Protein Analysis by Mass Spectrometry

Benjamin Taylor Barnhill
Salem, Virginia

B.S., Chemistry, Hampden-Sydney College, 2008

A Dissertation presented to the Graduate Faculty of the University of Virginia in Candidacy for the Degree of Doctor of Philosophy

Department of Chemistry

University of Virginia
August, 2017

Abstract

This dissertation describes two projects using mass spectrometry to analyze proteins. In project one we use LC-MS/MS and electron transfer dissociation to investigate the relationship between two *Arabidopsis thaliana* proteins, RGA and SPY, which are key plant growth regulators. SPY had previously been identified as an O-GlcNAc transferase that modified RGA, but our work led to the discovery that SPY is in fact a novel O-fucose transferase. Furthermore, we have established that O-GlcNAc and O-fucose modification have opposing effects on RGA. Modification with O-GlcNAc represses the growth inhibiting function of RGA, while O-fucose enhances it.

The second project describes the use of mass spectrometry to identify proteins that selectively bind to regions of DNA, called somatic hypermutation (SHM) enhancer regions, shown to be necessary for targeting SHM to the Immunoglobulin gene variable region in B cells. SHM is a key step in the immune system's generation of high affinity antibodies, but how the mutational process is confined to the immunoglobulin locus is not yet understood. Using a label free quantitation strategy we identify a number of nuclear proteins from the DT40 and Ramos B cell lines, in particular the transcription factors Ikaros and Aiolos, that preferentially bind to enhancer DNA vs. control DNA.

Acknowledgements

Five and a half years ago I met Don for the first time during a visit to UVA while I was trying to select a graduate program. He quietly described the work going on in the lab to me, and at the end of our meeting said something like, "well anyway... we have a place for you if you decide to come." In retrospect it's hard to believe one of the greatest gifts I've ever been given was delivered so casually (and barely audibly). Studying in Don's lab has given me the skills I need to pursue a career that is both interesting and meaningful, which is one of the greatest privileges one can have in life. Thank you Don for making a place for me at UVA, for creating the learning environment I grew so much in, and for your kind encouragement over the last five years.

As all Hunt lab alumni know, few, if any, of those skills would have been acquired without the help of Jeff Shabanowitz. Jeff has instilled in me an almost pathological need to ask and answer the question "but how does it work?" I think this may be the most valuable tool I'll take with me from the lab into my career and life. Thank you Jeff for your teaching, and for your friendship.

Nothing I've done over the course of my graduate career would have been possible without the support of so many past and present members of the Hunt group. Dina Bai is the hidden gem of our lab who somehow simultaneously keeps us sane and makes all our data analysis possible. Thank you Dina for your work and wise counsel. Much of my understanding of mass spec theory and instrumentation came from spontaneous discussions with Michelle English, often whether she wanted to have them or not, so thank you Michelle. Weihan Wang and I spent several years sharing a cubicle:

and a research project on O-GlcNAc that went nowhere but greatly enhanced my understanding of mass spec. Thank you Weihan for your quiet nature and excellent scientific work.

I never would have been able to perform a single experiment without training from Jenn Abelin, Lissa Anderson, and Stacy Malaker. Thank you all. I am especially grateful to Andrew Dawdy for his years of work on RGA and for trusting the next phase of the project to me. Amanda Wriston helped me a great deal with data analysis during early phases of the RGA project, and in my opinion is the true discoverer of O-Fucose.

I did not begin my tenure in Don's group alone, but in the company of two other excellent students: Paisley Trantham and Scott Ugrin. I feel very lucky to have had them both as colleagues and friends during my time here. Our dinners together with them and their spouses Jeff and Lindsey, in addition to the emotional rollercoaster of UVA basketball games with Scott, are largely responsible for my continued psychological health at as a graduate student. My friendship with Stephanie Lehman did not get off to an auspicious beginning when I pointedly ignored her questions during her prospective weekend at UVA in order to watch a basketball game on my phone while "leading a tour", but I'm glad she persisted. Our friendship with her and her husband Joel has added a lot to Sara's and my time in Charlottesville.

The work in this dissertation was performed using a brand new mass spectrometer, the Thermo Scientific Orbitrap Fusion, that arrived in May of 2014 in a big box and with a very steep learning curve. I never would have been able to collect a single piece of data if it weren't for the constant assistance of Chris Mullen and John

Syka at Thermo. In addition to helping me make the instrument run they taught me a lot along the way. Thank you both.

Both projects in this dissertation benefited from excellent scientific collaborations. Rodolfo Zentella and Tai-Ping Sun at Duke University conceived the RGA project from beginning to end, and it was a privilege to help them with it. David Schatz at Yale had no idea he was giving me the highlight of my graduate career when he emailed Don after a chance meeting at a conference several years ago to propose a new project. His graduate student Ravi Dinesh has been an excellent collaborator and for several years now the time stamps on his emails have been making me glad I'm getting a Ph.D. in analytical chemistry instead of immunology. Get some sleep Ravi, you deserve it.

My defense committee, Drs. David Brautigan, Linda Columbus, Ken Hsu, and Kevin Lehmann, have my heartfelt thanks for reading this document and providing feedback. In my entire academic career I have never completed a single written assignment without finding a severe typo minutes after turning it in. I'm sure this will be no exception, so please accept my apologies. Additional thanks to Drs. Columbus, Lehmann, and Jim Demas for their service on my candidacy exam committee.

After my third year at UVA I had the opportunity to spend a wonderful summer in the PAC-2 department at Genentech interning for Hunt lab alumna Feng Yang. Feng was an excellent mentor during my time at Genentech and has continued to help in my graduate career after I left. I'm very grateful to her for her time and encouragement. Additionally, I don't think I ever would have survived in the bay area if I hadn't been

taken in by my cousin Ginny and her husband Matt. The time I spent with them and their two children, Tazwell and Coco, that summer is one of my fondest memories as a graduate student.

I never would have attempted any of this in the first place if it weren't for the support of many people. Thank you to the chemistry faculty at Hampden-Sydney College: Bill Anderson, Kevin Dunn, Paul Mueller, Bill Porterfield, and Herb Sipe. I will never walk into a laboratory without remembering standing in the HSC chemistry labs performing my first experiments under your watchful gazes, and I'm grateful for your friendship in the years since I've graduated. Before graduate school I had the good fortune to find a position working in a laboratory (during a recession no less) at Pharmaceutical Product Development, in Richmond, VA. It was there that I developed my interest in biotechnology and my desire get a Ph.D., and I was very lucky to be encouraged along the way by my supervisors Hollie Barton and Diana Mathiasen. Thank you both.

I have always felt, and this has intensified as I've gotten older, that I won the parental lottery when I was born. I never would have finished this dissertation without the curiosity and perseverance I learned from my parents. I miss you Dad, I love you Mom, thank you both. To my older siblings, Tom and Jane, thank you for your love and support throughout my life.

Finally, I want to thank my wife Sara for supporting and encouraging me through graduate school. When we left Richmond in 2012 to move the 60 miles to Charlottesville it felt like such a big step. Now, for the second time in our lives, we're

beginning the summer in an apartment full of boxes waiting for the journey to our new home. If you had asked us five years ago to guess where that home would be, and given us one hundred chances, we couldn't have guessed the United Kingdom. I feel so grateful for the life we've had and continue to build together. Going back to school was the second best decision I've ever made after marrying you. I love you, and there's no one I'd rather go on this next adventure with.

Table of Contents

List of Figures, Tables, and Equations

Chapter 1:

List of Abbreviations

ºC:	Degrees celcius
•	Radical species
Å:	Angstroms
AA:	Amino acid
AC:	Alternating current
AGC:	Automatic gain control
Ala, A:	Alanine
amu:	Atomic mass units
API:	Atmospheric pressure ionization
Arg, R:	Arginine
Asn, N:	Asparagine
Asp, D:	Aspartic acid
c:	Centi ($1x10^{-2}$)
CAD:	Collision-activated dissociation
CID:	Collision induced dissociation
Cys, C:	Cysteine
Da:	Dalton
DC:	Direct current
DTT:	Dithiothreitol
ECD:	Electron capture dissociation
ESI:	Electrospray ionization

ETD:	Electron transfer dissociation
ETnoD:	Electron transfer with no dissociation
f:	femto ($1x10^{-15}$)
FETD:	Front-end electron transfer dissociation
FT:	Fourier transform
Gln, Q:	Glutamine
Glu, E:	Glutamic acid
Gly, G:	Glycine
HC:	Heavy chain
HCD:	Higher energy collisionally-activated dissociation
His, H:	Histidine
HPLC:	High-performance liquid chromatography
hr:	Hour
HV:	High voltage
Hz:	Hertz
i.d.:	Inner diameter
IgG:	Immunoglobulin gamma
Ile, I:	Isoleucine
IPC:	Isotope pattern calculator
kDa:	kilodalton
L:	Liter
LC:	Liquid chromatography
Leu, L:	Leucine

Lys, K:	Lysine
µ:	mico ($1x10^{-6}$)
m:	Milli ($1x10^{-3}$) or meter
M:	Mass or molar
m/z:	Mass-to-charge ratio
MALDI:	Matrix-assisted laser desorption ionization
Met, M:	Methionine
min:	Minute
mol:	Mole ($6.022x10^{23}$)
M:	Molar
MS:	Mass spectrometry
MS/MS:	Tandem mass spectrometry
MSn:	Tandem MS carried to the nth stage
MW:	Molecular weight
n:	Nano ($1x10^{-9}$)
NL:	Normalized level
o.d.:	Outer diameter
p:	Pico ($1x10^{-12}$)
PEEK:	Polyether(ethylene)ketide
Phe, F:	Phenylalanine
ppm:	Parts per million
Pro, P:	Proline
PTM:	Post-translational modification

R:	Resolution
RF:	Radio frequency
RPM:	Revolutions per minute
Ser, S:	Serine
SRIG:	Stacked ring ion guide
Thr, T:	Threonine
TIC:	Total ion current
Trp, W:	Tryptophan
Tyr, Y:	Tyrosine
Val, V:	Valine

"...for many purposes a theory whose consequences are easily followed is preferable to one which is more fundamental but also more unwieldy"

-J.J. Thompson, The Corpuscular Theory of Matter, 1907

Chapter 1: Introduction to the Dissertation

1.1 Overview

This dissertation describes two collaborative research projects in which mass spectrometry is used as a tool to identify and characterize proteins. In chapter two we investigate the relationship between two *Arabidopsis thaliana* proteins, RGA and SPY, which are key plant growth regulators. While these proteins have been the subject of significant study because of their implications in regulating the growth of modern high yield genetic variants of cereal grains (wheat, corn, rice), which are an indispensable part of the worldwide food supply, serious questions about their roles in the cell remained. It has long been suspected that SPY post-translationally modifies RGA, but the experimental approaches employed so far could not definitely answer that question. Our work yielded surprising new insights about the relationship between RGA and SPY that could not have been obtained without mass spectrometry, and these results are important for understanding not just plant growth, but cell signaling, and protein post-translational modification at large.

Chapter three describes the use of mass spectrometry to identify proteins that selectively bind to a region of DNA shown to be necessary for targeting somatic hypermutation (SHM) to the Immunoglobulin gene variable region in B cells. SHM is a key step in the immune system's generation of high affinity antibodies, which are necessary to protect from extracellular pathogens like viruses. Conversely, failure to properly regulate this mutational process is a possible source of B cell lymphomas. Here,

mass spectrometry was essential to distinguish the complicated array of proteins present in our samples.

These two projects differ in their goals, detailed characterization of a few proteins versus identification of many proteins, but the foundational mass spectrometry experiment employed is identical. This introductory chapter will begin by describing the general structure of proteins, how they are produced by the cell, and their importance in biology. It will then outline the mass spectrometry experiment applied to study proteins in these two very different research projects. The necessary theory and instrumentation background is presented as it relates to the specific instrument used for this work, The Thermo Scientific™ Orbitrap Fusion™ Tribrid™ (Fusion). The introduction closes with a discussion of recent instrumental advances incorporated into the Fusion system, and how those advances have affected the MS/MS experiment central to both research projects.

Chapter 2 is a testament to the impressive depth of analysis possible when studying protein post translational modification with the most advanced instrumentation. Chapter 3 highlights these advances as well, but also serves to articulate the challenges facing the field of protein MS. In a highly complex sample many of the components investigated by MS/MS are never identified, and some of these likely represent important peptides that could identify new proteins or proteo-forms. Before dealing with these challenges, however, we begin by addressing the question why study proteins at all?

1.2 Introduction to Proteins

A protein is a macromolecule composed of a chain of smaller molecules called amino acids. There are twenty common amino acids. Each has an identical chemical backbone but a unique functional side group. **Figure 1.1** shows the complete structures of all 20 common amino acids organized by the chemical properties of their side groups. The chain is held together by peptide bonds between the carboxyl end of one amino acid and the amine group of the next (**Figure 1.2**). A chain of amino acids has a free amine group at one end called the N-terminus, and a free carboxyl group at the other called the C-terminus. Each amino acid has an associated three letter abbreviation and 1 letter code (**Table 1.1**); protein sequences are typically represented using the 1-letter codes and displayed with the N-terminus on the left and the C-terminus on the right.

Proteins can vary tremendously in size; from fewer than fifty to thousands of amino acids in length. Long chains of amino acids are called *poly-peptides*, and this word is sometimes used interchangeably with *proteins*. Prior to mass spectrometry analysis purified proteins are often broken down into shorter groups of amino acids through enzymatic digestion. These small amino acid chains are called *peptides*.

A protein's amino acid sequence is also called its *primary structure*. The amino acid chain folds into higher order structures based on the chemical properties of the side chains in the primary structure; as examples, nonpolar side chains like valine and leucine will associate with one another through hydrophobic interactions, hydrogen bonds can form between acidic and basic side chains, proline binds onto its own backbone causing a sharp bend in the polypeptide chain, and cysteines form disulfide bonds with one another.

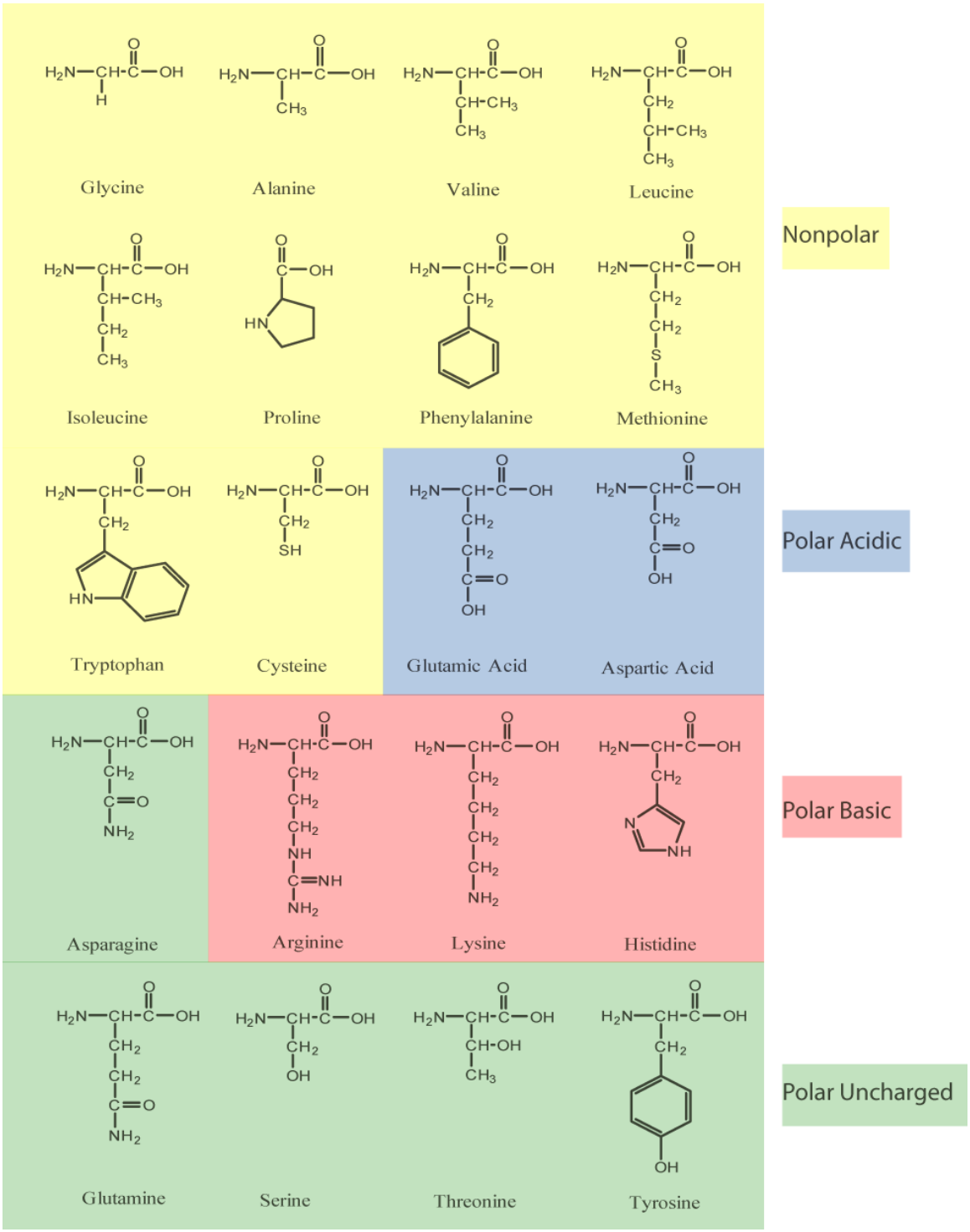

Figure 1.1: Structures of the twenty common amino acids.

The chemical properties of the *primary structure* generate organized *secondary structures* called alpha helices and beta sheets. These associate with one another and remaining disordered sections of the amino acid chain to create even more complex *tertiary structures*. Many proteins also join together into complexes. Whether made from a single protein or many, these unique three dimensional assemblies of amino acids can then perform highly sophisticated functions based on their shape and chemical properties.

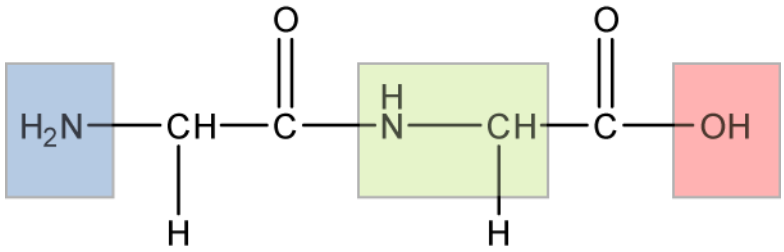

Figure 1.2: The peptide bond. Shown here is the dipeptide glycine-glycine. The peptide bond is highlighted in green, the amine terminus in blue, and the carboxyl terminus in red.

The order of the amino acids that compose a protein is dictated by the cell's genetic material, its DNA. DNA, or deoxyribonucleic acid, is stored in the cell's nucleus as a double stranded antiparallel chain repeating four nucleic acids, Adenine, thymine, guanine, and cytosine, joined together by phosphodiester bonds. The nucleic acids form complementary pairs based on hydrogen bonding, A to T and G to C, so that each DNA strand is paired with its inverse complement strand. The region of DNA that codes for a given protein is called a gene. Protein synthesis starts when the cell transcribes the sequence of DNA bases from one gene into another nucleic acid chain, messenger RNA

(mRNA). mRNA is similar in structure to DNA, but its backbone is ribose instead of deoxyribose, the base thymine has been replaced with uracil, and it is single stranded.

After the transcription of a gene into mRNA is complete, mRNA is transported out of the nucleus and onto a structure called the ribosome. On the ribosome mRNA is translated into a sequence of amino acids with the help of another nucleic acid, translational RNA (tRNA). Each molecule of tRNA pairs with a specific set of three mRNA nucleotides and carries with it the single amino acid indicated by that three letter code (**Table 1.2**).

The 20 Common Amino Acids			
Name	3-letter code	1-letter code	Chemical Composition
Glycine	Gly	G	C_2H_3NO
Alanine	Ala	A	C_3H_5NO
Serine	Ser	S	$C_3H_5NO_2$
Proline	Pro	P	C_5H_7NO
Valine	Val	V	C_5H_9NO
Threonine	Thr	T	$C_4H_7NO_2$
Cysteine	Cys	C	C_3H_5NOS
Leucine	Leu	L	$C_6H_{11}NO$
Isoleucine	Ile	I	$C_6H_{11}NO$
Asparagine	Asn	N	$C_4H_6N_2O_2$
Aspartic Acid	Asp	D	$C_4H_5NO_3$
Lysine	Lys	K	$C_6H_{12}N_2O$
Glutamine	Gln	Q	$C_5H_8N_2O_2$
Glutamic Acid	Glu	E	$C_5H_7NO_3$
Methonine	Met	M	C_5H_9NOS
Histidine	His	H	$C_6H_7N_3O$
Phenylalanine	Phe	F	C_9H_9NO
Arginine	Arg	R	$C_6H_{12}N_4O$
Tyrosine	Tyr	Y	C_9H_9NO
Tryptophan	Trp	W	$C_{11}H_{10}N_2O$

Table 1.1: Abbreviations of the twenty common amino acids

As new amino acids are brought into place by tRNA they are ligated onto the C-terminus of the growing amino acid chain. The formation of new peptide bonds is catalyzed by proteins within the ribosome structure. As the synthesis of its primary structure progresses the protein folds into secondary and tertiary structures. Once mRNA translation is complete and the ribosome releases the new protein it can finish forming its unique three dimensional structure.

The Genetic Code	
Amino Acid	mRNA codes
Glycine	GGU GGC GGA GGG
Alanine	GCU GCC GCA GCG
Serine	AGU AGC
Proline	CCU CCC CCA CCG
Valine	GUU GUC GUA GUG
Threonine	ACU ACC ACA ACG
Cysteine	UGU UGC
Leucine	CUU CUC CUA CUG
Isoleucine	AUU AUC AUA AUG
Asparagine	AAU AAC
Aspartic Acid	GAU GAC
Lysine	AAA AAG
Glutamine	CAA CAG
Glutamic Acid	GAA GAG
Methonine	AUG
Histidine	CAU CAC
Phenylalanine	UUU UUC
Arginine	AGA AGG
Tyrosine	UAU UAC
Tryptophan	UGG

Table 1.2: The genetic code. Amino acids in the left hand column are paired with molecules of tRNA complementary for one of the mRNA triplets list on the right. As tRNA associates with mRNA according to the code amino acids are ligated together into a poly-peptide.

Because a protein's amino acid sequence ultimately determines its structure and function, and this sequence is contained within the genetic data of the cell, one might expect DNA to provide all the information necessary to understand a protein; like having detailed architectural drawings for a building. Fortunately for the field of mass spectrometry this is not the case for three reasons.

First, DNA alone is not an accurate representation of a protein's final amino acid sequence. Before mRNA is trafficked out of the nucleus and translated into protein, it can be edited through a process called splicing. Proteins sometimes have multiple functional subunits, all of which may not be necessary at a given time. Splicing allows a single gene to code for a set of functional subunits that can be assembled into proteins with overlapping but distinct sequences and cellular functions [1].

Second, proteins can be chemically altered during or after their translation from mRNA. These changes are called post-translational modifications (PTMs) [2,3]. PTMs are covalent modifications to the side chains of amino acids, or to the free amino or carboxy termini of proteins. PTMs have no relationship to a protein's parent gene. Some common PTMs are shown in **Figure 1.3**. PTMs can be dynamic, meaning they come on and off of proteins rapidly (in seconds or less) during the course of specific cellular processes, or change the chemical structure of a protein for the duration of its existence. The specific amino acid residue on which a PTM occurs can be important to its effect, and a single protein can have many post-translational modifications to its primary structure. The many unique combinations of splice variants and post-translational modifications that can exist for a given protein are commonly referred to as proteoforms,

and two proteoforms of the same underlying protein may have different roles inside the cell.

Third, proteins are inherently interactive. Every protein must bind to another molecule at some point to perform its function. This means that contextual factors are as important as a protein's chemical properties in understanding its function. Where and when in the cell does a protein act, and what other molecules (proteins, lipids, carbohydrates, etc.) are present when it does?

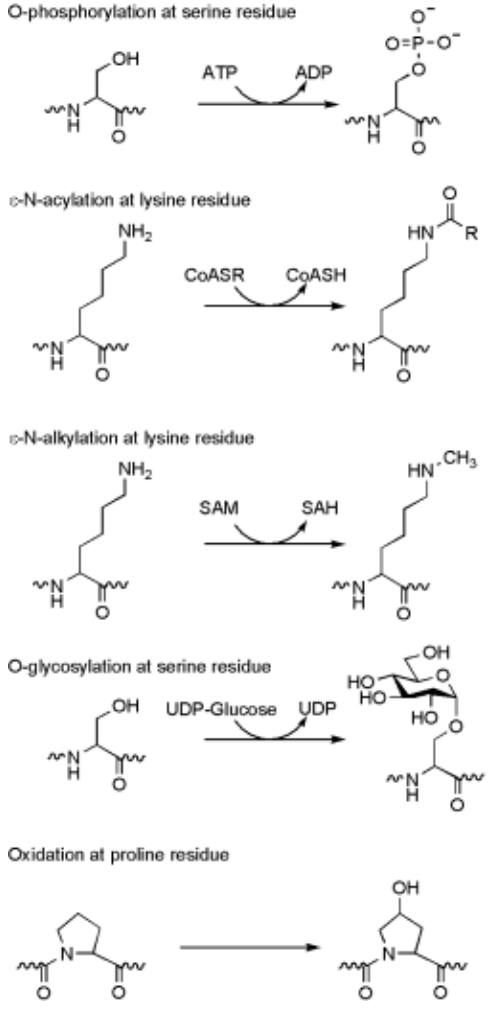

Figure 1.3: Examples of post-translational modification. Figure adapted from [2].

These three conditions mean that thorough study of a protein's function requires one to determine its primary amino acid sequence, and the type and location of any PTMs, in a way that is specific in space and time within the cell. Complete primary structure analysis of a protein captured from a specific cellular event should be the protein analytical chemist's goal. This is the only way to determine which proteoforms are at work in which cellular processes.

Mass spectrometry has the capability to sequence proteins and map post-translational modifications [4]. The high sensitivity of modern MS means that small amounts of protein purified from selected cellular compartments or at specific time points can be characterized. While it does not yet always provide comprehensive sequence analysis, mass spectrometry has become the dominant technique for protein sequencing, PTM mapping, and identification of proteins within complex samples. The next part of this chapter will describe in detail how mass spectrometry is currently applied to the analysis of proteins.

1.3 How Proteins are Studied by Mass Spectrometry

The protein identification and sequencing mass spectrometry method described in this section has been divided into phases; protein extraction and preparation, liquid chromatography, electrospray ionization, tandem mass spectrometry, and data analysis. Each phase is described generally with an emphasis on the purpose it plays in the overall mass spectrometry experiment. Specific parameters for the experiments in chapters two and three are provided in the methods sections of those chapters. Each section is presented in the order that it occurs experimentally with the greatest emphasis placed on

the tandem mass spectrometry experiment (MS/MS) that is the foundation of protein sequencing and PTM mapping.

1.3.1 Protein Extraction and Preparation

Both projects discussed in this dissertation followed a similar sample preparation strategy. Proteins were isolated from tens of millions of cells in a multistep protocol that involved mechanical cell lysis, followed by some form of targeted protein capture and removal of non-protein components. The objective was to eliminate the non-protein components of the cell (lipids, nucleic acids, etc.), and significantly reduce the number of proteoforms present so that the protein(s) of interest would have a higher relative abundance in the sample. Once the protein(s) of interest were purified they were broken into short peptides by proteolytic digestion with the enzyme trypsin. The purpose of digestion is the break the protein(s) down into peptides of a size amenable for chromatographic separation and MS/MS analysis.

1.3.2 Liquid Chromatography

Peptides were separated and concentrated in line with the mass spectrometer by nano-flow reverse phase liquid chromatography (n-RPLC) [5]. Peptides in aqueous solution were pressure loaded onto a homemade capillary liquid chromatography pre-column using a helium bomb (**Figure 1.4**). The column was then connected to an LC system and rinsed with 0.1% acetic acid in water for several column volumes to remove salts leftover from the sample preparation buffers. The pre-column was then connected with a teflon sleeve to a second analytical column that had an approximately 3 micrometer (μm) spray tip on one end created with a laser puller.

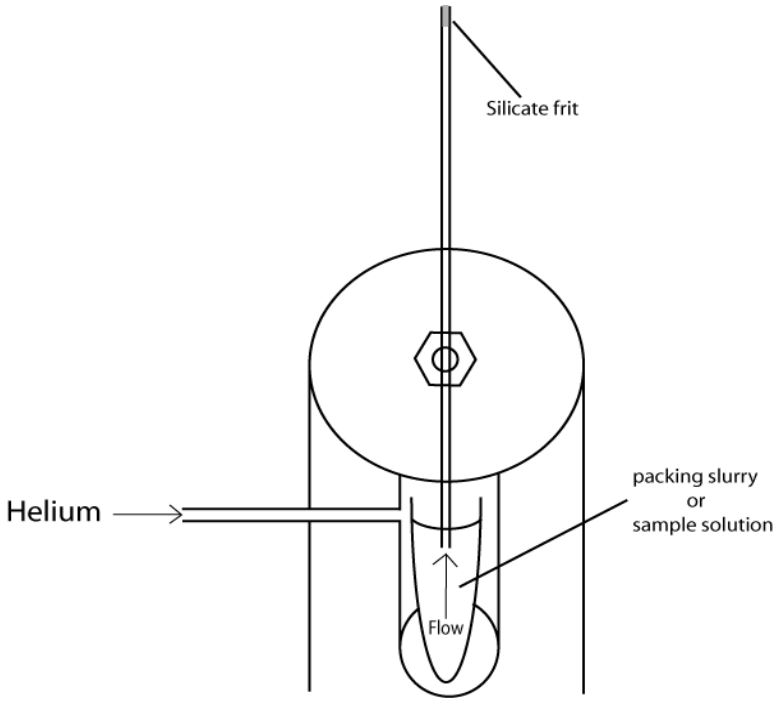

Figure 1.4: A helium capillary column pressure bomb. Helium is pumped into the bomb at 100-500psi. The increased pressure inside the bomb forces the solution inside the tube out of the bomb through the fused silica capillary. Druing column packing the solid phase bead accumulate behind the poros silicate frit blocking the end of the capillary. The same setup is used for sample loaded on columns already containing stationary phase.

