclease and peptidase activity, lengthening the half-life of the biomolecule within cells *(487,488)*. Indeed, we previously discussed a case for the HOTAIR:PRC2 complex was disrupted by a PNA both *in vitro* and then demonstrated the decreased viability of cancer cells both in cell culture as well as mouse xenograft models *(165)*.

Having established the HOTAIR binding arm as a significant contributor to binding KDM1A:CoREST, I envision production of PNAs 10-15 nt in length centered around the sequence motif I have potentially established. Which then can be incubated with HOTAIR before a binding assay to determine how effective each PNA is at disrupting binding. Subsequent competition assays as well as utilization of SPR or ITC will enable us to define the binding energy of the PNA *in vitro*. For *in vivo* work, to aid in the PNA crossing the cellular membrane, incorporation of a highly basic peptide such as the cell-penetrating peptide (RRRQRRKKR) onto a thiolated PNA will enable the PNA to reach the complex in cells to test both cell survival as well as mobility *(165)*. Furthermore, ideally in the future we will be able to test the effects of disruption in a solid tumor. Since this interaction is ubiquitous throughout cells, with a therapeutic approach in mind, we envision utilization of a pHILP peptide linkage to the PNA to ensure tumor-specific targeting *(165,489)*.

In addition to this method, I have also been focused on establishing a route to utilize small molecule inhibitors of this interaction. The Disney lab at Scripps Florida has designed a sequence-based program, Inforna, which is able to detect secondary structure elements, motifs, and pair each motif with a series of small molecules which would likely bind *(490,491)*. To accomplish this, the computational program is based on experimental design from a 2-Dimensional Combinatorial Screening system (2-DCS)

designed by the Disney lab *(492)*. The fitness of each potential interaction is scored through a computational method, Structure Activity Relationships Through Sequencing (START), which utilizes Z-scores to compare the frequency of that motif and how likely that interaction is against a pooled population *(493)*.

In collaboration with the Disney lab, we have utilized their program and protocol to establish motifs within HOTAIR as well as compounds which are likely to bind with high affinity and selectivity. The motif placement as well as their specific loop sequence are outlined in Figure 90.

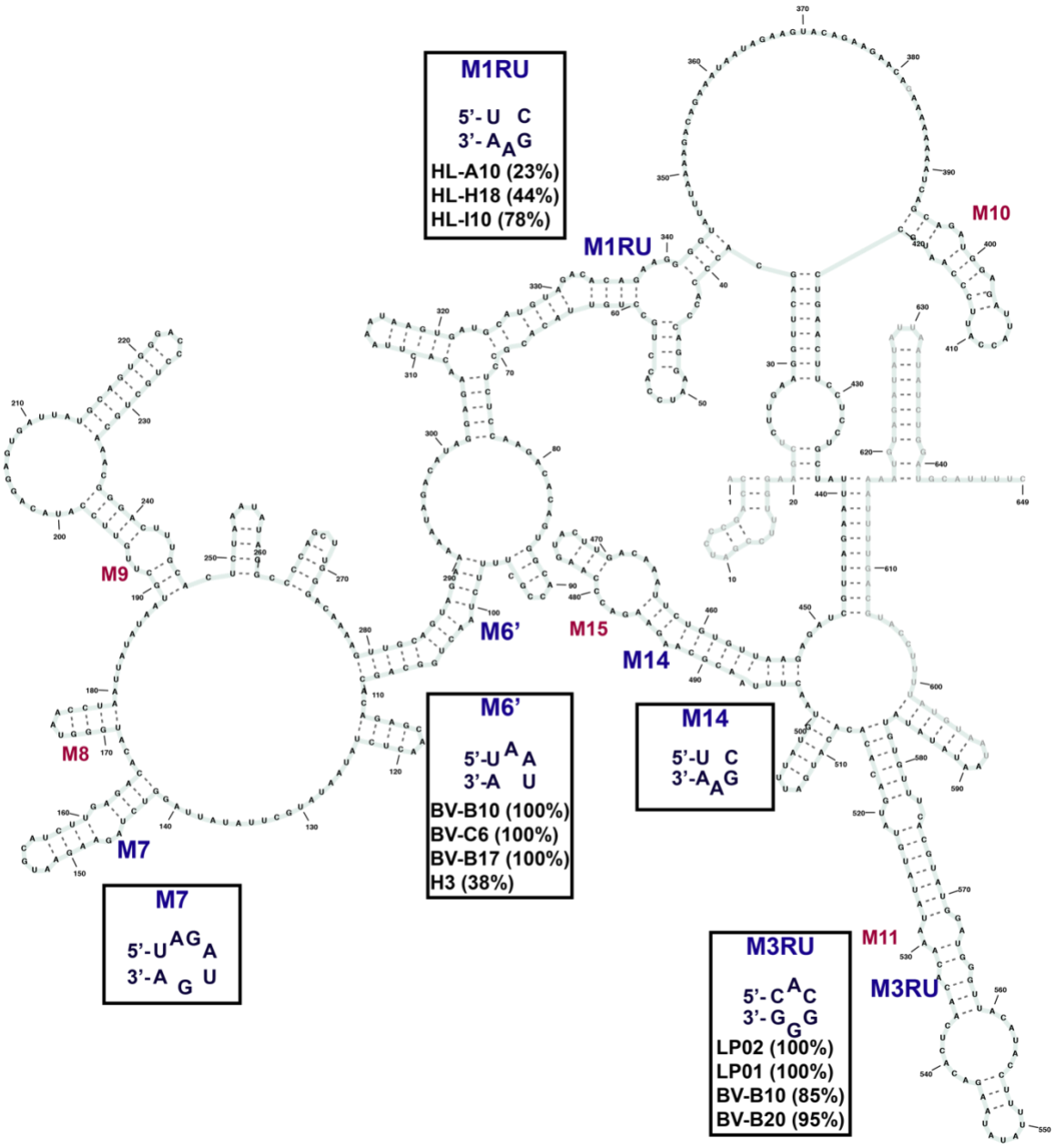


Figure 90: Utilization of Inforna to establish HOTAIR secondary structure motifs (red labels) to be targeted and the exact sequences (blue). We are extremely interested in M6′, M1RU and M2Ru as they lie along the HOTAIR binding arm.

Of the motifs assigned by Inforna, we were most interested in M1RU, M2RU, and M6′ as these lie along the binding arm of HOTAIR and could potentially disrupt enough of the interface or secondary structure of the binding arm to disrupt binding to KDM1A:CoREST. M1RU was assigned three small molecules and modest fitness scores, HL-A10 (23%), HL-H18 (44%), and HL-I10 (78%). M2RU was also assigned three small molecules, Paromycin A Az (100%), PL270-F6 (40%), and PL-270-H20 (43%). M2RU compounds overall have higher fitness scores than M1RU, which suggests they are more likely to bind selectively and tightly to the specific motif. Finally, M6′ pooled five small molecule hits: BV-B10 (100%), BV-B17 (100%), BV-C6 (100%), H8 (38%), and N10 (41%). M6′ hits also were assigned higher fitness scores than M1RU, and with three at 100%, these are of the highest interest to us. Structurally, each compound contains multiple aromatic groups trending towards heterocyclic. Eight of the 11 compounds contain a core reminiscent of a purine heterocyclic-ring system.

We have obtained these compounds from the Disney lab and envision previously established binding assays across a 100 μM to 0.1 μM concentration range to test the ability of the compounds to displace KDM1A:CoREST from HOTAIR. Upon establishing a compound of interest, we will be able to follow up with full competition assays in order to assess the inhibition constant as well as SHAPE-MaP to establish how the structure changes as a function of small molecule binding. Long-term we envision

utilizing these small molecules *in vivo* as previously described to assess the efficacy of the small molecule in a cellular environment.

Overall, we have described two methods, ASO PNAs and small molecule inhibitors, which will allow us to access RNA-targeted inhibitors of the KDM1A:CoREST:HOTAIR interaction interface. We have currently utilized Informa and obtained compounds and anticipating testing of the small molecule lead compounds towards establishing a small molecule readily able to disrupt this RNA:protein interface.Δ

5.6 Experimental methods

Selective 2′-hydroxyl acylation analyzed by primer extension with Mutational Profiling (SHAPE-MaP)

Freshly purified HOTAIR (8 pmol) was heated to 95 °C for 2 minutes, rapidly cooled to 4 °C for 2 minutes and then supplemented with folding buffer (10 mM HEPES (pH 8.0), 100 mM NaCl, 5 mM $MgCl_2$) and allowed to fold for 20 additional minutes at 37 °C. RNA was divided into (+/-) protein samples and incubated (37 °C, 20 minutes) with protein (50 μM/reaction), or nuclease-free water. Each of the two samples was further divided for the addition of either DMSO or 5-NIA to a final concentration of 2 mM and incubated (37 °C, 10 minutes). The reaction was then treated to a G-50 spin column clean up (GE healthcare, Illustra), 3-hour Proteinase K treatment (Table 29) at 37 °C, and a Trizol extraction. SSII (Thermo Fisher 18064014) was utilized in reverse

transcription of each reaction as previously reported by the Weeks lab *(401,402)* and prepared for next-generation sequencing via Nextera preparation per manufacturer instructions. A Nextera 500 V2 kit was utilized on a MiSeq NGS instrument.

Table 29: Proteinase K reaction buffer conditions

Reagent	Stock	Final
Tris (pH 8.0)	1 M	80 mM
NaCl	5 M	400 mM
SDS	10% (w/v)	3.0%
Nuclease-free Water		v/vt
Proteinase K	20 (μg/μL)	0.5 μg/μL

Data analysis

Samples were analyzed utilizing the SHAPEmapper program provided by the Weeks lab. Briefly, DMSO control samples were compared against 5-NIA treated samples in SHAPEmapper. The fold change between (-)-protein and (+)-protein samples was established, and a 1-fold cut off was used to established nucleotides of potential interest. Superfold arc comparison and Varna were used to produce graphics.

RNA:protein Chemical Crosslinking (RnP-MaP)

Sample preparation

Freshly purified HOTAIR (32 pmol) was heated to 95 °C for 2 minutes, rapidly cooled to 4 °C for 2 minutes and then supplemented with folding buffer (10 mM HEPES (pH 8.0), 100 mM NaCl, 5 mM $MgCl_2$) and allowed to fold for 20 additional minutes at

37 °C. RNA was divided into (+/-) protein samples and incubated (37 °C, 20 minutes) with protein (50 μM/reaction), or Nuclease-free water. Each of the two samples was further divided for the addition of either DMSO or SDA to a final concentration of 10 mM and incubated (37 °C, 10 minutes) and Tris-HCl (pH 8.0) was added to a final concentration of 100 mM. Each reaction was split across two wells of a 12-well plate and subjected to crosslinking (365 nm, max energy (9999), 3x, 4 °C). The reaction was subjected to an overnight Proteinase K treatment (Table 30) at 37 °C, and PCA extraction. SSII (Thermo Fisher 18064014) was utilized in reverse transcription of each reaction as previously reported by the Weeks lab (ref) and prepared for next-generation sequencing via Nextera preparation per manufacturer instructions. A Nextera 500 kit was utilized on a MiSeq NGS instrument.

Table 30: Proteinase K reaction buffer conditions

Reagent	Stock	Final
Tris (pH 8.0)	1 M	80 mM
NaCl	5 M	400 mM
SDS	10% (w/v)	3.0%
EDTA	0.5 M	40 mM
Nuclease-free Water		v/v_t
Proteinase K	20 (μg/μL)	0.5 μg/μL

Data analysis

Samples were analyzed utilizing the SHAPEmapper program provided by the Weeks lab. Briefly, DMSO control samples were compared against SDA treated samples in SHAPEmapper. The resulting files were utilized in RnP-MaPper 2 beta provided by the Weeks lab. The fold change between (-)-protein and (+)-protein samples was established, and a 1-fold cut off was used to established nucleotides of potential interest.

6. Conclusions

PTM of histone proteins plays a major role in the regulation of gene transcription. Epigenetic 'writer' and 'eraser' enzymes deposit and remove each PTM, respectively, in a site-specific and selective manner. The regulation of catalytic activity and substrate specificity is imparted to epigenetic enzymes through coregulatory protein binding. Furthermore, multimeric complexes containing coregulatory proteins with PTM 'reader' domains as well as DNA and RNA binding domains recruit the enzymes to specific gene loci in a temporal manner.

The molecular mechanisms for these functions are only beginning to be understood, but it is clear they play an important biological function. KDM1A is a lysine-specific mono- and di-demethylase of H3K4 and H3K9 residues *(32,65,146)*. KDM1A plays many important cellular functions including cellular differentiation and in the EMT *(75,76)*. Pathobiologically, KDM1A is also upregulated in a significant number of cancer phenotypes and subsequent inhibition and knockdown of KDM1A demonstrates decreased proliferation of the cancerous cells *(46,48,70,72,146)*.

The role of KDM1A within cells is largely dependent on its interactions with coregulatory proteins and complexes including the REST, NuRD, CtBP, SNAI1, Gfi1/Gfi1b, and NHR complexes *(57,72,73,89,111)*. Futhermore, lncRNA are being established as increasingly important in many cellular functions, including scaffolding

epigenetic complexes to target specific gene loci. KDM1A has been demonstrated previously to interact with HOTAIR, but there is little known about this binding interface *(154,163)*. The co-option of KDM1A by such a variety of complexes for different pathways and cellular processes suggests that pan-inhibition of the active site is not a viable approach. We hope that by understanding the molecular mechanisms of KDM1A within the context of coregulatory protein recruitment, complex-specific targeted inhibition of KDM1A will be achievable.

The specific aims of my thesis were to (1) interrogate the role of individual and multiple coregulatory partners in enzyme selectivity and specificity (2) establish tools to study the mechanisms of biochemical and biophysical of protein:protein and protein:lncRNA interactions and (3) elucidate key characteristics of protein:protein and protein:lncRNA interfaces towards targeted disruption.

Towards aim 1, I utilized cloning and mutagenesis methods to heterologously express and purify coregulatory partners of KDM1A in *E. coli*. I chose coregulatory partners found in a common catalytic core as well as several additional coregulatory proteins from a stable KDM1A-containing 5-mer complex. I have produced multiple constructs for four of these proteins to allow for multiple affinity purification routes as well as for future binding studies. I have further expressed each of these constructs and have made significant efforts towards the purification of each construct based on

solubility. The efforts I have made in this aim will drive future biophysical and catalytic activity work in the lab to not only establish mechanisms of binding between proteins but begin to produce increasingly intricate multimeric epigenetic complexes. These efforts are also currently being utilized to elucidate biochemical and biophysical characteristics of binding with lncRNA based on techniques I established in the lab.

Towards aim 2, I utilized HDX-MS as a method to characterize dynamic changes in proteins due to ligand binding. I noted that upon binding $CoREST_{286-482}$, the highest region of decreased deuterium uptake (over 20%) resides in the apical end of the tower domain. In agreement, we compared crystal structures and noted a 13 Å RMSD at the apical end of the tower domain between crystal structures (+/-) CoREST via structural alignment and PCA (Chapter 2). With purified proteins, such as those we obtained in Chapter 1, HDX-MS is a highly accessible and adaptable method that can be used in future work to establish dynamic changes in protein binding.

I also established several key biochemical and biophysical methods for our lab moving forward through which to explore lncRNA. In addition to optimizing the expression and purification of RNA *in vitro* (Appendix a), I performed EMSA assays with ^{32}P-HOTAIR fragments with KDM1A, $CoREST_{286-482}$, KDM1A:$CoREST_{286-482}$, and KDM1A:Linker. I found a requirement for the linker region of CoREST to be present for HOTAIR to show association to KDM1A at 1.29 ± 0.34 μM . Utilizing a computational

model, I have predicted a potential site of interaction along the tower domain of KDM1A (residues 429-442, **E**H**W**KKIV**K**T**Q**EEL**KE**) and the linker region of CoREST (residues 349-357, **KRQ**I**QN**I**KQ**). I hypothesized that the RNA was more likely to be double stranded, and potentially GC rich. I additionally demonstrated that only the first half of the HOTAIR domain D4 was required for binding to KDM1A:CoREST$_{286\text{-}482}$ at a similar, binding affinity of 2.97 ± 0.96 μM.

To assess the nt of HOTAIR D4 involved in binding to KDM1A:CoREST$_{286\text{-}482}$, I worked with the Weeks lab at UNC Chapel Hill to utilize their SHAPE-MaP and crosslinking of RNA to proteins technologies. Both cases required significant optimization to allow for recovery of samples; however, upon optimization, we were able to establish not only which nt changed in reactivity as a function of the proteins, but which were likely interacting with the proteins directly.

With SHAPE-MaP, we were able to establish a change in SHAPE reactivity, which is related to the flexibility and likelihood that the RNA is single- or double-stranded, due to the presence of KDM1A:CoREST$_{286\text{-}482}$. We revealed 83 nt with at least a 1.5–fold decrease in SHAPE-reactivity and 22 nt with a 2.0–fold decrease in SHAPE-reactivity. Upon plotting these nt on an experimentally driven free-energy secondary structure computational model, we noted that 44% of the significant nt were within a region of HOTAIR D4 that was primarily doublestranded.

RNP-MaP further pointed to an importance of the 'binding arm' of HOTAIR. Upon further investigation, I noted that not only were a majority of the nucleotides either guanine or cytosine (84.4%), but 62.5% of the nucleotides that were indicated to potentially bind KDM1A:CoREST$_{286-482}$ were also predicted to be base paired, which complemented the previous computational work. I also noted that there were six pairs in which both the nt in a base pair were RNP-activated, and all were G:C. Furthermore, four of the 6 were within the same binding arm I previously established with SHAPE-MaP.

Expanding out from the G:C sequence, I established that each of the nt were sequentially positioned next to another guanine or cytosine nt with an adenine directly adjacent. This potential binding motif suggests that, like other RNA binding proteins, KDM1A:CoREST$_{286-482}$ interact with a short RNA sequence. While the affinity to one of these motifs may be low, the high number of repeats allows for an overall higher binding affinity between RNA and protein.

Taken together, this suggests that while the affinity with KDM1A:CoREST$_{286-482}$ may only be modest, there are likely additional coregulatory proteins which may increase the affinity. It may also be that upon binding the nucleosome, there are several interactions of modest affinity which, in total, produce a higher affinity to allow for catalysis to occur. Interestingly, the SANT2 domain of CoREST has been demonstrated

to have only a modest 84 μM affinity for nucleosomal DNA. This may suggest that it is the combination of binding HOTAIR and the SANT2:DNA interactions, plus other yet unknown interactions, which drive KDM1A gene localization. With the additional coregulatory proteins from Chapter 1 in hand, this hypothesis will be able to begin to be tested.

Towards aim 3, I have also established multiple routes for approaches towards a targeted disruption of interactions. With Dr. Jennifer Link Schwabe and the Sullenger lab here at Duke University, we were able to establish SELEX as a method for production of high-affinity RNA aptamers in our lab. We demonstrated binding to the RNA library, which will be able to be further probed and explored in order to establish aptamers to stably disrupt protein:protein interactions. I also established a collaboration with the Disney lab at Scripps Florida in order to screen for small molecules through their Inforna database. We hope that the small molecules we have established utilizing our SHAPE-MaP and RNP-MaP data will allow us to disrupt this interface in a highly specific and competitive manner.

Appendix A: Optimization of in vitro RNA production

For the techniques we wished to utilize to elucidate the KDM1A:HOTAIR interface, we required high levels of RNA. The original protocol we obtained utilized a purchased T7 RNA Polymerase (NEB M0251) and buffer wherein the DNA template was incubated with the polymerase and buffer (37 °C, 8 hrs). To purify the RNA, we utilized a gel extraction from a denaturing acrylamide-urea gel which resulted in a modest μg yield.

When I decided to look towards optimization of the protocol I was utilizing, I saw three main areas of improvement, the efficiency of the T7 RNA Polymerase reaction, well as the extraction method, and the length of reaction and subsequent clean-up. I have been able to improve each of these three areas with high success, as detailed below.

To address the efficiency of the T7 RNA Polymerase reaction, I was gifted an aliquot of a T7 RNA Polymerase-containing expression plasmid from the Weeks lab at UNC Chapel Hill. I was able to utilize an IPTG-induced expression at a final IPTG concentration of 0.5 mM at an OD_{600} of 0.5 and 37 °C. The cells were mechanically lysed and purified via a batch elution method wherein the cells were washed in a 50 mM Tris-HCl (pH 7.8) buffered low imidazole buffer (20 mM, 700g, 4 °C, 10 min) and then eluted at a high imidazole buffer (285 mM, 700g, 4 °C, 10 min). We assessed the purification method via SDS-PAGE (Bio-Rad 4-20%, 120V, 65 min) with Coomassie staining (Figure

91) and noted a band at the correct MW of 99 kDa, though there were several contaminating bands. We assessed concentration via a Bradford assay to be 0.83 mg/mL at a total concentration of 7.5 mg.

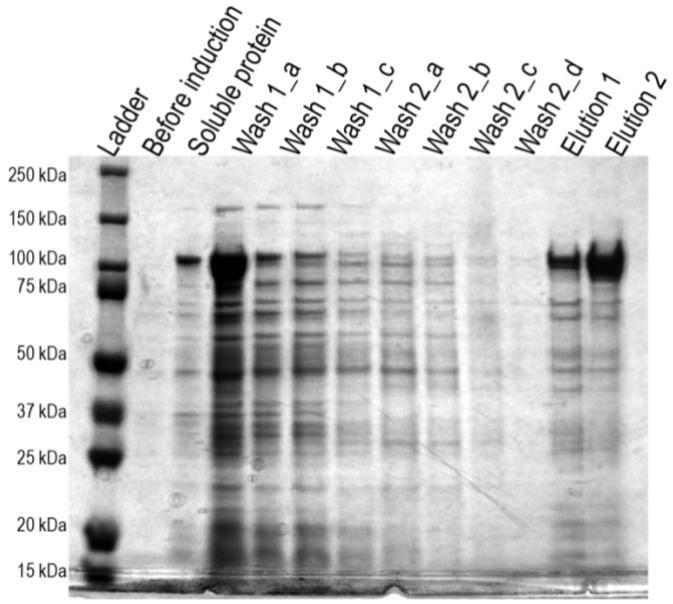


Figure 91: Purification of T7 RNA Polymerase. We obtained a T7 RNA Polymerase plasmid from the Weeks lab at UNC Chapel Hill. We were able to utilize standard expression (37 °C, OD_{600} 0.5, 0.5 mM IPTG) and batch wash methods with low imidazole (10 mM, 20 mM) and elution at 285 mM imidazole.

With purified T7 RNA Polymerase in-hand, we next performed a side-by-side comparison of the in-house purified (new) to purchased (old) T7 RNA Polymerase (NEB M0251) with the *in vitro* production of HOTAIR D4. The Weeks lab also provided a buffer they produced (Table 31), which we were able to additionally compare with both the new and old polymerases. We additionally sought to assess the effect of utilizing an

RNase inhibitor in the reactions. Each reaction was incubated at 37 °C for 8 hours and then run on a 6% denaturing acrylamide-urea gel (500 V, 3.5 hrs) (Figure 92). From this we were able to make several qualitative observations. Firstly, the purchased (old) buffer appears to be significantly less effective than the in-house (new) buffer. Several of the components such as spermidine are not stable in solution long-term, thus production of a buffer with fresh reagents likely increased the efficacy of the reactions utilizing the new buffer conditions. We also noted that the old polymerase is more high yielding than the new polymerase. The production of RNA appears very dependent on the presence of RNaseI, suggesting an inherent contamination, whereas the apparent RNA levels with the new polymerase do not appear dependent on the presence of RNaseI.