Both columns were constructed from polyimide coated fused silica capillaries of 75 μm inner diameter and 360 μm outer diameter. A solution of three parts potassium silicate and one part formamide was allowed to polymerize at the end of the capillary to form a porous barrier, behind which a methanol slurry of 5 μm porous silica beads functionalized with 18 carbon chains (C18) was packed into the capillary also using a helium pressure bomb.

Once inside the capillary peptides will preferentially partition between the hydrophobic stationary phase (C18 groups on the silica beads) and the liquid phase (solvent flowing through the capillary) based largely on their hydrophobicity, but also on

other factors not completely understood. Peptides are loaded onto the column in a solution of 0.1% acetic acid in water and are retained on the stationary phase. Once the two columns are connected to the LC and MS/MS analysis begins the solvent composition is gradually changed to increasing levels of acetonitrile. As the organic composition of the solvent increases peptides can move out of the stationary phase and into the mobile phase, elute off of the column, and pass into the mass spectrometer for analysis.

A purification method targeting a single protein will still capture lower amounts of hundreds of different proteins that are then digested into tens of thousands of distinct peptides that fall within the detection limits of the mass spectrometer. With the high sensitivity of modern mass spectrometers there is no such thing as a truly simple sample. Chromatographic separation is essential to reduce the number of peptides passing into the detection system at a given moment. Additionally, LC serves as a concentration step by focusing the many copies of a given peptide into a small liquid volume eluting off of the LC system over a few seconds to minutes. With the mass spectrometer serving as a detection system, liquid chromatography peak areas can also be used to collect relative quantitative information about the peptides eluting from the column.

The liquid chromatography used in this research is distinct in that it was conducted at flowrates of approximately 100 nanoliters per minute, while the vast majority of LC is performed at flow rates of greater than 100 microliters per minute. Nanoflow LC has become a popular choice for analyzing proteins of limited abundance by mass spectrometry because of its improved sensitivity over conventional LC [6,7].

The reason for this improvement is explained next in the discussion of electrospray ionization.

1.3.3 Electrospray Ionization

Because a mass spectrometer manipulates and detects only ions, the final step before MS/MS analysis must be to charge the peptides passing from the LC system towards the mass spectrometer via electrospray ionization (ESI).

During ESI solvent exits the LC system at high pressure from a narrow spray tip directed towards the inlet of the mass spectrometer while a high voltage (+2kV) is applied to the waste line after the split in the LC solvent system and ground is held at the inlet of the mass spectrometer (**Figure 1.5**). The high voltage gradient creates a concentration of positive ions in the solvent at the electrospray tip that causes the solvent to form a Taylor cone. The buildup of electric potential at the tip of the Taylor cone creates a jet of charged droplets of solvent moving from the tip of the column into the mass spectrometer [8]. Contained within these charged droplets are the peptides eluting off the LC column.

As the droplets pass into the mass spectrometer they experience intense heat and rapidly decreasing pressure. Solvent evaporates and the droplets shrink. The charges within the droplets become more concentrated. When the force of electrical repulsion within a droplet exceeds its surface tension the droplet has reached its Rayleigh limit and must reduce its charge. There are two proposed mechanisms for how this happens, and experimental evidence suggests that both occur [9].

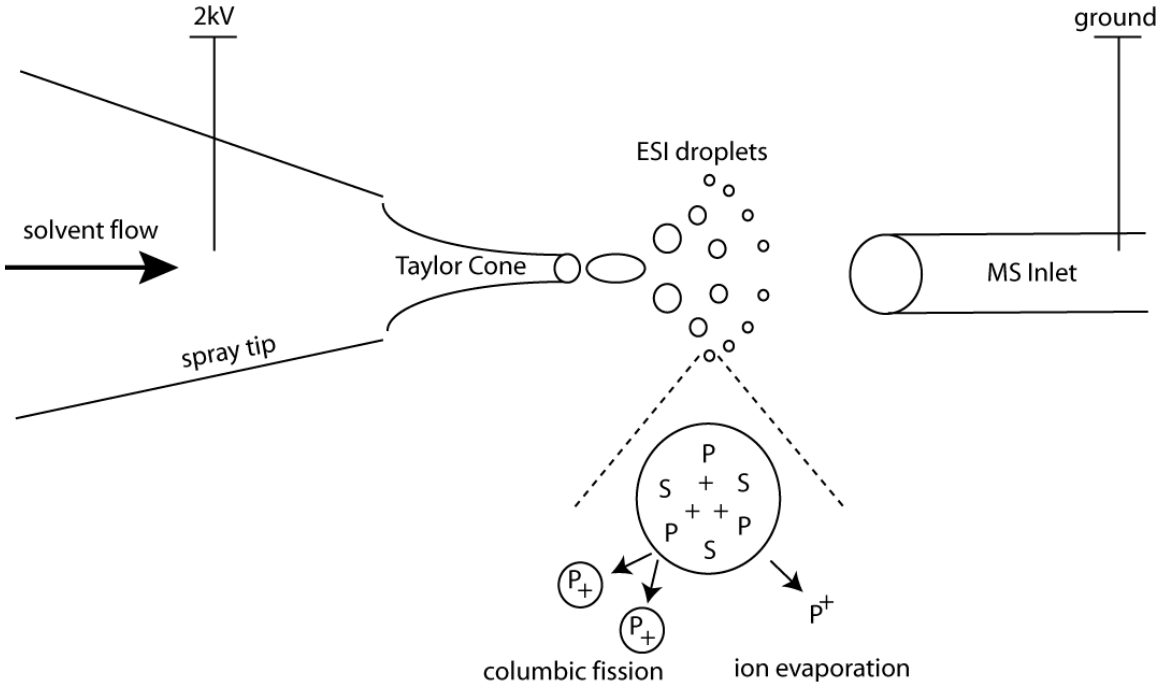

Figure 1.5: Electrospray ionization.

The first mechanism is columbic fission of the droplets. A droplet at its Rayleigh limit simply splits into two droplets, thereby increasing the ratio of surface area to charge and temporarily stabilizing the droplet until further solvent evaporation forces it to split again. After several generations of fission and evaporation the solvent is removed and only charged peptides remain. The second mechanism is charge ejection. Instead of splitting, the droplet ejects a peptide into the gas phase that carries some excess charge with it, thus reducing the overall charge of the droplet and temporarily stabilizing it.

Regardless of the specific mechanism one certainty is that ESI is a competitive process. For any given droplet traveling from the Taylor cone to the mass spectrometer

there are more molecules of solvent and peptide then there are excess protons, so not every molecule in the droplet can become an ion. Peptides compete with each other and the solvent molecules for protons based on their gas phase basicity [10], and only some become ionized. The ions are guided further into the instrument by electric fields where MS/MS analysis begins.

1.3.4 Tandem Mass Spectrometry Analysis

Tandem mass spectrometry, or MS/MS, is the foundational experiment in protein analysis by mass spectrometry [11]. Both chapters two and three of this dissertation describe the study of very different proteins under very different circumstances using the same MS/MS experimental structure. Every peptide MS/MS experiment has three parts: first ionized peptides have their intact mass determined by the mass spectrometer in an MS^1 scan, second, individual peptides are isolated and fragmented inside the mass spectrometer, and third, the masses of the fragments are determined by the MS^2 scan. The mass differences between its fragments, in conjunction with the intact mass, can be used to determine the sequence of the peptide and type and location of any post-translational modifications to its amino acids. Peptide sequences can then be linked back to their parent proteins through genetic databases of observed and predicted protein amino acid sequences. The mass spectrometer can only perform one MS/MS experiment at a time. The duration of each MS/MS event varies by the instrument and specific experimental parameters used, but in this work it varies between approximately 50 and 300 milliseconds. All of the liquid chromatography separations in this dissertation were

approximately 90 minutes long and over the course of each LC run tens of thousands of MS/MS experiments were performed.

All of the mass spectrometry experiments referenced in this dissertation were performed on a Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer (Fusion) [12]. This is a hybrid mass spectrometer that contains several distinct sets of devices: an ion optics system for the transmission of ions throughout the regions of the instrument, an Ion Routing Multipole for ion storage and higher energy collisional fragmentation, a Quadrupole Mass Filter for ion isolation, a Duel Cell Linear Ion Trap for ion storage, ion-ion reactions, low energy collisional fragmentation, and low resolution m/z measurement, and an Ultra High Field Orbitrap [13] for high resolution m/z measurement. A full diagram of the Fusion instrument can be seen in **Figure 1.6**. We are going to review the specific functions and operating principles of each sub-system of the Fusion instrument then describe their role in an MS/MS experiment.

1.3.4.1 Components of the Fusion Instrument

1.3.4.1.1 Ion Optics

Following electrospray ionization ions must be guided from the inlet of the mass spectrometer to the first ion storage chamber, the Ion Routing Multipole (IRM). The ion optics consist of the S-lens, Q-00, Active Beam Guide, MP1, MP 3, and all lenses. During transmission from the source to the IRM the Quadrupole Mass Filter and C-trap are also both part of the ion optics system.

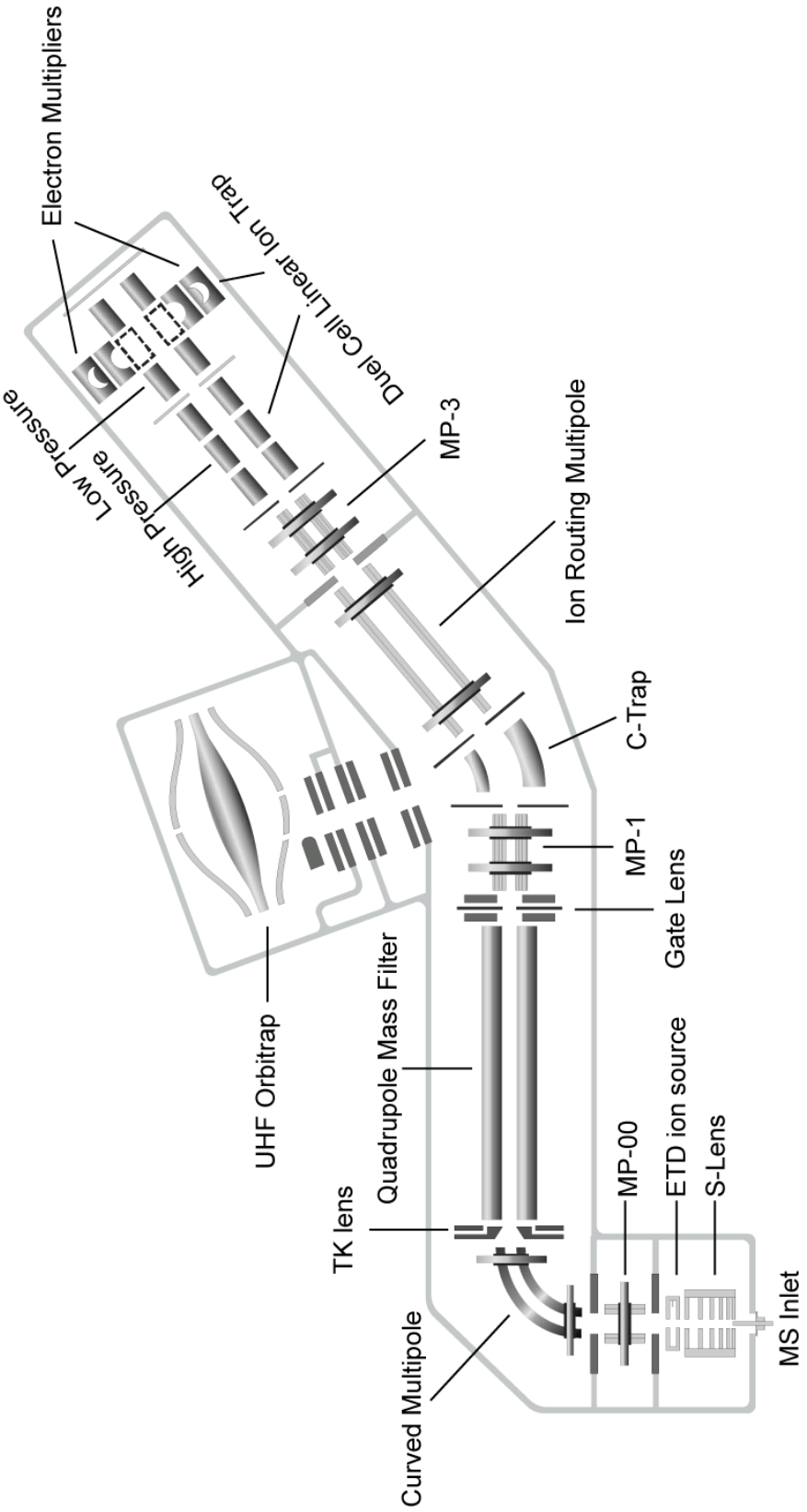

Figure 1.6: The Fusion instrument schematic. Figure adapted from [12].

Positive peptide ions are guided axially through the instrument by a static electric field of increasing negative potential until they reach the IRM, at which point positive potentials are applied to both ends of the device to contain the ion cloud. The field gradient is created by applying increasingly negative DC potentials to each component of the optics between the source and the IRM. The potentials vary based on the specific MS experimental parameters, but typically range from 0 to -30V and differ by only a few volts from one device to the next. Radially, ions traveling through the instrument are trapped by radio frequency AC potentials applied to multipole devices.

Within the ion optics system there are several specialized devices. The S-lens captures ions as they exit the capillary with high velocity and focus them into a concentrated beam that can be passed effectively into the rest of the instrument [14]. The curved multipole removes neutral species from the ion stream. Ions being guided by the DC and AC electric potentials should be carried through the curve, while neutral species being drawn into the instrument by the pressure differential should move in a straight path past the rods of the device and be pumped away. The gate lens regulates the passage of ions from source into the rest of the instrument. To stop the transmission of positive ions a high positive potential is applied to the lens. The Fusion system contains a specialized two part split gate lens that was developed to reduce transmission bias created by the different flight times of ions across the m/z range.

1.3.4.1.2 Ion Routing Multipole

The Ion Routing Multipole serves as an ion storage device and a fragmentation cell. The IRM is filled with nitrogen to a pressure of 8mTorr. During ion collection

peptide cations lose kinetic energy through collisions with the nitrogen bath gas as they enter the IRM. This collisional cooling allows more ions to be stably trapped than could be contained with electric fields alone. To fragment ions in the IRM, the segmented electrodes are used to generate a potential gradient much steeper than the one used during ion injection. This increases the energy of the collisions with the nitrogen to a level high enough to activate the peptides for fragmentation.

1.3.4.1.3 Quadrupole Mass Filter

The Quadrupole Mass Filter (QMF) consists of four parallel round rods (**Figure 1.7)** and has two operating modes; an AC only potential that allows a broad m/z range of ions (typically 150-2000m/z) to pass through the instrument to the IRM for collection, or a mass filter mode in which an AC and DC potential are both applied so that only a specific m/z range can pass through the quadrupole. The amplitude of the AC potential determines the center of the allowed m/z window, and the ratio of the DC to the AC potential determines the width of the window. The Fusion's QMF can filter windows from 1200m/z wide to 0.7m/z .

The QMF's operating principle is based on the stability equation for ions in a quadrupolar device, derived from the Mathieu equation (**Equation 1.1**) [15, 16] . The Mathieu equation is a second order linear differential equation developed by the French mathematician Mathieu while studying vibrating animal skins. Solutions to Mathieu's equation describe regions of stability in terms two unitless parameters, *a* and *q*, arising in the face of changing position (described by *u*) and the changing input ξ (**Equation 1.2**). Another way to state it is that the solutions to the Mathieu equation define conditions

under which the total change in displacement over time will be zero for a particle with a constantly changing position experiencing a changing but cyclical force.

Fortunately, this also perfectly describes the situation faced by a moving ion experiencing a changing electric field (AC potential). The parameters, *a* and *q*, can be related to the applied DC and AC potentials respectively (**Equations 1.3 & 1.4**) in a mass-to-charge (m/z) dependent manner. Solutions to the Mathieu equation, plotted in **Figure 1.8**, can then be used to define which m/z will be stable at particular voltages and frequencies.

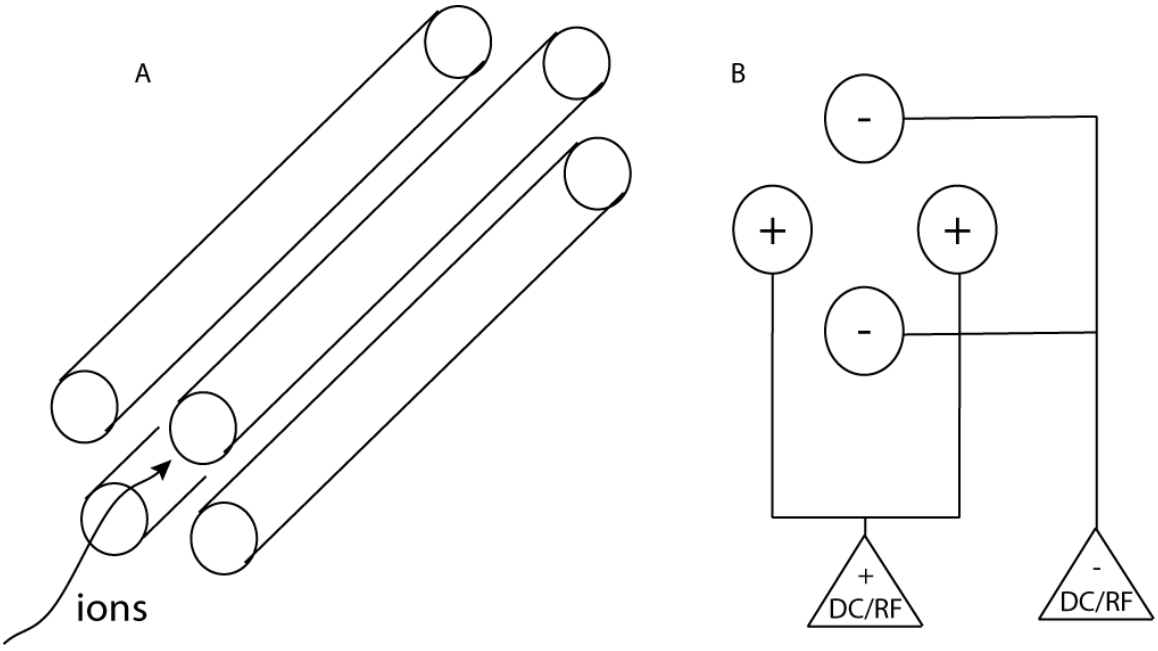

Figure 1.7: The Quadrupole Mass Filter. A) Ions travel through a series of four round rods spaced equally apart in the x and y axis. B) The x rods receive a positive DC bias and a radio frequency AC potential, while the y rods receive a negative DC bias and the same RF AC 180 degrees out of phase with the x rods.

$$\frac{\mathbf{d^2u}}{\mathbf{d\xi^2}} + (a_u - 2q_u \cos 2\xi)u = 0$$ **Equation 1.1: The canonical Mathieu equation.**

$$\xi = \frac{\Omega t}{2}$$ **Equation 1.2: The changing input in the Mathieu equation**

$$a_x = \frac{8eU}{mr_0^2\Omega^2}$$ **Equation 1.3: The a parameter**

$$q_x = \frac{-4eV}{mr_0^2\Omega_0^2}$$ **Equation 1.4: The q parameter**

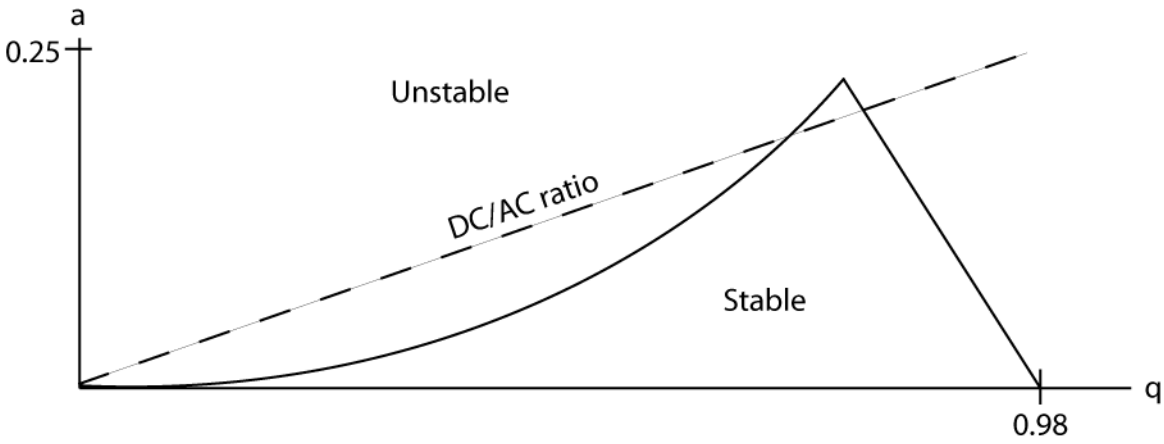

Figure 1.8: The Mathieu stability diagram in a/q space.

There is one particularly important caveat to the stability of ions in the QMF. The stability diagram assumes ions are entering the device with no prior momentum and positioned equidistant from all four rods. This is, of course, never the case. Ions in fact are entering the device with a great deal of kinetic energy having just passed out of the first stages of ion optics and are traveling in a cloud several millimeters in radius.

As a result, in addition to the stability diagram defined by Paul QMFs also have an "acceptance aperture" defined by their coordinates in the x/y plane of the rods and their momentum vector [17]. A consequence of this is that because ions of higher m/z have larger momentum vectors, they experience a smaller x/y acceptance aperture parameter when entering the QMF. As a result, the percent of ions that are successfully

transmitted from one end of the QMF to the other (the transmission efficiency) decreases as m/z increases.

1.3.4.1.4 Dual Cell Linear Ion Trap

The Linear Ion Trap (LIT), introduced by Syka *et al* in 2002, is a multipurpose device that can perform ion trapping, isolation, ion-ion reactions, collisional fragmentation, and low resolution m/z measurement (**Figure 1.9**) [18]. The linear ion trap consists of four parallel rods with hyperbolic inner surfaces spaced 4 mm apart in the *y* dimension and 5.5mm in the *x* dimension. Each rod has three sections that are electrically isolated from one another. The two end sections are equally sized at 12mm long, and the center section is 37mm long. The electrodes in the X dimension have narrow slits of 0.25 mm wide occupying the middle 30 mm of the center trap section. Outside both slits is a conversion dynode held at -15kV to focus positive ions exiting the trap onto the electron multiplier detection system. The trap is filled with a helium bath gas to a pressure of approximately 2.0×10^{-5} Torr.

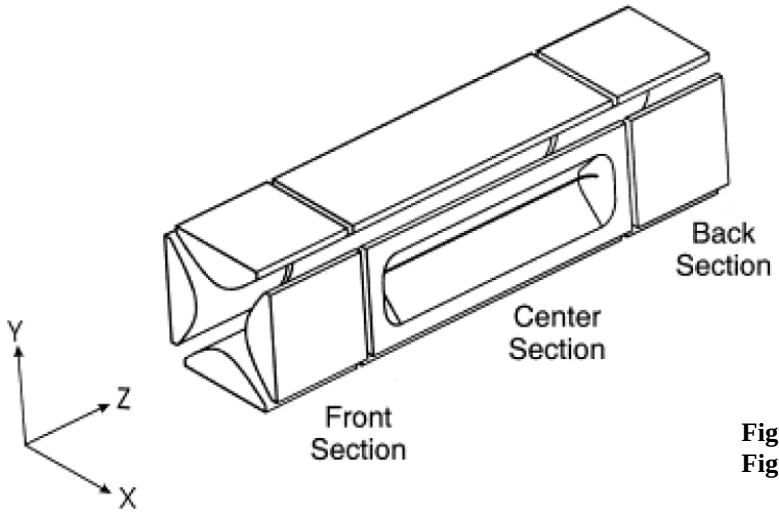

Figure 1.9: The linear ion trap. Figure adapted from [18].

The LIT confines ions based on the same principle as the QMF, and the Matthew equation is use to define the operating potentials on the rods. Notably though, the LIT is operated with no large DC offset so that the *a* value is effectively 0. Only the q value is used to determine the range of ions stably held in the trap. As a result the LIT can function as an ion storage device where many m/z values fall into the range of stable q values and can be confined within the trap. While the LIT stability equation suffers from the same limitation as the QMF, that ions are not entering the trap at rest and centered between the rods, collisional cooling from the He bath gas significantly reduces these effects so that the trapping efficiency is high across a wide m/z range.

To perform functions other than ion storage, the LIT takes advantage of a second phenomenon of ions in a quadrupolar AC electric field. In response to the AC trapping potential ions will orbit within the trap with frequencies in the *x* and *y* dimension that are functions of their m/z. By applying a small supplemental AC voltage, in the *x* dimension only and in resonance with the frequency of a particular m/z, the velocity and radius of oscillation for that species can be increased. This can be used to fragment ions through repeated high velocity collisions with the helium bath gas, eject ions from the trap by exciting them until they exceed the dimensions of the trap in order to isolate a particular m/z, or push ions through the slits in the *x* rods and into the electron multiplier detection system for m/z measurement.

The Fusion has a duel cell LIT. The ideal helium pressure for ion storage and fragmentation (higher pressure) differs slightly from the ideal pressure for m/z analysis (lower pressure), so a two part trap was introduced in 2009 [19]. The front and back

traps have identical dimensions, but the front helium pressure is $7x10^{-5}$ torr while the back is $1x10^{-5}$ torr.

An important advantage of the LIT is that an AC quadrupolar field will trap ions independent of the sign of their charge. In other words both positive and negative ions can be stably held in the trap at the same time. On the Fusion, the LIT is used for Electron Transfer Dissociation reactions. The properties of ETD fragmentation will be discussed later in this chapter. Before an ETD reaction, the peptide cation precursor chosen to be fragmented is held in the back section of the linear trap by DC offsets applied to the sectioned electrodes. Then radical anions of fluoranthene are generated in the first stage of the instrument ion source and directed to center section of the LIT. Finally, the DC potentials separating the cations and anions are removed, and an AC potential is applied to the end lenses of the trap so that cations and anions are stably trapped together in three dimensions. After a user defined period of time a negative DC offset is applied to the center section of the trap and positive potentials to the outer sections to remove the leftover fluoranthene anions while continuing to trap the peptide cation reaction products [20], thus ending the ETD reaction.

1.3.4.1.5 The Orbitrap and C-Trap

Introduced in 2000 by Makarov, the Orbitrap is a high resolution mass analyzer [21]. Prior to m/z measurement ions are passed into the C-trap, another quadrupolar based ion storage device created to aid ion injection into the Orbitrap. The C-trap is filled with nitrogen to a pressure of approximately 1 millitorr. Ions are collisionally cooled by the N_2 bath gas, and then squeezed into a compact packet by high voltages

applied to the four electrodes. After squeezing, the potentials on the electrodes at the entrance to the Orbitrap are removed and the ions are guided through a Z shaped path designed to prevent N_2 from entering the Orbitrap, which is kept at a pressure of $1x10^{-10}$ Torr or below. As ions pass into the Orbitrap a voltage ramp is applied to the central spindle electrode, eventually reaching -5kV. The ions assume a stable orbit around the central electrode. As the ions orbit the spindle electrode they also oscillate axially in the trap at a frequency that is a function of the electric field, the dimensions of the trap, and their m/z (**Equation 1.5**). As ions traverse the trap they induce a current detected by the two outer capping electrodes. This image current can then be Fourier transformed into ion frequencies and subsequently into a mass spectrum [22,23].

$$\omega = \sqrt{\left(\frac{z}{m}\right) k}$$ **Equation 1.5: Oscillation frequency in the orbitrap**

The resolution of the mass spectrum collected by the Orbitrap is determined by the number of periods of oscillation in the trap. Because frequency of oscillation decreases as m/z increases the resolution of the Orbitrap is inversely dependent on m/z. Resolution can be increased by allowing the ions more time to oscillate in the trap (collecting longer image currents). Resolution is defined by **Equation 1.6**, and its importance to the tandem MS experiment will be explained shortly. Transients on the Fusion system range from 32 to 1028ms for 15,000 and 480,000 resolution respectively at m/z 200.

$$\text{Resolution} = \frac{M_1}{M_1 - M_2}$$ **Equation 1.6: Definition of resolution in mass spectrometry**

1.3.4.1.6 Automatic Gain Control

All ion trap instruments, like the Ion Routing Multipole, Linear Ion Trap, and Orbitrap, have a limit to the number of elementary charges that can be stored before the cumulative field generated by those charges begins to distort the electric field created by the trapping electrodes. Additionally, too many charges can interfere with the ability of the ion optics to effectively transmit ions between sections of the instrument. As a result, the number of charges (and therefor ions) that enter the instrument is carefully regulated by a process known as automatic gain control. Prior to every MS scan the instrument allows ions to accumulate in the IRM for a very short period of time (<1ms). It then sends this small ion packet to the low pressure cell of the linear ion trap and ejects it into the electron multipliers. No mass spectrum is recorded, but an estimate of the ion current coming from the electrospray source (in charges per second) is made. The instrument then calculates the ion accumulation time necessary to reach the ideal number of charges, called the "target", and uses that time to accumulate ions for the next analytical scan [24]. One important consequence of this system is that as the charge states of peptides increase, the real number of those peptides accumulated for an MS experiment will go down. Now we will describe each individual step of the tandem MS experiment in more detail.

1.3.4.2 MS^1

The first step of the tandem MS experiment is acquisition of a high resolution mass spectrum of the intact peptides inside the Orbitrap. Peptide cations generated by electrospray are guided through the instrument by the ion optics and accumulated in the IRM to a target of $2x10^5$ charges. The ions are then passed into the C-Trap and injected into the Orbitrap for mass analysis.

Figure 1.10 is an MS^1 mass spectrum of tryptic peptides acquired in the Fusion's Orbitrap mass analyzer with a 250 millisecond transient. High resolution is of particular importance in the MS^1 scan during peptide analysis in order to determine the charge state and therefore the intact mass of a particular peptide [25]. Charge states of peptide ions are determined by the delta m/z between isotopic peaks for a particular peptide. Since 1.07% of all naturally occurring carbon atoms are the C^{13} isotope, and a given MS^1 scan contains hundreds if not thousands of each peptide species, some of the signal for a given peptide will come from molecules containing one or more C^{13} atoms and appear at a higher m/z. This range of peaks for a single molecular species is called the "isotopic envelope". The peaks in the isotopic envelope will differ by 1.008/z (the mass of a neutron divided by the number of charges on the peptide). Sufficient resolution makes it possible to determine the charge state of a peptide from its observed isotopic envelope, and then calculate the intact mass of the peptide.

Based on a single MS^1 scan multiple precursors will be targeted for MS^2 analysis. The Fusion software allows the user to target a predefined number of precursors (Ex. Top

10), or perform as many MS^2s as possible in a fixed time window (usually 2 to 3 seconds) before collecting a new MS^1 Spectrum.

1.3.4.3 Identification, Isolation, and Fragmentation of Specific m/z species

Once a high resolution MS^1 has been acquired, species appearing in it are targeted for fragmentation and MS^2 analysis in a data dependent manner. The instrument control software selects species appearing in the MS^1 scan in order of decreasing abundance for further study by MS^2. To increase the number of precursor ions selected for fragmentation a dynamic exclusion list is employed, meaning that once targeted for MS^2 analysis a particular species cannot be targeted again for a user defined period of time.

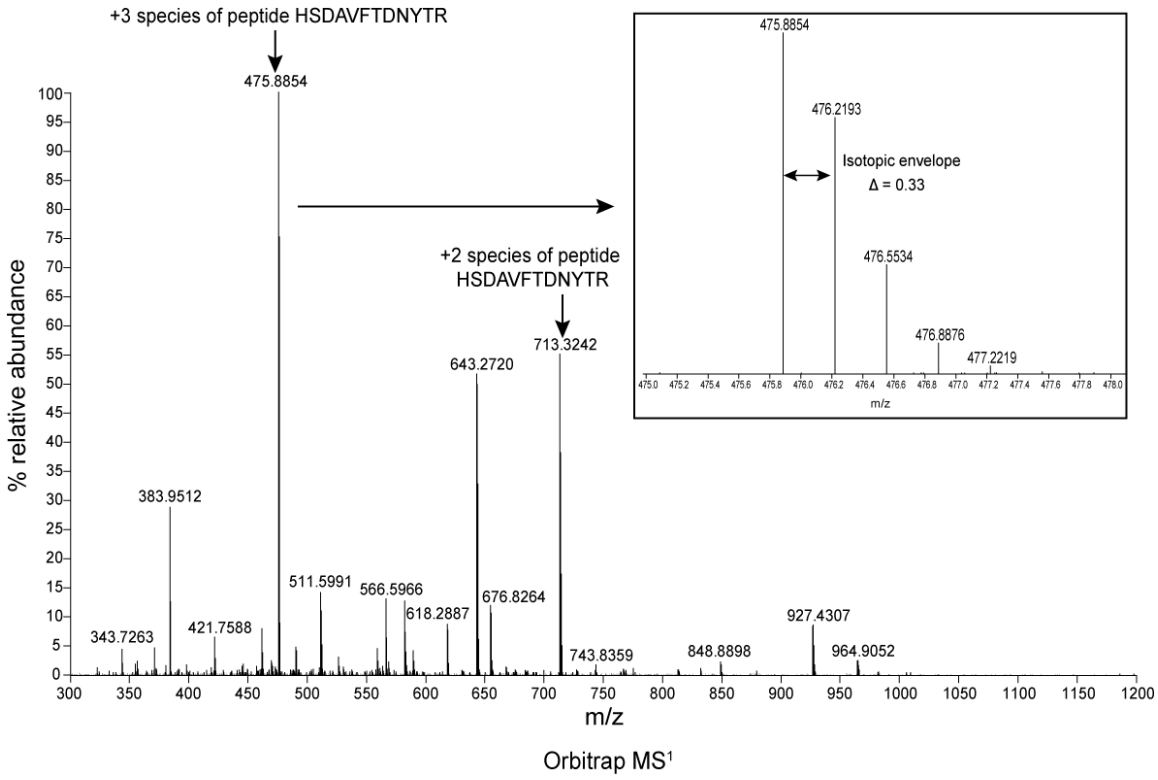

Figure 1.10 High Resolution MS^1. An Orbitrap spectrum of complex sample of tryptic peptides. The blown up portion shows the isotopic envelope around the +3 charge state of the peptide HSDAVFTDNYTR at 475.8846 m/z.

In the Fusion instrument, peptides of a specified m/z are isolated by the Quadrupole Mass Filter using a user defined m/z window and accumulated in the Ion Routing Multipole. The peptide of interest can then be fragmented by three mechanisms: high energy collisions with N_2 in the IRM (HCD), lower energy collisions with He in the high pressure cell of the Linear Ion Trap (CID), or reaction with radial fluoranthene ions in the high pressure cell of the Linear Ion Trap (ETD). Each option has advantages and disadvantages, and their effectiveness at producing informative sequence ions is often peptide specific.

Figure 1.11 shows the 6 types of backbone peptide fragments that can be used to sequence a peptide by mass spectrometry [4]. The collisional dissociation methods HCD and CID produce predominantly *b* and *y* type ions by fragmenting the peptide bond, while ETD produces predominantly *c* and *z* type ions by fragmenting the N-Cα bond. **Figure 1.14** shows an HCD, CID, and ETD MS^2 spectrum of the peptide **HSDAVFTDNYTR**.

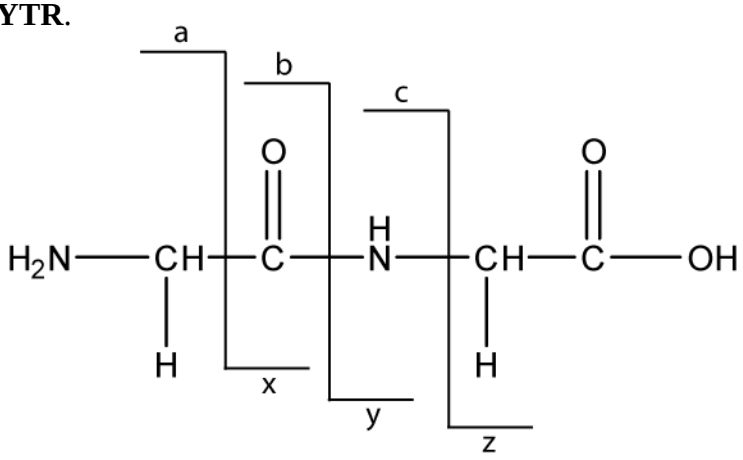

Figure 1.11: The six types of complementary fragment peptide ions.

For many peptides, HCD and CID produce nearly identical MS2 spectra, but there are some important differences in the two mechanisms of fragmentation. HCD takes place in the Ion Routing Multipole where the segmented electrodes create a high voltage potential gradient across the cell. As peptide cations are accelerated through the potential gradient they collide with the N_2 bath gas [26]. Some of the kinetic energy imparted to peptide ions by these collisions is converted to vibrational energy and facilitates peptide fragmentation. Because of the higher energy of the collisions, in addition to backbone cleavages HCD will also produce fragments of the amino acid side groups called immonium ions. During HCD fragments can also undergo additional collisions that cause secondary fragmentation events.

CID is a collisional fragmentation technique that uses the characteristic frequency of a particular m/z, as described during the discussion of the operating principle of the Linear Ion Trap, to resonantly excite peptide cations into collisions with the He bath gas of the LIT [27] . As in HCD the collisions impart kinetic energy that can be translated to vibrational energy and activate the peptide bond for fragmentation [28], but the collisions are of lower energy than HCD because of the lighter bath gas (He vs N_2), and lower velocity of the cations. Additionally, because CID relies on resonant excitation targeted at a particular m/z, fragments are much less likely to undergo secondary fragmentation events because they are out of resonance with the excitation voltage and no longer being accelerated into the He bath gas.

Both HCD and CID suffer from a critical disadvantage. They cause the loss of labile O-linked post translational modifications like phosphorylation and O-linked

Glycosylations, and example of an O-GlcNAc loss is seen in **Figure 1.12**. This loss can be used as a fingerprint for peptides containing an O-linked modification, but the loss makes it impossible to definitively map the modification to a single amino acid. Additionally, because the loss of the modification dominates the MS^2 spectrum there are fewer informative sequence ions. This makes it less likely that the spectra can be successfully used to determine an unknown peptide's sequence.

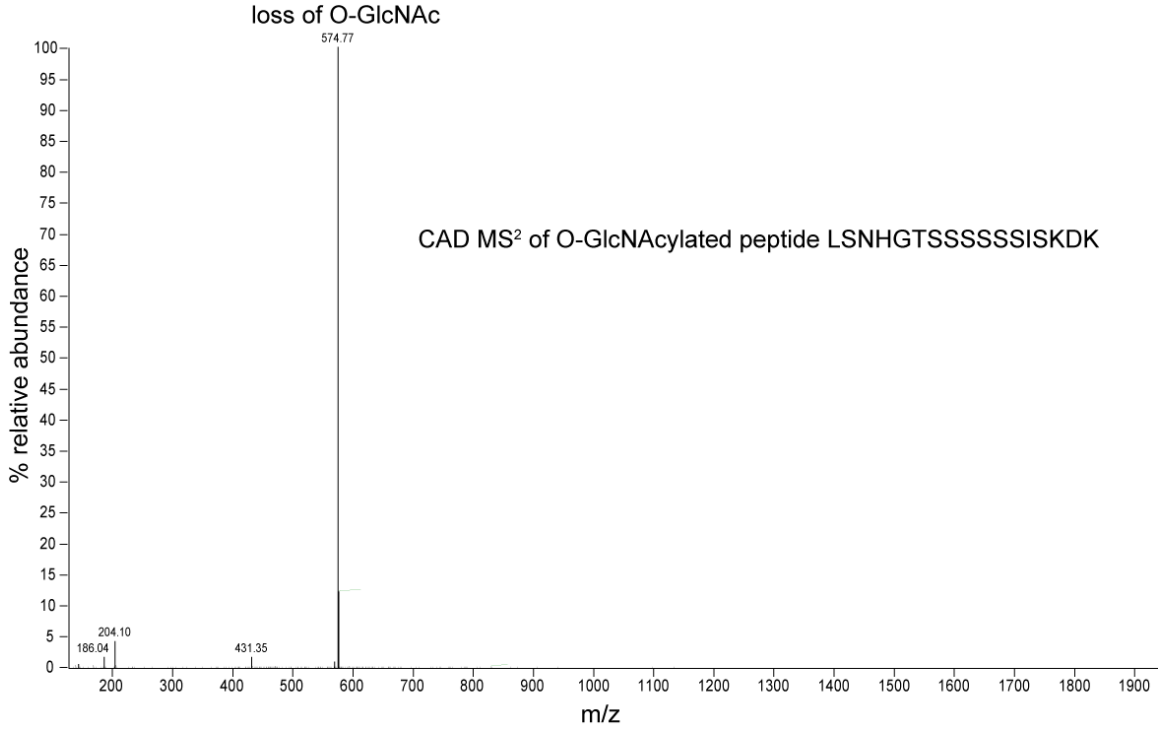

Figure 1.12: CAD of an O-GlcNAcylated peptide. The dominant peak in the MS2 spectrum is the loss of the neutral GlcNAc fragment from the peptide while there is essentially no backbone fragmentation to provide sequence information.

Electron Transfer Dissociation, developed in 2004 Hunt and colleagues, solves this problem [20]. During ETD a radical anion reagent is generated and simultaneously trapped with the peptide cations in the linear ion trap. Following the mechanism shown in **Figure 1.13** the radical electron is transferred from the anionic reagent to a molecular

orbital in the cationic peptide, then relaxes to one of the carbonyl carbons on the peptide backbone. The extra electron density makes the carbonyl carbon highly basic, and it abstracts a proton from a nearby charged site; likely the N-terminal amine or the side chain of a lysine or Arginine. The electron transfer is exothermic, and the intra-peptide proton transfer activates the N-Cα bond and facilitates fragmentation of the peptide. ETD produces *c* and *z* type ions, but can also result in side chain fragmentations [29].

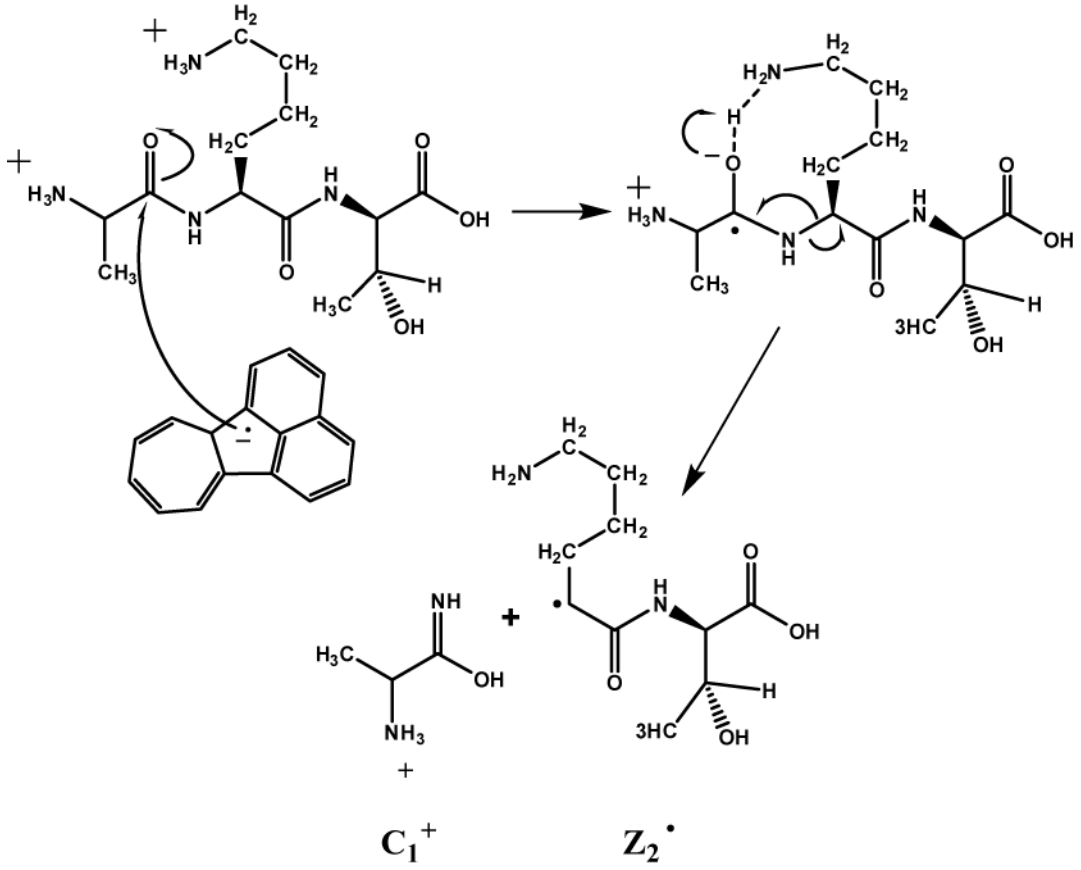

Figure 1.13: The ETD mechanism.

Importantly, not every electron transfer results in peptide cleavage. Termed ETnoD, these reactions create species in the MS^2 spectra with the same mass as the precursor but the charge reduced by 1 or more. One known cause of ETnoD is proline.

Because the N-Cα bond is bridged by the proline side chain its breakage does not result in two separate ions. Proline does not account for all instances of ETnoD, but the other causes are not yet well understood. It is known that there is a high correlation between charge density and ETD efficiency. Peptides with fewer than three charges, or m/z > 600 tend to not yield informative ETD spectra.

A key advantage of ETD is the analysis of peptides containing O-linked post translational modifications like phosphorylation. ETD does not cause loss of these modifications making it possible to assign them to specific amino acids.

1.3.4.4 MS^2

Following fragmentation by one of the possible mechanism the peptide fragment ions are sent to one of the mass analyzers (Orbitrap or LIT) for MS^2 analysis. High resolution MS2 spectra take more time to acquire and are generally less sensitive but can provide a higher degree of certainty in peptide identification. They are especially valuable for large highly charged peptides that yield more complex fragmentation spectra. It has been shown that high resolution MS^2s typically provide more successful peptide identifications from complex samples, while low resolution MS^2s are more effective for finding modified peptides of very low abundance.

1.3.5 Data Analysis

The final phase of the LC-MS/MS experiment is data analysis. In a single LC run of 90 minutes thousands of MS/MS experiments will be performed. Each of these experiments has its own data set that includes the intact mass of an unknown peptide, and fragment masses from that peptide. The experiment is successful if that data set can be

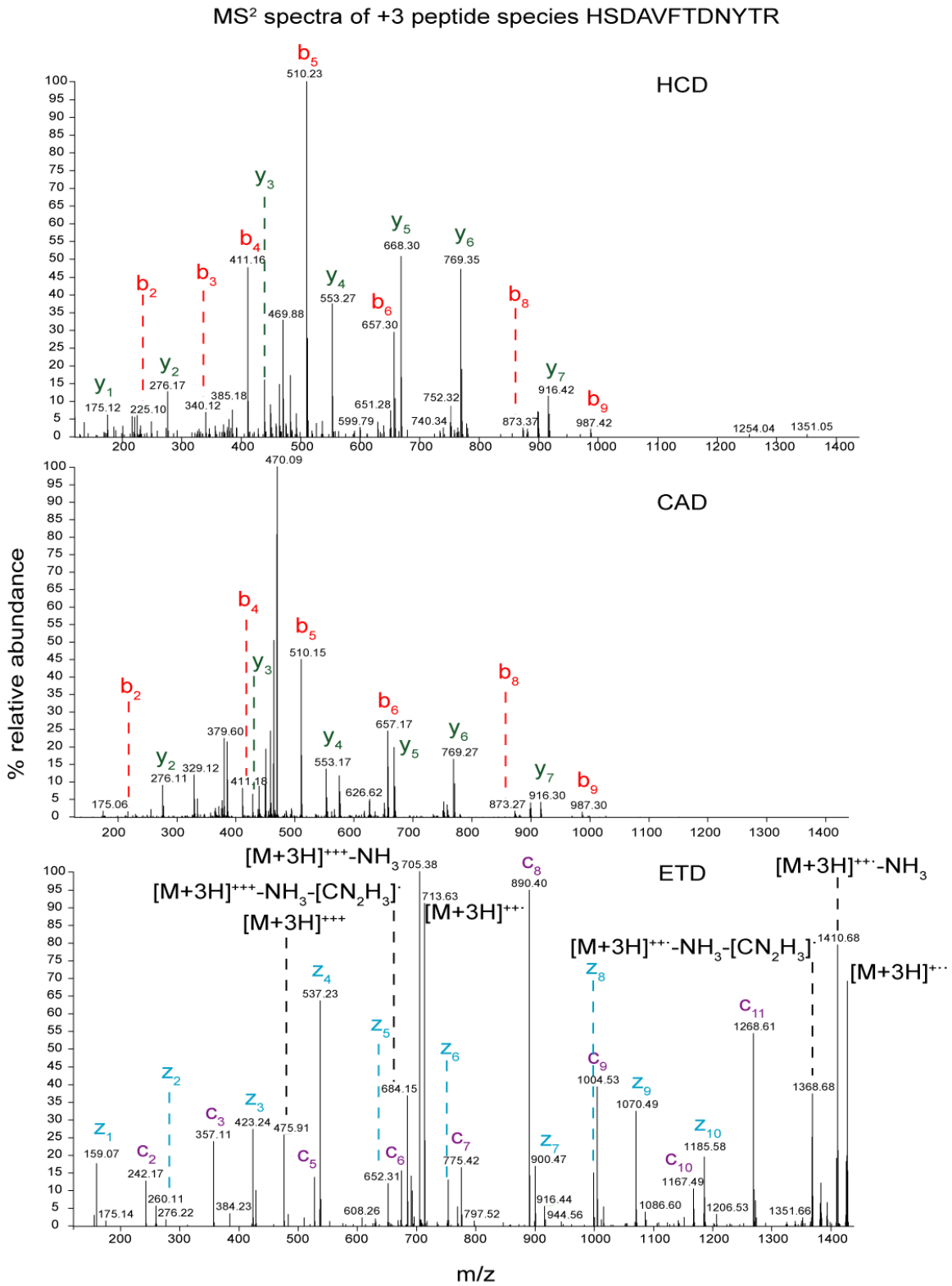

Figure 1.14: Three different fragmentation spectra of one peptide.

used to determine the sequence of the unknown peptide. This can be done manually by using the mass differences between the backbone fragments (b,y,c,or z ions) to determine the component amino acids of the peptide, or by a computer algorithm that compares the observed fragment masses against a set of theoretical backbone fragments generated from a database of possible peptides.

There are a number of commercially available MS/MS data search algorithms. Only two were applied to the data in this dissertation, MASCOT and SEQUEST. Both programs follow a similar scheme [30,31,32]. The software begins with a database of protein sequences provided by the user. Each protein in the database is broken into peptides based on a user specified enzyme (ex. Trypsin cleaves proteins after R and K). Each peptide has an intact mass, and a list of possible fragment masses. Theoretical peptide masses from the database are compared to the list of precursor masses from every MS/MS experiment. Matches are then compared on the fragment mass level. Database peptides with a high degree of correlation to observed MS/MS spectra are reported as "peptide spectral matches" (PSM).

The LC-MS/MS datasets used in this work were subjected to very different interpretation mechanisms. All of the peptides reported in chapter two were manually sequences from maximum accuracy of PTM site assignment. This was possible because the analysis was focused on a limited set of peptides from a single protein. In chapter three peptides were assigned to MS/MS spectra entirely by software. This was necessary because the samples contained more than 1,000 different proteins, and we had no prior knowledge of which were important to the results.

1.4 Advantages of MS/MS on the Fusion Platform

The outline of the tandem MS experiment used in this dissertation is always the same: MS^1 -> peptide selection -> isolation -> fragmentation -> MS^2. In a single LC run of 90 minutes thousands of these experiments will be performed. The greater the portion of unknown peptides selected for MS/MS by the instrument software, and the greater the number of MS/MS spectra that can be correctly matched with a peptide, the more comprehensive the analysis. In other words there are two properties of the LC-MS/MS experiment directly related to the mass spectrometer that determine its overall success. First, the rate at which the instrument can perform MS/MS, and second, the likelihood of those MS/MS to contain sufficient numbers of informative fragment ions to be interpretable.

The Fusion platform represents a significant advance in MS instrumentation in both these areas. The Fusion has a much higher MS/MS acquisition rate than previous generations of Orbitrap-Linear Ion Trap hybrid instruments, in part because of improvements in individual instrument components that reduce the time of each event in the MS/MS experiment, but largely because of its unique architecture [12]. In older instruments, the Linear Ion Trap was responsible for most of the subcomponents of the MS^2 scan event. Precursor Ions were accumulated in the LIT, Isolated in the LIT, Fragmented in the LIT, and mass analyzed in the LIT. As a result, each MS^2 event began at the completion of the prior event. The time to perform N MS^2 experiments was simply the sum of the length of each individual experiment. On the Fusion platform however, ion isolation and accumulation are now performed by the QMF and IRM respectively.

Ions are then passed to the LIT if CID or ETD fragmentation is taking place. MS^2 events can now overlap, with ion accumulation for the Nth event beginning as soon as the ions for the N-1 event have been passed to the LIT or Orbitrap.

A second advantage of the Fusion instrument is the flexible nature of its method control software. Combined with the three modes of fragmentation (HCD, CID, ETD) this gives the user the ability to treat precursor ions very differently from one scan to the next to maximize the quality of the MS/MS data. This flexibility will be discussed in more detail when it becomes relevant in chapter three.

1. 5 Conclusion

In the next two chapters we will see the same tandem MS experiment used to answer two very different biological questions. In chapter two we look for evidence that the Arabidopsis protein RGA is post translationally modified by the protein SPY, and in chapter three we try to identify the proteins selectively binding to a region of DNA in the Immunoglobulin locus of B cells, but in both projects we detect the proteins of interest by using MS/MS spectra to identify tryptic peptides.

Over the last several decades there have been dramatic improvements in MS instrument speed and sensitivity. These improvements have made the analysis of increasingly challenging samples possible. Some of the proteoforms identified in chapter two likely could not have been detected on previous generation instruments.

However, during the RGA analysis we knew exactly what we were looking for (modified forms of specific peptides from the protein RGA). As a result every aspect of the experiment was designed to maximize the chance of finding and characterizing those

peptides. Chapter two illustrates the power of protein MS when directed towards a narrow target, but chapter three articulates some of the biggest challenges remaining in protein MS.

During the SHM project our search was far less specific, and so the MS experiment and data analysis was designed to capture the widest possible range of proteins. Unfortunately, the vast majority of MS/MS experiments could not be successfully linked to a peptide sequence, and this is normal for experiments of this type. It is not clear whether this is due to a failure in the way peptides are captured, fragmented, and detected by the instrument, or a failure of the data interpretation software. Probably both are to blame. Data interpretation algorithms are beyond the scope of this research, but MS/MS experiment design is definitely not. A secondary goal of the work in chapter three then, is to identify trends in the MS/MS data that suggest areas for method improvement.

1.6 References

1. Jurica, M. S. & Moore, M. J. Pre-mRNA Splicing: Awash in a Sea of Proteins. *Mol. Cell* **12,** 5–14 (2003).

2. Uy, R. & Wold, F. Posttranslational covalent modification of proteins. *Science (80-.).* **198,** (1977).

3. Walsh, C. T., Garneau-Tsodikova, S. & Gatto, G. J. Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications. *Angew. Chemie Int. Ed.* **44,** 7342–7372 (2005).

4. Biemann, K. Mass Spectrometry of Peptides and Proteins. *Annu. Rev. Biochem.* **61,** 977–1010 (1992).

5. Susan E. Martin, †, Jeffrey Shabanowitz, Donald F. Hunt, *,‡ and & Marto, J. A. Subfemtomole MS and MS/MS Peptide Sequence Analysis Using Nano-HPLC Micro-ESI Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. (2000). doi:10.1021/AC000497V

6. Marginean, I. *et al.* Analytical Characterization of the Electrospray Ion Source in the Nanoflow Regime. *Anal. Chem.* **80,** 6573–6579 (2008).

7. Ficarro, S. B. *et al.* Improved Electrospray Ionization Efficiency Compensates for Diminished Chromatographic Resolution and Enables Proteomics Analysis of Tyrosine Signaling in Embryonic Stem Cells. *Anal. Chem.* **81,** 3440–3447 (2009).

8. Whitehouse, C. M., Dreyer, R. N., Yamashita, M. & Fenn, J. B. Electrospray interface for liquid chromatographs and mass spectrometers. *Anal. Chem.* **57,** 675–679 (1985).

9. Cech, N. B. & Enke, C. G. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrom. Rev.* **20,** 362–387 (2001).

10. Schnier, P. D., Gross, D. S. & Williams, E. R. On the maximum charge state and proton transfer reactivity of peptide and protein ions formed by electrospray ionization. *J. Am. Soc. Mass Spectrom.* **6,** 1086–1097 (1995).

11. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S. & Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci.* **83,** 6233–6237 (1986).

12. Senko, M. W. *et al.* Novel Parallelized Quadrupole/Linear Ion Trap/Orbitrap Tribrid Mass Spectrometer Improving Proteome Coverage and Peptide Identification Rates. *Anal. Chem.* **85,** 11710–11714 (2013).

13. Michalski, A. *et al.* Ultra high resolution linear ion trap Orbitrap mass spectrometer (Orbitrap Elite) facilitates top down LC MS/MS and versatile peptide fragmentation modes. *Mol. Cell. Proteomics* **11,** O111.013698 (2012).

14. Giles, K. *et al.* Applications of a travelling wave-based radio-frequency-only stacked ring ion guide. *Rapid Commun. Mass Spectrom.* **18,** 2401–2414 (2004).

15. Campana, J. E. Elementary theory of the quadrupole mass filter. *Int. J. Mass Spectrom. Ion Phys.* **33,** 101–117 (1980).

16. March, R. E. An Introduction to Quadrupole Ion Trap Mass Spectrometry. *J. Mass Spectrom.* **32,** 351–369 (1997).

17. Dawson, P. H. The acceptance of the quadrupole mass filter. *Int. J. Mass Spectrom. Ion Phys.* **17,** 423–445 (1975).

18. Schwartz, J. C., Senko, M. W. & Syka, J. E. . A two-dimensional quadrupole ion trap mass spectrometer. *J. Am. Soc. Mass Spectrom.* **13,** 659–669 (2002).

19. Olsen, J. V *et al.* A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol. Cell. Proteomics* **8,** 2759–69 (2009).

20. Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9528–33 (2004).

21. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **40,** 430–443 (2005).

22. Scigelova, M., Hornshaw, M., Giannakopulos, A. & Makarov, A. Fourier transform mass spectrometry. *Mol. Cell. Proteomics* **10,** M111.009431 (2011).

23. Lange, O., Damoc, E., Wieghaus, A. & Makarov, A. Enhanced Fourier transform for Orbitrap mass spectrometry. *Int. J. Mass Spectrom.* **369,** 16–22 (2014).

24. Schawartz, J. C., Zhou, X.-G. & Bier, M. E. Method and apparatus of increasing dynamic range and sensitivity of a mass spectrometer. (1996).

25. Mann, M. & Kelleher, N. L. Precision proteomics: the case for high resolution and high mass accuracy. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 18132–8 (2008).

26. Olsen, J. V *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4,** 709–712 (2007).

27. Louris, J. N. *et al.* Instrumentation, applications, and energy deposition in quadrupole ion-trap tandem mass spectrometry. *Anal. Chem.* **59,** 1677–1685

(1987).

28. Wysocki, V. H., Tsaprailis, G., Smith, L. L. & Breci, L. A. Mobile and localized protons: a framework for understanding peptide dissociation. *J. Mass Spectrom.* **35,** 1399–1406 (2000).

29. Good, D. M., Wirtala, M., McAlister, G. C. & Coon, J. J. Performance characteristics of electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics* **6,** 1942–51 (2007).