Table 31: In-house (new) buffer conditions provided by the Weeks lab at UNC Chapel Hill

Reagent	*Initial concentration*	*Final concentration*
rNTPs	25 mM	5 mM
RNaseI	40 U/µL	1 U
Tris (pH 8.0)	1 M	40 mM
$MgCl_2$	1 M	25 mM
Spermidine	100 mM	2.5 mM
Triton X-100	1%	0.01%
DTT	1 M	10 mM
New T7RNAP	0.8mg/mL	--
H2O	v/vt	v/vt
Template	20-100 ng/µL	150-750 ng
pyrophosphatase	0.2 U/µL	0.05 U

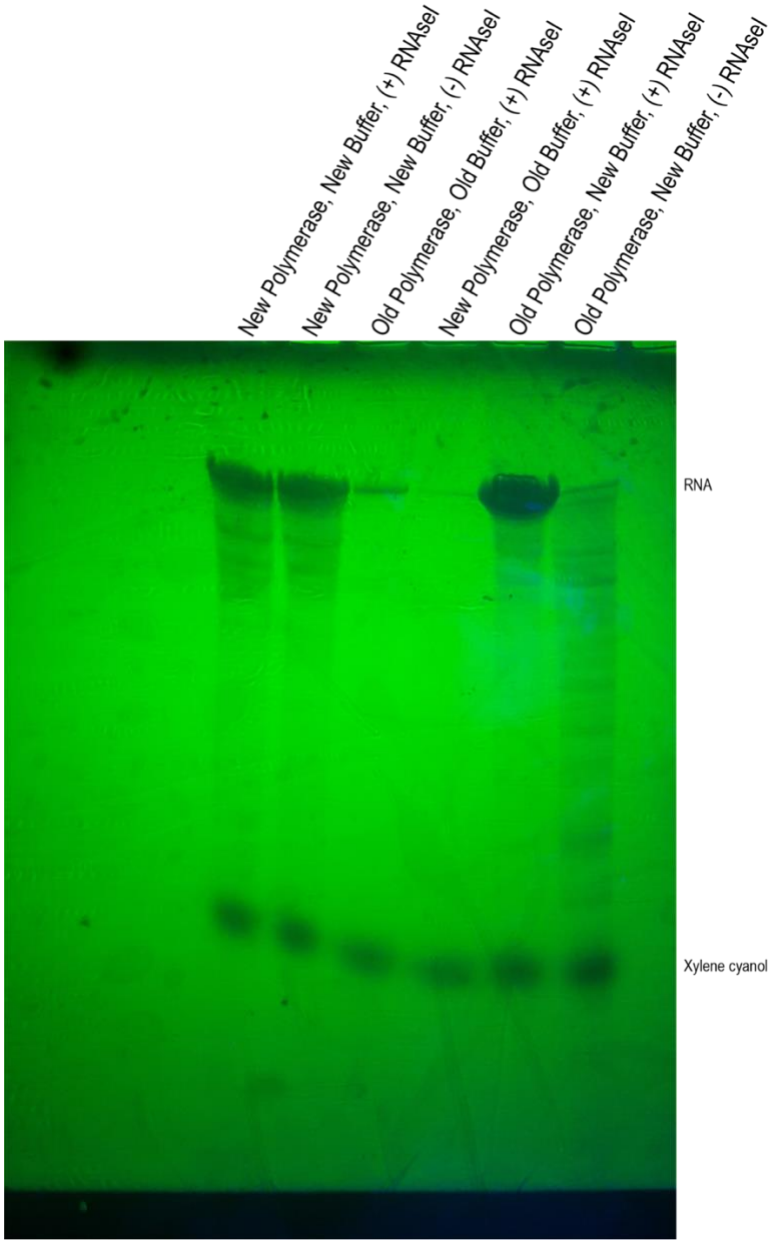


Figure 92: Acrylamide Urea gel of T7 RNA Polymerase test comparing purchased and in-house T7 RNA Polymerase as well as the utilization of an RNaseI. Each reaction was incubated at 37 °C for 8 hours and then run on a 6% denaturing acrylamide-urea gel (500 V, 3.5 hrs.). We noted an increase in qualitative yield of the in house T7 RNA Polymerase and buffer as well as the purchased T7 RNA Polymerase with the in-house buffer.

To quantitatively assess the amount of RNA that was produced in each of these reactions, we excised each band and performed a gel extraction. Briefly, the gel fragments were frozen to -80 °C and then incubated with TE buffer overnight at 4 °C. The resulting solution was filtered with a cellulose acetate syringe filter, washed three times and finally concentrated with TE in a swinging bucket rotor (3900g, 4 °C, 20 min). We quantified the levels of RNA with a sectrophotometer (Table 32). As we had noted visually, the incorporation of RNaseI into the reactions with new polymerase and buffer did not significantly affect the yield of the reactions. Surprisingly, while the yield of the old polymerase with the new buffer had appeared much greater than with the new polymerase, this was not reflected in the post-purification yields.

Table 32: Reaction yields comparing in-house versus purchased polymerase and buffer as well as inclusion of RNaseI

Sample	*Conc (ng/µL)*	*Conc (µg/µL)*
New Buffer, New Pol (+) RNaseI	1424.720	1.424
New Buffer, New Pol (-) RNaseI	1409.920	1.409
Old Buffer, Old Pol (+) RNaseI	118.960	0.118
Old Buffer, New Pol (+) RNaseI	48.880	0.0489
New Buffer, Old Pol (+) RNaseI	1038.320	1.038

Having reached our goal of a higher-yielding reaction and decreasing costs by eliminating the need to purchase T7 RNA Polymerase, we still sought additionally to decrease the total reaction and clean-up time as well as improve the purification protocol as the gel and gel extraction protocol took 22 hours in our hands and we

hypothesized we were losing yield in separating the RNA from the gel. Since 1995, an alternative method reported for purification of nucleic acids has been the use of carboxyl-coated solid-phase reverse immobilized (SPRI) paramagnetic beads. In a PEG-containing buffer, nucleic acids are desolvated and precipitate around the carboxylated beads. The beads are pelleted by use of a magnet to allow for removal of proteins, PEG, and other unwanted contaminants via extensive washing. The nucleic acids can then be resolubilized and eluted in a water or Tris-buffered solution and stored until use. While BeckmanCoulter has produced a line of these beads, AMPure, they are not cost effective. We were able to utilize Sera-Mag beads from GE Healthcare. Through initial wash steps with 0.1 M NaOH and 0.05M NaCl and then a final wash with 0.05 M NaCl to ensure the beads were RNase-contamination free, we could utilize just 1 mL of the SeraMag beads in a buffer system specific for RNA or DNA purification (Table 33) to produce 50 mL of SPRI paramagnetic beads.

Table 33: Buffers for in-house production of DNA and RNA SPRI paramagnetic beads

Reagent	*RNA SPRI beads (mL)*	*DNA SPRI beads (mL)*
NaCl, 5 M	25.000	25.00
RNase-free water	4.672	3.982
Trisodium citrate, 1 M	0.050	0.0
EDTA, 0.5 M	0.0	0.100
HCl, 1N	0.028	0.168
PEG 8000, 50% (w/v)	20.000 mL	20.00
Tween 20, 10% (v/v)	0.250	0.250
Sera-Mag bead suspension	1.00	1.00

We ran the same reactions as above comparing polymerase, buffer, and inclusion of RNaseI. Following completion of the reactions, we performed a SPRI bead purification with an incubation of 1.8x beads and subsequent washes with 70% ethanol (4x, 30 sec) and then elution in TE buffer. We quantified the levels of RNA with a spectrophotometer (Table 34). We noted that concentrations obtained in the SPRI paramagnetic bead clean-up procedure followed the same trends as the gel purification, as expected. Importantly, the purification with SPRI paramagnetic beads only takes 1.5 hours from start to finish, which is significantly improved over the 5-hour denaturing gel, followed by an overnight extraction and subsequent washing and concentration.

Table 34: Utilization of SPRI paramagnetic beads for cleanup of T7 RNA Polymerase reactions

Sample	*Conc (ng/µL)*	*Conc (µg/µL)*
New Buffer, New Pol (+) RNaseI	1602.810	1.602
New Buffer, New Pol (-) RNaseI	1586.160	1.586
Old Buffer, Old Pol (+) RNaseI	133.830	0.133
Old Buffer, New Pol (+) RNaseI	54.990	0.0549
New Buffer, Old Pol (+) RNaseI	1168.110	1.168

We next wanted to compare the overall yield of the reactions between the gel and bead purifications. Additionally, the reactions utilizing the old buffer were only 20 µL reactions whereas the reactions utilizing the new buffer conditions were 50 µL. We normalized the yields and concentrations across the reaction types and performed several comparisons as outlined (Table 35) with the fold change of total RNA yield shown. We noted that between the two polymerases, the buffer utilized did not produce a large fold change in the yield of RNA, with less than 1-fold change difference between any of the conditions, suggesting that the new buffer was superior to the old buffer for both polymerases, which is outlined in the 6.94 and 5.21-fold increase in yield with the new buffer system compared to the old. We also noted that, in 3 of the 5 comparisons, there was at least a 4-fold increase in yield between the gel and bead purifications. Most importantly, utilizing the new polymerase with the new buffer system compared to the original system, in a gel purification we were able to increase the yield 5.55-fold;

furthermore, with utilization of the SPRI paramagnetic beads as a purification system, we were able to increase this yield to a 29.9-fold increase.

Table 35: Fold-change comparisons of original and optimized *in vitro* RNA production and purification methods

	Gel	***SPRI paramagnetic beads***
New pol v old pol (old buffer)	0.799	0.411
New pol v old pol (new buffer)	1.06	1.37
(New pol) old buffer v new buffer	6.94	72.9
(Old pol) old buffer v new buffer	5.21	21.8
New pol new buffer v old pol old buffer	5.55	29.9

I have been able to establish an in-house method of T7 RNA polymerase and buffer purification system that results in a 5.55-fold increase in yield compared to utilization of a purchased polymerase and buffer. I have further optimized the purification system to further increase the yield to 29.9-fold compared to the original reaction conditions. Finally, I have optimized purification time to 1.5 hours over the original 22 hours. These optimization efforts enabled me to rapidly produce large quantities of materials that were utilized in Chapters 4 and 5.

Appendix B: Utilization of proteomics to assess the role of Chlamydia protease-like activity factor

According to the CDC, *C. trachomatis* is an obligate intracellular pathogen and the most commonly reported sexually transmitted disease (STD) in the United States. While there are treatments, there is no acquired-immunity and further there are an increasing number of drug-resistant cases reported, thus there is a need to establish new therapeutics to treat this STD *(496,497)*.

Our lab is interested in a virulence over viability approach wherein we target proteins and pathways which are critical for the infectious nature of the organism, but not the viability. This approach has been demonstrated to result in a slower mutation rate. One potential virulence factor that has been implicated is the protease Chlamydial protease-like activity factor (CPAF). Currently, there is debate within the literature as to the substrates of the protein *(498–501)*. The Valdivia lab at Duke University produced several strains of *Chlamydia trachomatis* which are isogenic except for the presence of CPAF: RST5 contains CPAF while RST17 does not *(502)*. In collaboration with Drs. Kenneth Maksimchuk and Tri Nguyen, who provided samples, we have worked in collaboration with the Duke Proteomics and Metabolomics Core Facility to determine which proteins are affected by the presence of CPAF specifically versus infection as a whole. To do this, we submitted samples that were uninfected, or infected with RST5, or

infected with RST17. We have been working to analyze the peptides to determine which are statistically significant and can be further explored for their role with CPAF.

Infection of A2EN Epithelial Cells and Analysis of Differential Proteolysis

The A2EN endocervical cell line was provided by the Valdivia lab. Cultures of A2EN cells were grown in Keratinocyte SFM media (Life Technologies) supplemented with 10% FBS (Sigma-Aldrich), bovine pituitary extract, and recombinant human epithelial growth factor (Life Technologies). For infection, A2EN cells were seeded in 6-well plates at $1.5x10^5$ cell/well and grown for 48 hours prior to infection. At an estimated $1.2x10^6$ cell/well at confluence, cells were infected at a MOI of 1.0 with either the RST5 or RST17 seed that was previously stored in 1x SPG buffer (0.25 M sucrose, 10 mM sodium phosphate, 5 mM L-glutamic acid; pH 7). Infections were synchronized by centrifugation at 2,000 rpm for 30 min. At both 18 and 48 hpi, cells were harvested by mechanical scraping in 8 M urea with 50 nM anti-CPAF peptide inhibitor (1 mL total per plate). Cells were pelleted, and samples were flash frozen with liquid nitrogen and submitted to the Duke Proteomics Core Facility for analysis.

The frozen samples were gently thawed on ice and were probe sonicated at power level 3, using 4 bursts for 5 seconds/burst, storing on ice between bursts. Let samples shake for 5 minutes at 30 °C to complete solubilization, then samples were centrifuged at 12,000 rcf for 5 minutes at 4 °C and stored on ice. Supernatant (3 μl) was

diluted 4-fold with AmBic, and protein concentrations were determined by mini-Bradford assay (Bio-Rad, Inc), and samples were normalized in fresh tubes to 20 μg protein in 106 μl. Samples (20 μg) from each condition were normalized and subjected to DTT reduction, iodoacetamide alkylation and trypsin digestion.

Following digestion, samples were acidified and diluted to 1 mL in 0.1% TFA. Solid-phase extraction (SPE) cleanup was performed using 1cc Sep-Pak C_{18} cartridges (Waters Corporation) and a vacuum manifold, per manufacturer recommendation. After SPE, samples were diluted to 20% water and lyophilized to dryness. Samples were resuspended in 40 μL 1/2/97 v/v/v TFA/MeCN/water containing 25 fmol/μL ADH1_YEAST MassPrep digest (Waters Corporation) as a surrogate standard. A portion (5 μl) of each sample was combined to create a QC pool to be run before the study for column conditioning and across the study queue.

Data Collection

Quantitative LC/MS/MS was performed on 0.6 μL (300 ng) protein digest per sample, using a nanoAcquity UPLC system (Waters Corp) coupled to a Q Exactive Plus Orbitrap mass spectrometer (Thermo Scientific) via a nanoelectrospray ionization source. Briefly, the sample was first trapped on a Symmetry C18 300 μm × 180 μm trapping column (5 μl/min at 99.9/0.1 v/v water/acetonitrile), after which the analytical separation was performed using a 1.7 μm Acquity HSS T3 C18 75 μm × 250 μm column (Waters Corp.) using a 90-min gradient of 5 to 40% acetonitrile with 0.1% formic acid at

a flow rate of 400 nL/min with a column temperature of 55 °C. Data collection on the Q Exactive Plus mass spectrometer was performed in a data-dependent MS/MS manner, using a 70,000 resolution precursor ion (MS1) scan followed by MS/MS (MS2) of the top 10 most abundant ions at 17,500 resolution. MS2 was accomplished using AGC target of 1x106 ions and max accumulation of 60 ms. MS2 used AGC target of 5x104 ions, 60 ms max accumulation, 2.0 m/z isolation window, 27V normalized collision energy, and 20 second dynamic exclusion. The total analysis cycle time for each sample injection was approximately 2 hours, and the experiment totaled 12 injections.

Samples were run in a randomized block fashion, with a QC pool analyzed at the beginning and the end of the study, and every third sample within the group. Following the analyses, data was imported into Rosetta Elucidator v3.3 (Rosetta Biosoftware, Inc), and all LC-MS files were aligned based on the accurate mass and retention time of detected ions features using a PeakTeller algorithm (Elucidator). The relative peptide abundance was calculated based on area-under-the-curve (AUC) of aligned features across all runs. The dataset had 68,804 quantified features and HCD fragmentation was performed to generate approximately 450,000 MS/MS spectra for sequencing by database searching. This MS/MS data was searched against the UniProtKB (www.uniprot.org) protein sequence database with Homo sapiens species taxonomy appended with the NCBI refseq database with *C. trachomatis* serovar 434/Bu taxonomy

(http://www.ncbi.nlm.nih.gov/protein), which also contained several surrogate standards sequences and common laboratory contamination proteins, as well as a reversed-sequence database for false discovery rate (FDR) determination.

Amino acid modifications allowed in database searching included fixed carbamidomethyl Cys (+57), and variable oxidation N-terminal acetylation (+42), and the data was searched with 5 ppm precursor, 0.02 Da product ion tolerance, and semi-tryptic enzyme specificity. Data was processed to the PeptideTeller data curation algorithm to determine FDR and was annotated at 0.9% peptide FDR.

Current progress

In order to establish which peptides and proteins were changing due to the function of CPAF, we needed to establish both statistical and biological cut-off points for our proteomic data. Furthermore, for peptide analysis, we would need to establish if the cleavage at both ends was tryptic (R or K in P1 position).

For peptide analysis, we first took the CV (%QC) of each peptide and plotted the full list (22801 and 29894 for 18 and 48 hours, respectively) in a box and whisker graph with JMP Pro 13. We utilized this to establish a cutoff of 16.57% (18-hour) and 11.90% (48-hour) at the 3rd quartile for variability in the %QC for both time points (Figure 93). We next set the minimum value to 1000000, the lower detection limit of the LC-MS/MS system, and log-2 transformed the data. We additionally removed any duplicate peptide sequences that were assigned to different proteins.

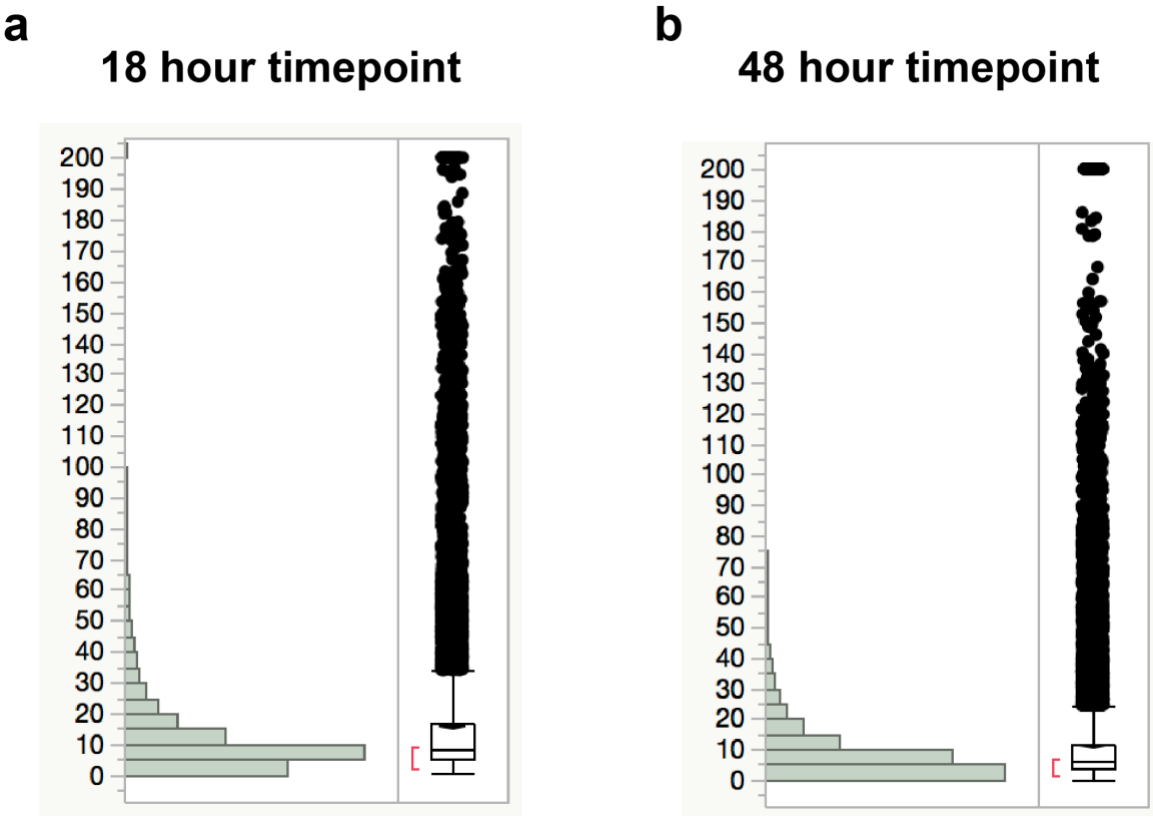


Figure 93: Box and whisker plots of the CV(%QC) of all peptides from the (a) 18-hour or (b) 48-hour time points

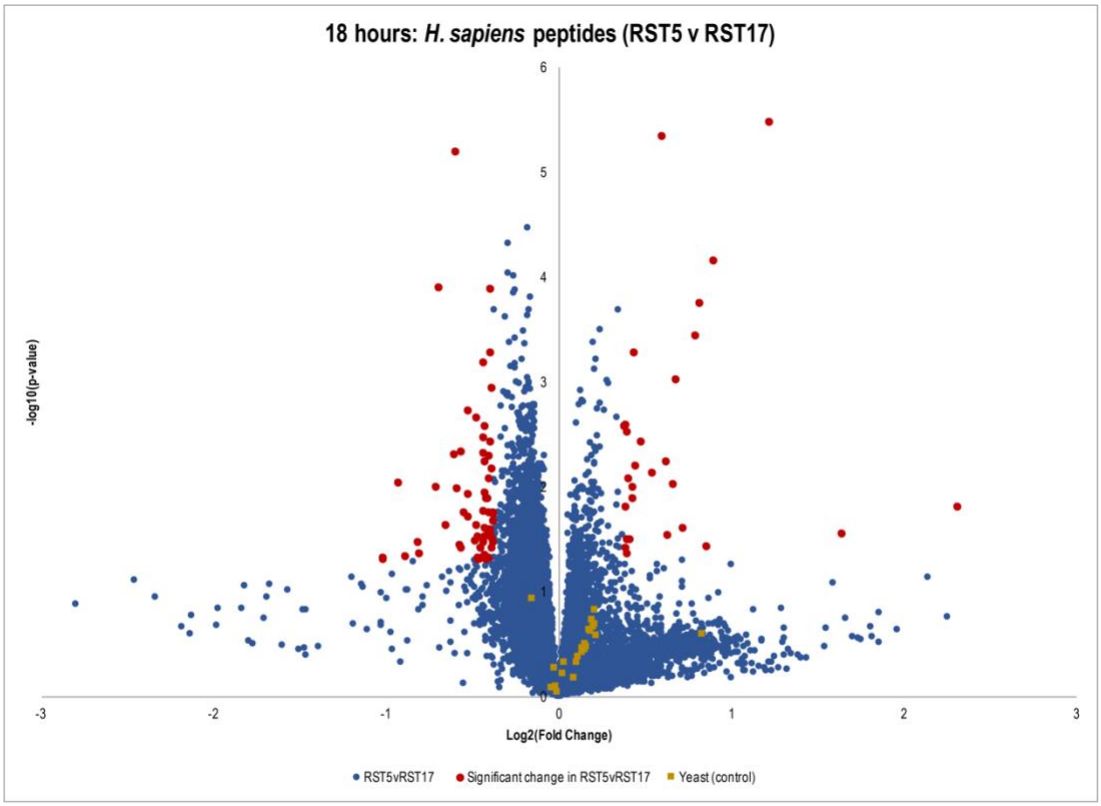
With the Elucidator list of peptides in hand, we next wished to establish if any of the peptides had been produced via CPAF cleavage. Utilizing Scaffold, we generated a list of peptides (90% protein threshold, 50% peptide threshold) which supplied the next and previous residues of each peptide sequence as well where the peptide resided within the total protein sequence. This allowed us to establish whether there were 0, 1, or 2 non-tryptic cleavages, suggesting a CPAF cleavage. If the initial peptide residue was 4 or less residues from the start of the protein sequence, we considered this N-terminal,

and thus not a CPAF cleavage. If there were no subsequent amino acids after the C-terminal of the peptide, these were labeled "C-terminal" and not considered a CPAF cleavage. We additionally removed the sequence duplicates from this dataset, additionally peptide sequences that could be assigned to multiple proteins.