30. Lewis Y. Geer, *,† *et al.* Open Mass Spectrometry Search Algorithm. (2004). doi:10.1021/PR0499491

31. Michael J. MacCoss, †,‡, Christine C. Wu, †,‡ and & John R. Yates, I. Probability-Based Validation of Protein Identifications Using a Modified SEQUEST Algorithm. (2002). doi:10.1021/AC025826T

32. Elias, J. E., Haas, W., Faherty, B. K. & Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2,** 667–675 (2005).

Chapter 2: Mapping a Novel O-Glycosylation on the Protein RGA

2.1 Introduction

This chapter describes our investigation into the interaction between two proteins, RGA and SPY. Both were previously known to be growth regulating proteins in *Arabidopsis thaliana*, and it was suspected that SPY post-translationally modified RGA, but the effects of SPY on RGA had never been directly studied. Our work led to a surprising discovery about SPY's function and a new post translational modification important to RGA, but to understand the importance of these results it is necessary to first explain how RGA came to be subject of our investigation.

2.1.1 Origin of Interest in the Protein RGA

During the 1960s and 1970s world food production outpaced population growth, particularly in the developing areas of South America, Asia, and Africa. This increase in agricultural production, since termed "The Green Revolution", prevented a worldwide food shortage and the subsequent mass starvation that had been widely predicted during the mid-20th century; most notably by the famous biologist Paul Erlich in his 1968 book "The Population Bomb" [1]. A major driver of this increase in agricultural production was the development of improved varieties of wheat, and later corn and rice, by plant geneticist Norman Borlaug, who was awarded the Nobel Peace Prize in 1970 [2]. These new plants exhibited a high yield semi-dwarf phenotype. They had increased grain mass, but also grew shorter stalks that required less water and were less likely to be damaged during storms.

The more productive and robust phenotypes of these new plants spurred research into the cellular signaling pathways mediating plant growth, and the initial focus of these investigations was on a plant hormone called gibberellin. Interest in gibberellin began in the early 20th century in Japan because of a disease called *bakanae* that caused rice plants to grow unusually tall [3]. The disease was eventually traced to a fungus named *Gibberella fujikuroi*. In 1938 two scientists at the University of Tokyo, Yabuta and Sumiki, were able to isolate growth stimulating compounds, which they named "gibberellins", from the fungus [3]. More than a decade passed before gibberellin caught the attention of western scientists, but its ability to enhance the growth of certain commercial crops led to the common use of gibberellin in the agricultural industry by the late 1950s [4].

Curiously, when tested on the new plants developed by Borlaug, gibberellin did not have its usual growth enhancing effect. These mutants of wheat, rice, and corn were gibberellin insensitive. This lack of response to gibberellin provided an important clue for researchers trying to understand the growth pathways that were altered in these plants. By generating a series of gibberellin insensitive mutants in different plant species and characterizing the function of the altered proteins through a combination of biochemical assays and sequence homology, researchers over the next several decades were able to form a near complete picture of how the gibberellin signaling pathway functioned.

While genetic experiments continued in corn, wheat, and rice plants, *Arabidopsis thaliana* also became a popular laboratory model for studying gibberellin signaling. Multiple Arabidopsis plants can be grown in a single petri dish, the plants grow well

under fluorescent lights, the full life cycle is only six weeks, and the genome is relatively small at 120mB divided into 5 chromosomes [5]. A number of Arabidopsis mutants with altered gibberellin responses and growth phenotypes were created in the 1980s and 1990s, and a few provided key insights into the gibberellin signaling pathway [6].

The Arabidopsis mutant *gai* was identified as a gibberellin insensitive dwarf, and the associated wild type protein GAI was determined to be a negative regulator of gibberellin signaling [7]. GAI was later shown to be an ortholog of RHT, D8, and SLR [18]. These proteins are mutated in the high-yield dwarf phenotypes of wheat, corn, and rice respectively that were created during the green revolution [9]. All of these proteins have nuclear localization signals and sequence homology to known transcription factors, and GAI had been shown to be a repressor of gibberellin responses.

GAI is one of a family of proteins named for a shared N-terminal amino acid sequence, DELLA [10]. DELLAs are putative transcription factors, with a nuclear localization signal, and are highly conserved throughout plant species (RHT, D8, and SLR1 are also DELLA proteins). In Arabidopsis there are five known DELLA proteins: GAI, RGA, RGL1, RGL2, and RGL3. All known DELLAs are negative regulators of Gibberellin signaling, but it was unclear whether DELLAs interact directly with DNA or suppress growth by binding to other transcription factors [11,12].

In 2005 a soluble gibberellin receptor protein was identified, GID1 [13]. GID1 was shown to enter the nucleus in the presence of bioactive gibberellin and bind to SLR1. Furthermore, another gibberellin insensitive mutant, *gid2*, was found to have a defective

subunit of an SCF E3 ubiquitin ligase, and SLR1 was shown to be degraded only in plants with wild type (WT) GID1 and GID2.

Based on data from these mutants and others it was hypothesized that gibberellin signaling functioned through the following mechanism; bioactive gibberellin bound to the soluble cytoplasmic gibberellin receptor protein GID1, which was then able to enter the nucleus and bind to DELLA proteins. This targeted DELLAs for degradation by the ubiquitin/proteasome pathway, and opened pro-growth genes to transcription (**Figure 2.1**) [14].

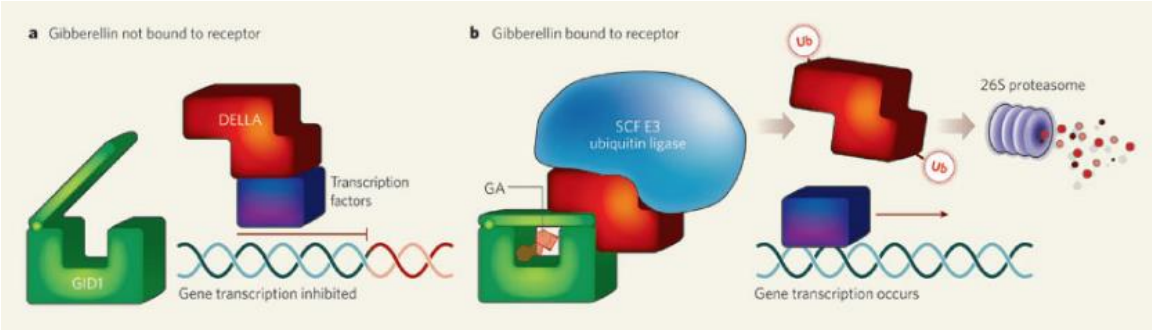

Figure 2.1: The current model of gibberellin signaling. a) In the absence of the gibberellin receptor GID1 DELLA proteins inhibit the transcription of growth related genes. b) When gibberellin binds to its receptor the complex can enter the nucleus and target the DELLLA proteins for degradation, thus allowing growth genes to be transcribed. Figure adapted from [14].

2.1.2 Unanswered Questions about Gibberellin Signaling

While the relationship between gibberellin, its receptor, and the DELLA proteins was becoming increasingly clear, the function of another protein modulator of gibberellin responses, SPY, remained a mystery. First identified in 1993, the *spy* (spindly) mutation caused a "gibberellin overdose" phenotype, in which the gibberellin signaling pathway seemed to be constitutively turned on [15,16]. The dwarf phenotype created by mutations in the synthesis pathway for gibberellin could be partially rescued by mutations

to SPY. SPY was clearly an actor in gibberellin signaling, but its relationship to the DELLAs was unknown. A breakthrough came in 1997 when the mammalian O-GlcNAc transferase (OGT) enzyme was identified [17]. Shortly afterwards SPY was tentatively identified as an OGT as well based on sequence homology with the mammalian protein. In 2002 a second OGT homolog in Arabidopsis, SEC (SECRET AGENT), was also identified [18,19]. Gibberellin researchers began operating under the assumption that SEC and SPY were O-GlcNAc transferases; and that O-GlcNAcylation of the DELLA proteins modulated their interaction with the gibberellin-receptor complex and subsequent degradation. However, there was no direct evidence of the O-GlcNAcylation of any of the DELLAs by SEC or SPY.

2.1.3 What is O-GlcNAc?

O-GlcNAc is O-linked N-Acetyl Glucosamine (**Figure 2.2**). An O-GlcNAc transferase (OGT) is an enzyme that catalyzes the addition of N-Acetyl Glucosamine (GlcNAc) to the hydroxyl group of serine and threonine residues in a protein. O-GlcNAc can be removed from proteins by a second enzyme, an O-GlcNAcase. O-GlcNAc is a post translational modification; altering the chemical properties of a protein in response to cellular processes without changing the amino acid sequence. The donor substrate for O-GlcNAcylation is UDP-GlcNAc, and its concentration is directly tied to glucose levels in the cell (**Figure 2.2**). As a result, O-GlcNAc modification has been proposed to be a nutrient sensing mechanism for cells.

First identified by Hart et al. in 1984 [20], O-GlcNAc is believed to be present in all eukaryotic cells and an essential component of many cellular processes [21].

O-GlcNAc is distinct from the branched O and N-linked glycans commonly seen in cells in that it is a dynamic single sugar modification.

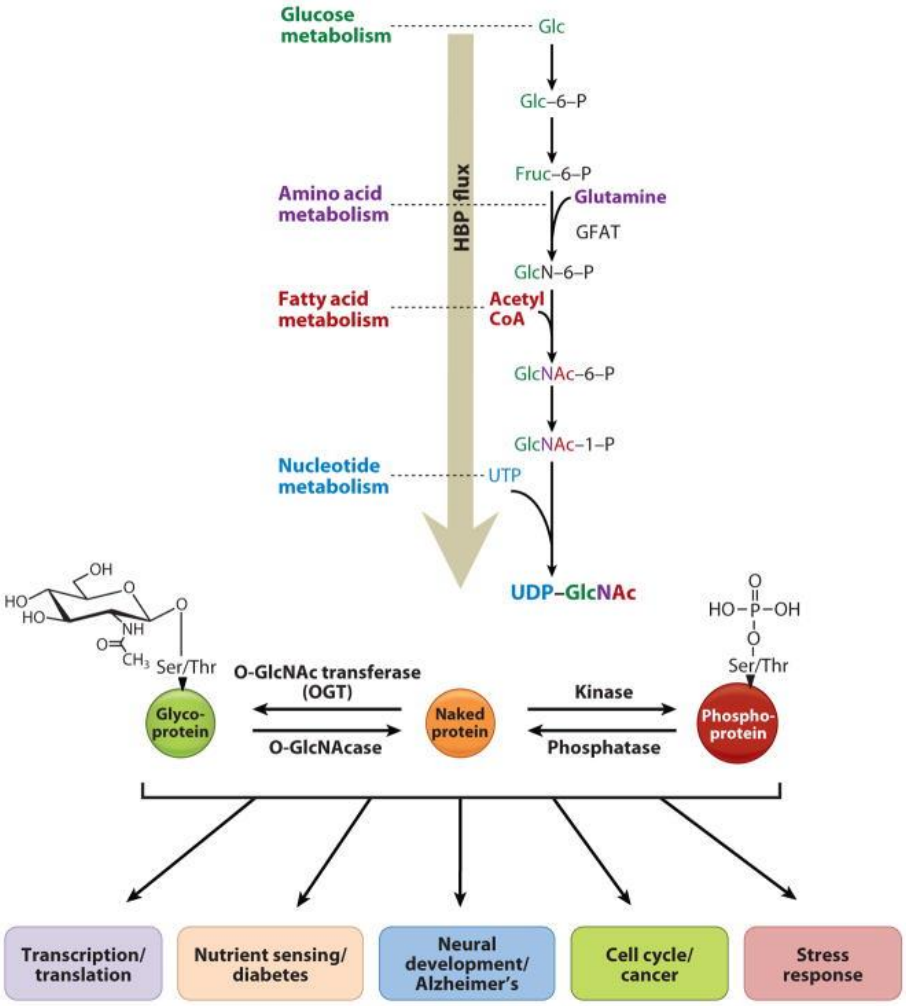

Figure 2.2: O-GlcNAc

Because of its dynamic nature O-GlcNAc is often compared to phosphorylation, another ubiquitous post-translational modification [22]. Some studies have demonstrated an interplay between O-GlcNAcylation and phosphorylation, and the modifications are known to share some amino acid sites [23].

LC-MS/MS is the most effective tool for mapping protein post-translational modifications, but several challenges exist when studying O-GlcNAc. The primary challenge is that, as described in chapter 1, O-linked modifications are labile under the collisional dissociation methods most commonly used for mass spectrometry based peptide sequencing. One approach to bypass this problem is a chemical reaction developed by the Hart lab. A beta-elimination/Michael addition (BEMAD) reaction is used to replace the O-link sugar with a dithiothreitol chemical tag that is not labile during collisional dissociation [24]. Unfortunately this approach suffers from several drawbacks. First, the chemical reaction poorly distinguishes between O-linked modifications on serine and threonine. While there appears to be a kinetic difference between the beta-elimination of O-GlcNAc and Phosphorylation, it is still impossible to guarantee only formerly O-GlcNAcylated residues will be tagged. There is also no data to suggest that the technique can distinguish between different O-linked sugars.

A second problem is that because of its dynamic nature, the relative concentration of any O-GlcNAcylated peptide is very low, making detection in complex samples difficult. The same challenge exists for phosphorylated peptides, but immobilized metal affinity chromatography can be used to enrich for phosphorylated peptides and raise their relative abundance in a complex sample. While several attempts have been made to enrich for O-GlcNAc [25,26], no reliable technique has been developed. To worsen matters, there is evidence that the ionization of glycosylated peptides is suppressed by their unmodified counterparts, making them even more difficult to detect [27].

2.1.4 Prior Mass Spectrometry Analysis of RGA

In 2008 the Hunt lab began collaborating with the Sun lab at Duke University to confirm the identity of SEC and SPY as O-GlcNAc transferases and to identify any sites of O-GlcNAc modification on DELLA proteins. O-GlcNAc is not labile when peptides are fragmented by electron transfer dissociation. During ETD the modification remains on the peptide being fragmented. This makes it possible to map the sugar(s) to a specific amino acid(s) even when multiple possible modification sites are present on a single peptide.

A two part approach was used to confirm that SEC O-GlcNAcylated RGA and determine the sites of modification. Samples of RGA expressed in tobacco plants with or without SEC, or RGA purified from Arabidopsis plants with mutations in SEC, were digested with trypsin or Asp-N and analyzed by LC-MS/MS using electron transfer dissociation (ETD). Initial experiments were done by Sushmit Maitra and Namrata Udeshi, but the bulk of the analysis was completed by Andrew Dawdy and described in his dissertation [28]. I completed some supplementary analysis to Andrew's work, and our findings, briefly summarized here, were published in [29].

RGA is extensively O-GlcNAcylated by SEC (**Figure 2.3**). The majority of modifications are found in several serine and/or threonine rich regions on the N-terminal half of the protein, although additional isolated sites were also identified albeit at much lower frequency (**Table 2.1**). Biological assays performed by the Sun lab indicated a role for O-GlcNAcylation of RGA by SEC in modulating Gibberellin signals during plant growth. O-GlycNAcylation of RGA by SEC blocks its association with other signaling

proteins and reduces it growth retarding effects. Importantly, mutations to SEC reduce plant growth.

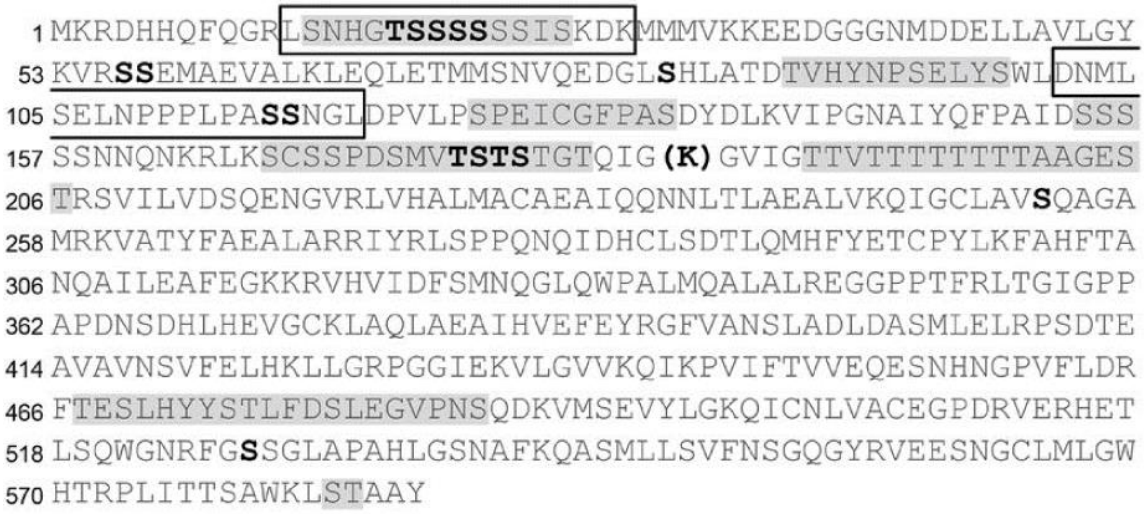

Figure 2.3: Sites of GlcNAcylation on RGA. Bolded residues indicate sites of O-GlcNAc modification confirmed by MS/MS. Highlighted areas are modified, but the exact site could not be determined. All sites were observed on RGA expressed with SEC in tobacco. Boxed regions were also observed on RGA-TAP purified from WT Arabidopsis. Figure adapted from [29].

Peptide	Abundance of Forms containing:							
	O-GlcNAc	Phosphate	O-Hexose	O-GlcNAc & Phosphate	O-GlcNAc & O-Hexose	Phosphate & O-Hexose	O-GlcNAc, Phosphate, & O-Hexose	Total
DHHQFQGRLSNHGTSSSSSSISK	74%	<1%	4%	3%	3%	<1%	<1%	84%
DELLAVLGYKVRSSEM	28%	0	1%					29%
DGLHLAT	6%	0	<1%					6%
DTVHYNPSELYSWL	<1%	0	0					<1%
DPVLPSPEICGFPAS	1%	<1%	0					1%
LSELNPPPLPASSNGL	61%	0	31%		2%			94%
VIPGNAIYQFPAIDSSSSSNNQNKR	1%	<1%	0					1%
SCSSPDSMVTSTSTGTQIGK	42%	5%	2%	9%	2%			60%
GVIGTTVTTTTTTTTAAGESTR	33%	0	1%		61%			94%
QIGCLAVSQAGAMR	<1%	0	0					<1%
FTESLHYYSTLFDSLEGVPNSQDK	<1%	0	0					<1%
FGSSGLAPAHLGSNAFK	1%	0	0					1%
LSTAAY	1%	0	0					1%

Table 2.1: Relative abundance of O-GlcNacylated RGA Asp-N peptides. These are Asp-N generated peptides from RGA expressed in tobacco plants with SEC. Figure adapted from [28].

A supplementary finding from the work on RGA and SEC was that in Arabidopsis mutants with mutations in SEC, no O-GlcNAcylation of RGA was observed. This suggested that SPY either did not modify RGA, or that it is not an O-GlcNAc transferase. Following the completion of the analysis of RGA +SEC we began a second project with the Sun lab to analyze the effect of the protein SPY on RGA. Those experiments and their surprising results are the focus of the rest of this chapter.

2.2 Materials, Equipment, and Instrumentation

Agilent Technologies (Palo Alto, CA)

1100 Series high performance liquid chromatograph

1100 Series vacuum degasser

Branson (Danbury, CT)

Branson 1200 ultrasonic bath

Eppendorf (Hauppauge, NY)

5414R Benchtop centrifuge

Honeywell (Morristown, NJ)

Burdick and Jackson® Acetonitrile, LC-MS grade

Labconco Corporation (Kansas City, MO)

Centrivap centrifugal vacuum concentrator

Molex (Lisle, IL)

Polymicro Technologies™ polyimide coated fused silica capillary

Sizes: 360 µm o.d. x 50, & 75 µm i.d.

PQ Corporation (Valley Forge, PA)

Kasil – Potassium silicate solution

Promega Corporation, (Madison, WI)

Sequencing grade modified trypsin

SGE Analytical Science (Melbourne, Australia)

PEEKsil tubing 1/16” o.d., 0.025 mm i.d.

Sigma Aldrich (St. Louis, MO)

2-propanol, LC-MS grade

Ammonium Acetate

Ammonium Hydroxide

Angiotensin I acetate salt hydrate, ≥99% purity (human)

1,4-Dithiothreitol, ≥97% purity

Glacial acetic acid, ≥99.9% purity

Iodoacetamide (Bioultra), ≥ 99% purity

Trichloroacetic acid

Vasoactive intestinal peptide fragment 1-12, ≥97% purity (human)

Sutter Instrument Co. (Novato, CA)

P-2000 microcapillary laser puller

Thermo Fisher Scientific (San Jose, CA/Bremen, Germany)

Calibration mixture

The Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer

Orbitrap Elite™ mass spectrometer (custom modified with front-end ETD)

Pierce® water, LC-MS grade

Urea

YMC Co., LTD (Kyoto, Japan)

ODS-AQ, C18 5 μm spherical silica particles, 120 Å pore size

ODS-AQ, C18 15 μm spherical silica particles, 120 Å pore size

Zeus Industrial Products (Orangeburg, SC)

Teflon tubing, 0.012" i.d. x 0.060" o.d.

2.3 Methods

2.3.1 Generation of RGA tryptic peptides from Tobacco, Ecoli, and Arabidopsis

Details for transgenic plant generation, growth, and protein harvesting, performed by R. Zentella, have been described previously [29,30]. Briefly, 6His-3XFLAG-RGA or RGA-TAP was tandem affinity purified from tobacco, Arabidopsis, or *E.coli.* onto agarose anti-3XFLAG or anti-Protein A beads respectively. 10% of the beads were transferred to a new tube, and protein was eluted with 1% SDS and analyzed by SDS-PAGE gel to estimate protein recovery. The remaining beads were treated with DTT to reduce protein disulfide bonds, alkylated with Iodoacetamide, and digested with trypsin. The supernatant containing RGA tryptic peptides was transferred to a new tube, dried under vacuum, and stored. Prior to MS analysis samples were reconstituted with 0.1% Acetic Acid in LCMS grade water to a concentration of 1pmol of RGA per μL based on the gel estimate.

2.3.2 HPLC Column Fabrication

Nano-Flow HPLC columns were fabricated in house. Pre-columns were generated by fritting the end of a 20cm length 360 μm o.d. x 75 μm i.d. Polymicro

Technologies™ polyimide coated fused silica capillary by wicking approximately 1cm of a solution of 3 parts potassium silicate and 1 part formamide into the end of the capillary, then baking in a lab over for approximately 16 hours at 60°C. After polymerization in the oven the columns were trimmed to that a frit of 2-4 mm remained. The column was then packed to a length of 12-13cm with ODS-AQ, C18 15 μm spherical silica particles, 120 Å pore size, in a methanol slurry, using a helium pressure bomb. Analytical columns were fritted by wicking approximately 3 cm of a solution of 3 parts potassium silicate and 1 part formamide into the end of the capillary. A soldering iron was then used to polymerize a 2-4 mm section of the silicate 3cm from the end of the capillary , and the rest of the solution was rinsed from the column. Analytical columns were packed with 12-13 cm of ODS-AQ, C18 5 μm spherical silica particles, 120 Å pore size, by the same method as the pre-columns. Columns were conditioned by several rounds of loading approximately 10pmol of internal standard peptides Angiotensin I and Vasoactive Intestinal Peptide fragment 1-12 and rinsing the column with an LC gradient of 0-100% solvent B in 17 minutes. (Solvent A: 0.1% acetic acid in water and Solvent B: 70% acetonitrile in 0.1% acetic acid in water). After conditioning analytical columns were modified with a nano-emitter tip for electrospray ionization. A 1 cm section of polyimide coating was removed from the capillary between the frit and the end of the column not containing packing material. A laser puller was used to form an emitter tip of approximately 3μm internal diameter.

2.3.3 Sample Loading and LC Separation

A 1 pmol fraction of sample (1µL) and 100fmol of internal standard peptides Angiotensin I and Vasoactive Intestinal Peptide fragment 1-12 (was pressure loaded onto a pre-column at a flow rate of <1 µL/min followed by a 15 min desalting rinse with 0.1M acetic acid at approximately 3 µL/min. The pre-column was butt-connected to the analytical column with a 2 cm Teflon sleeve (0.060 in o.d. x 0.012 in i.d). Tryptic RGA peptides were gradient eluted and electrosprayed into the mass spectrometer at a split-flow-generated rate of 100 nL/min by an Agilent1100 series binary LC pump using a linear LC gradient of 0-60% Solvent B in 60 min, 60-100% Solvent B in 8 min, hold 100% Solvent B for 2 min, 100%-0% Solvent B in 8 min, 100% Solvent A for 20 min (Solvent A: 0.1% acetic acid in water and Solvent B: 70% acetonitrile in 0.1% acetic acid in water).

2.3.4 Mass Spectrometric Data Acquisition Methods

All mass spectrometry analysis was performed on the The Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer. Full parameters for the acquisition method are contained with the mass spectrometry results files. Briefly, a full MS scan was performed from 300 to 1200 m/z in the Orbitrap at a resolution of 120,000 at m/z 200 and an AGC target of 2E5 charges and a maximum injection time of 100ms. Ions appearing in the MS^1 scan were selected for MS^2 analysis in a data-dependent manner in order of decreasing intensity isolated by the QMF with a 2 m/z window. Dynamic exclusion was turned on with a repeat count of 1, an exclusion duration of 10 seconds, and an exclusion window of ±10ppm. Precursors with charge states of 2-6 were

subjected to CAD fragmentation and MS^2 analysis in the linear ion trap at normal scan speed. Precursors with charge state 3-6 were also subjected to ETD fragmentation using calibrated reaction times and ion trap MS^2 analysis. Raw data was visualized using Thermo Xcalibur V 4.0.27.10.

2.4 Results

2.4.1 Mass Spectrometry Work Performed at UVA

The wild type Arabidopsis Thaliana protein RGA has 587 amino acids and a weight of 64kDa. For these analyses we relied largely on an altered version of the protein first employed during the RGA+SEC experiments. A 6-histidine 3xFLAG affinity tag was added to the protein's N-terminus to facilitate protein capture, and a lysine was inserted between G184 and G185 to add a trypsin cleavage site in a region of the protein of particular interest. The fusion protein was referred to as 6His3xFLAG-RGA-GKG and its full sequence with the affinity tags and lysine insertion marked can be seen in **Figure 2.3**. This fusion protein was expressed in tobacco plants because it was possible to generate a much larger portion of plant tissue in a short period of time relative to Arabidopsis. Tobacco purifications typically yielded RGA amounts >100 pmol from a single sample prep, as opposed to <10pmol when the protein was prepared from Arabidopsis tissue.

The histidine/3xFLAG affinity tag allowed for efficient two step purification as described in section 2.2.1; step one was a column purification using a nickel containing solid support with affinity for the histidine tag, step two was an on bead capture with anti-3xFLAG antibodies. Captured protein was then reduced, alkylated, and digested on

the beads with trypsin. Soluble peptides were separated from the beads, dried under vacuum, and then reconstituted in dilute acid prior to LC-MS/MS analysis. An in-silico

MRGSHHHHHHDYKDHDGDYKDHDIDYKDDDDKTDPMKRD
HHQFQGRLSNHGTSSSSSSISKDKMMMVKKEEDGGGNMD
DELLAVLGYKVRSSEMAEVALKLEQLETMMSNVQEDGLS
HLATDTVHYNPSELYSWLDNMLSELNPPPLPASSNGLDP
VLPSPEICGFPASDYDLKVIPGNAIYQFPAIDSSSSSNN
QNKRLKSCSSPDSMVTSTSTGTQIGKGVIGTTVTTTTTT
TTAAGESTRSVILVDSQENGVRLVHALMACAEAIQQNNL
TLAEALVKQIGCLAVSQAGAMRKVATYFAEALARRIYRL
SPPQNQIDHCLSDTLQMHFYETCPYLKFAHFTANQAILE
AFEGKKRVHVIDFSMNQGLQWPALMQALALREGGPPTFR
LTGIGPPAPDNSDHLHEVGCKLAQLAEAIHVEFEYRGFV
ANSLADLDASMLELRPSDTEAVAVNSVFELHKLLGRPGG
IEKVLGVVKQIKPVIFTVVEQESNHNGPVFLDRFTESLH
YYSTLFDSLEGVPNSQDKVMSEVYLGKQICNLVACEGPD
RVERHETLSQWGNRFGSSGLAPAHLGSNAFKQASMLLSV
FNSGQGYRVEESNGCLMLGWHTRPLITTSAWKLSTAAY

Figure 2.3: 6His-3xFLAG-RGA-GKG. This is the sequences of the modified RGA protein used in most experiments. The 6His tag is highlighted in red, the 3xFLAG tag in green, and the lysine insertion in blue. Regions of the protein typically observed in our mass spectrometry experiments are underlined.

trypsin digest of 6His3xFLAG-RGA-GKG can be seen in **Table 2.2**. The efficiency of the two step purification allowed us to generate samples where RGA was the dominant protein. A high relative abundance of RGA compared to contaminating species increased our ability to detect low level modified forms of RGA peptides.

Unless otherwise stated all samples were directly loaded onto and separated on homemade 20-25cm C18 75um i.d. reverse phase nanospray LC columns with a 90minute LC gradient and eluted into the mass spectrometer at flow rates of approximately 100nl/min. All samples also included the addition 100fmol of angiotension and vasoactive peptides as standards to evaluate run to run instrument performance. All samples were analyzed on the Fusion mass spectrometry system, described in detail in **Chapter 1**, using a standardized acquisition method described in section 2.2.5, but that included a high resolution MS1 scan in the Orbitrap followed by CID and ETD MS/MS spectra of precursors selected by the instrument computer based on intensity. Following instrumental analysis data files were subjected to manual interpretation and searching with the MASCOT algorithm.