To combine the information from both datasets, we next overlapped both sequences from a concatenation of the peptide sequence and the accession number of the protein associated with the peptide sequence which allowed us to eliminate sequences either not present in both datasets or assigned to a different protein in each dataset. To establish which peptides were statistically significant and had a significant fold change, based on power calculations we utilized p-value cut-off of 0.05 and a fold-change cut-off of +/-1.3-fold change (log2 (+/- 0.3785). The p-value was established using a standard 3-tailed student's t-test. We visualized the significant peptides via volcano plots as shown below in Figures 94a-96a for *H. sapiens* at 18- and 48-hour time-points for RST5 vs RST17 and *C. trachomatis* at the 48-hour time-point. There were no significant peptides at the 18-hour time point for *C. trachomatis*. Figure 94b-96b depict the peptides which met all of the cut-offs and additionally contained one non-tryptic cleavage. There were 3 peptides at 18 hours and 10 peptides at 48 hours which displayed one non-tryptic cleavage; however, none matched proteins we had hypothesized such as DNAJ or vimentin, which have previously been demonstrated to be cleaved by CPAF. It is likely more

sensitive methods will be required to fully establish which peptides are formed due to direct CPAF cleavage. We next sought to establish which proteins were impacted by the presence of CPAF

a

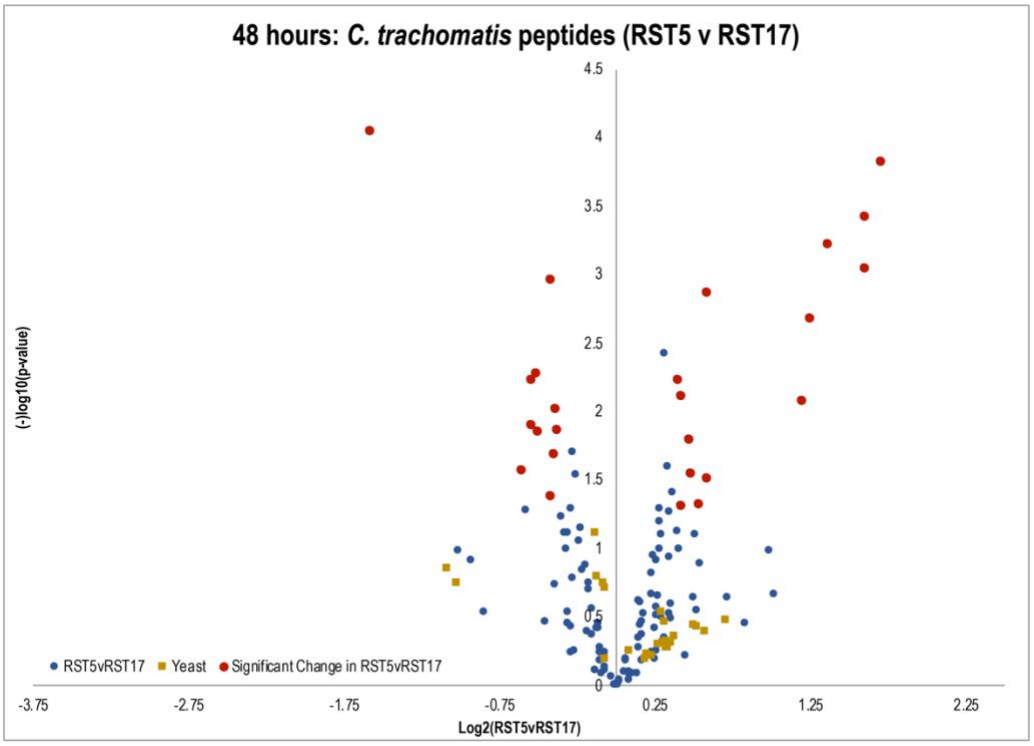


b

Modified Peptide Sequence	Primary Protein Name	Protein Description	%CV (QC)	p-value	Fold change
ISPDQLADLYK	ENOA_HUMAN	Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2	5.27	0.0483	-1.02
SISTSGGSFR	K2C5_HUMAN	Keratin, type II cytoskeletal 5 OS=Homo sapiens GN=KRT5 PE=1 SV=3	5.23	0.0500	-1.02
VLSLLALVKPE	PSME2_HUMAN	Proteasome activator complex subunit 2 OS=Homo sapiens GN=PSME2 PE=1 SV=3	12.2	0.0341	-0.813

Figure 94: Volcano plot for *H. sapiens* peptides at the 18-hour time point. After determining which peptides were present in both the Elucidator and Scaffold datasets, we (a) plotted all peptides present in both datasets (blue), highlighting those which demonstrated a significant p-value (0.05) and fold change (+/-0.3785) in red. We further assessed the number of tryptic cleavages for the significant (red) peptides and listed those with one non-tryptic cleavage, presumably due to CPAF.

a

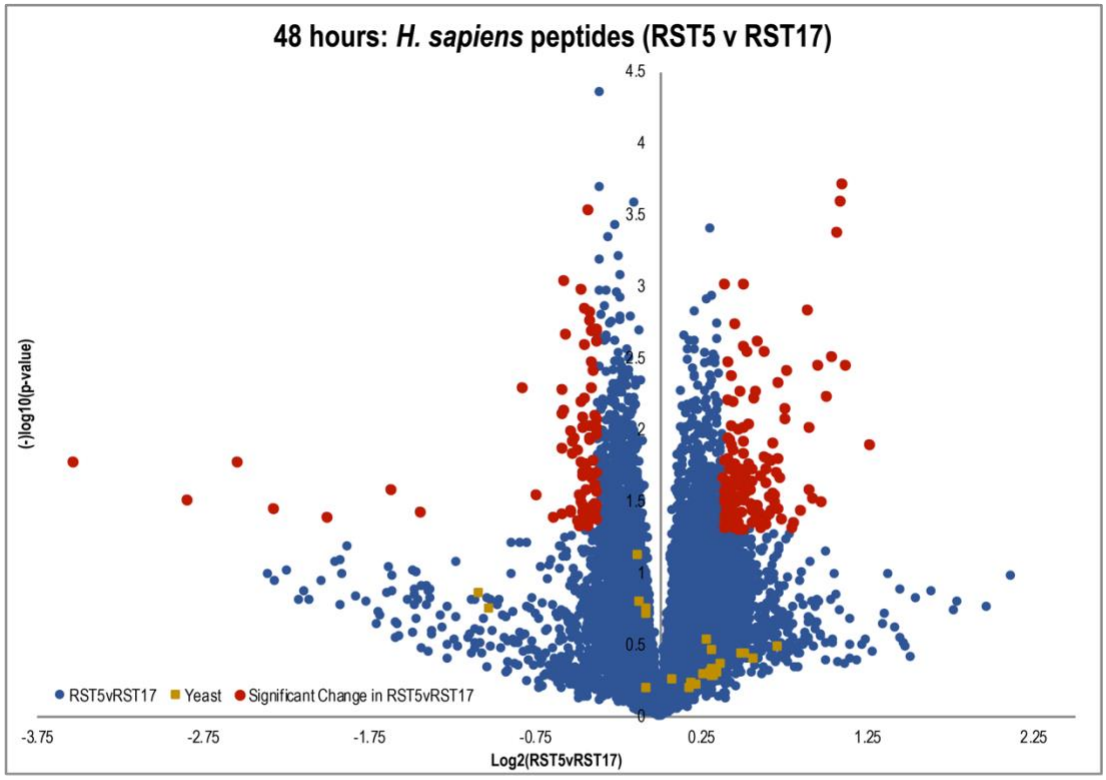


b

Modified Peptide Sequence	Primary Protein Name	Protein Description	%CV (QC)	p-value	Fold change
NITQGNILLQAR	478751174	gb|AGJ65172.1 inorganic pyrophosphatase [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546178	10.54240217	0.001345468	0.578190333

Figure 95: Volcano plot for *C. trachomatis* peptides at the 48-hour time point. After determining which peptides were present in both the Elucidator and Scaffold datasets, we (a) plotted all peptides present in both datasets (blue), highlighting those which demonstrated a significant p-value (0.05) and fold change (+/-0.3785) in red. We further assessed the number of tryptic cleavages for the significant (red) peptides and listed those with one non-tryptic cleavage, presumably due to CPAF.

a



b

Modified Peptide Sequence	Primary Protein Name	Protein Description	%CV (QC)	p-value	Fold change
PSSTQPALPK	NDUS7_HUMAN	NADH dehydrogenase [ubiquinone] iron-sulfur protein 7, mitochondrial OS=Homo sapiens GN=NDUFS7 PE=1 SV=3	9.42	0.00638	-0.470
VVNYTNR	SFXN1_HUMAN	Sideroflexin-1 OS=Homo sapiens GN=SFXN1 PE=1 SV=4	7.14	0.0103	-0.534
DAAQEPTGNNAEICLLPLDYGPCR	TFPI2_HUMAN	Tissue factor pathway inhibitor 2 OS=Homo sapiens GN=TFPI2 PE=1 SV=1	9.38	0.0243	0.456
ALFIRPF	ABCF2_HUMAN	ATP-binding cassette sub-family F member 2 OS=Homo sapiens GN=ABCF2 PE=1 SV=2	11.10	0.0305	0.9146
GGFCEVCKK	SAP_HUMAN	Proactivator polypeptide OS=Homo sapiens GN=PSAP PE=1 SV=2	4.54	0.0354	-0.406
NIKPQVDR	SAHH_HUMAN	Adenosylhomocysteinase OS=Homo sapiens GN=AHCY PE=1 SV=4	9.14	0.0373	-0.541
MQEFMILPVGAANFR	ENOA_HUMAN	Alpha-enolase OS=Homo sapiens GN=ENO1 PE=1 SV=2	10.07	0.0381	-1.44
AQEPTGNNAEICLLPLDYGPCR	TFPI2_HUMAN	Tissue factor pathway inhibitor 2 OS=Homo sapiens GN=TFPI2 PE=1 SV=1	9.32	0.0382	0.454
IIIFVPVPQLK	RS7_HUMAN	40S ribosomal protein S7 OS=Homo sapiens GN=RPS7 PE=1 SV=1	3.98	0.0407	-2.01
IRPEGQDSAFSWTDLLLK	SYMC_HUMAN	-tRNA synthetase, cytoplasmic OS=Homo sapiens GN=MARS PE=1	10.10	0.0476	-0.434

Figure 96: Volcano plot for *H. sapiens* peptides at the 48-hour time point. After determining which peptides were present in both the Elucidator and Scaffold datasets, we (a) plotted all peptides present in both datasets (blue), highlighting those which demonstrated a significant p-value (0.05) and fold change (+/-0.3785) in red. We further assessed the number of tryptic cleavages for the significant (red) peptides and listed those with one non-tryptic cleavage, presumably due to CPAF.

To establish which proteins were significant, as with the peptide analysis, the minimum intensity was set to 1000000. We opted to utilize the same CV(%QC) cut-offs

as the peptides established previously (16.56% at 18 hours, 11.90% for 48 hours), instead of looking at the number of peptides present for each protein to establish whether a protein would be included in the analysis. We chose this route to ensure incorporation of proteins with a low ability to be analyzed via MS/MS was low. Upon removing the proteins with a high %CV(QC), we performed a textbook ANOVA across all three conditions as well as utilizing a student's 2-tailed t-test to compare RST5 and RST17 conditions separately. When observing the results for the ANOVA for *H.* sapiens proteins, there were 650 proteins at 18 hours and 473 proteins at 48 hours and 49 proteins at 48 hours for *C. trachomatis* that met an ANOVA p-value of 0.05 or less, visualized by a heatmap in Figures 97-99. While many of the protein levels for *C. trachomatis* appeared to remain relatively the same between RST5 and RST17, the *H. sapiens* heatmaps revealed interesting trends. While there were distinct changes between uninfected compared to infected (RST5 and RST17), we noted that there were groupings between RST5 and RST17 which remained the same, while other groups of proteins responded oppositely, suggesting that these proteins were specifically affected by the presence of CPAF.

We next utilized a student's two-tailed t-test to establish which proteins had a significant p-value for RST5 vs RST17 and then further utilized a 1.3-fold cut off. There were 24 and 35 significant proteins for *H. sapiens* at the 18- and 48-hour time points,

respectively, and 9 significant proteins for *C. trachomatis* at the 48-hour time point as listed in Figures 100-102. We outlined which proteins decreased or increased in cellular level due to the presence of CPAF in red or green, respectively. In this analysis, if the level of protein decreased due to the presence of CPAF, this demonstrated a decreased fold change and is displayed in red. Conversely, if the level of protein increased due to the presence of CPAF, this demonstrated an increased fold change and is displayed in green. Again, it was not unexpected that there were no significant proteins for *C. trachomatis* at the 18-hour time point due to the lifecycle timeline of the infection.

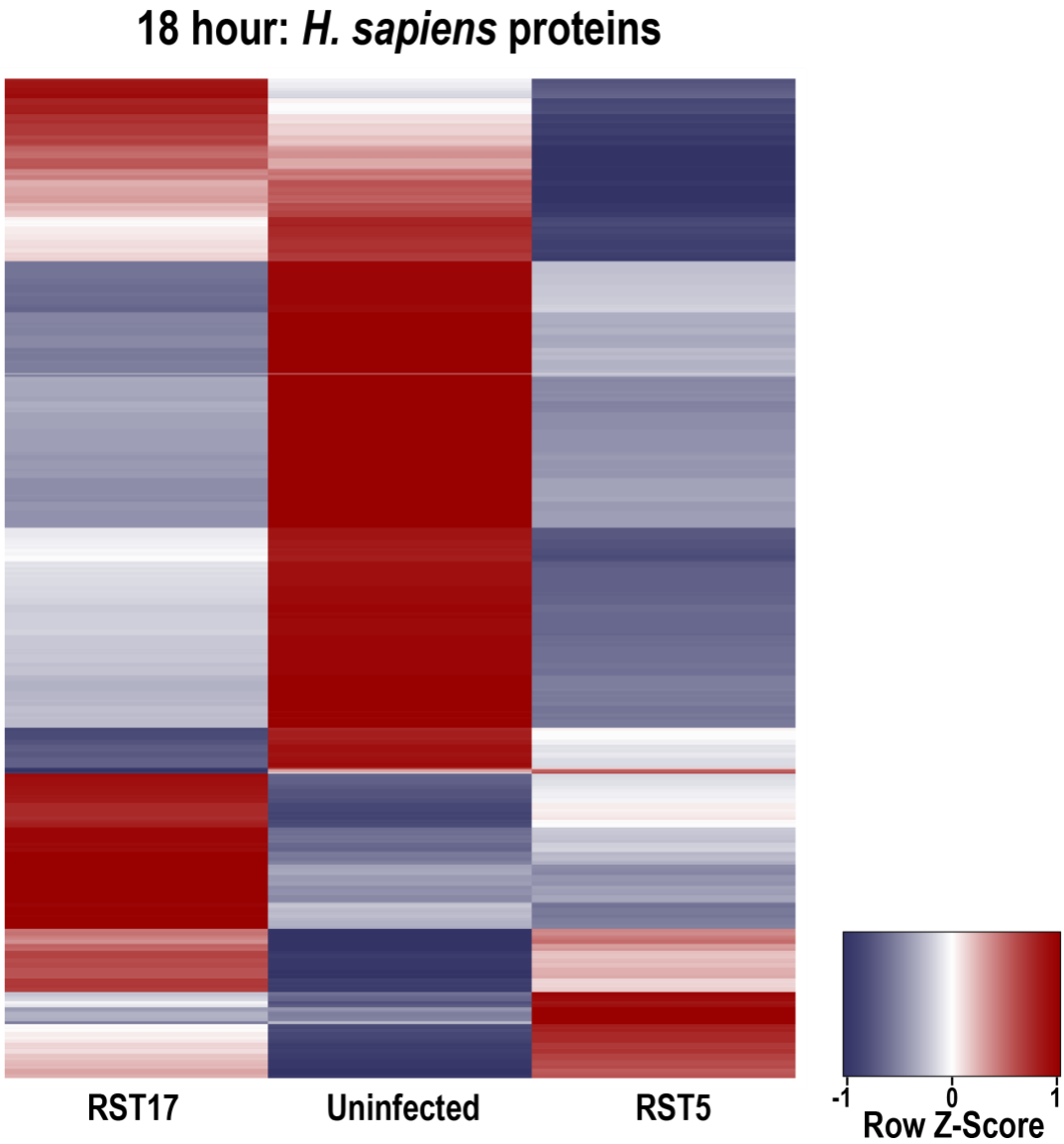


Figure 97: ***H. sapiens*** **proteins at 18 hours with an ANOVA p-value ≤ 0.05 across all three conditions. We were interested to note that while many of the proteins changed in a similar manner between uninfected and infected conditions (RST5 and RST17) there were also proteins which demonstrated an opposite fold-change between RST5 and RST17, suggesting that a series of proteins are affected by the presence of CPAF.**

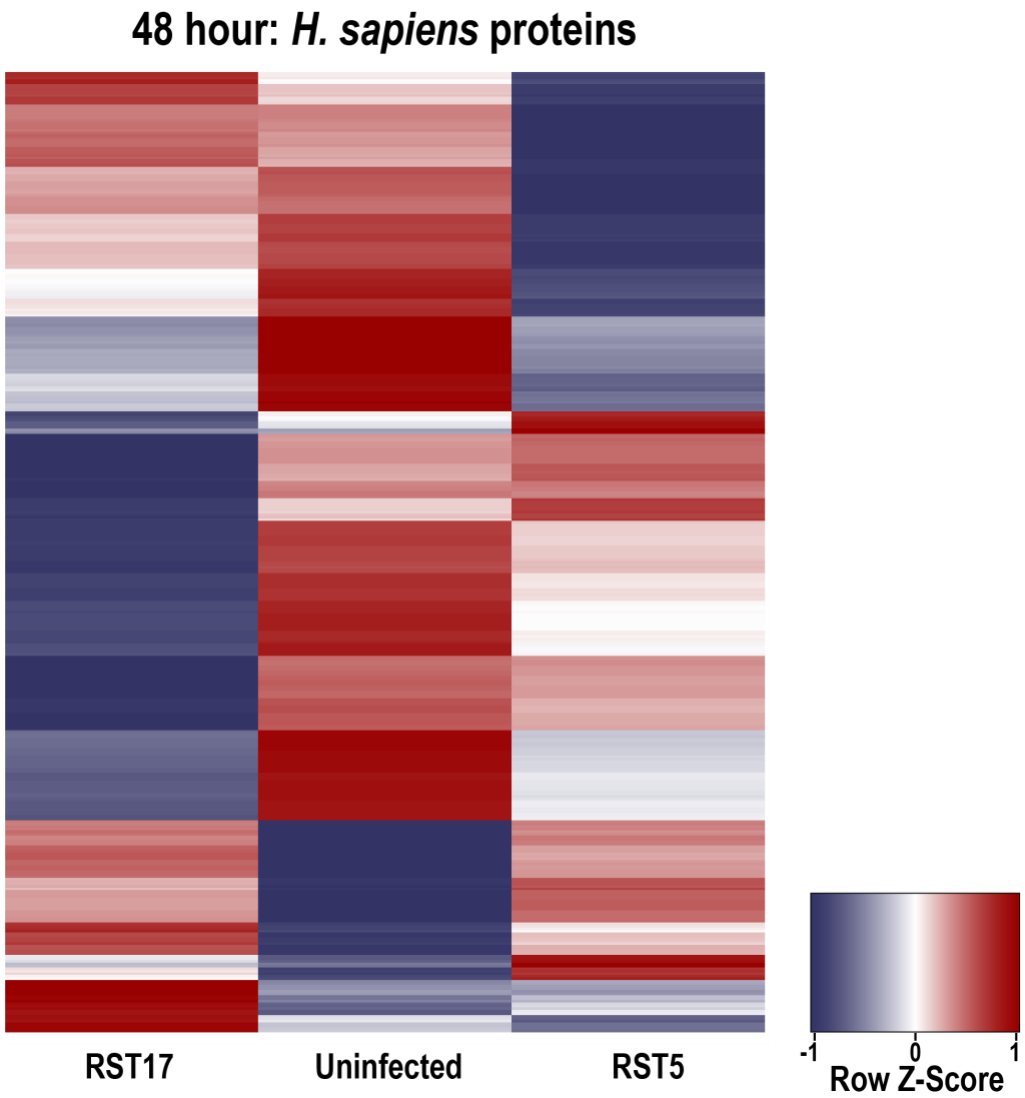


Figure 98: ***H. sapiens*** **proteins at 48 hours with an ANOVA p-value ≤ 0.05 across all three conditions. We were interested to note that while many of the proteins changed in a similar manner between uninfected and infected conditions (RST5 and RST17) there were also proteins which demonstrated an opposite fold-change between RST5 and RST17, suggesting that a series of proteins are affected by the presence of CPAF.**

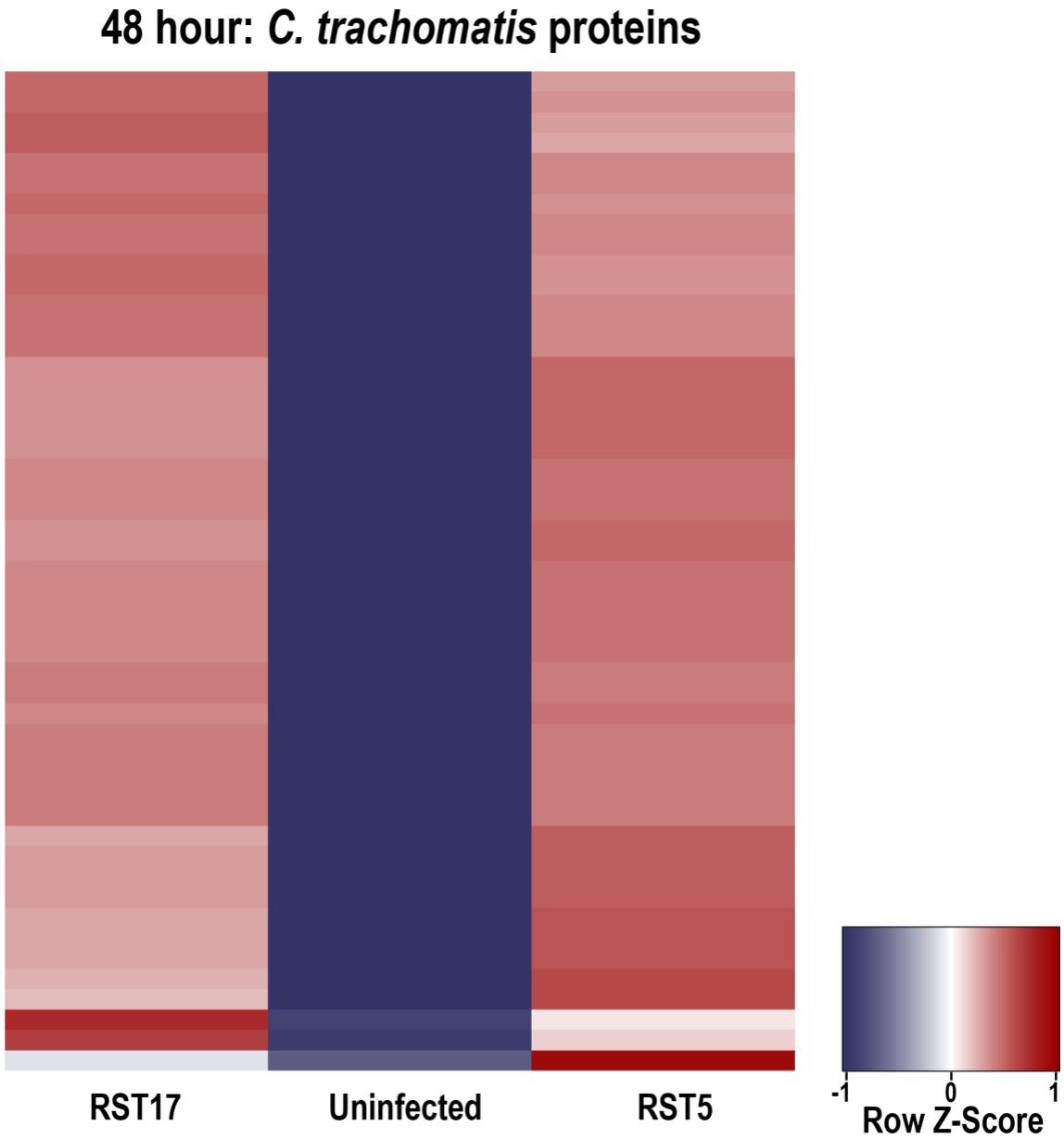


Figure 99: ***C. trachomatis*** **proteins at 48 hours with a significant ANOVA p-value ≤ 0.05 across all three conditions. We noted no significant changes between RST5 and RST17 for most significant proteins.**

Primary Protein Name	Protein Description	%CV (QC)	RST5vRST17 (p.value)	RST5vRST17 (F.C)
GRP3_HUMAN	Ras guanyl-releasing protein 3 OS=Homo sapiens GN=RASGRP3 PE=2 SV=1	10.61278	0.04401676	-0.684970414
ERR2_HUMAN	Steroid hormone receptor ERR1 OS=Homo sapiens GN=ESRRA PE=1 SV=3	10.05490	0.045535541	-0.658194456
PKP1_HUMAN	Plakophilin-1 OS=Homo sapiens GN=PKP1 PE=1 SV=2	15.97228	0.010547955	-0.581649078
MTA2_HUMAN	Metastasis-associated protein MTA2 OS=Homo sapiens GN=MTA2 PE=1 SV=1	9.58152	0.043674285	-0.562121968
TYDP2_HUMAN	Tyrosyl-DNA phosphodiesterase 2 OS=Homo sapiens GN=TDP2 PE=1 SV=1	7.12465	0.041792823	-0.49792465
PIR_HUMAN	Pirin OS=Homo sapiens GN=PIR PE=1 SV=1	6.02753	0.047197519	-0.470488987
NUDC3_HUMAN	NudC domain-containing protein 3 OS=Homo sapiens GN=NUDCD3 PE=1 SV=3	12.18751	0.024677459	-0.454404839
PURA1_HUMAN	Adenylosuccinate synthetase isozyme 1 OS=Homo sapiens GN=ADSSL1 PE=1 SV=1	3.37398	0.000652236	-0.437280448
KCMF1_HUMAN	E3 ubiquitin-protein ligase KCMF1 OS=Homo sapiens GN=KCMF1 PE=1 SV=2	11.19508	0.017750908	-0.435853827
KS6A3_HUMAN	Ribosomal protein S6 kinase alpha-3 OS=Homo sapiens GN=RPS6KA3 PE=1 SV=1	15.34517	0.030803476	-0.414766271
PPIC_HUMAN	Peptidyl-prolyl cis-trans isomerase C OS=Homo sapiens GN=PPIC PE=1 SV=1	5.04328	2.0457E-05	-0.38969297
PMVK_HUMAN	Phosphomevalonate kinase OS=Homo sapiens GN=PMVK PE=1 SV=3	6.06736	0.008650753	-0.387038532
C10_HUMAN	Protein C10 OS=Homo sapiens GN=C12orf57 PE=1 SV=1	14.86972	0.011075629	-0.382275177
EDC3_HUMAN	Enhancer of mRNA-decapping protein 3 OS=Homo sapiens GN=EDC3 PE=1 SV=1	10.30141	0.017779771	-0.379054626
TOIP1_HUMAN	Torsin-1A-interacting protein 1 OS=Homo sapiens GN=TOR1AIP1 PE=1 SV=2	7.06338	0.017190064	0.41044762
ARLY_HUMAN	Argininosuccinate lyase OS=Homo sapiens GN=ASL PE=1 SV=4	3.14929	0.000196797	0.414776687
INSC_HUMAN	Protein inscuteable homolog OS=Homo sapiens GN=INSC PE=1 SV=1	7.73021	0.003201586	0.418456324
BRD9_HUMAN	Bromodomain-containing protein 9 OS=Homo sapiens GN=BRD9 PE=1 SV=2	9.07362	0.010062517	0.426652895
CG050_HUMAN	Uncharacterized protein C7orf50 OS=Homo sapiens GN=C7orf50 PE=1 SV=1	9.13270	0.038171881	0.429404401
TRFE_HUMAN	Serotransferrin OS=Homo sapiens GN=TF PE=1 SV=2	1.95666	0.012828089	0.473427014
STK38_HUMAN	Serine/threonine-protein kinase 38 OS=Homo sapiens GN=STK38 PE=1 SV=1	1.60674	0.003323868	0.539641104
HBA_HUMAN	Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2	4.87572	0.000314205	0.771185299
APOE_HUMAN	Apolipoprotein E OS=Homo sapiens GN=APOE PE=1 SV=1	7.59138	0.000905175	0.789994894
SNPC2_HUMAN	snRNA-activating protein complex subunit 2 OS=Homo sapiens GN=SNAPC2 PE=1 SV=1	3.81523	0.030695348	1.225187236
ALBU_BOVIN	Serum albumin precursor - Bos taurus	4.64390	0.036537254	1.770024762

Figure 100: *H. sapiens* proteins at 18 hours which demonstrated a significant p-value via student's t-test for RST5 vs RST17 as well as a significant fold-change wherein green depicts and increase in protein level due to the presence of CPAF and red a decrease in protein level due to the presence of CPAF.

Primary Protein Name	Protein Description	%CV (QC)	RST5vRST17 (p.value)	RST5vRST17 (F.C)
MYPR_HUMAN	Myelin proteolipid protein OS=Homo sapiens GN=PLP1 PE=1 SV=2	6.1	0.018084056	-3.532928411
LRC53_HUMAN	Leucine-rich repeat-containing protein 53 OS=Homo sapiens GN=LRRC53 PE=4 SV=2	7.9	0.02971102	-0.870852873
GLNA_HUMAN	Glutamine synthetase OS=Homo sapiens GN=GLUL PE=1 SV=4	2.7	0.042013755	-0.760639129
ODBA_HUMAN	2-oxoisovalerate dehydrogenase subunit alpha, mitochondrial OS=Homo sapiens GN=BCKDHA PE=1 SV=2	5.8	0.014809861	-0.52790308
NDUS2_HUMAN	NADH dehydrogenase [ubiquinone] iron-sulfur protein 2, mitochondrial OS=Homo sapiens GN=NDUFS2 PE=1 SV=2	4.4	0.029340652	-0.472828341
IL1RA_HUMAN	Interleukin-1 receptor antagonist protein OS=Homo sapiens GN=IL1RN PE=1 SV=1	3.8	0.033698998	-0.472808479
MAP2_HUMAN	Microtubule-associated protein 2 OS=Homo sapiens GN=MAP2 PE=1 SV=4	9.0	0.045890204	-0.451251885
PSB9_HUMAN	Proteasome subunit beta type-9 OS=Homo sapiens GN=PSMB9 PE=1 SV=2	10.4	0.00128153	-0.442746706
SIAE_HUMAN	Sialate O-acetylesterase OS=Homo sapiens GN=SIAE PE=2 SV=1	10.0	0.042928989	-0.419171539
K2C6B_HUMAN	Keratin, type II cytoskeletal 6B OS=Homo sapiens GN=KRT6B PE=1 SV=5	3.7	0.006422776	-0.413605537
RIR2_HUMAN	Ribonucleoside-diphosphate reductase subunit M2 OS=Homo sapiens GN=RRM2 PE=1 SV=1	5.0	0.031101998	0.385547664
UCK2_HUMAN	Uridine-cytidine kinase 2 OS=Homo sapiens GN=UCK2 PE=1 SV=1	11.6	0.021909212	0.385677281
FINC_HUMAN	Fibronectin OS=Homo sapiens GN=FN1 PE=1 SV=3	1.1	0.020862219	0.392312415
DCNL5_HUMAN	DCN1-like protein 5 OS=Homo sapiens GN=DCUN1D5 PE=1 SV=1	6.2	0.03798931	0.447012493
UHRF1_HUMAN	E3 ubiquitin-protein ligase UHRF1 OS=Homo sapiens GN=UHRF1 PE=1 SV=1	6.6	0.024161754	0.461879165
DHX16_HUMAN	Putative pre-mRNA-splicing factor ATP-dependent RNA helicase DHX16 OS=Homo sapiens GN=DHX16 PE=1 SV=2	4.4	0.026574098	0.479681657
ELP3_HUMAN	Elongator complex protein 3 OS=Homo sapiens GN=ELP3 PE=1 SV=2	10.0	0.023420319	0.499122258
UBE2O_HUMAN	Ubiquitin-conjugating enzyme E2 O OS=Homo sapiens GN=UBE2O PE=1 SV=3	9.7	0.025724443	0.499703486
TF3C4_HUMAN	General transcription factor 3C polypeptide 4 OS=Homo sapiens GN=GTF3C4 PE=1 SV=2	5.3	0.039619348	0.523772491
NOC4L_HUMAN	Nucleolar complex protein 4 homolog OS=Homo sapiens GN=NOC4L PE=1 SV=1	10.9	0.024668378	0.530314626
PDD2L_HUMAN	Programmed cell death protein 2-like OS=Homo sapiens GN=PDCD2L PE=1 SV=1	7.4	0.026156485	0.544460484
RTCA_HUMAN	RNA 3'-terminal phosphate cyclase OS=Homo sapiens GN=RTCA PE=1 SV=1	2.8	0.003089432	0.549505272
CBLC_HUMAN	Signal transduction protein CBL-C OS=Homo sapiens GN=CBLC PE=1 SV=3	11.9	0.000308849	0.614836787
SPC25_HUMAN	Kinetochore protein Spc25 OS=Homo sapiens GN=SPC25 PE=1 SV=1	6.8	0.019237285	0.811170166
LRC71_HUMAN	Leucine-rich repeat-containing protein 71 OS=Homo sapiens GN=LRRC71 PE=2 SV=1	5.4	0.001387701	0.812263887
NF2IP_HUMAN	NFATC2-interacting protein OS=Homo sapiens GN=NFATC2IP PE=1 SV=1	10.2	0.0476305	0.878382883
NIN_HUMAN	Ninein OS=Homo sapiens GN=NIN PE=1 SV=3	0.3	0.008541829	0.900219181
LAMP2_HUMAN	Lysosome-associated membrane glycoprotein 2 OS=Homo sapiens GN=LAMP2 PE=1 SV=2	3.9	0.003323887	0.914501578
FETUA_HUMAN	Alpha-2-HS-glycoprotein OS=Homo sapiens GN=AHSG PE=1 SV=1	4.1	0.003033355	1.012331945
EOGT_HUMAN	EGF domain-specific O-linked N-acetylglucosamine transferase OS=Homo sapiens GN=EOGT PE=1 SV=1	6.6	0.001955418	1.332095033
IL27B_HUMAN	Interleukin-27 subunit beta OS=Homo sapiens GN=EBI3 PE=1 SV=2	6.5	0.00010846	1.472050278
HAUS4_HUMAN	HAUS augmin-like complex subunit 4 OS=Homo sapiens GN=HAUS4 PE=1 SV=1	6.0	0.000018345	2.007322311
KLD8A_HUMAN	Kelch domain-containing protein 8A OS=Homo sapiens GN=KLHDC8A PE=2 SV=2	8.1	2.81757E-05	2.149862319
E2AK3_HUMAN	Eukaryotic translation initiation factor 2-alpha kinase 3 OS=Homo sapiens GN=EIF2AK3 PE=1 SV=3	8.7	0.00488476	2.389050193
KAIN_HUMAN	Kallistatin OS=Homo sapiens GN=SERPINA4 PE=1 SV=3	7.6	0.000113205	2.596421168

Figure 101: *H. sapiens* proteins at 48 hours which demonstrated a significant p-value via student's t-test for RST5 vs RST17 as well as a significant fold-change wherein green depicts and increase in protein level due to the presence of CPAF and red a decrease in protein level due to the presence of CPAF.

Primary Protein Name	Protein Description	%CV (QC)	RST5vRST17 (p.value)	RST5vRST17 (F.C)
478751442	gb|AGJ65440.1 hypothetical protein CTLFINAL_02105 [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546446	8.2	0.041233032	-0.599705486
478751702	gb|AGJ65700.1 molecular chaperone DnaK [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546706	3.4	0.00381932	-0.421708081
478751292	gb|AGJ65290.1 hypothetical protein CTLFINAL_01350 [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546296	3.5	0.01601063	-0.391076622
478751535	gb|AGJ65533.1 3-ketoacyl-ACP reductase [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546539	7.9	0.046364425	0.384823414
478751822	gb|AGJ65820.1 glyceraldehyde-3-phosphate dehydrogenase [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546826	8.5	0.015517227	0.496779706
478751413	gb|AGJ65411.1 membrane protein [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546417	6.4	0.016757174	0.533218738
478751206	gb|AGJ65204.1 30S ribosomal protein S18 [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546210	8.5	0.049160493	0.545430132
478751174	gb|AGJ65172.1 inorganic pyrophosphatase [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546178	10.5	0.001345468	0.578190333
478751076	gb|AGJ65074.1 Major outer membrane porin [Chlamydia trachomatis L2/434/Bu(f)]; Duplicate proteins: gi|482546080	3.4	0.001875111	1.388672178

Figure 102: *C. trachomatis* proteins at 48 hours which demonstrated a significant p-value via student's t-test for RST5 vs RST17 as well as a significant fold-change wherein green depicts and increase in protein level due to the presence of CPAF and red a decrease in protein level due to the presence of CPAF.

References

(1) Waddington, C. H. (2012) The epigenotype. 1942. *Int. J. Epidemiol. 41*, 10–13.

(2) Felsenfeld, G. (2014) A brief history of epigenetics. *Cold Spring Harb. Perspect. Biol. 6*, a018200.

(3) Saunders, P. T. (2017) Epigenetics and Evolution. *Hum. Dev. 60*, 81–94.

(4) Nanney, D. L. (1958) EPIGENETIC CONTROL SYSTEMS. *Proc. Natl. Acad. Sci. 44*, 712–717.

(5) Saletore, Y., Meyer, K., Korlach, J., et al. (2012) The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol. 13*, 175.

(6) Butler, A. A., Webb, W. M., and Lubin, F. D. (2016) Regulatory RNAs and control of epigenetic mechanisms: Expectations for cognition and cognitive dysfunction. *Epigenomics 8*, 135–151.

(7) Stewart, K. R., Veselovska, L., and Kelsey, G. (2016) Establishment and functions of DNA methylation in the germline. *Epigenomics*.

(8) Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., et al. (2013) Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife 2013*, 523.

(9) Luger, K., Mäder, A. W., Richmond, R. K., et al. (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature 389*, 251–260.

(10) Peterson, C. L., and Laniel, M.-A. (2004) Histones and histone modifications. *Curr. Biol. 14*, R546–R551.

(11) Vaughn Jackson. (1990) In Vivo Studies on the Dynamics of Histone-DNA Interaction: Evidence for Nucleosome Dissolution during Replication and Transcription and a Low Level of Dissolution Independent of Both. *Biochemistry 29*, 719–731.

(12) Ramakrishnan, V. (1997) HISTONE STRUCTURE AND THE ORGANIZATION OF

THE NUCLEOSOME. *Annu. Rev. Biophys. Biomol. Struct. 26*, 83–112.

(13) Mazurkiewicz, J., Kepert, J. F., and Rippe, K. (2006) On the mechanism of nucleosome assembly by histone chaperone NAP1. *J. Biol. Chem. 281*, 16462–16472.

(14) Kepert, J. F., Mazurkiewicz, J., Heuvelman, G. L., et al. (2005) NAP1 Modulates Binding of Linker Histone H1 to Chromatin and Induces an Extended Chromatin Fiber Conformation * □ S.

(15) Mcbryant, S. J., Park, Y.-J., Abernathy, S. M., et al. (2003) Preferential Binding of the Histone (H3-H4) 2 Tetramer by NAP1 Is Mediated by the Amino-terminal Histone Tails*.

(16) Worcel, A., Han, S., and Wong, M. L. (1978) Assembly of newly replicated chromatin. *Cell 15*, 969–977.

(17) Jackson, V. (1987) Deposition of newly synthesized histones: new histones H2A and H2B do not deposit in the same nucleosome with new histones H3 and H4. *Biochemistry 26*, 2315–2325.

(18) Kornberg, R. D. (1974) Chromatin structure: A repeating unit of histones and DNA. *Science (80-.). 184*, 868–871.

(19) Jackson, V. (1988) Deposition of newly synthesized histones: hybrid nucleosomes are not tandemly arranged on daughter DNA strands. *Biochemistry 27*, 2109–2120.

(20) Nacheva, G. A., Guschin, D. Y., Preobrazhenskaya, O. V, et al. (1989) Change in the pattern of histone binding to DNA upon transcriptional activation. *Cell 58*, 27–36.

(21) Hansen, J. C., and Wolffe, A. P. (1992) Influence of Chromatin Folding on Transcription Initiation and Elongation by RNA Polymerase III. *Biochemistry 31*, 7977–7988.

(22) Saksouk, N., Simboeck, E., and Déjardin, J. (2015) Constitutive heterochromatin formation and transcription in mammals. *Epigenetics and Chromatin 8*.

(23) Zaret, K. S., Lerner, J., and Iwafuchi-Doi, M. (2016) Chromatin Scanning by Dynamic Binding of Pioneer Factors. *Mol. Cell 62*, 665–667.

(24) Raab, J. R., Smith, K. N., Spear, C. C., et al. (2019) SWI/SNF remains localized to chromatin in the presence of SCHLAP1. *Nat. Genet. 51*, 26–29.

(25) Fenley, A. T., Adams, D. A., and Onufriev, A. V. (2010) Charge state of the globular histone core controls stability of the nucleosome. *Biophys. J. 99*, 1577–1585.

(26) Tse, C., Sera, T., Wolffe, A. P., 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*, 4629–4638.

(27) Lee, D. Y., Hayes, J. J., Pruss, D., et al. (1993) A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell 72*, 73–84.

(28) Hebbes, T. R., Thorne, A. W., and Crane-Robinson, C. (1988) A direct link between core histone acetylation and transcriptionally active chromatin. *EMBO J. 7*, 1395–1402.

(29) Beisaw, A., Tsaytler, P., Koch, F., et al. (2017) BRACHYURY directs histone acetylation to target loci during mesoderm development. *EMBO Rep. 19*, e201744201.

(30) Ray, S. K., Li, H. J., Metzger, E., et al. (2014) CtBP and Associated LSD1 Are Required for Transcriptional Activation by NeuroD1 in Gastrointestinal Endocrine Cells. *Mol. Cell. Biol. 34*, 2308–2317.

(31) Lan, F., Collins, R. E., De Cegli, R., et al. (2007) Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature 448*, 718–722.

(32) Shi, Y., Lan, F., Matson, C., et al. (2004) Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. *Cell 119*, 941–953.

(33) Kim, A.-J., Park, J.-A., Kim, K.-C., et al. (2011) Deacetylation and methylation at histone H3 lysine 9 (H3K9) coordinate chromosome condensation during cell cycle progression. *Mol. Cells 31*, 343–349.

(34) Jeong, Y. S., Cho, S., Park, J. S., et al. (2010) Phosphorylation of serine-10 of histone H3 shields modified lysine-9 selectively during mitosis. *Genes to Cells 15*, 181–192.

(35) In, I., Chromatin, R., Inheritance, E., et al. (2008) Phosphorylation of H3S10 Blocks the Access of H3K9 by Specific Antibodies and Histone Methyltransferase. *J. Biol. Chem. 283*, 33585–33590.

(36) Strahl, B. D., and Allis, C. D. (2000, January 6) The language of covalent histone modifications. *Nature*. Nature Publishing Group.

(37) Yang, Z., Jones, A., Widschwendter, M., et al. (2015) An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. *Genome Biol. 16*, 140.

(38) Choo, K. B. (2011) Epigenetics in disease and cancer. *Malays. J. Pathol.*

(39) Zhang, Z., Lei, A., Xu, L., et al. (2017) Similarity in gene-regulatory networks suggests that cancer cells share characteristics of embryonic neural cells. *J. Biol. Chem. 292*, 12842–12859.

(40) Ropero, S., and Esteller, M. (2007) The role of histone deacetylases (HDACs) in human cancer. *Mol. Oncol.*

(41) Maiques-Diaz, A., and Somervaille, T. C. P. (2016) LSD1: Biologic roles and therapeutic targeting. *Epigenomics*.

(42) Licciardi, P. V, Kwa, F. A. A., Ververis, K., et al. (2012) Influence of natural and synthetic histone deacetylase inhibitors on chromatin. *Antioxid. Redox Signal. 17*, 340–54.

(43) Hanigan, T. W., Aboukhatwa, S. M., Taha, T. Y., et al. (2017) Negative Breast Cancer Cells Determines HDAC Inhibitor Binding and Selectivity Article Divergent JNK Phosphorylation of HDAC3 in Triple-Negative Breast Cancer Cells Determines HDAC Inhibitor Binding and Selectivity 1356–1367.

(44) Nakajima, N. I., Niimi, A., Isono, M., et al. (2017) Inhibition of the HDAC/Suv39/G9a pathway restores the expression of DNA damage-dependent major histocompatibility complex class I-related chain A and B in cancer cells. *Oncol. Rep. 38*, 693–702.

(45) Nalawansha, D. A., Zhang, Y., Herath, K., et al. (2018) HDAC1 Substrate Profiling

Using Proteomics-Based Substrate Trapping. *ACS Chem. Biol 13*, 3315–3324.

(46) Maes, T., Mascaro´, C., Tirapu, I., et al. (2018) ORY-1001, a Potent and Selective Covalent KDM1A Inhibitor, for the Treatment of Acute Leukemia. *Cancer Cell 33*, 495–511.

(47) Pishas, K. I., Drenberg, C. D., Taslim, C., et al. (2018) Therapeutic targeting of KDM1A/LSD1 in Ewing sarcoma with SP-2509 engages the endoplasmic reticulum stress response. *Mol. Cancer Ther. 17*.

(48) Burg, J. M., Link, J. E., Morgan, B. S., et al. (2015) KDM1 class flavin-dependent protein lysine demethylases. *Biopolymers 104*, 213–246.

(49) Schmidt, D. M. Z., and McCafferty, D. G. (2007) trans-2-phenylcyclopropylamine is a mechanism-based inactivator of the histone demethylase LSD1. *Biochemistry 46*, 4408–4416.

(50) Vasilatos, S. N., Katz, T. A., Oesterreich, S., et al. (2013) Crosstalk between lysine-specific demethylase 1 (LSD1) and histone deacetylases mediates antineoplastic efficacy of HDAC inhibitors in human breast cancer cells. *Carcinogenesis 34*, 1196–1207.

(51) Kouzarides, T. (2007, February 23) Chromatin Modifications and Their Function. *Cell*. Elsevier.

(52) Li, B., Carey, M., and Workman, J. L. (2007, February 23) The Role of Chromatin during Transcription. *Cell*. Elsevier.

(53) Forneris, F., Binda, C., Adamo, A., et al. (2007) Structural basis of LSD1-CoREST selectivity in histone H3 recognition. *J. Biol. Chem. 282*, 20070–20074.

(54) Yang, M., Culhane, J. C., Szewczuk, L. M., et al. (2007) Structural basis of histone demethylation by LSD1 revealed by suicide inactivation. *Nat. Struct. Mol. Biol. 14*, 535–539.

(55) Nicholson, T. B., and Chen, T. (2009) LSD1 demethylates histone and non-histone proteins. *Epigenetics 4*, 129–132.

(56) Yang, M., Culhane, J. C., Szewczuk, L. M., et al. (2007) Structural basis for the inhibition of the LSD1 histone demethylase by the antidepressant trans-2-phenylcyclopropylamine. *Biochemistry 46*, 8058–8065.

(57) Carnesecchi, J., Forcet, C., Zhang, L., et al. (2017) ERRα induces H3K9 demethylation by LSD1 to promote cell invasion. *Proc. Natl. Acad. Sci. 114*, 3909–3914.