M+H Mass	Location	Sequence	Peptide found
1391.6100	3-13	GSHHHHHHDYK	N
849.3373	14-20	DHDGDYK	N
905.3999	21-27	DHDIDYK	N
607.2205	28-32	DDDDK	N
591.2807	33-37	TDPMK	N
1024.4707	39-46	DHHQFQGR	Y
1478.7081	47-61	LSNHGTSSSSSSISK	Y
639.3027	64-68	MMMVK	N
2024.9117	70-88	EEDGGGNMDDELLAVLGYK	Y
1064.5292	91-100	SSEMAEVALK	Y
8141.8278	101-174	LEQLETMMSNVQEDGLSHLATDTVHYNPSELYSWLDNMLSELNPPPLPASSNGLDPVLPS	N
2551.2423	175-198	VIPGNAIYQFPAIDSSSSSNNQNK	Y
1973.8790	202-221	SCSSPDSMVTSTSTGTQIGK	Y
2127.0775	222-243	GVIGTTVTTTTTTTTAAGESTR	Y
1415.7488	244-256	SVILVDSQENGVR	Y
2664.4211	257-281	LVHALMACAEAIQQNNLTLAEALVK	Y
1404.7086	282-295	QIGCLAVSQAGAMR	Y
1211.6419	297-307	VATYFAEALAR	Y
451.2663	309-311	IYR	N
3321.5377	312-339	LSPPQNQIDHCLSDTLQMHFYETCPYLK	Y
1893.9493	340-356	FAHFTANQAILEAFEGK	Y
2738.4269	359-382	VHVIDFSMNQGLQWPALMQALALR	Y
860.4261	383-390	EGGPPTFR	Y
2157.0393	391-411	LTGIGPPAPDNSDHLHEVGCK	Y
1788.9279	412-426	LAQLAEAIHVEFEYR	Y
3745.8741	427-461	GFVANSLADLDASMLELRPSDTEAVAVNSVFELHK	Y
1039.6258	462-471	LLGRPGGIEK	Y
614.4235	472-477	VLGVVK	N
2766.4573	478-501	QIKPVIFTVVEQESNHNGPVFLDR	Y
2777.2940	502-525	FTESLHYYSTLFDSLEGVPNSQDK	Y
1025.5336	526-534	VMSEVYLGK	Y
1417.6562	535-547	QICNLVACEGPDR	Y
403.2299	548-550	VER	N
1227.5865	551-560	HETLSQWGNR	Y
1660.8441	561-577	FGSSGLAPAHLGSNAFK	Y
1757.8639	578-593	QASMLLSVFNSGQGYR	Y
2729.3538	594-617	VEESNGCLMLGWHTRPLITTSAWK	Y
625.3191	618-623	LSTAAY	N

Table 2.2: In silico trypsin digest of 6His-3xFLAG-RGA-GKG

Our sample prep and analysis scheme yielded 79% coverage of the fusion protein and covered 80/93 serines and threonines available for O-linked modification. Based on results from the analysis of O-GlcNAc modification of 6His3xFLAG-RGA-GKG by the Arabidopsis protein, and SPY homolog, SEC, we were expecting modifications by SPY to be concentrated in four regions of the protein that include S, T, or ST repeats. Because of the similarity between SPY and SEC it was assumed SPY was also an O-GlcNAc transferase, although there had been some previous evidence to call this assumption into question.

Early analysis of the LCMS data revealed some low level O-GlcNAcylated RGA peptides in the RGA+SPY tobacco samples that matched the most prominent O-GlcNAc sites observed in RGA+SEC tobacco experiments, however these modified peptides were also observed in the control samples and there was no increase above control for RGA+SPY. Manual inspection of the MS^1 spectra and sequencing of peaks not identified by the MASCOT search revealed a version of the RGA tryptic peptide LSNHGTSSSSSSISK shifted by mass +146 (**Figure 2.5**), which was determined to be an O-fucose modification (**Figure 2.6**). Importantly, the O-fucosylated peptides were only present when SPY was co-expressed with RGA. The discovery of this unexpected modification, never before observed on a nuclear protein, set the direction for the rest of our analysis. We operated on the assumption that SPY was a protein O-fucose transferase (POFUT) and not an OGT.

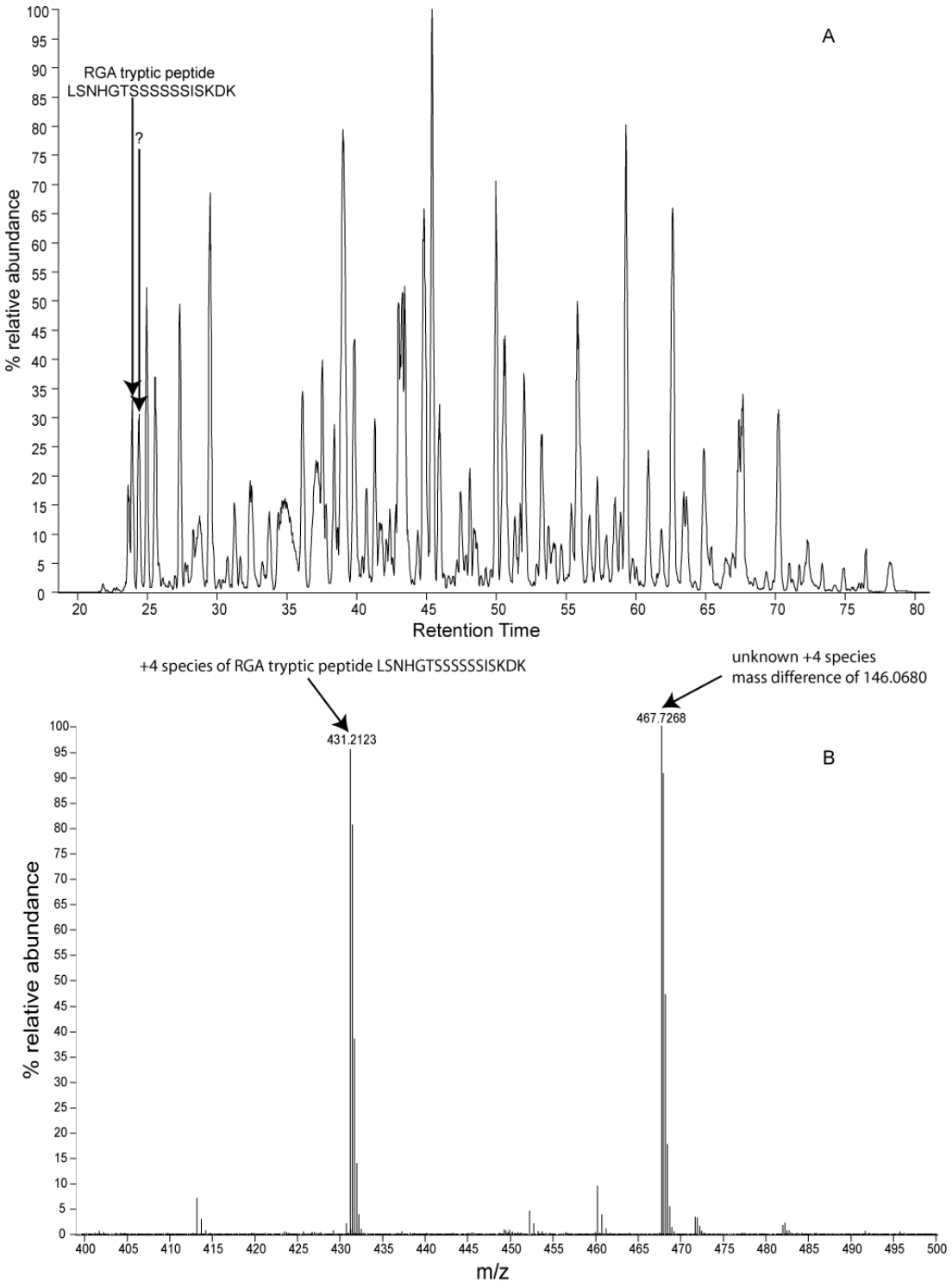

Figure 2.5: The discovery of O-fucosylated RGA. A) We discovered a significant peak in the LC chromatogram created by a peptide that did not match any unmodified RGA tryptic peptides or common contaminating tobacco proteins. B) The peak corresponded to a mass shift of 146.0680 from the unmodified RGA peptide LSNHGTSSSSSSISKDK, which is a product of a missed trypsin cleavage. This is the exact mass of an O-linked fucose.

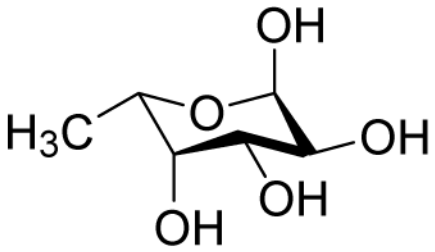

Figure 2.6: Fucose

We began by identifying as many O-Fucose residues on 6His3xFLAG-RGA-GKG as possible when co-expressed with SPY in tobacco. **Figure 2.7** is an ETD MS^2 spectrum of the mono-fucosylated peptide **LSNHGTSSSSSSISK**, and **Table 2.3** shows the fragment ion coverage map for the peptide. We were able to determine that the O-Fucose modification is evenly distributed across four sites in the peptide, and appears on both threonine and serine.

Overall, RGA is highly O-fucosylated by SPY, but on a much more limited number of residues compared to the O-GlcNAcylation by SEC. In addition to the first poly-S region already described, we identified O-fucosylation on three other regions of the protein, but no fucose residues were ever found outside of these regions in contract to O-GlcNAc. In addition to fucosylation, we also observed O-GlcNAcylation, phosphorylation, and a poorly understood O-linked Hexose (first seen by Andrew Dawdy in his dissertation work) in three of the four fucosylated regions. Our ability to observe combinatorial modifications was limited by the trypsin digestion, but we also were able to identify and in some cases precisely map multiple modifications within a single poly-S/T region. All identified modified peptide forms of the 6His3xFLAG-RGA-GKG are shown in **Figure 2.8**.

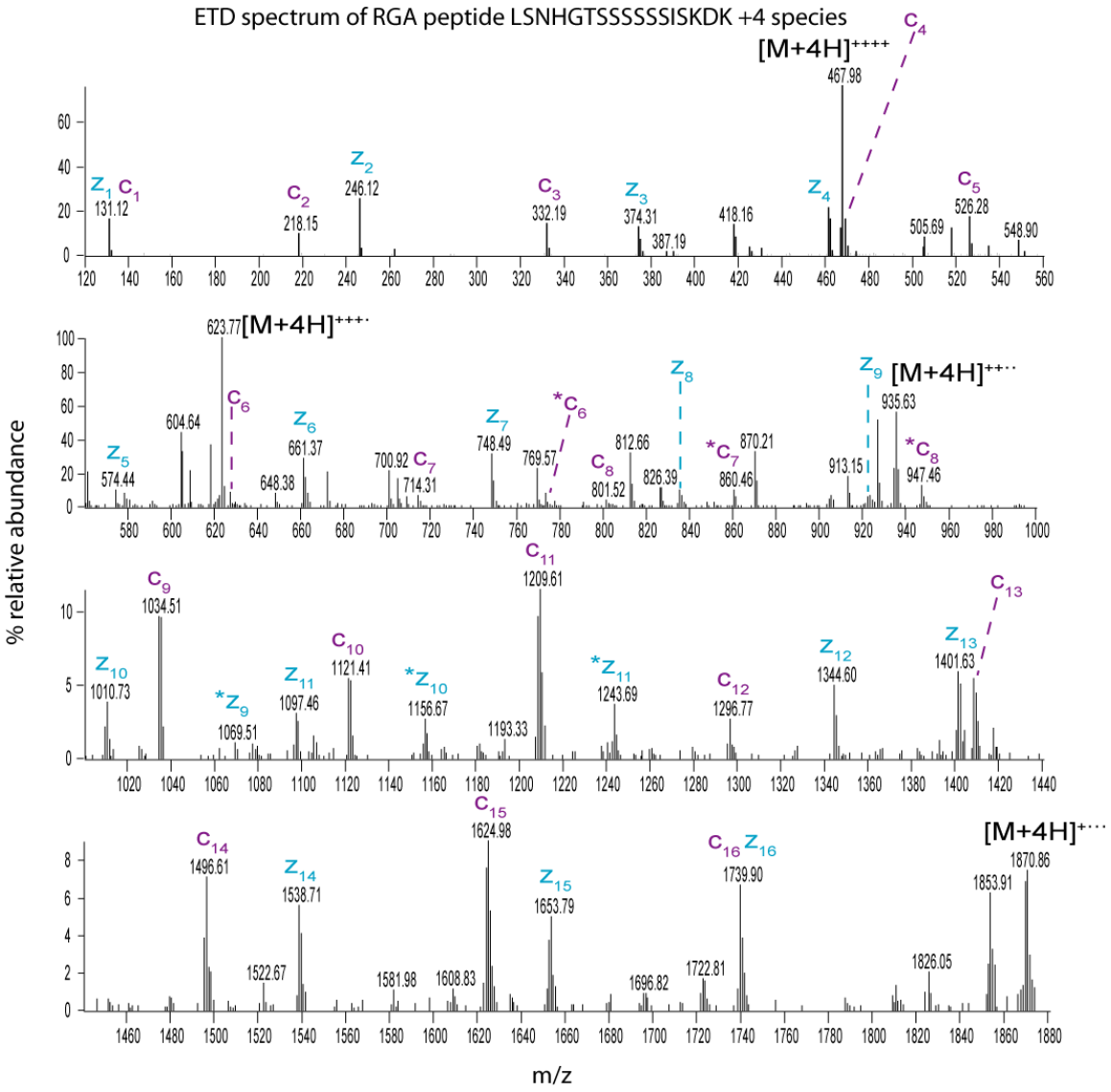

Figure 2.7: ETD MS^{2}spectrum of mono-Fucosylated peptide LSNHGTSSSSSSISKDK. C and Z type fragment ions are marked in purple and blue respectively, while unreacted precursor and charge reduced species from ETnoD are labeled in black. Some fragment ions exist with and without fucose because of the mixed modification sites. For fragments where two copies are present the heavier fucosylated version is marked with an *.

We assessed the relative abundance of modified and unmodified peptides from control and +SPY samples. Peptides were quantified by summing the ion current for any charge states and isotopic peaks >10% relative abundance at the center of the peptides chromatographic peak. The ion current for every peptide covering a given modified

ETD fragments from LSNHGTSSSSSSISKDK mono O-fucose						
C+1 ions w/ fucose	C+1 ions w/o fucose		Sequence		Z+1 ions w/o fucose	Z+1 ions w/ fucose
	131.12	1	L	17	1721.83	1867.89
364.21	218.15	2	S	16	1592.73	1738.79
478.25	332.19	3	N	15	1505.70	1651.75
615.31	469.25	4	H	14	1391.65	1537.71
672.33	526.27	5	G	13	1254.59	1400.65
773.38	627.32	6	T	12	1197.57	1343.63
860.41	714.35	7	S	11	1096.52	1242.58
947.44	801.39	8	S	10	1009.49	1155.55
1034.47	888.42	9	S	9	922.46	1068.52
1121.51	975.45	10	S	8	835.43	981.49
1208.54	1062.48	11	S	7	748.40	894.45
1295.57	1149.51	12	S	6	661.36	807.42
1408.66	1262.60	13	I	5	574.33	720.39
1495.69	1349.63	14	S	4	461.25	607.31
1623.78	1477.72	15	K	3	374.22	
1738.81	1592.75	16	D	2	246.12	
1867.89	1721.83	17	K	1	131.09	

Table 2.3: Fragment ion coverage map for RGA tryptic peptide LSNHGTSSSSSSISKDK. Observed fragments are highlighted in grey. Some ions appear with and without fucose because the modification is spread across multiple sites.

region was summed, and each individual modification form's relative abundance was expressed compared to the summed abundance of all detected forms including unmodified peptide. The relative abundance of all observed modified forms can be seen in **Table 2.4**.

LSNHGt$^{O\text{-}Fuc}$SSSSSSISKDK
LSNHGTs$^{O\text{-}Fuc}$SSSSSISKDK
LSNHGTSs$^{O\text{-}Fuc}$SSSSISKDK
LSNHGTSSs$^{O\text{-}Fuc}$SSSISKDK
LSNHGt$^{O\text{-}Hex}$SSSSSSISKDK
LSNHGTs$^{O\text{-}Hex}$SSSSSISKDK
LSNHGTSs$^{O\text{-}Hex}$SSSSISKDK
LSNHGTSSs$^{O\text{-}Hex}$SSSISKDK
LSNHGTs^{phos}SSSSSISKDK
LSNHGTSs^{phos}SSSSISKDK
LSNHGTSSs^{phos}SSSISKDK
LSNHGTSSSs^{phos}SSISKDK
LSNHGTSSSSs^{phos}SISKDK
LSNHGTSSSSSs^{phos}ISKDK
LSNHGt$^{O\text{-}Fuc}$SSs$^{O\text{-}Fuc}$SSSISKDK
LSNHGTSSs$^{O\text{-}Fuc}$SSs$^{O\text{-}Fuc}$ISKDK
LSNHGt$^{O\text{-}Hex}$SSs$^{O\text{-}Fuc}$SSSISKDK
LSNHGt$^{O\text{-}Fuc}$SSs$^{O\text{-}Hex}$SSSISKDK
LSNHGt$^{O\text{-}Fuc}$Ss$^{O\text{-}GlcNAc}$SSSSISK
LSNHGTSs$^{O\text{-}Fuc}$SSs^{phos}SISKDK
VIPGNAIYQFPAIDS[SSSS]$^{O\text{-}Fuc}$NNQNKR
ScS[SPDSMVTSTS]$^{O\text{-}GlcNAc}$TGTQIGK
LK[ScSSPDSMVTSTSTGT]$^{O\text{-}Fuc}$QIGK
LK[ScSSPDS]$^{O\text{-}Fuc}$MVTSTSTGTQIGK
GVIGTTVTTTT[TTT]$^{O\text{-}Fuc}$TAAGESTR
GVIGTTVTTT[TTTTT]$^{O\text{-}Fuc}$AAGESTR
GVIG[TTVTTTTTTTAAGEST]$^{O\text{-}Hex}$R
GVIGTTVTT[TTTT]$^{O\text{-}Glc}$TTAAGESTR
GVIGTTV[TTTTTTT]$^{O\text{-}Fuc+O\text{-}Glc}$AAGESTR

Figure 2.8: Observed modified RGA peptides from tobacco. This is a list of all observed modification sites for Fucose, Hexose, GlcNAc, and Phosphorylation observed in the tobacco samples. Sites that could not be mapped to a single residue are indicated by [ ].

LSNHGTSSSSSSISKDK or LSNHGTSSSSSSISK									
	unmodified	mono-fucose	mono-hexose	mono-GlcNAc	mono-phospho	di-fucose	fucose & hexose	fucose & GlcNAc	fucose & phospho
3xFlag-RGA(GKG) only	96.9%	0.3%	0.1%	2.5%	0.2%	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA(GKG) +SPY	45.5%	45.0%	4.3%	2.9%	0.4%	1.2%	0.3%	N.D.	0.4%
RGA(GKG) +SPY-2	96.6%	0.2%	0.1%	3.0%	0.1%	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-12	96.0%	0.2%	0.2%	3.6%	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-15	96.1%	0.5%	N.D.	2.7%	0.7%	N.D.	N.D.	N.D.	N.D.
RGA(GKG) + SPY-19	97.9%	0.2%	N.D.	2.0%	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) + 3tprSPY	27.8%	49.5%	8.7%	5.1%	N.D.	6.1%	1.4%	0.7%	0.6%
RGA-TAP Arabidopsis	96.9%	0.5%	N.D.	1.9%	0.7%	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA Arabidopsis	96.7%	0.4%	N.D.	2.7%	0.2%	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA Arabidopsis Sec3	98.7%	0.6%	N.D.	N.D.	0.7%	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA Arabidopsis Spy8	95.9%	0.2%	N.D.	3.7%	0.2%	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA E.Coli.	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA +Spy E.Coli.	99.7%	0.3%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA +Spy-2 E.Coli.	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.

VIPGNAIYQFPAIDSSSSSNNQNKR or VIPGNAIYQFPAIDSSSSSNNQNK									
	unmodified	mono-fucose	mono-hexose	mono-GlcNAc	mono-phospho	di-fucose	fucose & hexose	fucose & GlcNAc	fucose & phospho
3xFlag-RGA(GKG) only	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA(GKG) +SPY	96.2%	3.8%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-2	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-12	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-15	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) + SPY-19	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) + 3tprSPY	94.3%	5.7%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
RGA-TAP Arabidopsis	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA Arabidopsis	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA Arabidopsis Sec3	99.8%	0.2%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA Arabidopsis Spy8	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA E.Coli.	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA +Spy E.Coli.	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA +Spy-2 E.Coli.	100.0%	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.	N.D.

Table 2.4: Relative abundance of all RGA modified tryptic peptides.

LKScSSPDSMVTSTSTGTQIGK* or ScSSPDSMVTSTSTGTQICK*									
	unmodified	mono-fucose	mono-hexose	mono-GlcNAc	mono-phospho	di-fucose	fucose & hexose	fucose & GlcNAc	fucose & phospho
3xFlag-RGA(GKG) only	77.5%	0.9%	N.D.	1.2%	20.4%	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA(GKG) +SPY	80.2%	5.8%	N.D.	1.2%	12.8%	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-2	84.8%	1.1%	N.D.	2.1%	11.9%	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-12	84.6%	1.1%	N.D.	2.4%	12.0%	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-15	38.2%	N.D.	N.D.	N.D.	61.8%	N.D.	N.D.	N.D.	N.D.
RGA(GKG) + SPY-19	87.3%	0.5%	N.D.	1.3%	11.0%	N.D.	N.D.	N.D.	N.D.
RGA(GKG) + 3tprSPY	85.5%	3.0%	N.D.	1.3%	10.2%	N.D.	N.D.	N.D.	N.D.

GVIGTTVTTTTTTTTAAGESTR									
	unmodified	mono-fucose	mono-hexose	mono-GlcNAc	mono-phospho	di-fucose	fucose & hexose	fucose & GlcNAc	fucose & phospho
3xFlag-RGA(GKG) only	80.7%	0.5%	1.5%	17.3%	N.D.	N.D.	N.D.	N.D.	N.D.
3xFlag-RGA(GKG) +SPY	46.1%	48.2%	3.3%	1.0%	N.D.	N.D.	N.D.	1.4%	N.D.
RGA(GKG) +SPY-2	72.2%	0.9%	1.5%	25.4%	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-12	74.0%	1.0%	2.6%	22.4%	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) +SPY-15	90.5%	N.D.	N.D.	9.5%	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) + SPY-19	86.3%	N.D.	N.D.	13.7%	N.D.	N.D.	N.D.	N.D.	N.D.
RGA(GKG) + 3tprSPY	40.8%	45.5%	4.3%	5.6%	N.D.	N.D.	N.D.	3.8%	N.D.

Table 2.4: Continued from previous page.

To assess the influence, if any, of the Tobacco expression system on SPY's specificity and activity we also analyzed 6His3xFLAG-RGA-GKG expressed in *E. coli* with and without SPY. The extent of O-fucosylation was drastically reduced in the *E. coli* samples. We were only able to detect O-fucose within the N-terminal peptide LSNHGTSSSSSSISKDK. While the sites modified, T, S, S, S were identical, the modified peptide was <1% of the unmodified in relative abundance (**Table 2.4**).

To further explore the functionality of SPY, and the importance of various regions of the protein for its association with RGA and subsequent enzymatic activity we expressed 6His3xFLAG-RGA-GKG in tobacco with five SPY mutants: *spy-2*, *spy-12*, *spy-15*, *spy-19*, and *3TPR-spy*. The characteristics of the mutants are shown in **Figure 2.9**.

Mutants *spy-8*, *spy-12*, *spy-15*, and *spy-19* all showed dramatically reduced O-fucosylation activity, with modification levels falling to what was seen in the control samples. 3TPR-SPY, however, showed slightly increased modification activity over wild type SPY. Interestingly, the level of hexosylated peptide also increased in the 3TPR-SPY sample. In fact, although our analyses were not directed toward Hexosylation, there does appear to be a positive correlation in all samples between the level of O-fucosylated peptide and the level of O-hexosylated peptide. Levels of modification produced by the 5 SPY Mutants relative to control (no SPY co-expression) and wild type SPY can be seen in **Table 2.4**.

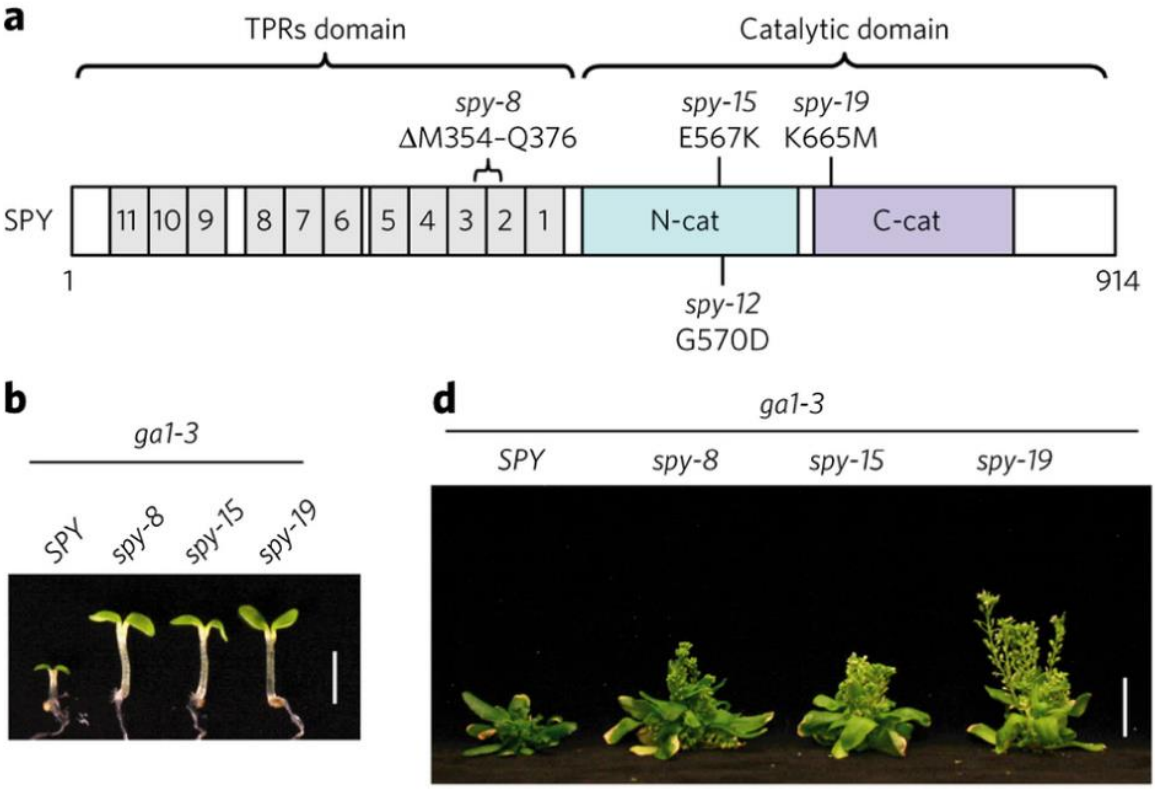

Figure 2.9: SPY mutants. a) Schematic of SPY protein structure. The spy-8 mutation is in the TPR protein-protein interaction domain, while spy-15 and spy-19 mutations are near the suspected active site of the enzyme. In a gibberellin deficient background (ga1-3) all three mutants studied by mass spectrometry rescue growth in the seedling (panel b) and adult (panel d) stage. **Figure adapted from [30].**

Following our experiments in tobacco and *E. coli* we began to investigate O-fucosylation directly in *Arabidopsis thaliana*. Our first Arabidopsis samples were generated from an RGA-TAP fusion protein clone that was already available in the Sun lab at Duke. The TAP tag is a large protein purification tag with two binding sites, a calmodulin binding peptide tag and a Protein A tag divided by a TEV protease cleavage site. TAP-tagged proteins are first captured with IgG (binds to Protein A), released using the TEV protease, and purified again using calmodulin. The RGA-TAP protocol did not

purify RGA as effectively as the histidine/immuno-purification scheme used with 6His-3xFLAG tag, but because the mutant was already available we elected to test it first before generating a 6His-3xFLAG RGA Arabidopsis clone. O-fucose was detectable in the RGA-TAP clones, but at much lower levels then the tobacco samples. Additionally, because the Arabidopsis clone did not have the lysine insertion present in the tobacco system, two of the four poly-S/T areas were not covered.

Once fucosylation had been confirmed in Arabidopsis, we generated a 6His-3xFLAG-RGA Arabidopsis clone and analyzed samples from WT, Spy8, and Sec3 Arabidopsis. Spy8 and Sec3 are both believed to be loss of function mutants. This RGA fusion protein did not contain the lysine insertion in its sequence, so we did not expect to observe the third a fourth poly S/T regions modified in tobacco. Modification sites identified in Arabidopsis were the same as those from tobacco, and relative levels of fucose and GlcNAc between the three Arabidopsis clone types can be seen in **Figure 2.10**.

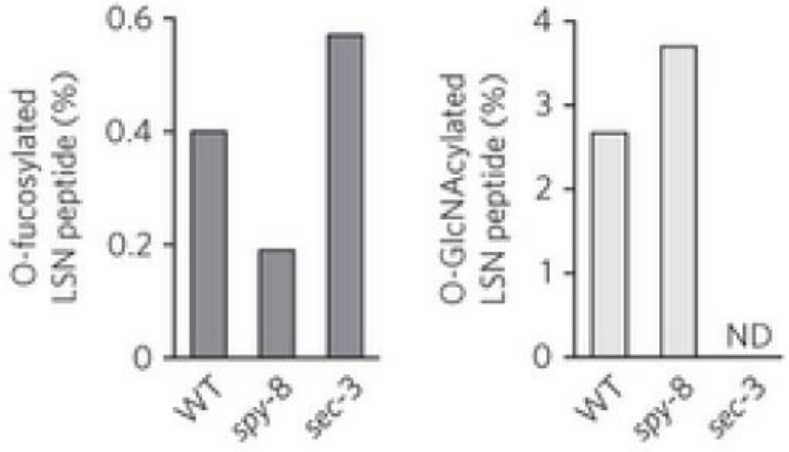

Figure 2.10: O-fucose activity in Arabidopsis. The two graphs show relative levels of fucosylation and GlcNAcylation on the peptide LSNHGTSSSSSSISKDK observed from Arabidopsis wild type and mutant samples. Figure adapted from [30].

Finally, we performed competition experiments between SPY and SEC by transiently expressing fixed amounts of SPY and varying amounts of SEC in tobacco. Increasing levels of SED relative to SPY led to decreased fucosylation and increased GlcNAcylation (**Figure 2.11**).