(58) Wissmann, M., Yin, N., Müller, J. M., et al. (2007) Cooperative demethylation by JMJD2C and LSD1 promotes androgen receptor-dependent gene expression. *Nat. Cell Biol. 9*, 347–353.

(59) Speranzini, V., Ciossani, G., Marabelli, C., et al. (2017) Probing the interaction of the p53 C-terminal domain to the histone demethylase LSD1. *Arch. Biochem. Biophys. 632*, 202–208.

(60) Baron, R., Binda, C., Tortorici, M., et al. (2011) Molecular mimicry and ligand recognition in binding and catalysis by the histone demethylase LSD1-CoREST complex. *Structure 19*, 212–220.

(61) Luka, Z., Pakhomova, S., Loukachevitch, L. V, et al. (2014) Crystal structure of the histone lysine specific demethylase LSD1 complexed with tetrahydrofolate. *Protein Sci. 23*, 993–998.

(62) Niwa, H., Sato, S., Hashimoto, T., et al. (2018) Crystal structure of LSD1 in complex with 4-[5-(Piperidin-4-ylmethoxy)-2-(p-tolyl) pyridin-3-yl]benzonitrile. *Molecules 23*, 1538.

(63) Stavropoulos, P., Blobel, G., and Hoelz, A. (2006) Crystal structure and mechanism of human lysine- specific demethylase-1. *Nat. Struct. Mol. Biol. 13*, 626–632.

(64) Gaweska, H., Pozzi, M. H., Schmidt, D. M. Z., et al. (2009) Use of pH and kinetic isotope effects to establish chemistry as rate-limiting in oxidation of a peptide substrate by LSD1. *Biochemistry 48*, 5440–5445.

(65) Forneris, F., Binda, C., Vanoni, M. A., et al. (2005) Histone demethylation catalysed by LSD1 is a flavin-dependent oxidative process. *FEBS Lett. 579*, 2203–2207.

(66) Marmorstein, R., Shiekhattar, R., Cairns, B. R., et al. (2006) Structure and function of the SWIRM domain, a conserved protein module found in chromatin regulatory complexes. *Proc. Natl. Acad. Sci. 103*, 2057–2062.

(67) Binda, C., Valente, S., Romanenghi, M., et al. (2010) Biochemical, structural, and biological evaluation of tranylcypromine derivatives as inhibitors of histone demethylases LSD1 and LSD2. *J. Am. Chem. Soc. 132*, 6827–6833.

(68) Fang, R., Chen, F., Dong, Z., et al. (2013) LSD2/KDM1B and its cofactor NPAC/GLYR1 endow a structural and molecular model for regulation of H3K4 demethylation. *Mol. Cell 49*, 558–570.

(69) Dorosz, J., Hyltoft Kristensen, L., Aduri, N. G., et al. (2019) Molecular architecture of the Jumonji C family histone demethylase KDM5B. *Sci. Rep. 9*.

(70) Li, Q., Shi, L., Gui, B., et al. (2011) Binding of the JmjC demethylase JARID1B to LSD1/NuRD suppresses angiogenesis and metastasis in breast cancer cells by repressing chemokine CCL14. *Cancer Res. 71*, 6899–6908.

(71) Yang, M., Gocke, C. B., Luo, X., et al. (2006) Structural Basis for CoREST-Dependent Demethylation of Nucleosomes by the Human LSD1 Histone Demethylase. *Mol. Cell 23*, 377–387.

(72) Wang, Y., Zhang, H., Chen, Y., et al. (2009) LSD1 Is a Subunit of the NuRD Complex and Targets the Metastasis Programs in Breast Cancer. *Cell 138*, 660–672.

(73) You, A., Tong, J. K., Grozinger, C. M., et al. (2001) CoREST is an integral component of the CoREST-human histone deacetylase complex. *Proc. Natl. Acad. Sci. U. S. A. 98*, 1454–1458.

(74) Iwase, S., Januma, A., Miyamoto, K., et al. (2004) Characterization of BHC80 in BRAF-HDAC complex, involved in neuron-specific gene repression. *Biochem. Biophys. Res. Commun. 322*, 601–608.

(75) Adamo, A., Sesé, B., Boue, S., et al. (2011) LSD1 regulates the balance between self-renewal and differentiation in human embryonic stem cells. *Nat. Cell Biol. 13*, 652–661.

(76) Jastrow, H., Vandesompele, J., Schulte, J. H., et al. (2012) Lysine-specific demethylase 1 restricts hematopoietic progenitor proliferation and is essential for terminal differentiation. *Leukemia 26*, 2039–2051.

(77) Whyte, W. A., Hoke, H. A., Young, R. A., et al. (2012) Enhancer decommissioning by LSD1 during embryonic stem cell differentiation. *Nature 482*, 221–225.

(78) Butler, J. S., and Dent, S. Y. R. (2013, April 18) The role of chromatin modifiers in normal and malignant hematopoiesis. *Blood*. American Society of Hematology.

(79) Voigt, P., Tee, W. W., and Reinberg, D. (2013, June 15) A double take on bivalent promoters. *Genes Dev.* Cold Spring Harbor Laboratory Press.

(80) Puri, D., Gala, H., Mishra, R., et al. (2015) High-wire act: The poised genome and cellular memory. *FEBS J. 282*, 1675–1691.

(81) Kerenyi, M. A., Shao, Z., Hsu, Y. J., et al. (2013) Histone demethylase Lsd1 represses hematopoietic stem and progenitor cell signatures during blood cell maturation. *Elife 2013*, e00633.

(82) Upadhyay, G., Chowdhury, A. H., Vaidyanathan, B., et al. (2014) Antagonistic actions of Rcor proteins regulate LSD1 activity and cellular differentiation. *Proc. Natl. Acad. Sci. 111*, 8071–8076.

(83) Jin, C., Merkurjev, D., Skola, D., et al. (2017) Histone demethylase LSD1 regulates hematopoietic stem cells homeostasis and protects from death by endotoxic shock. *Proc. Natl. Acad. Sci. 115*, E244–E252.

(84) Wei, X., Calvo-Vidal, M. N., Chen, S., et al. (2018) Germline lysine-specific demethylase 1 (lsd1/kdm1a) mutations confer susceptibility to multiple myeloma. *Cancer Res. 78*, 2747–2759.

(85) Magliulo, D., Bernardi, R., and Messina, S. (2018) Lysine-Specific Demethylase 1A as a Promising Target in Acute Myeloid Leukemia. *Front. Oncol. 8*.

(86) Liang, S., Liu, D., Chen, H., et al. (2013) Lysine-specific Demethylase 1 Represses THP-1 Monocyte-to-macrophage Differentiation. *Chinese Med. Sci. J. 28*, 82–87.

(87) Craene, B. De, and Berx, G. (2013) Regulatory networks defining EMT during cancer initiation and progression. *Nat. Rev. Cancer*.

(88) Ambrosio, S., Saccà, C. D., and Majello, B. (2017) Epigenetic regulation of epithelial to mesenchymal transition by the Lysine-specific demethylase LSD1/KDM1A. *Biochim. Biophys. Acta - Gene Regul. Mech. 1860*, 905–910.

(89) Lin, T., Ponn, A., Hu, X., et al. (2010) Requirement of the histone demethylase LSD1 in Snai1-mediated transcriptional repression during epithelial-mesenchymal transition. *Oncogene 29*, 4896–4904.

(90) Peinado, H., Olmeda, D., and Cano, A. (2007) Snail, ZEB and bHLH factors in tumour progression: An alliance against the epithelial phenotype? *Nat. Rev. Cancer*.

(91) Ford, M., Hu, C. Y., Weiss, S. J., et al. (2012) Canonical Wnt signaling regulates Slug activity and links epithelial-mesenchymal transition with epigenetic Breast Cancer 1, Early Onset (BRCA1) repression. *Proc. Natl. Acad. Sci. 109*, 16654–16659.

(92) Nishikawa, J. L., Ji, J.-Y., Mostoslavsky, R., et al. (2011) A SIRT1-LSD1 Corepressor Complex Regulates Notch Target Gene Expression and Development. *Mol. Cell 42*, 689–699.

(93) Alsaqer, S., Tashkandi, M., Kartha, V., et al. (2017) Pharmacological inhibition of LSD1 regulates oral cancer growth and metastasis. *Oncotarget 8*, Submitted.

(94) Wang, B., Zhao, B., Pang, L. P., et al. (2017) LPE-1, an orally active pyrimidine derivative, inhibits growth and mobility of human esophageal cancers by targeting LSD1. *Pharmacol. Res. 122*, 66–77.

(95) Takagi, S., Ishikawa, Y., Mizutani, A., et al. (2017) LSD1 inhibitor T-3775440 inhibits SCLC cell proliferation by disrupting LSD1 interactions with SNAG domain proteins INSM1 and GFI1B. *Cancer Res. 77*, 4652–4662.

(96) Kumar, M., Hemalatha, R., Kumar, R., et al. (2013) Epigenetics, Probiotic Metabolites and Colon Cancer Prevention: An Overview of Progress, Opportunities and Challenges. *Med. Epigenetics 1*, 60–69.

(97) Ambrosio, S., Saccà, C. D., Amente, S., et al. (2017) Lysine-specific demethylase

LSD1 regulates autophagy in neuroblastoma through SESN2-dependent pathway. *Oncogene 36*, 6701–6711.

(98) Schulte, J. H., Lim, S., Schramm, A., et al. (2009) Lysine-specific demethylase 1 is strongly expressed in poorly differentiated neuroblastoma: Implications for therapy. *Cancer Res.*

(99) Han, X., Gui, B., Xiong, C., et al. (2014) Destabilizing LSD1 by Jade-2 promotes neurogenesis: An antibraking system in neural development. *Mol. Cell*.

(100) Majello, B., Milazzo, G., Sorrentino, M. C., et al. (2015) Lysine-specific demethylase (LSD1/KDM1A) and MYCN cooperatively repress tumor suppressor genes in neuroblastoma. *Oncotarget*.

(101) Ding, J., Zhang, Z. M., Xia, Y., et al. (2013) LSD1-mediated epigenetic modification contributes to proliferation and metastasis of colon cancer. *Br. J. Cancer*.

(102) Chen, J., Ding, J., Wang, Z., et al. (2017) Identification of downstream metastasis-associated target genes regulated by LSD1 in colon cancer cells. *Oncotarget*.

(103) Park, B. H., Wang, W., Wu, Y., et al. (2012) Loss of LSD1 (lysine-specific demethylase 1) suppresses growth and alters gene expression of human colon cancer cells in a p53- and DNMT1(DNA methyltransferase 1)-independent manner. *Biochem. J.*

(104) Cusan, M., Cai, S. F., Mohammad, H., et al. (2018) LSD1 inhibition exerts its antileukemic effect by recommissioning PU.1- and C/EBPa-dependent enhancers in AML. *Blood 131*, 1730–1742.

(105) Sugino, N., Kawahara, M., Tatsumi, G., et al. (2017) A novel LSD1 inhibitor NCD38 ameliorates MDS-related leukemia with complex karyotype by attenuating leukemia programs via activating super-enhancers. *Leukemia 31*, 2303–2314.

(106) Fiskus, W., Sharma, S., Shah, B., et al. (2014) Highly effective combination of LSD1 (KDM1A) antagonist and pan-histone deacetylase inhibitor against human AML cells. *Leukemia 28*, 2155–2164.

(107) Zou, Z. K., Huang, Y. Q., Zou, Y., et al. (2017) Silencing of LSD1 gene modulates

histone methylation andacetylation and induces the apoptosis of JeKo-1 and MOLT-4 cells. *Int. J. Mol. Med. 40*, 319–328.

(108) Xu, X., Peng, W., Liu, C., et al. (2018) Flavone-based natural product agents as new lysine-specific demethylase 1 inhibitors exhibiting cytotoxicity against breast cancer cells in vitro. *Bioorganic Med. Chem. 27*, 370–374.

(109) Mazumdar, S., Arendt, L. M., Phillips, S., et al. (2015) CoREST1 promotes tumor formation and tumor stroma interactions in a mouse model of breast cancer. *PLoS One 10*, 1–16.

(110) Zhang, L., Carnesecchi, J., Cerutti, C., et al. (2018) LSD1-ERRα complex requires NRF1 to positively regulate transcription and cell invasion. *Sci. Rep. 8*, 10041.

(111) Regufe da Mota, S., Bailey, S., Strivens, R. A., et al. (2018) LSD1 inhibition attenuates androgen receptor V7 splice variant activation in castration resistant prostate cancer models. *Cancer Cell Int. 18*.

(112) Singh, M. M., Johnson, B., Venkatarayan, A., et al. (2015) Preclinical activity of combined HDAC and KDM1A inhibition in glioblastoma. *Neuro. Oncol. 17*, 1463–1473.

(113) Hakimi, M.-A., Bochar, D. A., Chenoweth, J., et al. (2002) A core-BRAF35 complex containing histone deacetylase mediates repression of neuronal-specific genes. *Proc. Natl. Acad. Sci. 99*, 7420–7425.

(114) Ray, S. K., Li, H. J., Metzger, E., et al. (2014) CtBP and Associated LSD1 Are Required for Transcriptional Activation by NeuroD1 in Gastrointestinal Endocrine Cells. *Mol. Cell. Biol. 34*, 2308–2317.

(115) Barrios, A. P., Gomez, A. V., Saez, J. E., et al. (2014) Differential Properties of Transcriptional Complexes Formed by the CoREST Family. *Mol. Cell. Biol. 34*, 2760–2770.

(116) Kim, S. A., Chatterjee, N., Jennings, M. J., et al. (2015) Extranucleosomal DNA enhances the activity of the LSD1/CoREST histone demethylase complex. *Nucleic Acids Res. 43*, 4868–4880.

(117) Hwang, S., Schmitt, A. A., Luteran, A. E., et al. (2011) Thermodynamic characterization of the binding interaction between the histone demethylase LSD1/KDM1 and CoREST. *Biochemistry 50*, 546–557.

(118) Pilotto, S., Speranzini, V., Tortorici, M., et al. (2015) Interplay among nucleosomal DNA, histone tails, and corepressor CoREST underlies LSD1-mediated H3 demethylation. *Proc. Natl. Acad. Sci. 112*, 2752–2757.

(119) Lunyak, V., Burgess, R., Prefontaine, G., et al. (2002) Corepressor-Dependent Silencing of Chromosomal Regions Encoding Neuronal Genes. *Science (80-.). 298*, 1747–1752.

(120) Marmorstein, L. Y., Kinev, A. V, Chan, G. K. T., et al. (2001) A human BRCA2 complex containing a structural DNA binding component influences cell cycle progression. *Cell 104*, 247–257.

(121) Rodriguez-Paredes, M., Garcia-Gutierrez, P., Reyes, J. C., et al. (2012) Control of neuronal differentiation by sumoylation of BRAF35, a subunit of the LSD1-CoREST histone demethylase complex. *Proc. Natl. Acad. Sci. 109*, 8085–8090.

(122) Cowger, J. J. M., Zhao, Q., Isovic, M., et al. (2007) Biochemical characterization of the zinc-finger protein 217 transcriptional repressor complex: Identification of a ZNF217 consensus recognition sequence. *Oncogene 26*, 3378–3386.

(123) Baron, R., and Vellore, N. A. (2012) LSD1/CoREST reversible opening-closing dynamics: Discovery of a nanoscale clamp for chromatin and protein binding. *Biochemistry 51*, 3151–3153.

(124) Gocke, C. B., and Yu, H. (2008) ZNF198 stabilizes the LSD1-CoREST-HDAC1 complex on chromatin through its MYM-type zinc fingers. *PLoS One 3*.

(125) Quinlan, K. G. R., Nardini, M., Verger, A., et al. (2006) Specific Recognition of ZNF217 and Other Zinc Finger Proteins at a Surface Groove of C-Terminal Binding Proteins. *Mol. Cell. Biol. 26*, 8159–8172.

(126) Kwok, A., Crossley, M., Perdomo, J., et al. (2006) Role of the C-Terminal Binding Protein PXDLS Motif Binding Cleft in Protein Interactions and Transcriptional Repression. *Mol. Cell. Biol. 26*, 8202–8213.

(127) Castet, A., Boulahtouf, A., Versini, G., et al. (2004) Multiple domains of the receptor-interacting protein 140 contribute to transcription inhibition. *Nucleic Acids Res. 32*, 1957–1966.

(128) Fernandes, I., Bastien, Y., Wai, T., et al. (2003) Ligand-dependent nuclear receptor corepressor LCoR functions by histone deacetylase-dependent and -independent mechanisms. *Mol. Cell 11*, 139–150.

(129) Ueda, J., Tachibana, M., Ikura, T., et al. (2006) Zinc finger protein Wiz links G9a/GLP histone methyltransferases to the co-repressor molecule CtBP. *J. Biol. Chem. 281*, 20120–20128.

(130) Subramanian, T., Vijayalingam, S., Zhou, Y., et al. (2007) Role of the PLDLS-Binding Cleft Region of CtBP1 in Recruitment of Core and Auxiliary Components of the Corepressor Complex. *Mol. Cell. Biol. 28*, 269–281.

(131) Basta, J., and Rauchman, M. (2017) The Nucleosome Remodeling and Deacetylase Complex in Development and Disease, in *Translating Epigenetics to the Clinic*, pp 37–72. Mosby.

(132) Lin, Y., Wu, Y., Li, J., et al. (2010) The SNAG domain of snail1 functions as a molecular hook for recruiting lysine-specific demethylase 1. *EMBO J. 29*, 1803–1816.

(133) Van Der Meer, L. T., Jansen, J. H., and Van Der Reijden, B. A. (2010) Gfi1 and Gfi1b: key regulators of hematopoiesis. *Leukemia 24*, 1834–1843.

(134) Sakane, N., Kwon, H. S., Pagans, S., et al. (2011) Activation of hiv transcription by the viral tat protein requires a demethylation step mediated by lysine-specific demethylase 1 (LSD1/KDM1). *PLoS Pathog. 7*.

(135) De Colibus, L., and Mattevi, A. (2006, December 1) New frontiers in structural flavoenzymology. *Curr. Opin. Struct. Biol.* Elsevier Current Trends.

(136) Rivero, S., Ceballos-Chávez, M., Bhattacharya, S. S., et al. (2015) HMG20A is required for SNAI1-mediated epithelial to mesenchymal transition. *Oncogene 34*, 5264–5276.

(137) Phelan, J. D., Shroyer, N. F., Cook, T., et al. (2010) Gfi1-Cells & Circuits: Unraveling transcriptional networks of development and disease. *curr opin hematol 17*, 300–307.

(138) Zhang, Y., Chang, W., Hu, Z., et al. (2009) The Growth Factor Independence-1 (Gfi1) Is Overexpressed in Chronic Myelogenous Leukemia. *Acta Haematol. 123*, 1–5.

(139) Vassen, L., Khandanpour, C., Ebeling, P., et al. (2009) Growth factor independent 1b (Gfi1b) and a new splice variant of Gfi1b are highly expressed in patients with acute and chronic leukemia. *Int. J. Hematol. 89*, 422–430.

(140) Xu, W., and Kee, B. L. (2007) Growth factor independent 1B (Gfi1b) is an E2A target gene that modulates Gata3 in T-cell lymphomas. *Blood 109*, 4406–4414.

(141) Chowdhury, A. H., Ramroop, J. R., Upadhyay, G., et al. (2013) Differential Transcriptional Regulation of meis1 by Gfi1b and Its Co-Factors LSD1 and CoREST. *PLoS One 8*.

(142) Saleque, S., Kim, J., Rooke, H. M., et al. (2007) Epigenetic Regulation of Hematopoietic Differentiation by Gfi-1 and Gfi-1b Is Mediated by the Cofactors CoREST and LSD1. *Mol. Cell 27*, 562–572.

(143) Millard, C. J., Watson, P. J., Celardo, I., et al. (2013) Class I HDACs share a common mechanism of regulation by inositol phosphates. *Mol. Cell 51*, 57–67.

(144) Smits, A. H., Jansen, P. W. T. C., Poser, I., et al. (2013) Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Res. 41*.

(145) Zhou, Y., Bolton, E. C., and Jones, J. O. (2015, February) Androgens and androgen receptor signaling in prostate tumorigenesis. *J. Mol. Endocrinol.*

(146) Günther, T., Schneider, R., Schüle, R., et al. (2005) LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription. *Nature 437*, 436–439.

(147) Cai, C., He, H. H., Chen, S., et al. (2011) Androgen Receptor Gene Expression in Prostate Cancer Is Directly Suppressed by the Androgen Receptor Through

Recruitment of Lysine-Specific Demethylase 1. *Cancer Cell 20*, 457–471.

(148) Thomas, C., and Gustafsson, J.-Å. (2011) The different roles of ER subtypes in cancer biology and therapy. *Nat. Publ. Gr. 11*, 597–608.

(149) Hervouet, E., Cartron, P. F., Jouvenot, M., et al. (2013, March 27) Epigenetic regulation of estrogen signaling in breast cancer. *Epigenetics*. Taylor & Francis.

(150) Liu, T., Khanna, K. M., Chen, X., et al. (2000) DNA Oxidation as Triggered by H3K9me2 Demethylation Drives Estrogen-Induced Gene Expression. *Science (80-.). 343*, 593.

(151) Garcia-Bassets, I., Kwon, Y. S., Telese, F., et al. (2007) Histone Methylation-Dependent Mechanisms Impose Ligand Dependency for Gene Activation by Nuclear Receptors. *Cell 128*, 505–518.

(152) Zhao, W., Geng, D., Li, S., et al. (2018) LncRNA HOTAIR influences cell growth, migration, invasion, and apoptosis via the miR-20a-5p/HMGA2 axis in breast cancer. *Cancer Med. 7*, 842–855.

(153) Zhang, J., Zhang, P., Wang, L., et al. (2014) Long non-coding RNA HOTAIR in carcinogenesis and metastasis. *Acta Biochim. Biophys. Sin. (Shanghai).*

(154) Gupta, R. A., Shah, N., Wang, K. C., et al. (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature 464*, 1071–1076.

(155) Xiao, Z., Qu, Z., Chen, Z., et al. (2018) LncRNA HOTAIR is a Prognostic Biomarker for the Proliferation and Chemoresistance of Colorectal Cancer via MiR-203a-3p-Mediated Wnt/ß-Catenin Signaling Pathway. *Cell. Physiol. Biochem. 46*, 1275–1285.

(156) Wang, X., Liu, W., Wang, P., et al. (2018) RNA interference of long noncoding RNA HOTAIR suppresses autophagy and promotes apoptosis and sensitivity to cisplatin in oral squamous cell carcinoma. *J. Oral Pathol. Med. 47*, 930–937.

(157) Ma, J., Fan, Y., Feng, T., et al. (2017) HOTAIR regulates HK2 expression by binding endogenous miR- 125 and miR-143 in oesophageal squamous cell carcinoma progression *8*, 86410–86422.

(158) Cheng, D., Deng, J., Zhang, B., et al. (2018) LncRNA HOTAIR epigenetically suppresses miR-122 expression in hepatocellular carcinoma via DNA methylation. *EBioMedicine 36*, 159–170.

(159) Wu, H., Shang, X., Shi, Y., et al. (2016) Genetic variants of lncRNA HOTAIR and risk of epithelial ovarian cancer among Chinese women *7*.

(160) Dong, T., Peng, Y., Song, N., et al. (2019) HOX transcript antisense RNA is elevated in gastric carcinogenesis and regulated by the NF-κB pathway. *J. Cell. Biochem.*

(161) Da, M., Ma, J., Zhang, Y., et al. (2017) High expression level of long non-coding RNA HOTAIR is associated with poor overall survival in gastric cancer patients: evidence from meta-analysis. *J. BUON. 22*, 911–918.