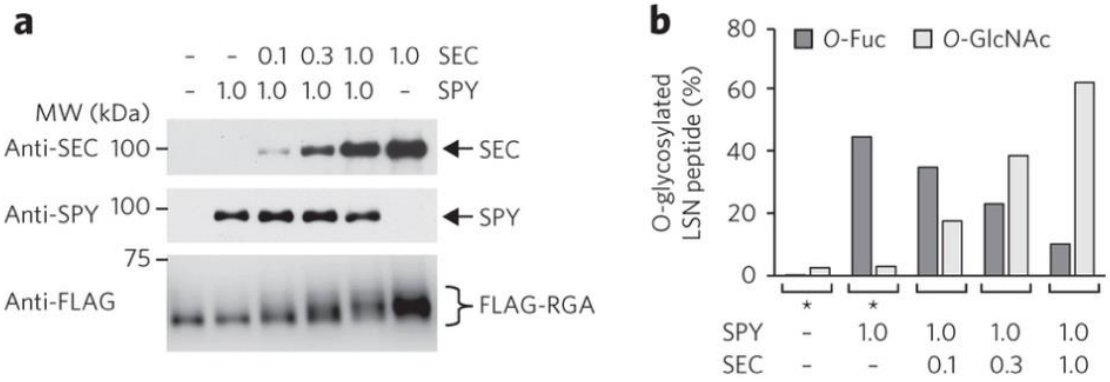

Figure 2.11: SPY and SEC competition experiment. a) An immunoblot assay shows stable levels of SPY expression as SEC is increased. FLAG-RGA mobility decreases as SEC increases likely due to increasing levels of O-GlcNAc modification. b) Mass spectrometry analysis of modified peptides showed decreasing fucosylation and increasing GlcNAcylation correlating with the increase in SEC as determined by peak area ratios between the modified peptides. Asterisk, data included from previous control experiments where RGA was expressed without SPY or SEC. Figure adapted from [30].

2.4.2 Biological Assays Performed at Duke

We performed in vitro assays to assess the direct activity of SPY on RGA peptides. Two RGA peptides shown to be modified in plant samples by LC-MS/MS were incubated with 3-TPR SPY (a truncated mutant that showed no loss of activity in our plant assays) or SEC, and GDP-fucose, the presumed donor substrate for Spy O-fucosyltransferase activity. Peptides were then analyzed by Matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS).

As shown in **Figure 2.12**, 88% of peptide1 was modified in the presence of 3TPR-SPY and GDP-fucose, but no modification of peptide 1 was found in the absence of either. Peptide 2 was also O-fucosylated by Spy, but at much lower levels than peptide 1. Additionally, incubation of RGA peptide 1 with GDP-fucose and SEC did not yield any detectable O-fucosylated peptide.

To evaluate the donor substrate tolerance of SPY four other nucleotide-sugars were incubated with 3TPR-SPY and peptide 1. These sugars were UDP-GlcNAc, GDP-mannose, UDP-galactose, UDP-glucose. No O-fucosyltransferase activity was detected by MALDI-MS in the presence of sugars other than GDP-fucose.

To further characterize the POFUT activity of SPY, we used a malachite green-coupled reaction described in detail in . Briefly, the glycosyltransferase reaction is coupled with a phosphatase (ectonucleoside triphosphate diphosphohydrolase, ENTPD) that releases the β-phosphate of GDP, which can then be detected by malachite green reagents. Using this assay we determined that SPY activity is pH sensitive, with the highest activity at pH 8.2 Two reactions were performed, one with a fixed GDP-fucose concentration at 800uM, and one with a fixed peptide concentration at 312.5uM. For 3TPR-SPY, the K_m for RGA peptide1 was 8.23± 0.10μM, with a k_{cat} of 0.50±0.02 sec^{-1}; The K_m for GDP-fucose was determined to be 50.48±3.90 μM, with kcat of 0.27± 0.01sec^{-1} (**Figure 2.12**).

We assessed the phenotypic effects of the spy mutants assessed earlier through LCMS/MS in a Gibberellin deficient mutant *ga1-3* background. Previous work has

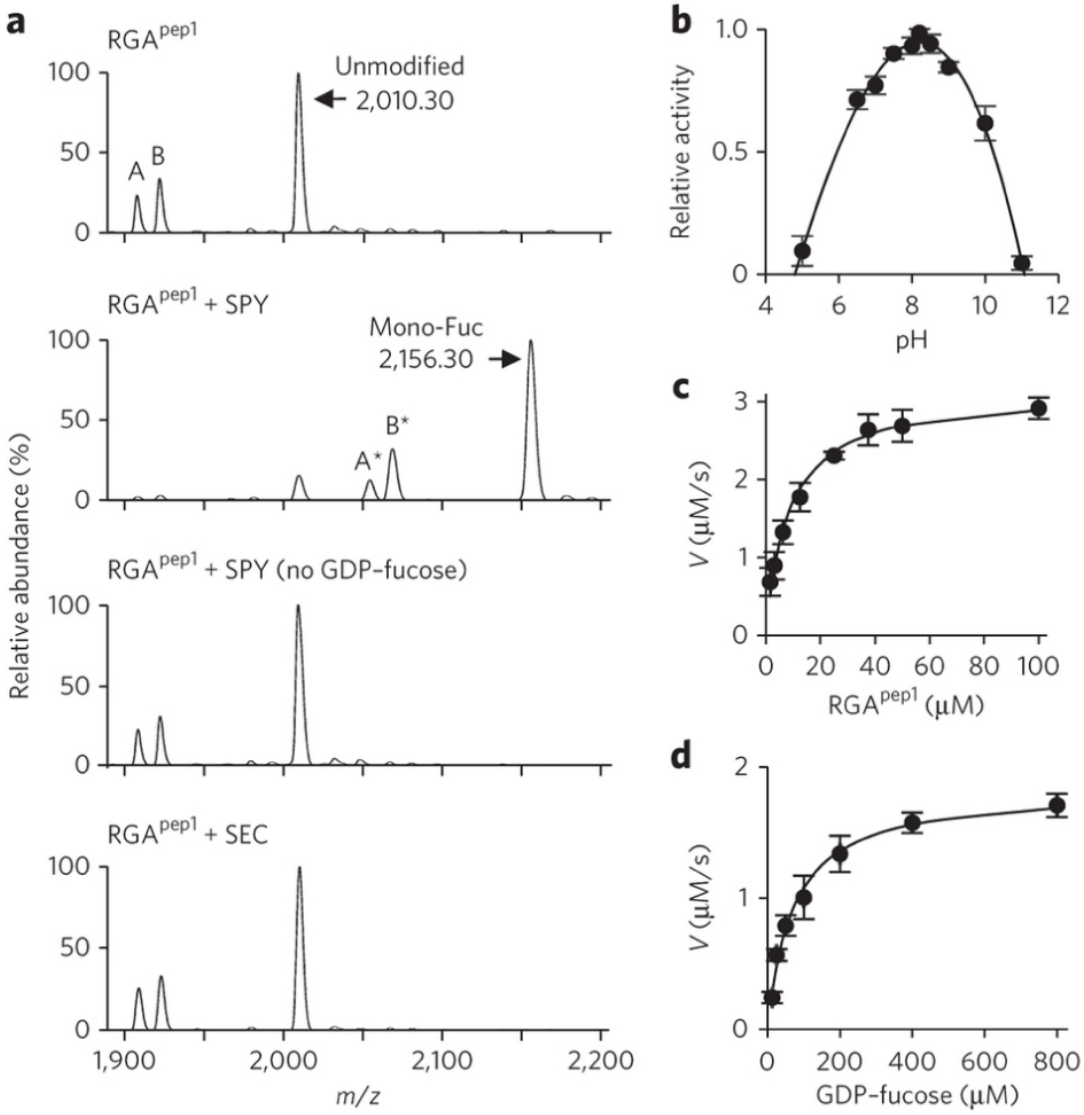

Figure 2.12: SPY in vitro activity and kinetics. Figure adapted from [30].

shown that spy-12, spy-15, and spy-19 show more severe fertility defects and earlier flowering in Arabidopsis than spy-8. In this work spy-8 rescued the hypocotyl growth of ga1-3 in the seedling stage to a similar level as observed in spy-15 and spy-19 mutants. At the adult stage, spy-19 rescued the stem growth defect of ga1-3 more efficiently than spy-8 and spy-15 (**Figure 2.9**).

We repeated the in vitro malachite green assay with purified spy mutants (spy-8, spy-15, and spy-19) to assess their remaining enzymatic activity. Spy-8 POFUT activity was 7.3% relative to the wild type protein, but spy-15 and spy-19 had no detectable POFUT activity.

Spy has a demonstrable effect on the function of RGA, but does not promote its degradation or restrict its nuclear localization. Another possible pathway for Spy's effect on RGA, is that O-fucosylation regulates RGA's interaction with other nuclear proteins. We investigated this possibility by performing pulldown assays with three known RGA interacting proteins, BZR1, PIF3, and PIF4, expressed in E.coli as Glutathione S-transferase fusion proteins. FLAG-RGA purified from WT Arabidopsis showed stronger interaction with the three proteins than FLAG-RGA from spy-8 Arabidopsis. If fucosylation of RGA enhances its interaction with BZR1, PIF3, and PIF4, then SPY mutation should enhance the transcription of genes normally induced by those transcription factors. We performed RT-qPCR and found that transcript levels of IAA9 and PRE1 were increased in spy-8 and spy-19 mutants.

Furthurmore, ga1-3 spy seedlings had longer hypocotyls than those of ga1-3 seedlings, an effect known to be promoted by PIF3, PIF4 and PIF5 (refs. 37,38 from NCB paper). In the presence of a BR-biosynthesis inhibitor, spy-8 enhanced the BR response in hypocotyl elongation. This suggests that O-fucosylation of RGA enhances its activity and negatively regulates GA-, BR-, and PIF-dependant pathways.

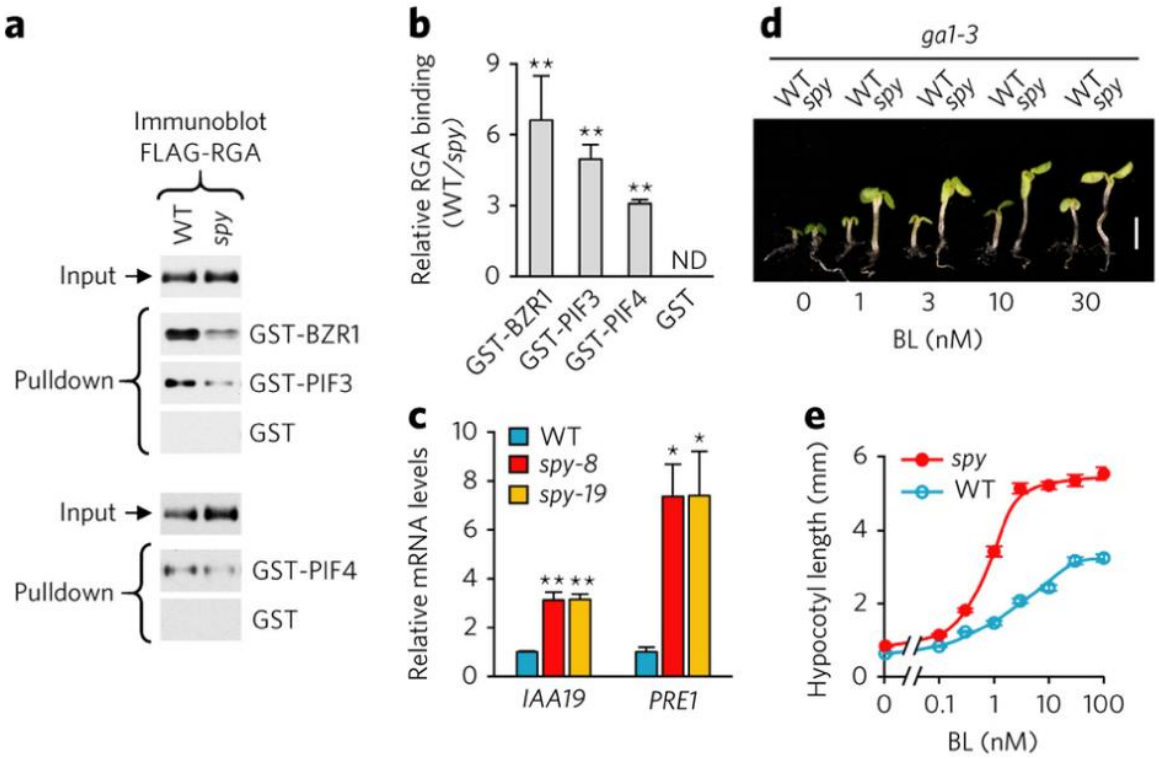

Figure 2.13: RGA-transcription factor interaction experiments. Figure adapted from [30].

2.5 Conclusions

These experiments show that SPY is a protein O-fucosyl Transferase enzyme (POFUT), not an O-GlcNAc transferase as previously suspected. Fucose is a common component of the O and N-linked glycans that decorate many proteins, but O-linked mono fucose has not been extensively detected. Two protein O-fucosyltransferases have been identified in humans, and the donor substrate for both is GDP-Fucose. Known examples of O-fucosylation occur on the membrane bound signaling protein notch, but here appears to be little commonality between previously known O-fucosyltransferases and SPY, with the exception of the shared substrate GDP-fucose.

The Arabidopsis proteins SEC and SPY, and their respective activities as an O-GlcNAc transferase and O-fucose transferase, are now known to have opposing rather than overlapping functions in DELLA regulation. We propose a new model for the regulation of RGA by SPY and SEC (**Figure 2.14)** where GlcNAcylation promotes a closed, less active form of RGA that has reduced interactions with transcription factors and therefore reduced growth inhibiting activity while O-fucosylation by SPY stabilizes an open form of RGA that can bind to transcription factors to regulate growth.

2.6 Future Work

We have immediate plans for two new lines of investigation into RGA and SPY. We have recently generated an Arabidopsis line containing the 6His-3xFLAG-RGA-GKG protein that provides sequence coverage of all known modified sites on RGA. We have already purified samples of this protein from Arabidopsis and are working to fully map sites of O-fucose on RGA in Arabidopsis. Additionally, we hope to use these

samples to investigate the interplay between O-GlcNAc, O-fucose, and phosphorylation on RGA in Arabidopsis. Additionally, it remains unclear whether other DELLA proteins are also modified by SPY. A preliminary analysis showed some evidence for fucosylation of RGL1 by SPY in tobacco, but the results are not definitive and more work needs to be done.

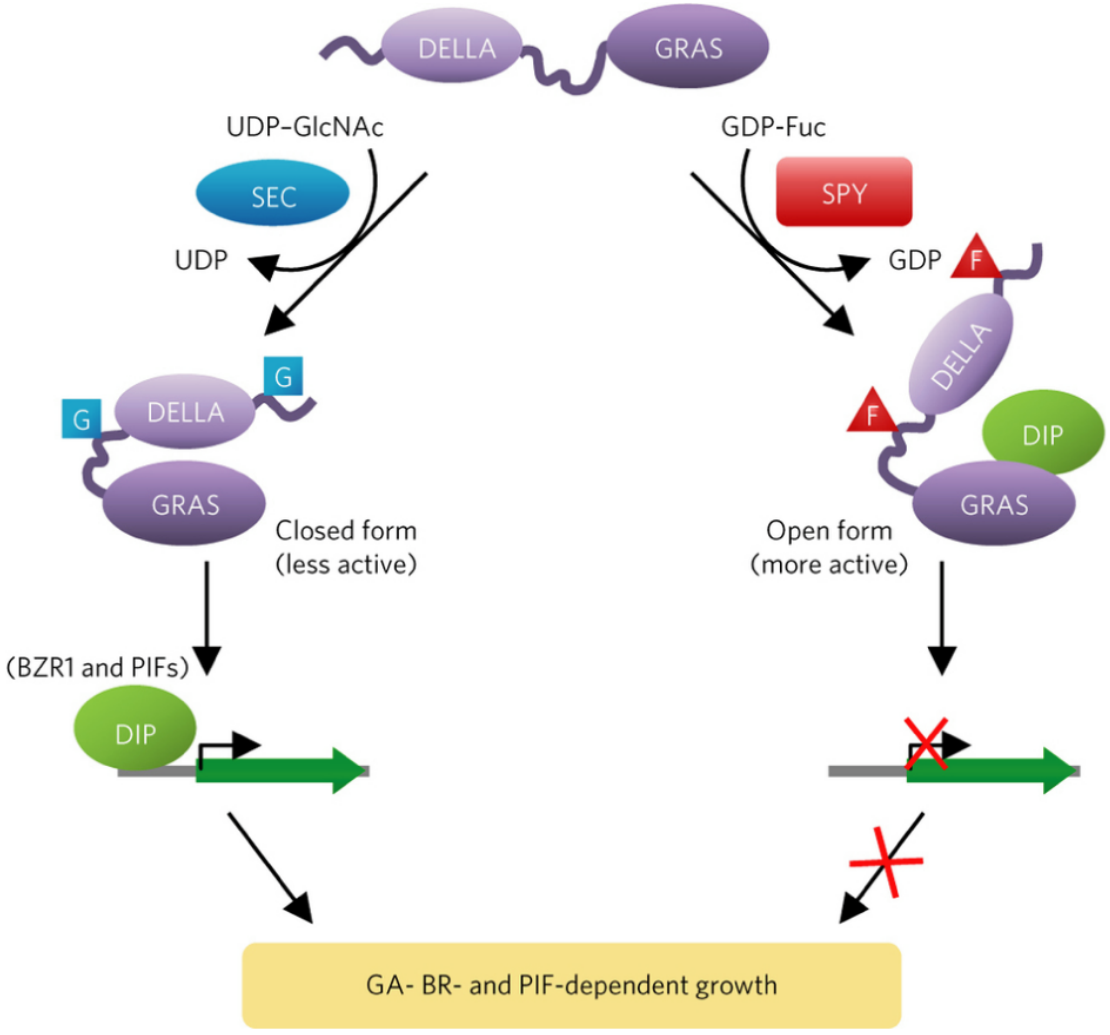

Figure 2.14: A new model for DELLA regulation by SPY/SEC. Figure adapted from [30].

2.7 References

1. Erlich, P. R. *The Population Bomb*. (Rivercity Press, 1968).
2. Borlaug, N. E. THIRD INTERNATIONAL WHEAT GENETICS SYMPOSIUM Wheat Breeding and its Impact on World Food Supply Wheat Breeding and its Impact on World Food Supply. in *Pmc. 3rd Int. Wheat Genet. Symp. Canberra* 1–36 (1968).
3. Stowe, B. B. & Yamaki, T. The History and Physiological Action of the Gibberellins. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **8,** 181–216 (1957).
4. Merritt, J. Gibberellins for Agriculture. *J. Agric. Food Chem.* **6,** 184–187 (1958).
5. Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D. & Koornneef, M. Arabidopsis thaliana: A Model Plant for Genome Analysis. *Science (80-.).* **282,** (1998).
6. Koorneef, M. *et al.* A gibberellin insensitive mutant of Arabidopsis thaliana. *Physiol. Plant.* **65,** 33–39 (1985).
7. Peng, J. *et al.* The Arabidopsis GAI gene defines a signaling pathway that negatively regulates gibberellin responses. *Genes Dev.* **11,** 3194–205 (1997).
8. Ikeda, A. *et al.* slender rice, a constitutive gibberellin response mutant, is caused by a null mutation of the SLR1 gene, an ortholog of the height-regulating gene GAI/RGA/RHT/D8. *Plant Cell* **13,** 999–1010 (2001).
9. Harberd, N. P. *et al.* 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* **400,** 256–261 (1999).
10. Sun, T. & Gubler, F. MOLECULAR MECHANISM OF GIBBERELLIN

SIGNALING IN PLANTS. *Annu. Rev. Plant Biol.* **55,** 197–223 (2004).

11. Silverstone, A. L., Ciampaglio, C. N. & Sun, T. The Arabidopsis RGA gene encodes a transcriptional regulator repressing the gibberellin signal transduction pathway. *Plant Cell* **10,** 155–69 (1998).
12. Fleet, C. M. & Sun, T. A DELLAcate balance: the role of gibberellin in plant morphogenesis. *Curr. Opin. Plant Biol.* **8,** 77–85 (2005).
13. Griffiths, J. *et al.* Genetic characterization and functional analysis of the GID1 gibberellin receptors in Arabidopsis. *Plant Cell* **18,** 3399–414 (2006).
14. Hedden, P. Plant biology: Gibberellins close the lid. *Nature* **456,** 455–456 (2008).
15. Jacobsen, S. E. & Olszewski, N. E. Mutations at the SPINDLY locus of Arabidopsis alter gibberellin signal transduction. *Plant Cell* **5,** 887–96 (1993).
16. Jacobsen, S. E., Binkowski, K. A., Olszewski, N. E. & Phinney, B. 0. SPINDLY, a tetratricopeptide repeat protein involved in gibberellin signal transduction inArabidopsis (plant hormones/gibberellin response mutants). *Plant Biol.* **93,** 9292–9296 (1996).
17. Kreppel, L. K., Blomberg, M. A. & Hart, G. W. Dynamic glycosylation of nuclear and cytosolic proteins. Cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats. *J. Biol. Chem.* **272,** 9308–15 (1997).
18. Thornton, T. M., Swain, S. M. & Olszewski, N. E. Gibberellin signal transduction presents …the SPY who O-GlcNAc'd me. *Trends Plant Sci.* **4,** 424–428 (1999).
19. Hartweck, L. M., Scott, C. L. & Olszewski, N. E. Two O-Linked N-

Acetylglucosamine Transferase Genes of Arabidopsis thaliana L. Heynh. Have Overlapping Functions Necessary for Gamete and Seed Development. *Genetics* **161,** (2002).

20. Torres, C. R. & Hart, G. W. Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. *J. Biol. Chem.* **259,** 3308–3317 (1984).
21. Wells, L., Vosseller, K. & Hart, G. W. Glycosylation of Nucleocytoplasmic Proteins: Signal Transduction and O-GlcNAc. *Science (80-.).* **291,** (2001).
22. Hart, G. W., Housley, M. P. & Slawson, C. Cycling of O-linked β-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature* **446,** 1017–1022 (2007).
23. Hart, G. W., Slawson, C., Ramirez-Correa, G. & Lagerlof, O. Cross Talk Between O-GlcNAcylation and Phosphorylation: Roles in Signaling, Transcription, and Chronic Disease. *Annu. Rev. Biochem.* **80,** 825–858 (2011).
24. Wells, L. *et al.* Mapping sites of O-GlcNAc modification using affinity tags for serine and threonine post-translational modifications. *Mol. Cell. Proteomics* **1,** 791–804 (2002).
25. Vosseller, K. *et al.* O-linked N-acetylglucosamine proteomics of postsynaptic density preparations using lectin weak affinity chromatography and mass spectrometry. *Mol. Cell. Proteomics* **5,** 923–34 (2006).
26. Wang, Z. *et al.* Enrichment and site mapping of O-linked N-acetylglucosamine by a combination of chemical/enzymatic tagging, photochemical cleavage, and electron transfer dissociation mass spectrometry. *Mol. Cell. Proteomics* **9,** 153–60

(2010).

27. Wang, Z. & Hart, G. W. Glycomic Approaches to Study GlcNAcylation: Protein Identification, Site-mapping, and Site-specific O-GlcNAc Quantitation. *Clin. Proteomics* **4,** 5–13 (2008).

28. Dawdy, A. Characterization of o-Glycosylation and Phosphorylation on Nuclear Protein by Mass Spectrometry. (University of Virginia, 2013).

29. Zentella, R. *et al.* O-GlcNAcylation of master growth repressor DELLA by SECRET AGENT modulates multiple signaling pathways in Arabidopsis. *Genes Dev.* **30,** (2016).

30. Zentella, R. *et al.* The Arabidopsis O-fucosyltransferase SPINDLY activates nuclear growth repressor DELLA. *Nat. Chem. Biol.* (2017). doi:10.1038/nchembio.2320

Chapter 3: Identification of SHM Enhancer Binding Proteins

3.1 Introduction

This chapter describes the use of tandem mass spectrometry to identify proteins involved in targeting B cell somatic hypermutation to the immunoglobulin gene locus. Chapter 2 of this dissertation demonstrated how the tandem mass spectrometry experiment detailed in chapter 1 could be applied to the in depth analysis of a single protein. By performing many MS/MS experiments on a relatively simple sample dominated by a single protein we were able to identify rare modified forms of that protein and discover a new post-translational modification. In this chapter, we will use the same foundational tandem MS experiment to identify *many* proteins within complex samples. Unfortunately, we will see that this breadth of analysis comes at a cost. A tryptic digest of a sample containing thousands of proteins yields far more peptides, and therefore potential targets for our MS/MS experiments, than even the most advanced instrument can analyze during the course of a liquid chromatography gradient. To worsen matters, many of the MS/MS experiments we do perform will not be correctly matched to a peptide sequence. As a result, our sequence coverage of the proteins we find is far lower than that seen on RGA in chapter 2 (>80%), and we do not gain any appreciable information about post-translational modifications. Nevertheless, we were able to identify several proteins of interest for further study by our collaborators.

3.1.1 The Function of Somatic Hypermutation

B cell somatic hypermutation is part of a larger process, called antibody affinity maturation, which is essential to the production of high affinity antibodies [1,2]. Antibodies are soluble proteins, secreted by B cells, that bind to pathogens like viruses and bacteria and facilitate their elimination or destruction by other parts of the immune system. The higher the affinity of the antibody for its complementary pathogen the more effective it is at preventing or eliminating an infection. The immune system's challenge is that there is a constantly changing and near infinite array of possible pathogen molecular structures antibodies will need to recognize.

There are five classes of antibody (IgA, IgD, IgE, IgG, and IgM) active in different tissues or at different points in the immune response, but they are all produced by B cells and all recognize antigen by the same mechanism. The core antibody unit is two roughly 25 kD identical light chains and two 50 kD identical heavy chains joined in a tetramer. Both light and heavy chains are characterized by variable and constant regions, with the variable regions responsible for antigen binding and the constant region of the heavy chain responsible for effector functions.

It is estimated that a human will produce over $1x10^9$ different antibody sequences in their lifetime, while the human genome is estimate to contain only approximately 25,000 genes . This diverse antibody repertoire cannot, therefore, possibly be contained in the germline DNA of B cells. Instead, there is a sophisticated two part process for generating novel antibody sequences; part one is V(D)J rearrangement, and part two is affinity maturation (**Figure 3.1**).

B cells begin development in the bone marrow. During their early stages of differentiation there is no single gene coding for the entire antibody protein. Instead the protein is divided into DNA segments V, J, and C for the light chain, and V, D, J, and C for the heavy chain. There are multiple copies of each segment. The V, D, and J segments code for the variable regions of the antibody that interact with antigen, and the C regions determine the class of the antibody (ex. IgG, IgA, etc.). Specialized proteins, notable RAG1 and RAG2 randomly assemble V, D, J, and C segments into a functional antibody gene [3]. While the segments are being joined new DNA bases are randomly inserted in the joints, altering the gene even further. When V(D)J rearrangement is complete, for both light and heavy chains, the B cell has a functional antibody gene. A membrane bound version of the antibody, called the B cell receptor (BCR), is displayed on the surface of the B cell and tested for self-reactivity. B cells that successfully generate a non-self-reactive BCR leave the bone marrow and enter a lymph node.

During an infection B cells waiting in the bone marrow are activated when their BCR binds to a pathogen passing through the lymph node. Activated B cells begin dividing rapidly. Some clones begin secreting low affinity antibodies immediately to combat the infection, while others enter the affinity maturation process. These B cells form a cluster within the lymph node called a germinal center that also contains specialized dendritic cells and helper T cells [4].

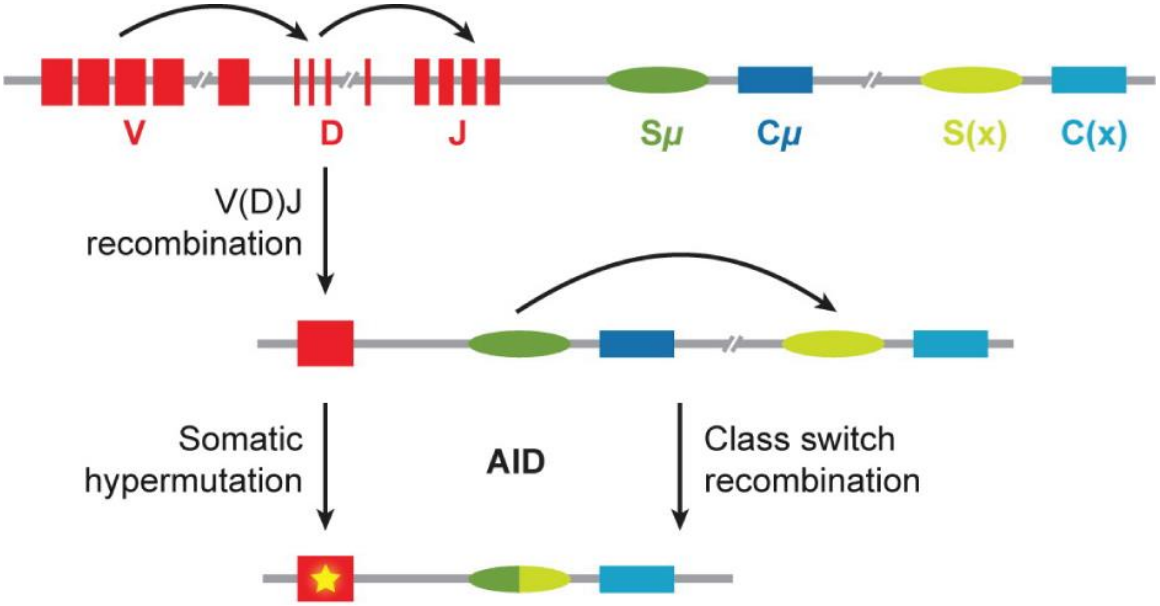

Figure 3.1: V(D)J recombination and somatic hypermutation drive antibody diversity. An additional process called class switch recombination, in which antibodies change constant regions to alter their effector function, shares some mechanistic features with SHM. Figure adapted from [1].

Within the germinal center the activated B cells cycle between periods of rapid proliferation during which they alter their BCR, and antigen capture from dendritic cells and interaction with helper T cells. B cell clones compete for available antigen, then endocytose and digest it. Fragments of captured antigen are displayed again on the B cell surface bound within MHC Class II proteins. Helper T cells recognize these antigen/MHC II complexes and through them form stable interactions with the B Cells. During this interaction the T cell provides critical stimulatory signals to the B cell [5].

Antigen capture is competitive, and T cell help is dependent on antigen capture. B cells that do not receive sufficient T cell help will die, while those that do continue through cycles of proliferation and BCR modification. This survival pressure selects for B cell clones that have improved the affinity of their BCR for the pathogen, and discards changes that do not improve affinity. This is the affinity maturation process, and it

increases the K_d of antibody for target pathogen by several orders of magnitude over the first several days of an immune response [6]. Once the B cells have reached a threshold affinity, they become antibody secreting plasma cells and distribute their high affinity antibodies throughout the body to help eliminate the infection (**Figure 3.2**).

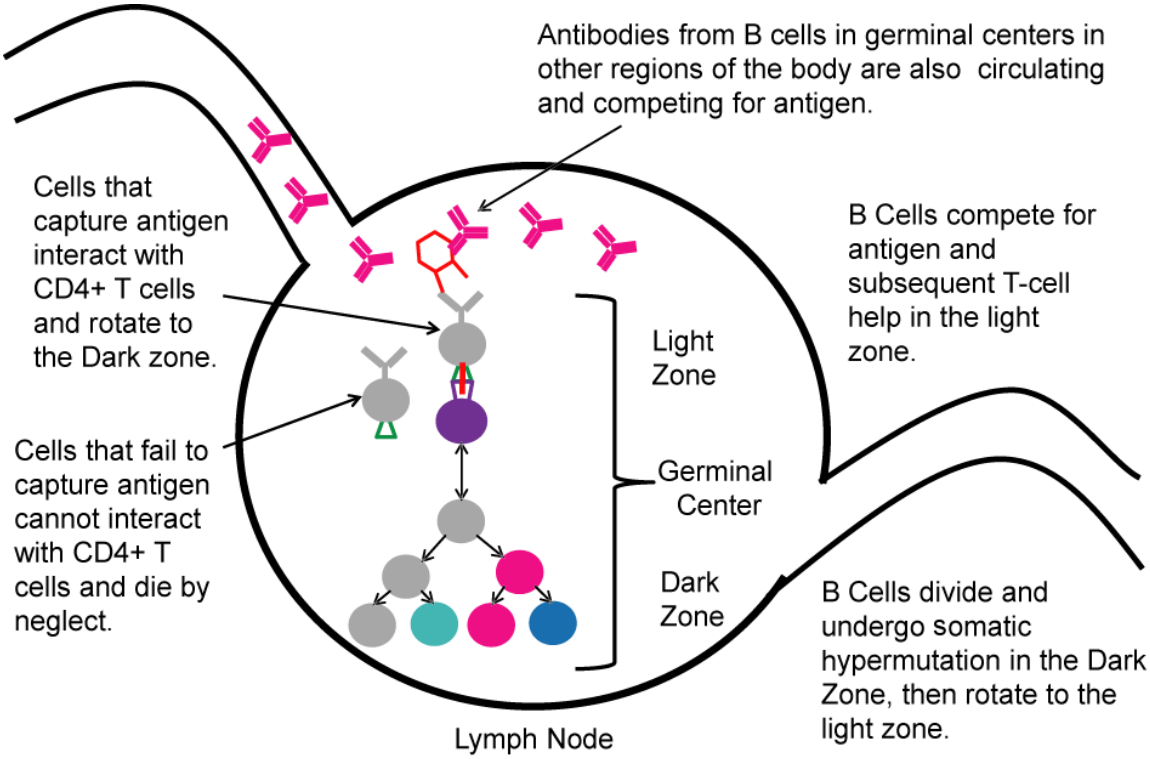

Figure 3.2: SHM and the germinal center.

The affinity of an antibody for pathogen is determined by the amino acids in the variable regions of the protein that interact with antigen. Affinity maturation changes this interaction by changing amino acids within the variable regions. As described in chapter one, a proteins amino acid sequence is determined by DNA according to the genetic code **(Table 1.2**). B cell somatic hypermutation refers to the step during affinity maturation when B cells alter the DNA bases that code for the hypervariable region of the

antibody protein, thereby changing the amino acids responsible for binding to the pathogen and the antibody's affinity for its target. Most importantly, these mutations are well confined to the DNA coding for the variable region of the antibody protein.

3.1.2 Biochemistry of Somatic Hypermutation

Somatic Hypermutation begins with the enzyme Activation-Induced Cytidine Deaminase (AID) [7]. AID deaminates cytidines to uracils, which creates a base pair mismatch between the DNA coding and noncoding strand. Mice and Humans deficient in AID do not undergo somatic hypermutation, proving that AID is an essential actor in SHM. AID only acts on single stranded DNA, and the rate of mutation is correlated with the rate of transcription. AID has a hotspot motif, WRCY (W = A/T, R = A/G, Y = C/T), but not all hotspots are deamidated and not all deamidations occur in a hotspot.

AID introduces mismatches between C and G only, but 60% of SHM mutations are between A-T base pairs. Additionally, when constitutively expressed in non-B Cells AID readily acts across the entire genome. These two facts mean that while AID initiates SHM, it alone is not responsible for the mutations or their targeting to the Ig locus. Current models for SHM view it as a two-step process. Step one is introduction of a G-U mismatch by AID. This occurs randomly in any transcriptionally active single stranded DNA. Step two is repair of that mismatch, and it is the second step that is now believed to be specifically targeted to the Ig locus.

Once the G-U mismatch is created there are three possible fates for the DNA lesion. The first is replication across the mismatch, which would result in a G to A and C to T mutation at the site of deamidation. The second is Base Excision Repair in which

the uracil is removed by the Uracil-DNA glycosylase enzyme UNG and the abasic site is filled with a new nucleotide by a DNA polymerase. The third pathway is recognition of the G-U mismatch by the mismatch recognition protein heterodimer MSH2/MSH6. This triggers the mismatch repair pathway where a short section of DNA is excised and entirely rewritten.

There is strong evidence that all three of these pathways are active and responsible for mutations during SHM. Mice deficient in UNG and MSH2/MSH6 still undergo SHM, but only acquire the G to A and C to T mutations characteristic of replication across the U-G mismatch. Mice deficient in MSH2/MSH6 are unable to mutate at A-T base pairs near deamidation sites. **Figure 3.3** illustrates the three G-U mismatch processing pathways active in Somatic Hypermutation and the types of mutations they can confer [2,8].

One of the biggest unanswered questions about the SHM process centers around the BER and Mismatch repair pathways. Cells regularly utilize these mechanisms during DNA replication, and under normal circumstances they have an extremely low error rate. The overall error rate for DNA replication is 1 in 10^9 base pairs, while the error rate during SHM is roughly 1 in every 1000 base pairs. Why do these DNA repair mechanisms become error prone during SHM, and how is that activity confined to the Immunoglobulin gene variable region? A number of lower fidelity DNA polymerases

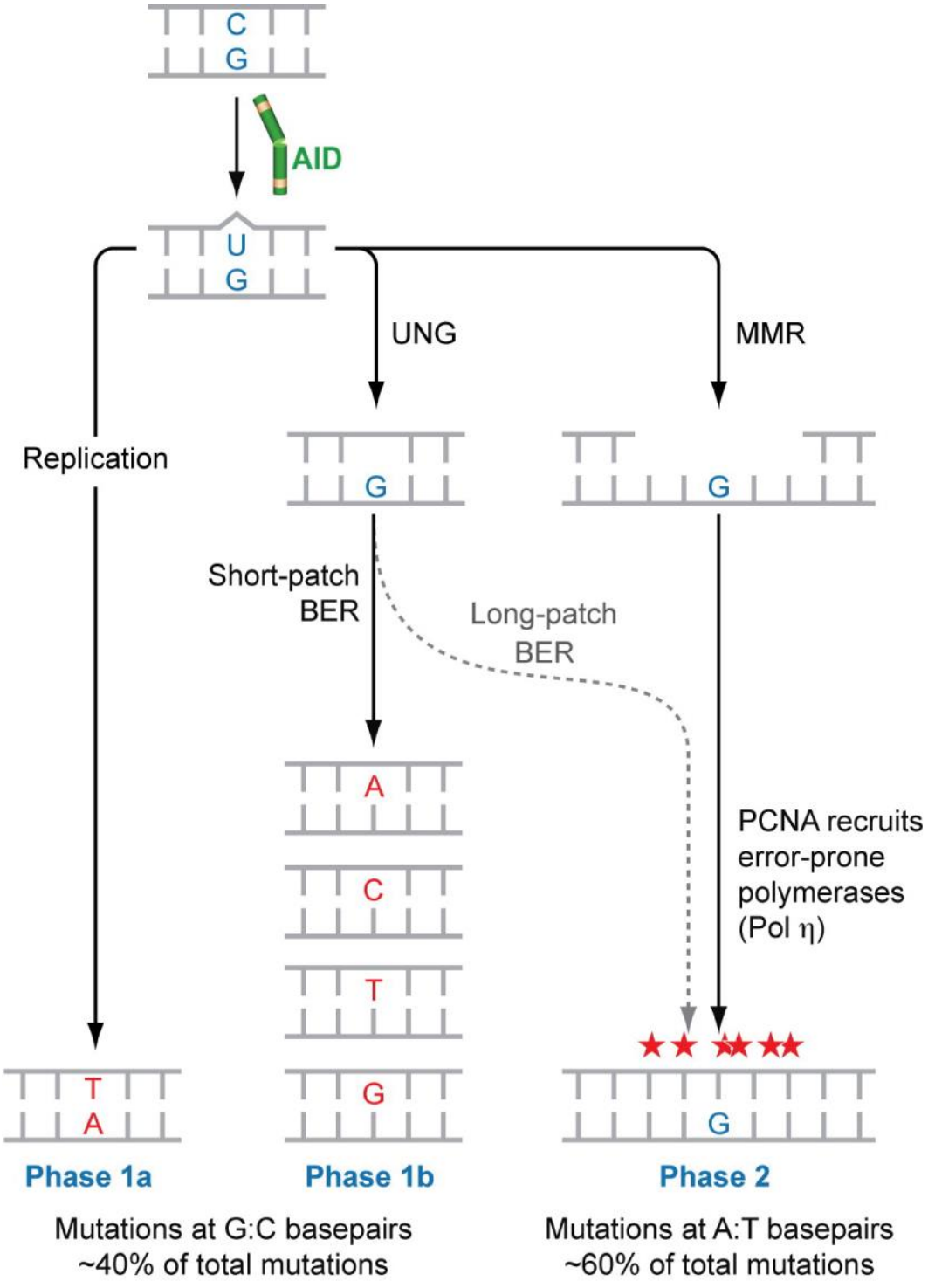

Figure 3.3: Post AID mutational pathways. Adapted from [7].

have been discovered, and several have been shown to affect SHM rates when deleted [9,10], but currently there is no clear evidence for how these low fidelity polymerases might be targeted to the variable region of the Immunoglobulin protein gene.

3.1.3 Identification of Somatic Hypermutation Enhancing DNA Sequences

In addition to the coding segments that dictate the sequence of a protein, DNA also contains regulatory segments that facilitate the interaction of proteins involved in DNA transcription, replication, and repair [11]. In recent years there has been growing evidence that a region of DNA near the antibody coding segment is responsible for targeting SHM [12,13], likely by recruiting specialized proteins involved in the mutation process.

In 2009 a 10kB region near the chicken Igλ light chain locus was shown to have SHM enhancing effects, but the relative importance of specific nucleotide sequences within this larger region remained a mystery [14]. Recently our collaborators, The Schatz lab at Yale Medical School, have developed a highly sensitive green fluorescent protein (GFP) based assay to detect mutation rates for a single gene, and used it to investigate the chicken and human immunoglobulin gene locus in more detail [15].

They have identified much shorter regions of DNA containing multiple transcription factor binding motifs that when inserted near the GFP gene have significant effects on rates of SHM, and term these DNA segments “SHM enhancers”. Point mutations within the transcription factor binding motifs of the enhancers reduce mutation rates. Importantly, parts of these sequences are conserved between the chicken, murine, and human genome. We propose that they act by recruiting proteins necessary for

targeting the error prone repair step of SHM. Our goal is to use a selected set of the chicken and human enhancer sequences, those shown to have the highest SHM enhancing effect, as bait to capture proteins that preferentially bind to these DNA regulatory elements, and then identify those proteins bound to the DNA by LC-MS/MS.

3.1.4 Our Proposed Experimental Approach

Biotinylated enhancer or control DNA is bound to streptavidin coated magnetic beads. DT40 or Ramos cells are lysed, and the intact nuclei captured. Nuclei are lysed and the soluble portion separated. Soluble nuclear lysates are then incubated with the DNA coated magnetic beads. Proteins bound to the beads are digested with trypsin and the resulting peptides analyzed by LC-MS/MS. Database search software is used to correlate MS/MS spectra with tryptic peptides and map those tryptic peptides back to their parent protein. LC peak areas are used to relatively quantitate peptides between enhancer and control samples and identify proteins that bind preferentially to the enhancer DNA.

3.2 Materials, Equipment, and Instrumentation

Agilent Technologies (Palo Alto, CA)

1100 Series high performance liquid chromatograph

1100 Series vacuum degasser

Branson (Danbury, CT)

Branson 1200 ultrasonic bath

Eppendorf (Hauppauge, NY)

5414R Benchtop centrifuge

Honeywell (Morristown, NJ)

Burdick and Jackson® Acetonitrile, LC-MS grade

Labconco Corporation (Kansas City, MO)

Centrivap centrifugal vacuum concentrator

Molex (Lisle, IL)

Polymicro Technologies™ polyimide coated fused silica capillary

Sizes: 360 µm o.d. x 50, & 75 µm i.d.

PQ Corporation (Valley Forge, PA)

Kasil – Potassium silicate solution

Promega Corporation, (Madison, WI)

Sequencing grade modified trypsin

SGE Analytical Science (Melbourne, Australia)

PEEKsil tubing 1/16" o.d., 0.025 mm i.d.

Sigma Aldrich (St. Louis, MO)

2-propanol, LC-MS grade

Ammonium Acetate

Ammonium Hydroxide

Angiotensin I acetate salt hydrate, ≥99% purity (human)

1,4-Dithiothreitol, ≥97% purity

Glacial acetic acid, ≥99.9% purity

Iodoacetamide (Bioultra), ≥ 99% purity

Trichloroacetic acid

Vasoactive intestinal peptide fragment 1-12, ≥97% purity (human)

Sutter Instrument Co. (Novato, CA)

P-2000 microcapillary laser puller

Thermo Fisher Scientific (San Jose, CA/Bremen, Germany)

Calibration mixture

The Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer

Streptavidin coated M-280 Dynabeads (magnetic)

Orbitrap Elite™ mass spectrometer (custom modified with front-end ETD)

Pierce® water, LC-MS grade

Urea

YMC Co., LTD (Kyoto, Japan)

ODS-AQ, C18 5 μm spherical silica particles, 120 Å pore size

ODS-AQ, C18 15 μm spherical silica particles, 120 Å pore size

Zeus Industrial Products (Orangeburg, SC)

Teflon tubing, 0.012" i.d. x 0.060" o.d.

3.3 Methods

Our nuclear protein extraction, DNA capture, and on bead digestion protocols were adapted from [16–18]. The detailed procedure is in **Appendix A**. Four DNA sequences were used for protein capture based on [15], Chicken Enhancer, Chicken Control, Human Enhancer, Human Control. Annotated DNA sequences can be seen in **Appendix B**.

Capillary LC column fabrication and liquid chromatography gradients were identical to those described in **Chapter 2.**

Several different MS acquisition methods were used for sample analysis. All mass spectrometry analysis was performed on the The Thermo Scientific™ Orbitrap Fusion™ Tribrid™ mass spectrometer. Full parameters for the acquisition method are contained with the mass spectrometry results files. For acquisition methods a full MS scan was performed from 300 to 1200 m/z in the Orbitrap, at a resolution of 120,000 at m/z 200, an AGC target of $3E^5$ charges, and a maximum injection time of 100ms. Ions appearing in the MS1 scan with charge states between 2 and 6 were selected for MS2 analysis in a data-dependent manner in order of decreasing intensity and isolated by the QMF with a 2 m/z window. Dynamic exclusion was turned on with a repeat count of 1, an exclusion duration of 30 seconds, and an exclusion window of ±10ppm. Methods differed in their MS^2 acquisition parameters.

HCD fragmentation was performed in the IRM at a pressure of 8 millitor and a normalized collision energy of 25% on a target of $5E^4$ charges. HCD MS^2 spectra were acquired in the Orbitrap at a resolution of 15,000 at m/z 200. CID fragmentation took place in the high pressure cell of the linear ion trap at normalized collision energy of 30% on a target of $1E^4$ charges. CID MS^2 spectra were acquired in the ion trap at the normal scan rate. ETD fragmentation took place in the high pressure cell of the linear ion trap using calibrated charge state dependent reaction times, a precursor target of $1E^4$ charges and a fluoranthene reagent target of $2E^5$ charges.

MS result files were evaluated using Thermo Scientific Proteome Discoverer Beta V2.2.0.336. HCD and CID spectra were searched using the SEQUEST algorithm and ETD using the MASCOT algorithm against either the Chicken TrEMBL database or the Human Swissprot database downloaded from uniprot. PD generated decoy databases, and peptide spectral matches were filtered to a 1% false discovery rate by the Percolator algorithm. Match peptides were assigned to proteins, and proteins quantitated based on the summed peak areas of all their assigned peptides. Protein abundances were normalized for each sample against the total measured peptide content of that sample, then averaged across any technical and biological replicates. Final ratios of protein abundance between enhancer and control samples were calculated.

3.4 Results and Discussion

3.4.1 DT40 Samples

We performed three primary protein capture experiments. The first experiment identified and relatively quantified proteins captured from the nuclei of DT40 cells using either the chicken enhancer or chicken enhancer control sequence described above. The second experiment identified and relatively quantified proteins captured from the nuclei of Ramos cells using either the Human enhancer or Human enhancer control sequence, and the third experiment used Ramos but chicken Enhancer and control sequences.

All protein capture experiments were performed in biological triplicate, and samples from the first experiment were analyzed in technical triplicates with three different MS2 acquisition schemes: low resolution ion trap ETD MS^2, low resolution ion trap CID MS^2, or High Resolution Orbitrap HCD MS^2. Data from the ETD analysis of

the third control replicate was omitted because a clog in the LC system interfered with data acquisition. Samples from the Ramos/Human and Ramos/Chicken experiments were analyzed only once using the same High Resolution Orbitrap HCD MS^2 method.

Proteome Discoverer Beta version 2.2.0.336 was chosen in part because of its unique label free quantitation features. Each protein was quantified by summing the integrated LC peak area for all peptides assigned to that protein. Relative protein abundances in each sample were normalized against the total measured peptide content of that sample, averaged across all 9 replicates for the Enhancer samples and 8 replicates for the control samples, and ratios of each protein's abundance in Enhancer vs Control samples was calculated.

Across all biological and technical replicates (17 LC-MS/MS Runs) from experiment 1 there were 144,651 individual MS/MS experiments, 33,860 of which were matched with peptide sequences from the chicken TrEMBL database for an overall PSM rate of just over 23%. Many of the MS/MS events represented multiple targeting of the same peptide, and so the PSMs were condensed into 6,083 unique peptide hits. These peptides were assigned to 1850 proteins.

Figure 3.4 shows the protein abundance distributions in all samples before and after normalization. **Figure 3.5** shows the overlap between peptides identified between biological replicates and activation type, and **Figure 3.6** performs the same comparison on the protein level. **Figure 3.7** shows the total protein identification overlap from all fragmentation types across all samples. From the DT40 nuclei 89 proteins were found with at least 2 PSMs and at least a 2X increase in signal in the DIVAC vs Control

samples (**Table 3.1**). These proteins vary in relative abundance by more than 3 orders of magnitude. There are also large differences in the number of PSMs, individual peptides identified, and sequence coverage for each protein that scale roughly with relative abundance.

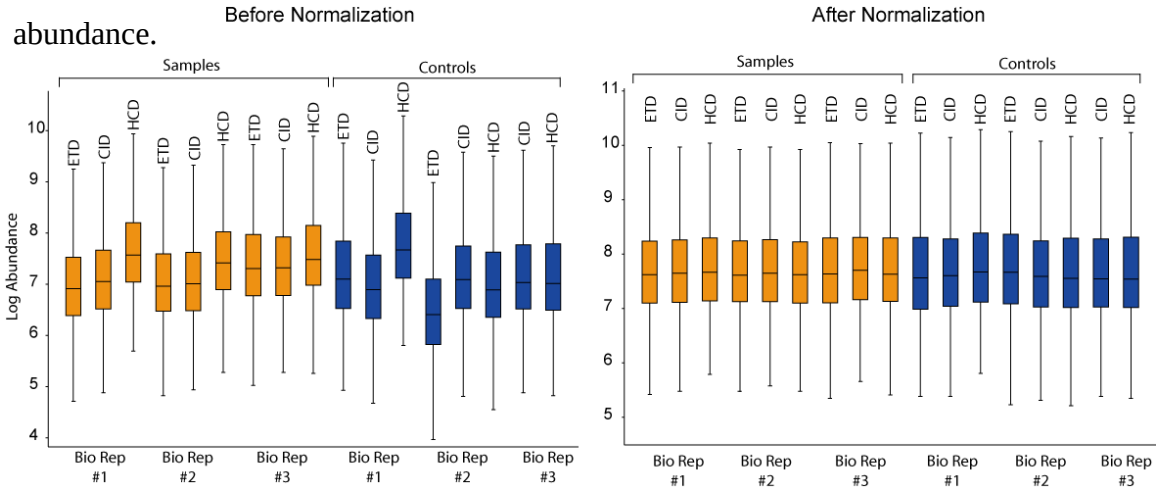

Figure 3.4: Protein abundances from all DT40 replicates before and after normalization.

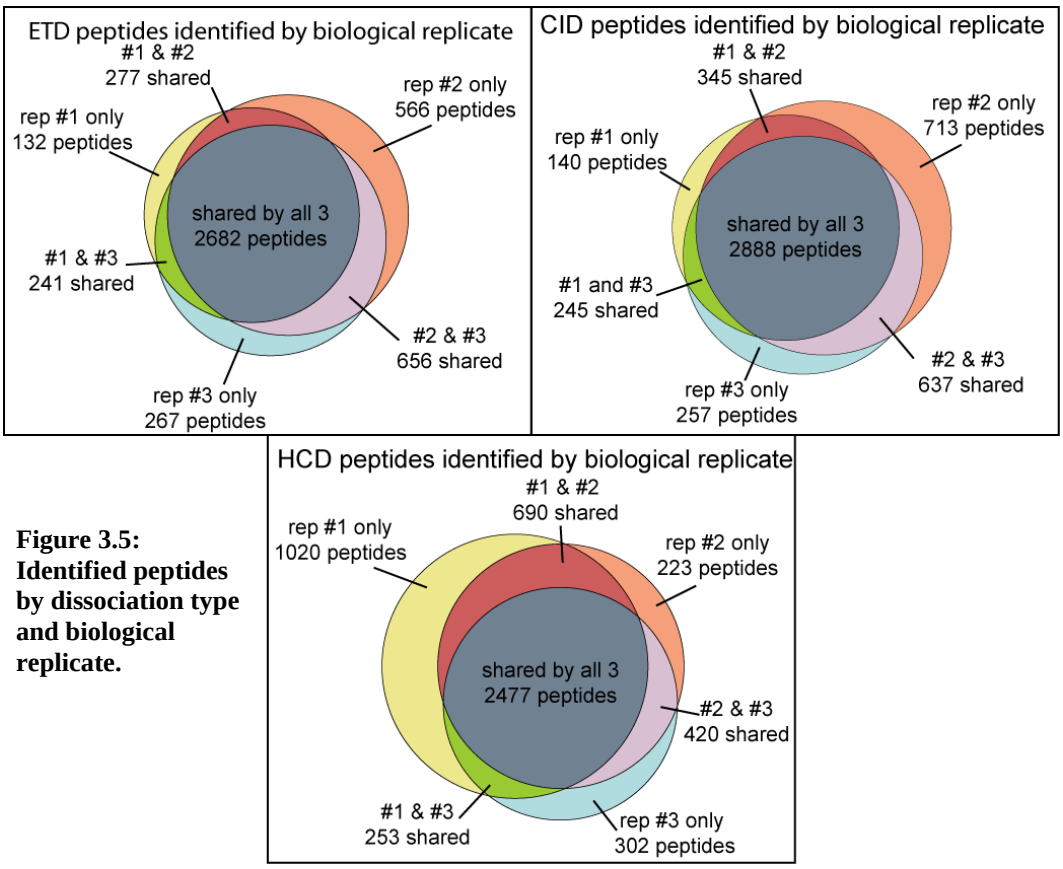

Figure 3.5: Identified peptides by dissociation type and biological replicate.

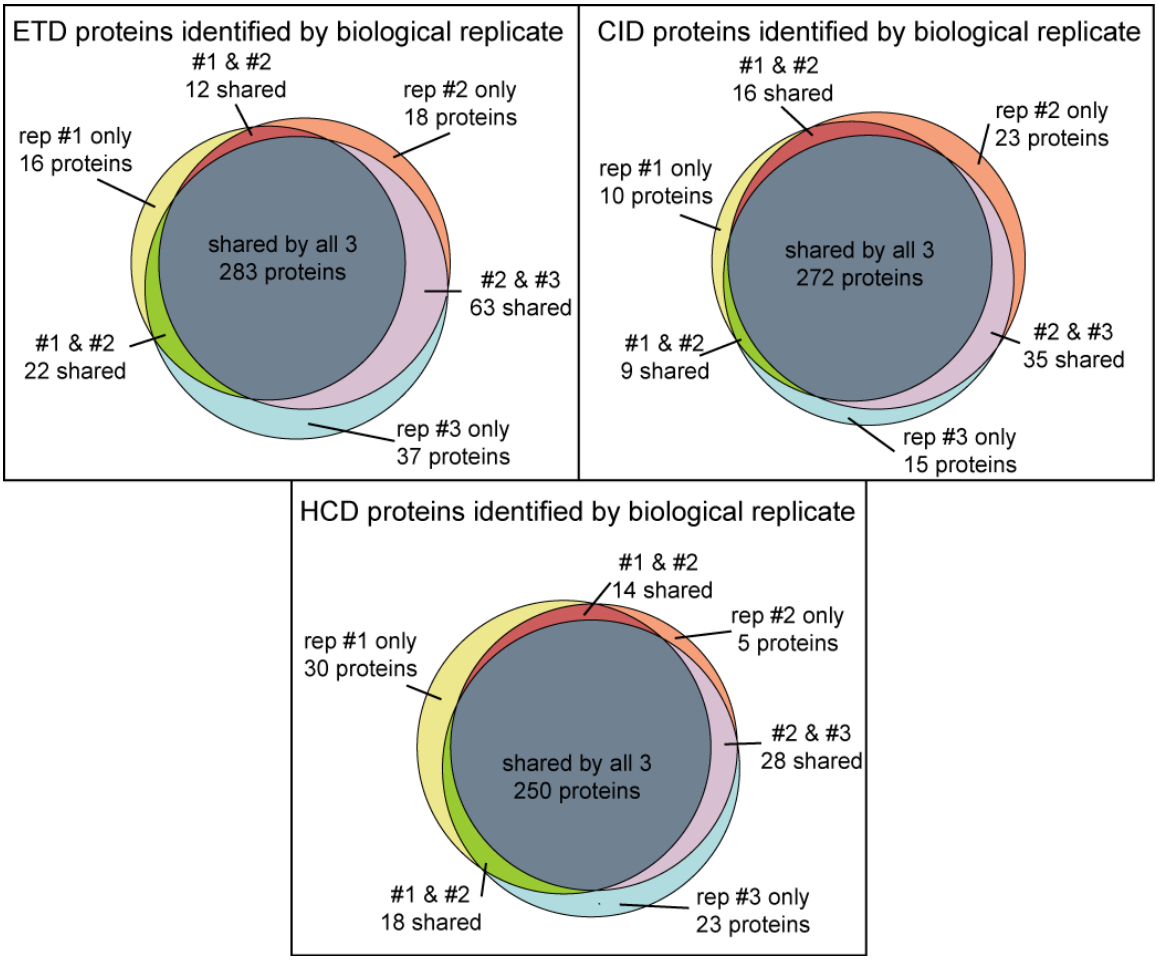

Figure 3.6: Identified peptides by dissociation type and biological replicate.

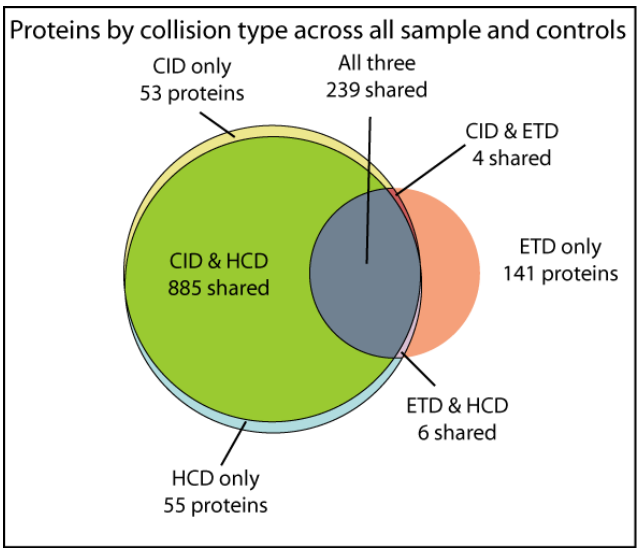

Figure 3.7: Identified proteins by collision type across all replicates.