(162) Ma, M. zhe, Li, C. xiao, Zhang, Y., et al. (2014) Long non-coding RNA HOTAIR, a c-Myc activated driver of malignancy, negatively regulates miRNA-130a in gallbladder cancer. *Mol. Cancer 13*.

(163) Tsai, M., Manor, O., Wan, Y., et al. (2010) Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science (80-.). 329*, 689–693.

(164) Ren, Y., Wang, Y., Zhang, J., et al. (2019) Targeted design and identification of AC1NOD4Q to block activity of HOTAIR by abrogating the scaffold interaction with EZH2 1–16.

(165) Özeş, A. R., Wang, Y., Zong, X., et al. (2017) Therapeutic targeting using tumor specific peptides inhibits long non-coding RNA HOTAIR activity in ovarian and breast cancer. *Sci. Rep. 7*, 1–11.

(166) Porro, A., Feuerhahn, S., and Lingner, J. (2014) TERRA-Reinforced Association of LSD1 with MRE11 Promotes Processing of Uncapped Telomeres. *Cell Rep. 6*, 765–776.

(167) Hirschi, A., Martin, W. J., Luka, Z., et al. (2016) G-quadruplex RNA binding and recognition by the lysine-specific histone demethylase-1 enzyme. *RNA 22*, 1250–1260.

(168) Yao, T., Song, L., Jin, J., et al. (2008) Distinct Modes of Regulation of the Uch37

Deubiquitinating Enzyme in the Proteasome and in the Ino80 Chromatin-Remodeling Complex. *Mol. Cell*.

(169) VanderLinden, R. T., Hemmis, C. W., Schmitt, B., et al. (2015) Structural Basis for the Activation and Inhibition of the UCH37 Deubiquitylase. *Mol. Cell 57*, 901–911.

(170) Qian, C., and Zhou, M. M. (2006) SET domain protein lysine methyltransferases: Structure, specificity and catalysis. *Cell. Mol. Life Sci. 63*, 2755–2763.

(171) Xiao, B., Jing, C., Kelly, G., et al. (2005) Specificity and mechanism of the histone methyltransferase Pr-Set7. *Genes Dev. 19*, 1444–1454.

(172) Ketel, C. S., Andersen, E. F., Vargas, M. L., et al. (2005) Subunit Contributions to Histone Methyltransferase Activities of Fly and Worm Polycomb Group Complexes. *Mol. Cell. Biol. 25*, 6857–6868.

(173) Moritz, L. E., and Trievel, R. C. (2017) Structure, mechanism, and regulation of polycomb repressive complex 2. *J. Biol. Chem.*

(174) Chen, X., and Walters, K. J. (2015) Structural plasticity allows UCH37 to be primed by RPN13 or locked down by INO80G. *Mol. Cell 57*, 767–768.

(175) Hobert, O., Jallal, B., and Ullrich, A. (1996) Interaction of Vav with ENX-1, a putative transcriptional regulator of homeobox gene expression. *Mol. Cell. Biol. 16*, 3066–3073.

(176) Laible, G., Wolf, A., Dorn, R., et al. (1997) Mammalian homologues of the Polycomb-group gene Enhancer of zeste mediate gene silencing in Drosophila heterochromatin and at S. cerevisiae telomeres. *EMBO J. 16*, 3219–3232.

(177) Tschiersch, B., Hofmann, A., Krauss, V., et al. (1994) The protein encoded by the Drosophila position-effect variegation suppressor gene Su(var)3-9 combines domains of antagonistic regulators of homeotic gene complexes. *EMBO J. 13*, 3822–3831.

(178) Cao, R., and Zhang, Y. (2004) The functions of E(Z)/EZH2-mediated methylation of lysine 27 in histone H3. *Curr. Opin. Genet. Dev.*

(179) Cao, R., Wang, L., Wang, H., et al. (2002) Role of histone H3 lysine 27 methylation in polycomb-group silencing. *Science (80-.). 298*, 1039–1043.

(180) Pasini, D., Bracken, A. P., Jensen, M. R., et al. (2004) Suz12 is essential for mouse development and for EZH2 histone methyltransferase activity. *EMBO J. 23*, 4061–4071.

(181) Cao, R., and Zhang, Y. (2004) SUZ12 is required for both the histone methyltransferase activity and the silencing function of the EED-EZH2 complex. *Mol. Cell 15*, 57–67.

(182) Kirmizis, A., Bartley, S. M., Kuzmichev, A., et al. (2004) Silencing of human polycomb target genes is associated with methylation of histone H3 Lys 27. *Genes Dev. 18*, 1592–1605.

(183) Koontz, J. I., Soreng, a L., Nucci, M., et al. (2001) Frequent fusion of the JAZF1 and JJAZ1 genes in endometrial stromal tumors. *Proc. Natl. Acad. Sci. U. S. A. 98*, 6348–53.

(184) Weinmann, A. S., Bartley, S. M., Zhang, T., et al. (2001) Use of Chromatin Immunoprecipitation To Clone Novel E2F Target Promoters. *Mol. Cell. Biol. 21*, 6820–6832.

(185) Yamamoto, K., Sonoda, M., Inokuchi, J., et al. (2004) Polycomb Group Suppressor of Zeste 12 Links Heterochromatin Protein 1α and Enhancer of Zeste 2. *J. Biol. Chem. 279*, 401–406.

(186) Sewalt, R. G., van der Vlag, J., Gunster, M. J., et al. (1998) Characterization of interactions between the mammalian polycomb-group proteins Enx1/EZH2 and EED suggests the existence of different mammalian polycomb-group protein complexes. *Mol. Cell. Biol. 18*, 3586–3595.

(187) Margueron, R., Justin, N., Ohno, K., et al. (2009) Role of the polycomb protein EED in the propagation of repressive histone marks. *Nature 461*, 762–767.

(188) Van Der Vlag, J., and Otte, A. P. (1999) Transcriptional repression mediated by the human polycomb-group protein EED involves histone deacetylation. *Nat. Genet. 23*, 474–478.

(189) Montgomery, N. D., Yee, D., Montgomery, S. A., et al. (2007) Molecular and Functional Mapping of EED Motifs Required for PRC2-Dependent Histone Methylation. *J. Mol. Biol. 374*, 1145–1157.

(190) Wu, H., Zeng, H., Dong, A., et al. (2013) Structure of the catalytic domain of EZH2 reveals conformational plasticity in cofactor and substrate binding sites and explains oncogenic mutations. *PLoS One 8*, e83737.

(191) Justin, N., Zhang, Y., Tarricone, C., et al. (2016) Structural basis of oncogenic histone H3K27M inhibition of human polycomb repressive complex 2. *Nat. Commun. 7*, 11316.

(192) Ciferri, C., Lander, G. C., Maiolica, A., et al. (2012) Molecular architecture of human polycomb repressive complex 2. *Elife 2012*, 1–22.

(193) Kasinath, V., Faini, M., Poepsel, S., et al. (2018) Structures of human PRC2 with its cofactors AEBP2 and JARID2. *Science (80-.). 944*, 1–10.

(194) Brooun, A., Gajiwala, K. S., Deng, Y.-L., et al. (2016) Polycomb repressive complex 2 structure with inhibitor reveals a mechanism of activation and drug resistance. *Nat. Commun. 7*, 11384.

(195) Beekman, R., Valkhof, M. G., Sanders, M. A., et al. (2012) Sequential gain of mutations in severe congenital neutropenia progressing to acute myeloid leukemia. *Blood 119*, 5071–5077.

(196) Score, J., Hidalgo-Curtis, C., Jones, A. V., et al. (2012) Inactivation of polycomb repressive complex 2 components in myeloproliferative and myelodysplastic/myeloproliferative neoplasms. *Blood 119*, 1208–1213.

(197) Yang, C. Y., and Wang, S. (2017) Allosteric Inactivation of Polycomb Repressive Complex 2 (PRC2) by Inhibiting Its Adapter Protein: Embryonic Ectodomain Development (EED). *J. Med. Chem. 60*, 2212–2214.

(198) Lingel, A., Sendzik, M., Huang, Y., et al. (2017) Structure-Guided Design of EED Binders Allosterically Inhibiting the Epigenetic Polycomb Repressive Complex 2 (PRC2) Methyltransferase. *J. Med. Chem. 60*, 415–427.

(199) Huang, Y., Zhang, J., Yu, Z., et al. (2017) Discovery of First-in-Class, Potent, and Orally Bioavailable Embryonic Ectoderm Development (EED) Inhibitor with Robust Anticancer Efficacy. *J. Med. Chem. 60*, 2215–2226.

(200) Yao, T., Song, L., Xu, W., et al. (2006) Proteasome recruitment and activation of the Uch37 deubiquitinating enzyme by Adrm1. *Nat. Cell Biol. 8*, 994–1002.

(201) Burgie, S. E., Bingman, C. A., Soni, A. B., et al. (2012) Structural characterization of human Uch37. *Proteins 80*, 649–54.

(202) Yao, T., Song, L., Jin, J., et al. (2008) Distinct Modes of Regulation of the Uch37 Deubiquitinating Enzyme in the Proteasome and in the Ino80 Chromatin-Remodeling Complex. *Mol. Cell 31*, 909–917.

(203) Shi, Y. J., Matson, C., Lan, F., et al. (2005) Regulation of LSD1 histone demethylase activity by its associated factors. *Mol. Cell 19*, 857–864.

(204) Burg, J. M., Makhoul, A. T., Pemble, C. W., et al. (2015) A rationally-designed chimeric KDM1A/KDM1B histone demethylase tower domain deletion mutant retaining enzymatic activity. *FEBS Lett. 589*, 2340–2346.

(205) Vermeulen, M., Eberl, H. C., Matarese, F., et al. (2010) Quantitative Interaction Proteomics and Genome-wide Profiling of Epigenetic Histone Marks and Their Readers. *Cell 142*, 967–980.

(206) Fu, J., Yang, Z., Wei, J., et al. (2006) Nuclear protein NP60 regulates p38 MAPK activity. *J. Cell Sci. 119*, 115–23.

(207) Kim, H., Kang, K., and Kim, J. (2009) AEBP2 as a potential targeting protein for Polycomb Repression Complex PRC2. *Nucleic Acids Res. 37*, 2940–2950.

(208) Li, G., Margueron, R., Ku, M., et al. (2010) Jarid2 and PRC2, partners in regulating gene expression. *Genes Dev. 24*, 368–380.

(209) Pasini, D., Cloos, P. A. C., Walfridsson, J., et al. (2010) JARID2 regulates binding of the Polycomb repressive complex 2 to target genes in ES cells. *Nature 464*, 306–310.

(210) Sanulli, S., Justin, N., Teissandier, A., et al. (2015) Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation. *Mol. Cell 57*, 769–783.

(211) Son, J., Shen, S. S., Margueron, R., et al. (2013) Nucleosome-binding activities within JARID2 and EZH1 regulate the function of PRC2 on chromatin. *Genes Dev. 27*, 2663–2677.

(212) Wang, X., Paucek, R. D., Gooding, A. R., et al. (2017) Molecular analysis of PRC2 recruitment to DNA in chromatin and its inhibition by RNA. *Nat. Struct. Mol. Biol. 24*, 1028–1038.

(213) Zhang, Q., Vo, N., and Goodman, R. H. (2000) Histone binding protein RbAp48 interacts with a complex of CREB binding protein and phosphorylated CREB. *Mol. Cell. Biol. 20*, 4970–8.

(214) Burg, J. M., Gonzalez, J. J., Maksimchuk, K. R., et al. (2016) Lysine-Specific Demethylase 1A (KDM1A/LSD1): Product Recognition and Kinetic Analysis of Full-Length Histones. *Biochemistry 55*, 1652–1662.

(215) Yang, H., Salz, T., Zajac-Kaye, M., et al. (2014) Overexpression of histone deacetylases in cancer cells is controlled by interplay of transcription factors and epigenetic modulators. *FASEB J. 28*, 4265–4279.

(216) Mann, B. S., Johnson, J. R., Cohen, M. H., et al. (2007) FDA Approval Summary: Vorinostat for Treatment of Advanced Primary Cutaneous T-Cell Lymphoma. *Oncologist 12*, 1247–1252.

(217) Millard, C. J., Watson, P. J., Celardo, I., et al. (2013) Class I HDACs share a common mechanism of regulation by inositol phosphates. *Mol Cell 51*, 57–67.

(218) Pflum, M. K. H., Tong, J. K., Lane, W. S., et al. (2001) Histone Deacetylase 1 Phosphorylation Promotes Enzymatic Activity and Complex Formation. *J. Biol. Chem. 276*, 47733–47741.

(219) Lee, H., Rezai-Zadeh, N., and Seto, E. (2004) Negative regulation of histone deacetylase 8 activity by cyclic AMP-dependent protein kinase A. *Mol. Cell. Biol. 24*, 765–73.

(220) Dowling, D. P., Gantt, S. L., Gattis, S. G., et al. (2008) Structural Studies of Human Histone Deacetylase 8 and Its Site-Specific Variants Complexed with Substrate and Inhibitors Structural Studies of Human Histone Deacetylase 8 and Its Site-Specific Variants Complexed with Substrate and Inhibitors †,‡. *Biochemistry 47*, 13554–13563.

(221) Dowling, D. P., Gattis, S. G., Fierke, C. A., et al. (2010) Structures of metal-substituted human histone deacetylase 8 provide mechanistic inferences on biological function. *Biochemistry 49*, 5048–5056.

(222) Wolfson, N. A., Pitcairn, C. A., Sullivan, E. D., et al. (2014) An enzyme-coupled assay measuring acetate production for profiling histone deacetylase specificity. *Anal. Biochem. 456*, 61–69.

(223) Gantt, S. L., Gattis, S. G., and Fierke, C. A. (2006) Catalytic Activity and Inhibition of Human Histone Deacetylase 8 Is Dependent on the Identity of the Active Site Metal Ion †. *Biochemistry 45*, 6170–6178.

(224) Castañeda, C. A., Wolfson, N. A., Leng, K. R., et al. (2017) HDAC8 substrate selectivity is determined by long- and short-range interactions leading to enhanced reactivity for full-length histone substrates compared with peptides. *J. Biol. Chem. 292*, 21568–21577.

(225) Martin, C., Cao, R., and Zhang, Y. (2006) Substrate preferences of the EZH2 histone methyltransferase complex. *J. Biol. Chem. 281*, 8365–8370.

(226) da Rocha, S. T., Boeva, V., Escamilla-Del-Arenal, M., et al. (2014) Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. *Mol. Cell 53*, 301–316.

(227) Wu, L., Murat, P., Matak-Vinkovic, D., et al. (2013) (Supplement) The binding interaction between long non-coding RNA HOTAIR and PRC2 proteins. *Biochemistry 52*, 9519–27.

(228) Lian, Y., Li, Z., Fan, Y., et al. (2017) The lncRNA-HOXA-AS2 / EZH2 / LSD1 oncogene complex promotes cell proliferation in pancreatic cancer. *Am J Transl Res 9*, 5496–5506.

(229) Zheng, W., and Yu, A. (2018) EZH2-mediated suppression of lncRNA-LET promotes cell apoptosis and inhibits the proliferation of post-burn skin fibroblasts. *Int. J. Mol. Med.*

(230) Shi, Y., Sawada, J. I., Sui, G., et al. (2003) Coordinated histone modifications mediated by a CtBP co-repressor complex. *Nature 422*, 735–738.

(231) Chinnadurai, G. (2009) The transcriptional corepressor CtBP: A foe of multiple tumor suppressors. *Cancer Res. 69*, 731–734.

(232) Byun, J. S., and Gardner, K. (2013) C-terminal binding protein: A molecular link between metabolic imbalance and epigenetic regulation in breast cancer. *Int. J. Cell Biol. 2013*.

(233) Boulding, T., McCuaig, R. D., Tan, A., et al. (2018) LSD1 activation promotes inducible EMT programs and modulates the tumour microenvironment in breast cancer. *Sci. Rep. 8*, 1–18.

(234) Borra, M. T., Langer, M. R., Slama, J. T., et al. (2004) Substrate specificity and kinetic mechanism of the Sir2 family of NAD+-dependent histone/protein deacetylases. *Biochemistry 43*, 9877–9887.

(235) Horton, J. R., Upadhyay, A. K., Qi, H. H., et al. (2010) Enzymatic and structural insights for substrate specificity of a family of jumonji histone lysine demethylases. *Nat. Struct. Mol. Biol. 17*, 38–44.

(236) Peng, H., Prokop, J., Karar, J., et al. (2017) Familial and Somatic BAP1 Mutations Inactivate ASXL1/2-Mediated Allosteric Regulation of BAP1 Deubiquitinase by Targeting Multiple Independent Domains. *Cancer Res.* canres.2876.2017.

(237) Misaghi, S., Ottosen, S., Izrael-Tomasevic, A., et al. (2009) Association of C-Terminal Ubiquitin Hydrolase BRCA1-Associated Protein 1 with Cell Cycle Regulator Host Cell Factor 1. *Mol. Cell. Biol. 29*, 2181–2192.

(238) Daou, S., Hammond-Martel, I., Mashtalir, N., et al. (2015) The BAP1/ASXL2 histone H2A deubiquitinase complex regulates cell proliferation and is disrupted in cancer. *J. Biol. Chem. 290*, 28643–28663.

(239) Patnaik, D., Chin, H. G., Esteve, P.-O., et al. (2004) Substrate specificity and kinetic mechanism of mammalian G9a histone H3 methyltransferase. *J. Biol. Chem. 279*, 53248–53258.

(240) Berndsen, C. E., Tsubota, T., Lindner, S. E., et al. (2008) Molecular functions of the histone acetyltransferase chaperone complex Rtt109–Vps75. *Nat. Struct. Mol. Biol. 15*, 948.

(241) Jones, S. E., Olsen, L., and Gajhede, M. (2018) Structural Basis of Histone Demethylase KDM6B Histone 3 Lysine 27 Specificity. *Biochemistry 57*, 585–592.

(242) Dutto, I., Scalera, C., and Prosperi, E. (2017, April 23) CREBBP and p300 lysine acetyl transferases in the DNA damage response. *Cell. Mol. Life Sci.* Springer International Publishing.

(243) Ogryzko, V. V, Schiltz, R. L., Russanova, V., et al. (1996) The Transcriptional Coactivators p300 and CBP Are Histone Acetyltransferases. *Cell 87*, 953–959.

(244) Kuo, Y. M., and Andrews, A. J. (2013) Quantitating the Specificity and Selectivity of Gcn5-Mediated Acetylation of Histone H3. *PLoS One 8*.

(245) Tao, Y., Zhong, C., Zhu, J., et al. (2017) Structural and mechanistic insights into regulation of HBO1 histone acetyltransferase activity by BRPF2. *Nucleic Acids Res. 45*, 5707–5719.

(246) Han, J., Lachance, C., Ricketts, M. D., et al. (2018) The scaffolding protein JADE1 physically links the acetyltransferase subunit HBO1 with its histone H3/H4 substrate. *J. Biol. Chem. 293*, 4498–4509.

(247) Su, Z., Wang, F., Lee, J. H., et al. (2016) Reader domain specificity and lysine demethylase-4 family function. *Nat. Commun. 7*, 1–15.

(248) Yang, Z., Qiu, Q., Chen, W., et al. (2018) Structure of the Arabidopsis JMJ14-H3K4me3 complex provides insight into the substrate specificity of KDM5 subfamily histone demethylases. *Plant Cell 30*, 167–177.

(249) Ho, M.-C., Wilczek, C., Bonanno, J. B., et al. (2013) Structure of the arginine methyltransferase PRMT5-MEP50 reveals a mechanism for substrate specificity.

PLoS One 8, e57008.

(250) Angrand, P. O., Vennin, C., Le Bourhis, X., et al. (2015) The role of long non-coding RNAs in genome formatting and expression. *Front. Genet.*

(251) Somarowthu, S., Legiewicz, M., Chillón, I., et al. (2015) HOTAIR Forms an Intricate and Modular Secondary Structure. *Mol. Cell 58*, 353–361.

(252) Wu, L., Murat, P., Matak-Vinkovic, D., et al. (2013) Binding interactions between long noncoding RNA HOTAIR and PRC2 proteins. *Biochemistry*.

(253) Ren, Y., Wang, Y., Zhang, J., et al. (2019) Targeted design and identification of AC1NOD4Q to block activity of HOTAIR by abrogating the scaffold interaction with EZH2. *Clin. Epigenetics 11*, 1–16.

(254) Portoso, M., Ragazzini, R., Brenčič, Ž., et al. (2017) PRC2 is dispensable for *HOTAIR* -mediated transcriptional repression. *EMBO J. 36*, 981–994.

(255) Zhao, J., Ohsumi, T. K., Kung, J. T., et al. (2010) Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq. *Mol. Cell 40*, 939–953.

(256) Long, Y., Bolanos, B., Gong, L., et al. (2017) Conserved RNA-binding specificity of polycomb repressive complex 2 is achieved by dispersed amino acid patches in EZH2. *Elife 6*, 1–23.

(257) Davidovich, C., Zheng, L., Goodrich, K. J., et al. (2013) Promiscuous RNA binding by Polycomb repressive complex 2. *Nat. Struct. Mol. Biol. 20*, 1250–1257.

(258) Duan, Y.-C., Ma, Y.-C., Qin, W.-P., et al. (2017) Design and synthesis of tranylcypromine derivatives as novel LSD1/HDACs dual inhibitors for cancer treatment. *Eur. J. Med. Chem.*

(259) Abdel-Magid, A. F. (2017) Lysine-Specific Demethylase 1 (LSD1) Inhibitors as Potential Treatment for Different Types of Cancers. *ACS Med. Chem. Lett. 8*, 1134–1135.

(260) Milelli, A., Marchetti, C., Turrini, E., et al. (2018) Novel polyamine-based Histone deacetylases-Lysine demethylase 1 dual binding inhibitors. *Bioorg. Med. Chem. Lett.*

15–18.

(261) Kalin, J. H., Wu, M., Gomez, A. V., et al. (2018) Targeting the CoREST complex with dual histone deacetylase and demethylase inhibitors. *Nat. Commun. 9*.

(262) Inui, K., Zhao, Z., Yuan, J., et al. (2017) Stepwise assembly of functional C-terminal REST/NRSF transcriptional repressor complexes as a drug target. *Protein Sci. 26*, 997–1011.

(263) Humphrey, G. W., Wang, Y., Russanova, V. R., et al. (2001) Stable histone deacetylase (HDAC1/2) complexes distinguished by presence of SANT domain proteins CoREST/kiaa0071 or Mta-L1 Running title: Stable histone deacetylase complexes Downloaded from. *JBC 276*, 6817–6824.

(264) Neissi, A., Shapouri, M. S. A., Najafabadi, M. G., et al. (2013) Using rice flour for purification of maltose binding fusion proteins expressed in Escherichia coli. *Jundishapur J. Microbiol. 6*, 227–229.

(265) Kellermann, O. K., and Ferenci, T. (1982) Maltose-binding protein from Escherichia coli, in *Methods in Enzymology*, pp 459–463. Academic Press.

(266) Behura, S. K., and Severson, D. W. (2013) Codon usage bias: Causative factors, quantification methods and genome-wide patterns: With emphasis on insect genomes. *Biol. Rev. 88*, 49–61.

(267) de Castro, I. J., Amin, H. A., Vinciotti, V., et al. (2017) Network of phosphatases and HDAC complexes at repressed chromatin. *Cell Cycle 16*, 2011–2017.

(268) Halkidou, K., Gaughan, L., Cook, S., et al. (2004) Upregulation and Nuclear Recruitment of HDACl in Hormone Refractory Prostate Cancer. *Prostate 59*, 177–189.

(269) Wang, L., Charroux, B., Kerridge, S., et al. (2008) Atrophin recruits HDAC1/2 and G9a to modify histone H3K9 and to determine cell fates. *EMBO Rep. 9*, 555–562.

(270) Gantt, S. L., Gattis, S. G., and Fierke, C. A. (2006) Catalytic activity and inhibition of human histone deacetylase 8 is dependent on the identity of the active site metal ion. *Biochemistry 45*, 6170–6178.

(271) Liu, X., Chen, Y., and Fierke, C. A. (2017) Inner-Sphere Coordination of Divalent Metal Ion with Nucleobase in Catalytic RNA. *J. Am. Chem. Soc. 139*, 17457–17463.

(272) Kuo, A. J., Song, J., Cheung, P., et al. (2012) The BAH domain of ORC1 links H4K20me2 to DNA replication licensing and Meier-Gorlin syndrome. *Nature 484*, 115–119.

(273) Zhao, D., Zhang, X., Guan, H., et al. (2016) The BAH domain of BAHD1 is a histone H3K27me3 reader. *Protein Cell 7*, 222–226.

(274) Walz, T., Buchberger, J. R., Moazed, D., et al. (2008) Role of the Conserved Sir3-BAH Domain in Nucleosome Binding and Silent Chromatin Assembly. *Mol. Cell 28*, 1015–1028.

(275) Yang, N., and Xu, R. M. (2013, May 27) Structure and function of the BAH domain in chromatin biology. *Crit. Rev. Biochem. Mol. Biol.* Taylor & Francis.

(276) Millard, C. J., Fairall, L., and Schwabe, J. W. R. (2014) Towards an understanding of the structure and function of MTA1. *Cancer Metastasis Rev. 33*, 857–867.

(277) Li, D. Q., Pakala, S. B., Nair, S. S., et al. (2012) Metastasis-associated protein 1/nucleosome remodeling and histone deacetylase complex in cancer. *Cancer Res.*

(278) Harper, S., and Speicher, D. W. (2011) Purification of Proteins Fused to Glutathione S-Transferase. *Methods Mol. Biol. 681*, 259–280.

(279) Wu, M., Hayward, D., Kalin, J. H., et al. (2018) Lysine-14 acetylation of histone h3 in chromatin confers resistance to the deacetylase and demethylase activities of an epigenetic silencing complex. *Elife 7*, e37231.

(280) McCammon, J. A., Baron, R., Mattevi, A., et al. (2011) Molecular Mimicry and Ligand Recognition in Binding and Catalysis by the Histone Demethylase LSD1-CoREST Complex. *Structure 19*, 212–220.

(281) Skjærven, L., Jariwala, S., Yao, X.-Q., et al. (2016) Online interactive analysis of protein structure ensembles with Bio3D-web. *Bioinformatics 32*, 3510–3512.

(282) Long, S. B., Long, M. B., White, R. R., et al. (2008) Crystal structure of an RNA aptamer bound to thrombin. *RNA 14*, 2504–2512.

(283) Rusconi, C. P., Scardino, E., Layzer, J., et al. (2002) RNA aptamers as reversible antagonists of coagulation factor IXa. *Nature 419*, 90–94.

(284) Irani Shemirani, A., Weinstein, Z. B., Zaman, M. H., et al. (2016) Controlling uncertainty in aptamer selection. *Proc. Natl. Acad. Sci. 113*, 12076–12081.

(285) Wang, J., Rudzinski, J. F., Gong, Q., et al. (2012) Influence of Target Concentration and Background Binding on In Vitro Selection of Affinity Reagents. *PLoS One 7*, 43940.

(286) Nitsche, A., Kurth, A., Dunkhorst, A., et al. (2007) One-step selection of Vaccinia virus-binding DNA aptamers by MonoLEX. *BMC Biotechnol. 7*.

(287) Lauridsen, L. H., Shamaileh, H. A., Edwards, S. L., et al. (2012) Rapid one-step selection method for generating nucleic acid aptamers: Development of a DNA Aptamer against α-bungarotoxin. *PLoS One* (Antopolsky, M., Ed.) *7*, e41702.

(288) Mosing, R. K., and Bowser, M. T. (2009) Isolating aptamers using capillary electrophoresis-SELEX (CE-SELEX). *Methods Mol. Biol. 535*, 33–43.

(289) Consortium, T. E. P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature 489*, 57–74.

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

(291) Feingold, E. A., Good, P. J., Guyer, M. S., et al. (2004, October 22) The ENCODE (ENCyclopedia of DNA Elements) Project. *Science (80-.).* American Association for the Advancement of Science.

(292) Jiang, Y., Mo, H., Luo, J., et al. (2018) HOTAIR Is a Potential Novel Biomarker in Patients with Congenital Heart Diseases. *Biomed Res. Int. 2018*, 1–7.

(293) Tsai, M., Manor, O., Wan, Y., et al. (2010) Modification Complexes. *Science (80-.).*

329, 689–693.

(294) Palazzo, A. F., and Lee, E. S. (2015) Non-coding RNA: What is functional and what is junk? *Front. Genet. 5*, 2.

(295) Blobel, G., and Potter, V. R. (2005) Studies on free and membrane-bound ribosomes in rat liver. *J. Mol. Biol. 26*, 279–292.

(296) Waldron, C., and Lacroute, F. (1975) Effect of Growth Rate on the Amounts of Ribosomal and Transfer Ribonucleic Acids in Yeast. *J. Bacteriol.*

(297) Hangauer, M. J., Vaughn, I. W., and McManus, M. T. (2013) Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. *PLoS Genet. 9*, 1003569.

(298) Li, L. J., Leng, R. X., Fan, Y. G., et al. (2017, December) Translation of noncoding RNAs: Focus on lncRNAs, pri-miRNAs, and circRNAs. *Exp. Cell Res.*

(299) Huang, Y., Jia Shen, X., Zou, Q., et al. (2011) Biological functions of microRNAs: a review. *J. Physiol Biochem 67*, 129–139.

(300) Cai, Y., Yu, X., Hu, S., et al. (2009, December 1) A Brief Review on the Mechanisms of miRNA Regulation. *Genomics, Proteomics Bioinforma.* Elsevier.

(301) Hosseinahli, N., Aghapour, M., Duijf, P. H. G., et al. (2018, August 1) Treating cancer with microRNA replacement therapy: A literature review. *J. Cell. Physiol.* John Wiley & Sons, Ltd.

(302) Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell 75*, 843–54.

(303) Chalfie, M., Horvitz, H. R., and Sulston, J. E. (1981) Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell 24*, 59–69.

(304) Ambros, V., and Horvitz, H. R. (1987) The lin-14 locus of Caenorhabditis elegans controls the time of expression of specific postembryonic developmental events. *Genes Dev. 1*, 398–414.

(305) Hong, Y., Lee, R. C., and Ambros, V. (2000) Structure and function analysis of LIN-14, a temporal regulator of postembryonic developmental events in Caenorhabditis elegans. *Mol. Cell. Biol. 20*, 2285–95.

(306) PH, O., and VR, A. (1999) The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initation of translation. *Dev. Biol. 216*, 671–680.

(307) Tijsterman, M., and Plasterk, R. H. A. (2004, April 2) Dicers at RISC: The mechanism of RNAi. *Cell*. Cell Press.

(308) Cai, L., Chen, Q., Fang, S., et al. (2018) MicroRNA-329 inhibits cell proliferation and tumor growth while facilitates apoptosis via negative regulation of KDM1A in gastric cancer. *J. Cell. Biochem. 119*, 3338–3351.

(309) Chen, L., Wang, X., Huang, W., et al. (2017) MicroRNA-137 and its downstream target LSD1 inversely regulate anesthetics-induced neurotoxicity in dorsal root ganglion neurons. *Brain Res. Bull.*

(310) Huang, C., Zhang, X., Cheng, J., et al. (2014) Upregulation of miR-137 protects anesthesia-induced hippocampal neurodegeneration. *Int. J. Clin. Exp. Pathol. 7*, 5000–5007.

(311) Smalheiser, N. R. (2012) The search for endogenous siRNAs in the mammalian brain. *Exp. Neurol. 235*, 455–463.

(312) Watanabe, T., Totoki, Y., Toyoda, A., et al. (2008) Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature 453*, 539–543.

(313) Pratt, A. J., and MacRae, I. J. (2009) The RNA-induced silencing complex: a versatile gene-silencing machine. *J. Biol. Chem. 284*, 17897–901.

(314) Dana, H., Mahmoodi Chalbatani, G., Mahmoodzadeh, H., et al. (2017) Molecular Mechanisms and Biological Functions of siRNA. *Int. J. Biomed. Sci. 13*, 48–57.

(315) Kurzynska-Kokorniak, A., Koralewska, N., Pokornowska, M., et al. (2015) The

many faces of Dicer: The complexity of the mechanisms regulating Dicer gene expression and enzyme activities. *Nucleic Acids Res. 43*, 4365–4380.

(316) Achawanantakun, R., Chen, J., Sun, Y., et al. (2015) LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. *Bioinformatics 31*, 3897–3905.

(317) Verheggen, K., Volders, P. J., Mestdagh, P., et al. (2017) Noncoding after All: Biases in Proteomics Data Do Not Explain Observed Absence of lncRNA Translation Products. *J. Proteome Res. 16*, 2508–2515.

(318) Ji, Z., Song, R., Regev, A., et al. (2015) Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife 4*.

(319) KOTAKE, Y., GOTO, T., NAEMURA, M., et al. (2017) Long Noncoding RNA PANDA Positively Regulates Proliferation of Osteosarcoma Cells. *Anticancer Res. 37*, 81–86.

(320) Ceribelli, M., Dolfini, D., Merico, D., et al. (2008) The Histone-Like NF-Y Is a Bifunctional Transcription Factor. *Mol. Cell. Biol. 28*, 2047–2058.

(321) Kim, C. G., and Sheffery, M. (1990) Physical characterization of the Purified CCAAT Transcription Factor, a-CP1*. *J. Biol. Chem. 265*, 13362–13369.

(322) Karsenty, G., Golumbek, P., and De Crombrugghe, B. (1988) Point Mutations and Small Substitution Mutations in Three Different Upstream Elements Inhibit the Activity of the Mouse a2(I) Collagen Promoter*. *J. Biol. Chem. 263*, 13909–13915.

(323) Dorn, A., Bollekens, J., Staub, A., et al. (1987) A multiplicity of CCAAT box-binding proteins. *Cell 50*, 863–872.

(324) Mantovani, R. (1998) A survey of 178 NF-Y binding CCAAT boxes CCAAT BOXES AND CCAAT BOX BINDING PROTEINS. *Nucleic Acids Res. 26*, 1135–1143.

(325) Vogels, R., Charité, J., de Graaff, W., et al. (1993) Proximal cis-acting elements cooperate to set Hoxb-7 (Hox-2.3) expression boundaries in transgenic mice. *Development 118*, 71–82.

(326) Wiebe, M. S., Wilder, P. J., Kelly, D., et al. (2000) Isolation, characterization, and

differential expression of the murine Sox-2 promoter. *Gene 246*, 383–393.

(327) Niimi, T., Hayashi, Y., Futaki, S., et al. (2004) SOX7 and SOX17 regulate the parietal endoderm-specific enhancer activity of mouse laminin α1 gene. *J. Biol. Chem. 279*, 38055–38061.

(328) Kao, C. Y., Tanimoto, A., Arima, N., et al. (1999) Transactivation of the human cdc2 promoter by adenovirus E1A. E1A induces the expression and assembly of a heteromeric complex consisting of the CCAAT box binding factor, CBF/NF-Y, and a 110-kDa DNA-binding protein. *J. Biol. Chem. 274*, 23043–23051.

(329) Manni, I., Mazzaro, G., Gurtner, A., et al. (2001) NF-Y Mediates the Transcriptional Inhibition of the cyclin B1, cyclin B2, and cdc25C Promoters upon Induced G2Arrest. *J. Biol. Chem. 276*, 5570–5576.

(330) Hu, Q., Bhattacharya, C., and Maity, S. N. (2002) CCAAT binding factor (CBF) binding mediates cell cycle activation of topoisomerase IIα: Conventional CBF activation domains are not required. *J. Biol. Chem. 277*, 37191–37200.

(331) Hung, T., Wang, Y., Lin, M. F., et al. (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters, in *Nature Genetics*, pp 621–629.

(332) Puvvula, P. K., Desetty, R. D., Pineau, P., et al. (2014) Long noncoding RNA PANDA and scaffold-attachment-factor SAFA control senescence entry and exit. *Nat. Commun. 5*, 5323.

(333) Peng, C., Hu, W., Weng, X., et al. (2017) Over Expression of Long Non-Coding RNA PANDA Promotes Hepatocellular Carcinoma by Inhibiting Senescence Associated Inflammatory Factor IL8. *Sci. Rep. 7*.

(334) Mourtada-Maarabouni, M., Pickard, M. R., Hedge, V. L., et al. (2009) GAS5, a non-protein-coding RNA, controls apoptosis and is downregulated in breast cancer. *Oncogene 28*, 195–208.

(335) Kino, T., Hurt, D. E., Ichijo, T., et al. (2010) Noncoding RNA Gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci. Signal. 3*, ra8.

(336) Lipfert, J., Skinner, G. M., Keegstra, J. M., et al. (2014) Double-stranded RNA under force and torque: Similarities to and striking differences from double-stranded DNA. *Proc. Natl. Acad. Sci. 111*, 15408–15413.

(337) Wang, M., Guo, C., Wang, L., et al. (2018) Long noncoding RNA GAS5 promotes bladder cancer cells apoptosis through inhibiting EZH2 transcription. *Cell Death Dis. 9*, 238.

(338) Teplova, M., Song, J., Gaw, H. Y., et al. (2010) Structural Insights into RNA Recognition by the Alternate-Splicing Regulator CUG-Binding Protein 1. *Structure 18*, 1364–1377.

(339) Romero-Barrios, N., Legascue, M. F., Benhamed, M., et al. (2018) Splicing regulation by long noncoding RNAs. *Nucleic Acids Res. 46*, 2169–2184.

(340) Laurent, B., Ruitu, L., Murn, J., et al. (2015) A Specific LSD1/KDM1A Isoform Regulates Neuronal Differentiation through H3K9 Demethylation. *Mol. Cell 57*, 957–970.

(341) Yamazaki, T., Souquere, S., Chujo, T., et al. (2018) Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. *Mol. Cell 70*, 1038–1053.e7.

(342) Nakagawa, S., Yamazaki, T., and Hirose, T. (2018) Molecular dissection of nuclear paraspeckles: Towards understanding the emerging world of the RNP milieu. *Open Biol.*

(343) Naganuma, T., Nakagawa, S., Tanigawa, A., et al. (2012) Alternative 3′-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J. 31*, 4020–4034.

(344) Cooper, D. R., Carter, G., Li, P., et al. (2014) Long non-coding RNA NEAT1 associates with SRp40 to temporally regulate PPARγ2 splicing during adipogenesis in 3T3-L1 cells. *Genes (Basel). 5*, 1050–1063.

(345) Barry, G., Briggs, J. A., Vanichkina, D. P., et al. (2014) The long non-coding RNA Gomafu is acutely regulated in response to neuronal activation and involved in schizophrenia-associated alternative splicing. *Mol. Psychiatry 19*, 486–494.

(346) Malakar, P., Shilo, A., Mogilevsky, A., et al. (2017) Long noncoding RNA MALAT1 promotes hepatocellular carcinoma development by SRSF1 upregulation and mTOR activation. *Cancer Res. 77*, 1155–1167.

(347) Kong, J., Sun, W., Li, C., et al. (2016) Long non-coding RNA LINC01133 inhibits epithelial–mesenchymal transition and metastasis in colorectal cancer by interacting with SRSF6. *Cancer Lett. 380*, 476–484.

(348) Wang, X., Sehgal, L., Jain, N., et al. (2016) lncRNA MALAT1 promotes development of mantle cell lymphoma by associating with ezh2. *J. Transl. Med. 14*, 346.

(349) Zang, C., Nie, F., Wang, Q., et al. (2016) Long non-coding RNA LINC01133 represses KLF2, P21 and E-cadherin transcription through binding with EZH2, LSD1 in non small cell lung cancer. *Oncotarget 7*, 11696–11707.

(350) Denzler, R., Agarwal, V., Stefano, J., et al. (2014) Assessing the ceRNA Hypothesis with Quantitative Measurements of miRNA and Target Abundance. *Mol. Cell 54*, 766–776.

(351) Abdollahzadeh, R., Daraei, A., Mansoori, Y., et al. (2018) Competing endogenous RNA (ceRNA) cross talk and language in ceRNA regulatory networks: A new look at hallmarks of breast cancer. *J. Cell. Physiol.* 1–21.

(352) Shuwen, H., Qing, Z., Yan, Z., et al. (2018, March 1) Competitive endogenous RNA in colorectal cancer: A systematic review. *Gene*. Elsevier.

(353) Pan, Y., Li, C., Chen, J., et al. (2016) The Emerging Roles of Long Noncoding RNA ROR (lincRNA-ROR) and its Possible Mechanisms in Human Cancers. *Cell. Physiol. Biochem.* Karger Publishers.

(354) Wang, Y., Xu, Z., Jiang, J., et al. (2013) Endogenous miRNA Sponge lincRNA-RoR Regulates Oct4, Nanog, and Sox2 in Human Embryonic Stem Cell Self-Renewal. *Dev. Cell 25*, 69–80.

(355) Weeks, K. M., Baker, A. R., Wooten, J. S., et al. (2018) Functional classification of long non-coding RNAs by k-mer content. *Nat. Genet. 50*, 1474–1482.

(356) Lappin, T. R. J., Grier, D. G., Thompson, A., et al. (2006, January) HOX genes: Seductive science, mysterious mechanisms. *Ulster Med. J.* Ulster Medical Society.

(357) Duboule, D. (1992, June 1) The vertebrate limb: A model system to study the Hox/hom gene network during development and evolution. *BioEssays*. John Wiley & Sons, Ltd.

(358) Wilson, R., Ainscough, R., Anderson, K., et al. (1994) 2.2 Mb of contiguous nucleotide sequence from chromosome III of C. elegans. *Nature 368*, 32–38.

(359) Quinonez, S. C., and Innis, J. W. (2014, January 1) Human HOX gene disorders. *Mol. Genet. Metab.* Academic Press.

(360) Grier, D. G., Thompson, A., Kwasniewska, A., et al. (2005, January 1) The pathophysiology of HOX genes and their role in cancer. *J. Pathol.* John Wiley & Sons, Ltd.

(361) Abate-Shen, C. (2002, October 1) Deregulated homeobox gene expression in cancer: Cause or consequence? *Nat. Rev. Cancer*. Nature Publishing Group.

(362) Wang, K. C., and Chang, H. Y. (2012) Molecular mechanisms of long noncoding RNAs. *Mol Cell 43*, 904–914.

(363) He, S., Liu, S., and Zhu, H. (2011) The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evol. Biol. 11*, 102.

(364) Schorderet, P., and Duboule, D. (2011) Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS Genet.* (Bickmore, W. A., Ed.) *7*, e1002071.

(365) Schorderet, P., and Duboule, D. (2011) Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS Genet. 7*, 1002071.

(366) Amândio, A. R., Necsulea, A., Joye, E., et al. (2016) Hotair Is Dispensible for Mouse Development. *PLoS Genet. 12*.

(367) Pan, Y., Wu, Y., Hu, J., et al. (2018) Long noncoding RNA HOTAIR promotes renal

cell carcinoma malignancy through alpha-2, 8-sialyltransferase 4 by sponging microRNA-124. *Cell Prolif.* e12507.

(368) Li, H., Tang, X.-M., Liu, Y., et al. (2017) Association of Functional Genetic Variants of HOTAIR with Hepatocellular Carcinoma (HCC) Susceptibility in a Chinese Population. *Cell. Physiol. Biochem. 44*, 447–454.

(369) Yang, T., He, X., Chen, A., et al. (2018) LncRNA HOTAIR contributes to the malignancy of hepatocellular carcinoma by enhancing epithelial-mesenchymal transition via sponging miR-23b-3p from ZEB1. *Gene 670*, 114–122.

(370) Wu, L., Zhang, L., and Zheng, S. (2017) Role of the long non-coding RNA HOTAIR in hepatocellular carcinoma (review). *Oncol. Lett. 14*, 1233–1239.

(371) Xavier-Magalhães, A., Oliveira, A. I., de Castro, J. V., et al. (2017) Effects of the functional HOTAIR rs920778 and rs12826786 genetic variants in glioma susceptibility and patient prognosis. *J. Neurooncol. 132*, 27–34.

(372) Liu, L., Cui, S., Wan, T., et al. (2018) Long non-coding RNA HOTAIR acts as a competing endogenous RNA to promote glioma progression by sponging miR-126-5p. *J. Cell. Physiol. 233*, 6822–6831.

(373) Lu, M., Liao, Y., Chen, P., et al. (2017) Targeting LncRNA HOTAIR suppresses cancer stemness and metastasis in oral carcinomas stem cells through modulation of *8*, 98542–98552.

(374) Xiang, S. T., Zou, P. L., Tang, Q., et al. (2018) HOTAIR-mediated reciprocal regulation of EZH2 and DNMT1 contribute to polyphyllin I-inhibited growth of castration-resistant prostate cancer cells in vitro and in vivo. *Biochim. Biophys. Acta - Gen. Subj.*

(375) Chiyomaru, T., Yamamura, S., Fukuhara, S., et al. (2013) Genistein Inhibits Prostate Cancer Cell Growth by Targeting miR-34a and Oncogenic HOTAIR. *PLoS One 8*, 70372.

(376) Tao, S., He, H., and Chen, Q. (2015) Estradiol induces HOTAIR levels via GPER-mediated miR-148a inhibition in breast cancer. *J. Transl. Med. 13*.

(377) Huang, X., and Lu, S. (2017) MicroR-545 mediates colorectal cancer cells proliferation through up-regulating epidermal growth factor receptor expression in HOTAIR long non-coding RNA dependent. *Mol. Cell. Biochem. 431*, 45–54.

(378) Wang, L., Dong, P., Wang, W., et al. (2017) Gemcitabine treatment causes resistance and malignancy of pancreatic cancer stem-like cells via induction of lncRNA HOTAIR. *Exp. Ther. Med. 14*, 4773–4780.

(379) Zhang, Y., Zhou, Y., Xu, T., et al. (2018) Clinical Value of Long Noncoding RNA HOTAIR as a Novel Biomarker in Digestive Cancers: A Meta-Analysis. *Technol. Cancer Res. Treat. 17*, 1–9.

(380) Zhang, M., Song, Y., and Zhai, F. (2018) ARFHPV E7 oncogene, lncRNA HOTAIR, miR-331-3p and its target, NRP2, form a negative feedback loop to regulate the apoptosis in the tumorigenesis in HPV positive cervical cancer. *J. Cell. Biochem.*

(381) Qiu, H., Liu, Q., Li, J., et al. (2016) Analysis of the association of HOTAIR single nucleotide polymorphism (rs920778) and risk of cervical cancer. *Apmis 124*, 567–573.

(382) Kim, H. J., Lee, D. W., Yim, G. W., et al. (2015) Long non-coding RNA HOTAIR is associated with human cervical cancer progression. *Int. J. Oncol. 46*, 521–530.

(383) Zhang, Y., Cheng, X., Liang, H., et al. (2018) Chemico-Biological Interactions Long non-coding RNA HOTAIR and STAT3 synergistically regulate the cervical cancer cell migration and invasion. *Chem. Biol. Interact. 286*, 106–110.

(384) Zhang, Y., Cheng, X., Liang, H., et al. (2018) Long non-coding RNA HOTAIR and STAT3 synergistically regulate the cervical cancer cell migration and invasion. *Chem. Biol. Interact. 286*, 106–110.

(385) Bayram, S., Ülger, Y., Sümbül, A. T., et al. (2015) A functional HOTAIR rs920778 polymorphism does not contributes to gastric cancer in a Turkish population: a case–control study. *Fam. Cancer 14*, 561–567.

(386) Pan, W., Liu, L., Wei, J., et al. (2016) A functional lncRNA HOTAIR genetic variant contributes to gastric cancer susceptibility. *Mol. Carcinog. 55*, 90–96.

(387) Yang, G., Fu, Y., Lu, X., et al. (2018, October 1) LncRNA HOTAIR/miR-613/c-met axis modulated epithelial-mesenchymal transition of retinoblastoma cells. *J. Cell. Mol. Med.* John Wiley & Sons, Ltd (10.1111).

(388) Zhang, X., Zhou, L., Fu, G., et al. (2014) The identification of an ESCC susceptibility SNP rs920778 that regulates the expression of lncRNA HOTAIR via a novel intronic enhancer. *Carcinogenesis 35*, 2062–2067.

(389) Cifuentes-Rojas, C., Hernandez, A. J., Sarma, K., et al. (2014) Regulatory Interactions between RNA and Polycomb Repressive Complex 2. *Mol. Cell 55*, 171–185.

(390) Betancur, J. G., and Tomari, Y. (2015) Cryptic RNA-binding by PRC2 components EZH2 and SUZ12. *RNA Biol. 12*, 959–965.

(391) Saldaña-Meyer, R., Reinberg, D., Umezawa, A., et al. (2014) Interactions between JARID2 and Noncoding RNAs Regulate PRC2 Recruitment to Chromatin. *Mol. Cell 53*, 290–300.

(392) Li, X., Kazan, H., Lipshitz, H. D., et al. (2014) Finding the target sites of RNA-binding proteins. *Wiley Interdiscip. Rev. RNA 5*, 111–130.

(393) Fernandez, M., Kumagai, Y., Standley, D. M., et al. (2011) Prediction of dinucleotide-specific RNA-binding sites in proteins. *BMC Bioinformatics 12*, S5.

(394) El-Manzalawy, Y., Caragea, C., Terribilini, M., et al. (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics 13*.

(395) Yi, P., Wang, Z., Feng, Q., et al. (2015) Structure of a Biologically Active Estrogen Receptor-Coactivator Complex on DNA. *Mol. Cell 57*, 1047–1058.

(396) Iwakiri, J., Tateishi, H., Chakraborty, A., et al. (2012) Dissecting the protein-RNA interface: The role of protein surface shapes and RNA secondary structures in protein-RNA recognition. *Nucleic Acids Res. 40*, 3299–3306.

(397) Wang, X., Vukovic, L., Koh, H. R., et al. (2015) Dynamic profiling of double-stranded RNA binding proteins. *Nucleic Acids Res. 43*, 7566–7576.

(398) Leulliot, N., Quevillon-Cheruel, S., Graille, M., et al. (2004) A new α-helical extension promotes RNA binding by the dsRBD of Rnt1p RNAse III. *EMBO J. 23*, 2468–2477.

(399) Barraud, P., Emmerth, S., Shimada, Y., et al. (2011) An extended dsRBD with a novel zinc-binding motif mediates nuclear retention of fission yeast Dicer. *EMBO J. 30*, 4223–4235.

(400) Vukovic, L., Koh, H. R., Myong, S., et al. (2014) Substrate recognition and specificity of double-stranded RNA binding proteins. *Biochemistry 53*, 3457–3466.

(401) Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006) Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): Quantitative RNA structure analysis at single nucleotide resolution. *Nat. Protoc.*

(402) Busan, S., and Weeks, K. M. (2018) Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA 24*, 143–148.

(403) Vendeix, F. A. P., Munoz, A. M., and Agris, P. F. (2009) Free energy calculation of modified base-pair formation in explicit solvent: A predictive model. *RNA 15*, 2278–2287.

(404) Vakonakis, I., and LiWang, A. C. (2004) N1⋯N3 Hydrogen Bonds of A:U Base Pairs of RNA Are Stronger than Those of A:T Base Pairs of DNA. *J. Am. Chem. Soc. 126*, 5688–5689.

(405) Acharya, P., Cheruku, P., Chatterjee, S., et al. (2004) Measurement of Nucleobase pKa Values in Model Mononucleotides Shows RNA-RNA Duplexes to Be More Stable than DNA-DNA Duplexes. *J. Am. Chem. Soc. 126*, 2862–2869.

(406) Swart, M., Guerra, C. F., and Bickelhaupt, F. M. (2004) Hydrogen bonds of RNA are stronger than those of DNA, but NMR monitors only presence of methyl substituent in uracil/thymine. *J. Am. Chem. Soc. 126*, 16718–16719.

(407) Riml, C., Glasner, H., Rodgers, M. T., et al. (2015) On the mechanism of RNA phosphodiester backbone cleavage in the absence of solvent. *Nucleic Acids Res. 43*, 5171–5181.

(408) Lipkin, D., Talbert, P. T., and Cohn, M. (1954) The Mechanism of the Alkaline Hydrolysis of Ribonucleic Acids. *J. Am. Chem. Soc. 76*, 2871–2872.

(409) Pyle, A. M. (2002) Metal ions in the structure and function of RNA. *J. Biol. Inorg. Chem. 7*, 679–690.

(410) Sigel, R. K. O., and Pyle, A. M. (2007) Alternative Roles for Metal Ions in Enzyme Catalysis and the Implications for Ribozyme Chemistry. *Chem. Rev 107*, 97–113.

(411) Abouhaidar3, M. G., and Ivanovb, I. G. (1999) Non-Enzymatic RNA Hydrolysis Promoted by the Combined Catalytic Activity of Buffers and Magnesium Ions. *Z. Naturforsch*.

(412) Bashkin, J. K., and Jenkins, L. A. (1994) The Role of Metals in the Hydrolytic Cleavage of DNA and RNA. *Comments Inorg. Chem. 16*, 77–93.

(413) Forconi, M., and Herschlag, D. (2009) Metal Ion-Based RNA Cleavage as a Structural Probe, in *Methods in Enzymology*, pp 91–106. Academic Press.

(414) Klug, A., Brown, R. S., and Dewan, J. C. (1985) Crystallographic and Biochemical Investigation of the Lead(II)-Catalyzed Hydrolysis of Yeast Phenylalanine tRNA. *Biochemistry 24*, 4785–4801.

(415) Williams, A. (1966) The mechanism of action of ribonuclease. *Chem. Commun.* 590–591.

(416) Butzow, J. J., and Eichhorn, G. L. (1965) Interaction of Metal Ions with Nucleic Acids and Related Compounds. XVII. On the Mechanism of Degradation of Polyribonucleotides and Oligoribonucleotides by Zinc (II) Ions*. *Biochemistry 10*, 2019–2027.

(417) Westheimer, F. H. (1968) Pseudo-Rotation in the Hydrolysis of Phosphate Esters. *Acc. Chem. Res. 1*, 70–78.

(418) Holley, R. W., and Lazar, V. A. (1961) Metal Content of " Soluble '' -fraction Ribonucleic Acids. *J. Biol. Chem. 236*, 1446–1448.

(419) Belousoff, M. J., Ung, P., Forsyth, C. M., et al. (2009) New macrocyclic terbium(III) complex for use in RNA footprinting experiments. *J. Am. Chem. Soc. 131*, 1106–1114.

(420) Zagórowska, I., Kuusela, S., and Lönnberg, H. (1998) Metal ion-dependent hydrolysis of RNA phosphodiester bonds within hairpin loops. A comparative kinetic study on chimeric ribo/2′-O-methylribo oligonucleotides. *Nucleic Acids Res. 26*, 3392–3396.

(421) Leonarski, F., D'Ascenzo, L., and Auffinger, P. (2019) Nucleobase carbonyl groups are poor Mg2+ inner-sphere binders but excellent monovalent ion binders-a critical PDB survey. *RNA 25*, 173–192.

(422) Ts'o, P. O. P., Kondo, N. S., Schweizer, M. P., et al. (1969) Studies of the Conformation and Interaction in Dinucleoside Mono- and Diphosphates By Proton Magnetic Resonance. *Biochemistry 8*, 997–1029.

(423) Soliva, R., Luque, F. J., Alhambra, C., et al. (1999) Role of sugar re-puckering in the transition of a and b forms of dna in solution. a molecular dynamics study. *J. Biomol. Struct. Dyn. 17*, 89–99.

(424) Zhang, N., Zhang, S., and Szostak, J. W. (2012) Activated ribonucleotides undergo a sugar pucker switch upon binding to a single-stranded RNA template. *J. Am. Chem. Soc. 134*, 3691–3694.

(425) Saenger, W., Hunter, W. N., and Kennard, O. (1986) DNA conformation is determined by economics in the hydration of phosphate groups. *Nature 324*, 385–388.

(426) Foloppe, N., and MacKerell, A. D. (1998) Conformational Properties of the Deoxyribose and Ribose Moieties of Nucleic Acids: A Quantum Mechanical Study. *J. Phys. Chem. B 102*, 6669–6678.

(427) Darré, L., Ivani, I., Dans, P. D., et al. (2016) Small Details Matter: The 2′-Hydroxyl as a Conformational Switch in RNA. *J. Am. Chem. Soc. 138*, 16355–16363.

(428) Spitale, R. C., Flynn, R. A., Zhang, Q. C., et al. (2015) Structural imprints in vivo decode RNA regulatory mechanisms. *Nature 519*, 486–490.

(429) Mortimer, S. A., and Weeks, K. M. (2009) C2′-endo nucleotides as molecular timers suggested by the folding of an RNA domain. *Proc Natl Acad Sci U S A 106*, 15622–15627.

(430) Ussery, D. W. (2002) DNA Structure: A-, B- and Z-DNA Helix Families, in *Encyclopedia of Life Sciences*.

(431) Holbrook, S. R., and Kim, S.-H. (1997) RNA crystallography. *Biopolymers 44*, 3–21.

(432) Rohs, R., West, S. M., Sosinsky, A., et al. (2009) The role of DNA shape in protein-DNA recognition. *Nature 461*, 1248–1253.

(433) Nekludova, L., and Pabo, C. O. (1994) Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. *Proc Natl Acad Sci USA 91*, 6948–6952.

(434) Tanaka, Y., Fujii, S., Hiroaki, H., et al. (1999) A′-form RNA double helix in the single crystal structure of r(UGAGCUUCGGCUC). *Nucleic Acids Res. 27*, 949–955.

(435) Thapar, R., Denmon, A. P., and Nikonowicz, E. P. (2014) Recognition modes of RNA tetraloops and tetraloop-like motifs by RNA-binding proteins. *Wiley Interdiscip. Rev. RNA*.

(436) Varani, G. (1995) Exceptionally Stable Nucleic Acid Hairpins. *Annu. Rev. Biophys. Biomol. Struct. 24*, 379–404.

(437) Wimberly, B., Varani, G., and Tinoco, I. (1993) The Conformation of Loop E of Eukaryotic 5S Ribosomal RNA. *Biochemistry 32*, 1078–1087.

(438) Szewczak, A. A., Moore, P. B., Chang, Y. L., et al. (1993) The conformation of the sarcin/ricin loop from 28S ribosomal RNA. *Proc. Natl. Acad. Sci. 90*, 9581–9585.

(439) Endo, Y., Glück, A., and Wool, I. G. (1991) Ribosomal RNA identity elements for ricin A-chain recognition and catalysis. *J. Mol. Biol. 221*, 193–207.

(440) Larsen, N., and Zwieb, C. (1991) SRP-RNA sequence alignment and secondary structure. *Nucleic Acids Res. 19*, 209–215.

(441) Zuker, M., and Sankoff, D. (1984) Rna Secondary Prediction and Their Prediction. *Bull. Math. Biol. 46*, 591–621.

(442) Condon, D. E., Kennedy, S. D., Mort, B. C., et al. (2015) Stacking in RNA: NMR of four tetramers benchmark molecular dynamics. *J. Chem. Theory Comput. 11*, 2729–2742.

(443) Jaeger, J. A., and Tinoco, I. (1993) An NMR Study of the HIV-1 TAR Element Hairpin. *Biochemistry 32*, 12522–12530.

(444) Fürtig, B., Richter, C., Wöhnert, J., et al. (2003) NMR spectroscopy of RNA. *ChemBioChem 4*, 936–962.

(445) Marchanka, A., Simon, B., Althoff-Ospelt, G., et al. (2015) RNA structure determination by solid-state NMR spectroscopy. *Nat. Commun. 6*.

(446) Scott, L. G., and Hennig, M. (2008) RNA structure determination by NMR. *Methods Mol. Biol. 452*, 29–61.

(447) Wärmländer, S., Sen, A., and Leijon, M. (2000) Imino proton exchange in DNA catalyzed by ammonia and trimethylamine: Evidence for a secondary long-lived open state of the base pair. *Biochemistry 39*, 607–615.

(448) Snoussi, K., and Leroy, J. L. (2001) Imino proton exchange and base-pair kinetics in RNA duplexes. *Biochemistry 40*, 8898–8904.

(449) Robertus, J. D., Ladner, J. E., Finch, J. T., et al. (1974) Structure of yeast phenylalanine tRNA at 3 A resolution. *Nature 250*, 546–551.

(450) Yonuschot, G., Helman, D., Mushrush, G., et al. (1978) Terbium as a solid-state probe for RNA. *Bioinorg. Chem. 8*, 405–418.

(451) Harris, D. A., Todd, G. C., and Walter, N. G. (2014) Terbium(III) Footprinting as a Probe of RNA Structure and Metal Binding Sites, in *Handbook of RNA Biochemistry: Second, Completely Revised and Enlarged Edition*, pp 255–268.

(452) Hargittai, M. R., Hargittai, M. R. S., and Musier-forsyth, K. (2000) Use of terbium as a probe of tRNA tertiary structure and folding Use of terbium as a probe of tRNA

tertiary structure and folding 1672–1680.

(453) Tijerina, P., Mohr, S., and Russell, R. (2007) Dms footprinting of structured rnas and rna-protein complexes. *Nat. Protoc. 2*, 2608–2623.

(454) Peattie, D. A., and Gilbert, W. (1980) Chemical probes for higher-order structure in RNA. *Proc. Natl. Acad. Sci. 77*, 4679–4682.

(455) Brookes, P., and Lawley, P. (1961) The reaction of mono- and di-functional alkylating agents with nucleic acids. *Biochem. J. 80*, 496–503.

(456) Lawley, P. ., and Brookes, P. (1957) The Alkylation of Guanosine and Guanylic Acid. *Proc. Chem. SOC 228*, 478.

(457) Lamoureux, G., and Agüero, C. (2009) A comparison of several modern alkylating agents. *Arkivoc 2009*, 251–264.

(458) LAWLEY, P. D., and BROOKES, P. (1963) FURTHER STUDIES ON THE ALKYLATION OF NUCLEIC ACIDS AND THEIR CONSTITUENT NUCLEOTIDES. *Biochem. J. 89*, 127–38.

(459) Watters, K. E., Yu, A. M., Strobel, E. J., et al. (2016) Characterizing RNA structures in vitro and in vivo with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Methods 103*, 34–48.

(460) McGinnis, J. L., Dunkle, J. A., Cate, J. H. D., et al. (2012) The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc. 134*, 6617–6624.

(461) Deigan, K. E., Li, T. W., Mathews, D. H., et al. (2009) Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. 106*, 97–102.

(462) Gherghe, C. M., Shajani, Z., Wilkinson, K. A., et al. (2008) Strong correlation between SHAPE chemistry and the generalized NMR order parameter (S2) in RNA. *J. Am. Chem. Soc. 130*, 12244–12245.

(463) Busan, S., and Weeks, K. M. (2018) Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA 24*, 143–148.

(464) Siegfried, N. A., Busan, S., Rice, G. M., et al. (2014) RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods 11*, 959–965.

(465) Underwood, J. G., Uzilov, A. V, Katzman, S., et al. (2010) FragSeq: Transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods 7*, 995–1001.

(466) Weeks, K. M. (2011) RNA structure probing dash seq. *Proc. Natl. Acad. Sci. 108*, 10933–10934.

(467) Lucks, J. B., Mortimer, S. A., Trapnell, C., et al. (2011) Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. 108*, 11063–11068.

(468) Busan, S., and Weeks, K. M. (2018) Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. *RNA 24*, 143–148.

(469) Smola, M. J., Rice, G. M., Busan, S., et al. (2015) Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nat. Protoc. 10*, 1643–1669.

(470) Lange, S. J., Maticzka, D., Möhl, M., et al. (2012) Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res. 40*, 5215–5226.

(471) Cheong, H.-K., Kim, N.-K., and Cheong, C. (2015) RNA Structure: Tetraloops. *eLS* 1–6.

(472) Sheehy, J. P., Davis, A. R., and Znosko, B. M. (2010) Thermodynamic characterization of naturally occurring RNA tetraloops. *RNA 16*, 417–429.

(473) Hao, Y., and Kieft, J. S. (2016) Three-way junction conformation dictates self-association of phage packaging RNAs. *RNA Biol. 13*, 635–645.

(474) Klostermeier, D., and Millar, D. P. (2002) Energetics of hydrogen bond networks in RNA: Hydrogen bonds surrounding G+1 and U42 are the major determinants for the tertiary structure stability of the hairpin ribozyme. *Biochemistry 41*, 14095–14102.

(475) Lescoute, A., and Westhof, E. (2006) Topology of three-way junctions in folded RNAs. *RNA 12*, 83–93.

(476) De, M., Pen˜a, L. A., Pen˜a, P., et al. (2009) Three-way RNA junctions with remote tertiary contacts: A recurrent and highly versatile fold. *RNA 15*, 1949–1964.

(477) Clote, P., Ferré, F., Kranakis, E., et al. (2005) Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA 11*, 578–591.

(478) Jones, S., Van Heyningen, P., Berman, H. M., et al. (1999) Protein-DNA interactions: A structural analysis. *J. Mol. Biol. 287*, 877–896.

(479) Chao, J. A., Patskovsky, Y., Patel, V., et al. (2010) ZBP1 recognition of β-actin zipcode induces RNA looping. *Genes Dev. 24*, 148–158.

(480) SAUNDERS, L. R., and BARBER, G. N. (2003) The dsRNA binding protein family: critical roles, diverse cellular functions. *FASEB J. 17*, 961–983.

(481) Pollum, M., Jockusch, S., and Crespo-Hernández, C. E. (2015) Increase in the photoreactivity of uracil derivatives by doubling thionation. *Phys. Chem. Chem. Phys. 17*, 27851–27861.

(482) Nakamura, S., Kawabata, H., Muramatsu, H., et al. (2016) Effect of 5-Substitution of Uracil Base in DNA Photocrosslinking Using 3-Cyanovinylcarbazole. *Chem. Lett. 45*, 887–889.

(483) Masliah, G., Barraud, P., and Allain, F. H. T. (2013) RNA recognition by double-stranded RNA binding domains: A matter of shape and sequence. *Cell. Mol. Life Sci. 70*, 1875–1895.

(484) Mackereth, C. D., and Sattler, M. (2012, June 1) Dynamics in multi-domain protein recognition of RNA. *Curr. Opin. Struct. Biol.* Elsevier Current Trends.

(485) Annemieke, A.-R., and Van Ommen, G.-J. B. (2007) Antisense-mediated exon skipping: A versatile tool with therapeutic and research applications. *RNA 13*, 1609–1624.

(486) Muntoni, F., Aartsma-Rus, A., Hemmings, R., et al. (2017) Development of Exon Skipping Therapies for Duchenne Muscular Dystrophy: A Critical Review and a Perspective on the Outstanding Issues. *Nucleic Acid Ther. 27*, 251–259.

(487) Plyte, S. E., and Kneale, G. G. (2001) Limited proteolysis of protein-nucleic acid complexes. *Methods Mol. Biol. 148*, 315–21.

(488) Pellestor, F., and Paulasova, P. (2004) The peptide nucleic acids (PNAs), powerful tools for molecular genetics and cytogenetics. *Eur. J. Hum. Genet. 12*, 694–700.

(489) Andreev, O. A., Wyatt, L. C., Crawford, T., et al. (2018) Peptides of pHLIP family for targeted intracellular and extracellular delivery of cargo molecules to tumors. *Proc. Natl. Acad. Sci. 115*, E2811–E2818.

(490) Disney, M. D., Winkelsas, A. M., Velagapudi, S. P., et al. (2016) Inforna 2.0: A Platform for the Sequence-Based Design of Small Molecules Targeting Structured RNAs. *ACS Chem. Biol. 11*, 1720–1728.

(491) Velagapudi, S. P., Gallo, S. M., and Disney, M. D. (2014) Sequence-based design of bioactive small molecules that target precursor microRNAs. *Nat. Chem. Biol. 10*, 291–297.

(492) Disney, M. D., Labuda, L. P., Paul, D. J., et al. (2008) Two-dimensional combinatorial screening identifies specific aminoglycoside-RNA internal loop partners. *J. Am. Chem. Soc. 130*, 11185–11194.

(493) Velagapudi, S. P., Seedhouse, S. J., and Disney, M. D. (2010) Structure-activity relationships through sequencing (StARTS) defines optimal and suboptimal RNA motif targets for small molecules. *Angew. Chemie - Int. Ed. 49*, 3816–3818.

(494) Pishas, K. I., Lessnick, S. L., Pishas, K. I., et al. (2018) Ewing sarcoma resistance to SP-2509 is not mediated through KDM1A/LSD1 mutation. *Oncotarget 9*, 36413–36429.

(495) Velinder, M., Singer, J., Bareyan, D., et al. (2016) GFI1 functions in transcriptional control and cell fate determination require SNAG domain methylation to recruit LSD1. *Biochem. J. 473*, 3355–3369.

(496) Panzetta, M. E., Valdivia, R. H., and Saka, H. A. (2018, December 12) Chlamydia persistence: A survival strategy to evade antimicrobial effects in-vitro and in-vivo. *Front. Microbiol.* Frontiers.

(497) Páez-Canro, C., Alzate, J. P., González, L. M., et al. (2019) Antibiotics for treating urogenital Chlamydia trachomatis infection in men and non-pregnant women. *Cochrane Database Syst. Rev.*

(498) Grieshaber, S. S., and Grieshaber, N. A. (2014) The role of the chlamydial effector CPAF in the induction of genomic instability. *Pathog. Dis. 72*, 5–6.

(499) Häcker, G. (2014) The chlamydial protease CPAF: Important or not, important for what? *Microbes Infect. 16*, 367–370.

(500) Bednar, M. M., Jorgensen, I., Valdivia, R. H., et al. (2011) Chlamydia Protease-like Activity Factor (CPAF): Characterization of proteolysis activity in vitro and development of a nanomolar affinity CPAF zymogen-derived inhibitor. *Biochemistry 50*, 7441–7443.

(501) Johnson, K. A., Lee, J. K., Chen, A. L., et al. (2015) Induction and inhibition of CPAF activity during analysis of Chlamydia-infected cells. *Pathog. Dis. 73*, 1–8.

(502) Saka, H. A., Thompson, J. W., Chen, Y. S., et al. (2015) Chlamydia trachomatis infection leads to defined alterations to the lipid droplet proteome in epithelial cells. *PLoS One 10*, 1–27.

Biography

Meghan Lawler received Bachelor's of Science in Chemistry, ACS Certified and a Bachelor's of Science in Biology with a concentration in Molecular Biology, both with honors, concurrently Western Carolina University (WCU) May 2014. From August 2012 through May 2014, she was involved in a research project through the Chemistry and Physics department with Dr. William R. Kwochka exploring the stoichometry of boroxine-amine interactions. From Aug 2013 through May 2014 she established a new mechanism through which to directly amplify mitochondrial DNA from a buccal swab with Dr. Mark Wilson through the Forensic Science program at Western Carolina University. She also participated in an NSF-funded REU at Virginia Commonwealth (VCU) with Dr. Ashton Cropp exploring bio-orthogonal methods to incorporate unnatural amino acids *in vivo*.

After joining the Chemistry department at Duke University Fall of 2014, Meghan has been with the McCafferty lab since January 2015 working on establishing the role of coregulatory proteins and lncRNA in epigenetic enzyme functions. During this time, she has been awarded the NIH-funded Pharmacological Sciences Training Program (PSTP) grant from August 2015 through August 2017 as well as the Department of Education funded GAANN fellowship through the Chemistry department, enabling her to expand on her teaching skills towards her academic career goals.