Table Index	Accession	Protein Name	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
1	Q9PU55	Lymphoid transcription factor (Aiolos)	5.9	28	47	67	37	541	IKZF3
2	F1NT33	DNA-binding protein Ikaros	5.4	50	48	67	33	522	IKZF1
3	F1NT33	DNA-binding protein Ikaros	5.2	50	49	69	31	459	IKZF1
4	R4GIE6	Uncharacterized protein	9.3	33	30	38	24	192	TFAP4 Homolog
5	E1BYY8	Uncharacterized protein	3.8	12	54	32	41	125	Not named
6	F1NIE9	Transcription Factor 12	4.4	7	24	66	20	105	TCF12
7	F2Z895	Paired box protein Pax-5A (Fragment)	2.4	26	56	82	3	82	PAX5
8	Q5ZM50	Uncharacterized protein	2.1	5	16	75	8	57	CBFB
9	Q70IK3	Helix-loop-helix protein E12	4.7	4	15	22	11	55	TCF3
10	A0A1D5P6W8	Uncharacterized protein	2.6	6	53	29	17	50	KDM1A
11	F2Z896	Paired box protein Pax-5B (Fragment)	4.8	1	77	64	2	42	PAX5
12	E1BWG6	Uncharacterized protein	2.2	16	23	24	9	38	RCOR1
13	E1C717	Uncharacterized protein	19.0	38	176	37	12	34	ELF2
14	F1NDM6	Uncharacterized protein	2.7	39	33	44	10	30	PCGF6
15	A0A1D6UPS9	Uncharacterized protein	2.1	7	29	14	14	28	SF3B3
16	Q5ZKP9	Transcription factor max-like protein X	4.7	1	1	36	6	28	MLX
17	E1C979	Uncharacterized protein	3.5	7	13	37	10	27	E2F6
18	Q5ZLC2	Uncharacterized protein	2.4	56	72	19	10	26	L3MBTL2
19	F1NXG3	Ephrin-A2	13.4	42	227	28	13	26	ELF1
20	Q5ZMG4	Uncharacterized protein	13.4	42	227	28	13	26	ELF1
21	Q5ZM49	Uncharacterized protein	4.6	8	84	48	6	25	E2F6
22	E1BYY8	Uncharacterized protein	2.0	19	57	6	14	24	MGA
23	F1P3E8	Uncharacterized protein	3.1	4	7	28	6	24	TFDP1
24	E1BVU2	Uncharacterized protein	12.9	51	41	21	12	22	MLXIP
25	Q5ZMM8	Uncharacterized protein	3.6	9	46	19	7	20	LOC416354; CTBPL
26	E1BRU0	Uncharacterized protein	2.3	64	60	17	10	19	MAD1L1
27	F1NP51	Lamin-B2	2.1	2	61	11	7	18	LMNB2
28	Q9YHW8	Ets domain protein (Fragment)	1000.0	8	X	24	6	16	ETV6
29	F1NRC2	Uncharacterized protein	1000.0	8	X	20	6	16	ETV6
30	Q5ZIZ6	Uncharacterized protein	3.4	4	11	22	7	15	CTBP1

Table 3.1: Proteins with at least a 2x abundance increase in enhancer vs control DT40 samples. (X) Some coefficients of variance could not be calculated because the protein was not detected in at least three replicates. Highlighted proteins were also identified in the Ramos samples.

Table Index	Accession	Protein Name	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
31	Q5ZI54	Uncharacterized protein	3.4	49	93	22	8	14	SF3B3
32	F1NJF0	Uncharacterized protein	3.7	67	60	38	3	12	Not named
33	F1NKI9	circadian locomoter output cycles protein kaput	5.6	70	110	12	7	12	CLOCK
34	F1NKI8	Circadian locomoter output cycles protein kaput	5.6	70	110	12	7	12	CLOCK
35	H9KYX5	Uncharacterized protein	2.5	70	92	11	3	12	LOC100857912; RFX5
36	F1NSF2	Uncharacterized protein	2.7	1	47	18	5	11	FOXK1
37	F1NP75	Uncharacterized protein	2.4	16	88	14	2	10	RFXAP
38	Q5ZLB1	Transcription Factor 12	4.7	42	29	83	3	10	TCF12
39	E1C1C5	Aryl hydrocarbon receptor nuclear translocator-like protein 2	19.6	1	131	14	6	9	ARNTL2
40	A0A1D5P5H2	Uncharacterized protein	3.5	58	89	2	2	8	MBTD1
41	E1BZS3	Uncharacterized protein	3.5	58	89	2	2	8	MBTD1
42	A0A1D5PV94	Uncharacterized protein	3.5	58	89	2	2	8	MBTD1
43	O42388	Ubiquitin-60S ribosomal protein L40	38.8	176	X	15	2	8	UBA52
44	F1NV33	DNA mismatch repair protein	2.1	15	107	9	6	7	MSH2
45	Q76FQ4	DNA mismatch repair protein	2.1	15	107	8	5	6	MSH2
46	F1NKX5	Neurogenic differentiation factor	3.2	40	25	5	2	6	NEUROD1
47	F1NNS0	Uncharacterized protein	3.3	30	51	9	3	6	FOXK2
48	A0A1D5PT56	Uncharacterized protein	3.3	30	51	11	3	6	FOXK2
49	E1BX21	Uncharacterized protein	4.2	60	118	4	3	6	SIN3A
50	F1NEF6	Uncharacterized protein	1000.0	19	X	14	5	5	ACAD9
51	F1NB13	Uncharacterized protein	40.6	102	107	6	2	5	CDC7
52	E1BUG1	Uncharacterized protein	28.6	21	X	2	2	5	XPC
53	F1NX56	Uncharacterized protein	5.9	37	2	5	3	5	OGT
54	F1P1F3	DNA topoisomerase	1000.0	19	X	2	1	4	LOC416518
55	A0A1D5NVH0	Uncharacterized protein	58.1	28	26	3	2	4	MITF
56	Q90ZM5	sterol regulatory element binding protein 1	2.5	37	31	3	3	4	SREBF1
57	A0A1D5PYG1	Uncharacterized protein	5.6	64	64	3	2	4	SFMBT1
58	Q6LEK2	Uncharacterized protein (Fragment)	1000.0	76	X	41	2	4	NPM1
59	B6E281	Microphthalmia-associated transcription factor isoform M	58.1	28	26	4	2	4	MITF
60	F1NR90	Uncharacterized protein	2.8	12	10	6	2	4	E2F3

Table 3.1 continued.

Table Index	Accession	Protein Name	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
61	F1NHX4	REST corepressor 3	2.9	54	152	5	1	4	RCOR3
62	Q9IAU0	Microphthalmia protein	58.1	28	26	4	2	4	MITF
63	O73871	Microphthalmia-associated transcription factor	58.1	28	26	3	2	4	MITF
64	Q9IAI1	BHLH/PAS transcription factor Clock	6.2	21	3	6	3	4	CLOCK
65	R4GLV5	Uncharacterized protein	2.5	37	31	3	3	4	SREBF1
66	E1BTX5	Uncharacterized protein	5.6	64	64	3	2	4	SFMBT1
67	E1C7L0	Uncharacterized protein	4.7	34	53	4	2	4	CTBP2
68	A0A1D5P6U0	DNA (cytosine-5)- methyltransferase	23.5	189	X	1	1	3	Not named
69	C0LUU5	E2F2 transcription factor (Fragment)	2.8	14	10	6	1	3	E2F2
70	Q5W4T6	RAD50 protein (Fragment)	19.1	121	122	1	1	3	RAD50
71	F1NV53	Uncharacterized protein	4.7	34	26	4	3	3	FBXO11
72	R4GHN9	Uncharacterized protein	3.2	70	30	3	3	3	LUZP1
73	E1BUI5	Uncharacterized protein	36.8	8	X	3	3	3	MED1
74	A0A1D5PCZ5	Uncharacterized protein	5.8	59	51	22	2	3	Gga.1264
75	A0A1D5PXC5	Uncharacterized protein	5.8	59	51	17	2	3	Gga.1264
76	F1NEG1	Uncharacterized protein	19.1	121	122	1	1	3	RAD50
77	A0A1D5P9R0	Uncharacterized protein	43.7	33	X	16	2	2	Not named
78	E1BZV0	Neuronal PAS domain-containing protein 2	5.6	71	3	2	2	2	NPAS2
79	Q90979	Nuclear factor erythroid 2-related factor 1	1000.0	7	X	6	2	2	NRF1
80	A0A1D5NX41	Uncharacterized protein	1000.0	X	X	2	1	2	YEATS2
81	F1NRI4	Uncharacterized protein	28.2	75	5	12	1	2	DYNLL2
82	A0A1D5PNQ1	Neuronal PAS domain-containing protein 2	5.6	71	3	2	2	2	NPAS2
83	A0A1D5NY09	Uncharacterized protein	1000.0	X	X	2	1	2	YEATS2
84	F1NV79	Uncharacterized protein	1000.0	14	X	4	2	2	SETDB2
85	F1P0P8	Uncharacterized protein	1000.0	57	X	4	1	2	NFYC
86	F1N8L5	Uncharacterized protein	28.2	75	5	12	1	2	DYNLL1; LOC100859081
87	E1C1L2	Uncharacterized protein	342.2	73	143	8	2	2	TFCP2L1
88	A0A1D5PY06	Uncharacterized protein	1000.0	X	X	2	1	2	YEATS2
89	Q1T768	Dsn1 protein	5.0	1	423	5	1	2	DSN1

Table 3.1 continued.

3.4.2 Ramos Samples

The twelve LC-MS/MS data files from experiments two and three were analyzed as a single set by PD using the same processing and consensus workflows employed for the orbitrap HCD replicates from experiment one, with only the proteins database changed. For the 12 samples there were 124,798 individual MS/MS experiments, 42,434 of which were matched with peptide sequences from the human Swissprot database for an overall PSM rate of 33.62%. These PSMs were condensed into 6,653 unique peptide hits which were then assigned to 2,590 proteins.

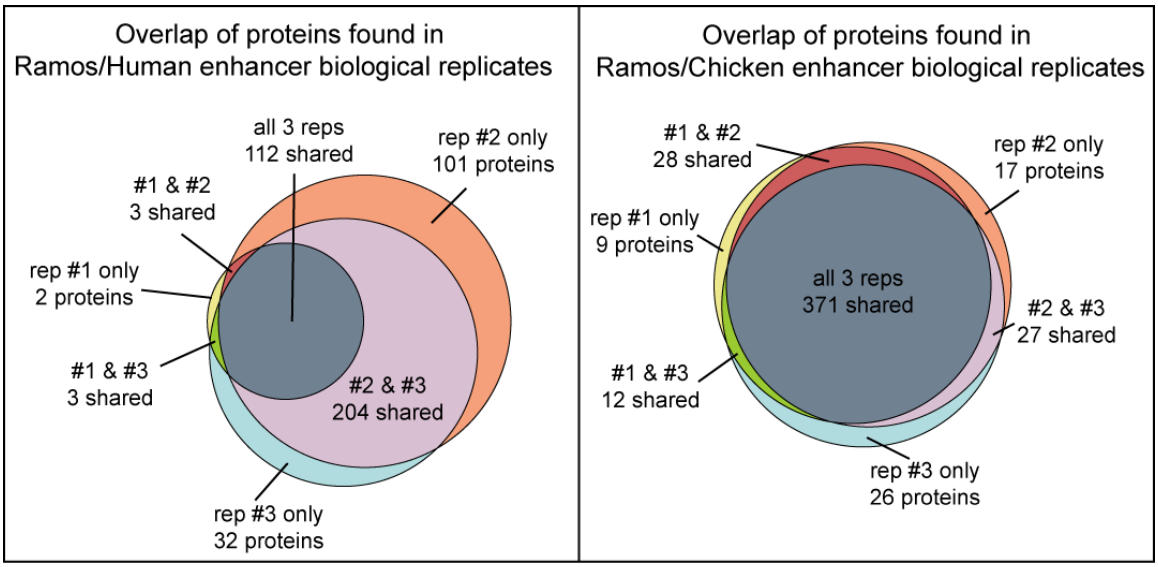

Figure 3.8: Overlap of proteins identified in human and chicken enhancer Ramos samples by biological replicate.

Figure 3.8 shows the overlap between proteins identified from Ramos cells between biological replicates with either the Human or Chicken enhancer sequences. **Figure 3.9** shows the protein abundance distributions in all samples before and after normalization. **Table 3.2** lists the 199 proteins for which there was a 2X signal increase

over control in the Ramos/human samples, and **Table 3.3** shows the 184 proteins for the Ramos/Chicken Samples. There were 88 proteins, highlighted in grey, that were enriched at least 2X over control by both the chicken and human enhancer sequences. These results show a similar distribution of protein properties as the DT40 samples.

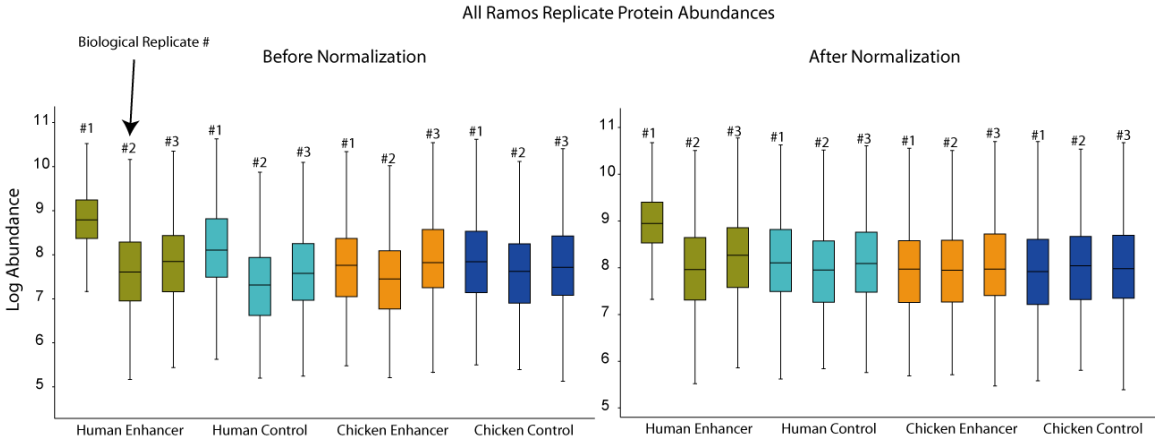

Figure 3.9: Protein abundances for all Ramos replicates.

3.4.3 Ikaros, Aiolos, and Helios

Among all three experiments the protein most consistently enriched in the enhancer samples was Ikaros. Ikaros is a zinc finger DNA binding protein with well-established links to transcriptional regulation and B cell development. The chicken and human canonical Ikaros sequences are 519 and 518 amino acids in length respectively, both have 6 zinc finger domains, and are 85% identical. The protein Ikaros is a member of the Ikaros family of zinc finger transcription factors which also includes Aiolos, Helios, Eos, and Pegasus. There are 8 known splice variants of Ikaros, Aiolos has 16, Helios 8, Eos 2, but Pegasus only 1. Aiolos and Helios were also identified by

the data processing algorithm as being enriched in the DIVAC samples in all three experiments. Figure 3.x shows the alignment of the canonical sequences for Ikaros, Aiolos, and Helios. Regions of each protein potential observed by mass spectrometry are highlighted. While there are substantial numbers of peptides unique to Ikaros or Aiolos, the only peptides assigned to Helios by the Proteome Discoverer search algorithm are also part of the sequence of Ikaros and Aiolos. Due to its lack of unique peptides, it is unlikely that Helios is actually enriched in the enhancer samples.

Figure 3.x shows the 8 human Ikaros isoform sequences aligned and regions of the protein observed by mass spectrometry are highlighted. Unfortunately, because of the extensive sequence overlap among Ikaros isoforms it is impossible to determine which ones are present in our sample. Additionally, there is no discernable pattern to the abundance in peptides across the Ikaros sequence to suggest the dominance of one Isoform over another.

Table Index	Accession	Description	Abundance Ratio Sample/Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
1	O60832-1	H/ACA ribonucleoprotein complex subunit 4	2.2	1	31	58	30	293	DKC1
2	O60832-2	Isoform 3 of H/ACA ribonucleoprotein complex subunit 4	2.2	19	35	52	19	215	DKC1
3	Q13422	DNA-binding protein Ikaros	2.2	53	58	46	20	167	IKZF1
4	Q13422-7	Isoform Ik7 of DNA-binding protein Ikaros	2.2	39	60	44	18	159	IKZF1
5	Q13422-2	Isoform Ik2 of DNA-binding protein Ikaros	3.0	66	51	45	16	158	IKZF1
6	Q13422-3	Isoform Ik3 of DNA-binding protein Ikaros	2.2	18	44	44	16	155	IKZF1
7	P49711-1	Transcriptional repressor CTCF	3.0	82	20	35	18	133	CTCF
8	Q9UKT9-4	Isoform 4 of Zinc finger protein Aiolos	2.3	53	9	43	18	113	IKZF3
9	P49711-2	Isoform 2 of Transcriptional repressor CTCF	4.2	55	23	32	11	105	CTCF
10	Q9UKT9-6	Isoform 6 of Zinc finger protein Aiolos	2.3	31	16	42	16	102	IKZF3
11	Q9UKT9-3	Isoform 3 of Zinc finger protein Aiolos	2.3	31	16	39	16	102	IKZF3
12	Q02447-1	transcription factor Sp3	6.3	23	8	10	8	98	SP3
13	Q02447-3	Isoform 3 of Transcription factor Sp3	6.3	23	8	16	8	98	SP3
14	Q02447-5	Isoform 5 of Transcription factor Sp3	6.3	23	8	11	8	98	SP3
15	Q9UKT9-8	Isoform 8 of Zinc finger protein Aiolos	2.3	29	23	37	15	91	IKZF3
16	Q8NFW8	N-acylneuraminate cytidylyltransferase	3.1	81	4	42	15	90	CMAS
17	Q9UKT9-14	Isoform 14 of Zinc finger protein Aiolos	4.1	34	7	43	10	71	IKZF3
18	P36873-2	Isoform Gamma-2 of Serine/threonine-protein phosphatase PP1-gamma catalytic subunit	2.6	10	74	31	8	66	PPP1CC
19	P36873-1	serine/threonine-protein phosphatase PP1-gamma catalytic subunit	2.6	10	74	32	8	66	PPP1CC
20	Q01664	Transcription factor AP-4	2.6	86	21	33	9	61	TFAP4
21	P56270-2	Isoform 2 of Myc-associated zinc finger protein	3.1	68	107	26	9	51	MAZ
22	P25490	Transcriptional repressor protein YY1	5.3	47	88	19	9	48	YY1
23	P62750	60S ribosomal protein L23a	2.1	21	42	51	10	47	RPL23A
24	P56270-1	Myc-associated zinc finger protein	3.4	42	112	36	9	45	MAZ
25	P56270-3	Isoform 3 of Myc-associated zinc finger protein	3.4	42	112	37	9	45	MAZ
26	Q13422-8	Isoform Ikx of DNA-binding protein Ikaros	3.0	10	113	41	8	42	IKZF1
27	P26583	High mobility group protein B2	4.4	19	140	29	7	39	HMGB2
28	Q02080	Myocyte-specific enhancer factor 2B	2.5	144	60	27	7	37	MEF2BNB-MEF2B
29	P39019	40S ribosomal protein S19	8.3	35	85	43	8	36	RPS19
30	Q9H9B1	histone-lysine N-methyltransferase EHMT1	4.7	74	32	8	7	30	EHMT1

Table 3.2: Proteins with at least a 2x greater abundance in Ramos/Human enhancer samples vs control. Proteins highlighted in grey are common between the Ramos/human and Ramos/chicken samples.

Table Index	Accession	Description	Abundance Ratio Sample/Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
31	Q9H9B1-3	Isoform 3 of Histone-lysine N-methyltransferase EHMT1	4.7	74	32	8	7	30	EHMT1
32	Q9NZI8	Insulin-like growth factor 2 mRNA-binding protein 1	7.3	40	19	19	8	28	IGF2BP1
33	Q96MM3	Zinc finger protein 42 homolog	3.3	54	51	10	3	27	ZFP42
34	O15391	transcription factor YY2	5.3	13	62	11	4	27	YY2
35	Q9HBE1-1	POZ-, AT hook-, and zinc finger-containing protein 1	2.2	16	134	13	7	25	PATZ1
36	Q9HBE1-2	Isoform 2 of POZ-, AT hook-, and zinc finger-containing protein 1	2.2	16	134	17	7	25	PATZ1
37	Q9HBE1-3	Isoform 3 of POZ-, AT hook-, and zinc finger-containing protein 1	2.2	16	134	14	7	25	PATZ1
38	Q12873-1	Chromodomain-helicase-DNA-binding protein 3	2.2	104	73	4	6	21	CHD3
39	Q12873-2	Isoform 2 of Chromodomain-helicase-DNA-binding protein 3	2.2	104	73	4	6	21	CHD3
40	Q5T5X7	BEN domain-containing protein 3	2.8		X	27	14	21	BEND3
41	Q12873-3	Isoform 3 of Chromodomain-helicase-DNA-binding protein 3	2.2	104	73	4	6	21	CHD3
42	Q8IWS0-2	Isoform 2 of PHD finger protein 6	3.1	215	67	14	4	19	PHF6
43	Q9H9B1-4	Isoform 4 of Histone-lysine N-methyltransferase EHMT1	23.5	117	643	6	4	19	EHMT1
44	Q9HBE1-4	Isoform 4 of POZ-, AT hook-, and zinc finger-containing protein 1	3.6	13	136	12	5	18	PATZ1
45	P15923	Transcription factor E2-alpha	6.0	196	82	20	6	17	TCF3
46	P15923-3	Isoform 3 of Transcription factor E2-alpha	6.0	196	82	14	5	16	TCF3
47	Q9NPE3	H/ACA ribonucleoprotein complex subunit 3	4.4	25	14	52	3	15	NOP10
48	Q9NSU2-1	Three-prime repair exonuclease 1	4.3	100	20	22	5	15	TREX1
49	Q15007	Pre-mRNA-splicing regulator WTAP	2.8	45	24	16	4	13	WTAP
50	Q9UKS7-1	Zinc finger protein Helios	4.1	18	87	3	2	11	IKZF2
51	Q9UKS7-2	Isoform 2 of Zinc finger protein Helios	4.1	18	87	3	2	11	IKZF2
52	Q8IZL8	Proline-, glutamic acid- and leucine-rich protein 1	2.3	1	84	6	4	10	PELP1
53	Q96QD9-1	UAP56-interacting factor	4.1	131	50	8	2	9	FYTTD1
54	Q9Y3B2	Exosome complex component csl4	37.8	25	X	24	3	9	EXOSC1
55	Q15427	Splicing factor 3b subunit 4	2.1	3	61	9	2	9	SF3B4
56	O95232-1	Luc7-like protein 3	1000.0	X	X	8	3	8	LUC7L3
57	Q86U86-1	Protein polybromo-1	30.5	33	X	5	5	8	PBRM1
58	P62847-2	Isoform 2 of 40S ribosomal protein S24	16.3	22	X	21	2	7	RPS24
59	Q6NW34	Nucleolus and neural progenitor protein	5.9	72	87	4	2	7	C3orf17; NEPRO
60	P16632-1	40S ribosomal protein S24	16.3	22	X	20	2	7	RPS24

Table 3.2 continued.

Table Index	Accession	Description	Abundance Ratio Sample/Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
61	Q9Y2S0-1	DNA-directed RNA polymerases I and III subunit RPAC2	26.5	136	2	8	1	7	POLR1D
62	P62847-4	Isoform 4 of 40S ribosomal protein S24	16.3	22	X	9	2	7	RPS24
63	P62847-3	Isoform 3 of 40S ribosomal protein S24	16.3	22	X	20	2	7	RPS24
64	Q9P1U0	DNA-directed RNA polymerase I subunit RPA12	34.1	1	X	13	1	7	ZNRD1
65	Q6NW34-2	Isoform 2 of Nucleolus and neural progenitor protein	5.9	72	87	5	2	7	C3orf17; NEPRO
66	P08048-3	Isoform 3 of Zinc finger Y-chromosomal protein	70.1	104	30	5	2	6	ZFY
67	P16989-2	Isoform 2 of Y-box-binding protein 3	35.8	83	X	9	2	6	CSDA; YBX3
68	P08048-2	Isoform 2 of Zinc finger Y-chromosomal protein	70.1	104	30	5	2	6	ZFY
69	P17010-3	Isoform 3 of Zinc finger X-chromosomal protein	70.1	104	30	4	2	6	ZFX
70	P08048	Zinc finger Y-chromosomal protein	70.1	104	30	4	2	6	ZFY
71	P17010	Zinc finger X-chromosomal protein	70.1	104	30	4	2	6	ZFX
72	O94906-1	Pre-mRNA-processing factor 6	1000.0	293	X	6	4	6	PRPF6
73	P17010-2	Isoform 2 of Zinc finger X-chromosomal protein	70.1	104	30	6	2	6	ZFX
74	O94906-2	Isoform 2 of Pre-mRNA-processing factor 6	1000.0	293	X	6	4	6	PRPF6
75	P16989-1	Y-box-binding protein 3	35.8	83	X	7	2	6	CSDA; YBX3
76	Q9Y483-2	Isoform 2 of Metal-response element-binding transcription factor 2	45.4	390	X	22	3	6	MTF2
77	Q2VY69	zinc finger protein 284	2.4	X	X	5	1	6	ZNF284
78	Q96C00	zinc finger and BTB domain-containing protein 9	1000.0	X	X	9	2	6	ZBTB9
79	Q13610-2	Isoform 2 of Periodic tryptophan protein 1 homolog	1000.0	51	X	6	1	5	PWP1
80	Q8NI51-11	Isoform 11 of Transcriptional repressor CTCFL	8.6	20	530	7	3	5	CTCFL
81	P45973	chromobox protein homolog 5	157.2	81	X	13	2	5	CBX5
82	Q8NI51-5	Isoform 5 of Transcriptional repressor CTCFL	8.6	20	530	5	3	5	CTCFL
83	Q8NI51-7	Isoform 7 of Transcriptional repressor CTCFL	8.6	20	530	3	3	5	CTCFL
84	Q8NI51-10	Isoform 10 of Transcriptional repressor CTCFL	8.6	20	530	5	3	5	CTCFL
85	Q8NI51	Transcriptional repressor CTCFL	8.6	20	530	3	3	5	CTCFL
86	Q8NAV1-1	Pre-mRNA-splicing factor 38A	35.6	58	X	4	1	5	PRPF38A
87	O14519	Cyclin-dependent kinase 2-associated protein 1	37.8	77	X	9	1	5	CDK2AP1
88	Q8NI51-6	Isoform 6 of Transcriptional repressor CTCFL	8.6	20	530	5	3	5	CTCFL
89	Q8NI51-3	Isoform 3 of Transcriptional repressor CTCFL	8.6	20	530	3	3	5	CTCFL
90	Q6PI98-1	INO80 complex subunit C	2.1	187	30	16	2	5	INO80C

Table 3.2 continued.

Table Index	Accession	Description	Abundance Ratio Sample/Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
91	Q8NI51-2	Isoform 2 of Transcriptional repressor CTCFL	8.6	20	530	4	3	5	CTCFL
92	O43660-1	Pleiotropic regulator 1	3.1	30	22	9	2	5	PLRG1
93	O14519-2	Isoform 2 of Cyclin-dependent kinase 2-associated protein 1	37.8	77	X	11	1	5	CDK2AP1
94	P43246-1	DNA mismatch repair protein MSH2	13.1	11	19	5	3	4	MSH2
95	P37275-4	Isoform 4 of Zinc finger E-box-binding homeobox 1	1000.0	X	X	3	3	4	ZEB1
96	Q8NI51-8	Isoform 8 of Transcriptional repressor CTCFL	8.6	20	530	2	2	4	CTCFL
97	Q96EZ8-3	Isoform 3 of Microspherule protein 1	1000.0	X	X	7	3	4	MCRS1
98	Q96EZ8	Microspherule protein 1	1000.0	X	X	7	3	4	MCRS1
99	P61244-3	Isoform 3 of Protein max	1000.0	X	X	18	2	4	MAX
100	P17483	homeobox protein Hox-B4	28.7	46	X	8	1	4	HOXB4
101	P37275-2	Isoform 2 of Zinc finger E-box-binding homeobox 1	1000.0	X	X	3	3	4	ZEB1
102	Q9NU22	Midasin	1000.0	X	X	1	4	4	MDN1
103	Q9Y2D9	Zinc finger protein 652	1000.0	64	X	5	2	4	ZNF652
104	Q15723-5	ETS-related transcription factor Elf-2	1000.0	X	X	7	2	4	ELF2
105	O75940	Survival of motor neuron-related-splicing factor 30	12.8	136	87	25	3	4	SMNDC1
106	P37275-3	Isoform 3 of Zinc finger E-box-binding homeobox 1	1000.0	X	X	3	3	4	ZEB1
107	Q9H6W3	Bifunctional lysine-specific demethylase and histidyl-hydroxylase NO66	1000.0	344	X	5	2	4	C14orf169; RIOX1
108	Q15723-2	Isoform 2 of ETS-related transcription factor Elf-2	1000.0	X	X	8	2	4	ELF2
109	Q13573	SNW domain-containing protein 1	1000.0	X	X	7	2	4	SNW1
110	Q96EZ8-2	Isoform 2 of Microspherule protein 1	1000.0	X	X	7	3	4	MCRS1
111	Q15723-3	Isoform 3 of ETS-related transcription factor Elf-2	1000.0	X	X	8	2	4	ELF2
112	P37275-1	Zinc finger E-box-binding homeobox 1	1000.0	X	X	3	3	4	ZEB1
113	P41182-2	Isoform 2 of B-cell lymphoma 6 protein	5.1	X	X	4	2	4	BCL6
114	Q15723-1	Isoform 1 of ETS-related transcription factor Elf-2	1000.0	X	X	7	2	4	ELF2
115	P61244-4	Isoform 4 of Protein max	1000.0	X	X	14	2	4	MAX
116	O94805	Actin-like protein 6B	35.8	X	X	3	1	4	ACTL6B
117	P41182-1	B-cell lymphoma 6 protein	5.1	X	X	4	2	4	BCL6
118	P37275-5	Isoform 5 of Zinc finger E-box-binding homeobox 1	1000.0	X	X	3	3	4	ZEB1
119	Q8TF20-1	Zinc finger protein 721	2.4	X	X	2	1	3	ZNF721
120	P52434-2	Isoform 2 of DNA-directed RNA polymerases I, II, and III subunit RPABC3	1000.0	X	X	13	1	3	POLR2H

Table 3.2 continued.

Table Index	Accession	Description	Abundance Ratio Sample/Control	CV [%] Sample	CV [%] Control	Coverage [%]	#Peptides	#PSMs	Gene Symbol
122	Q8N2I2-4	Isoform 4 of Zinc finger protein 619	2.4	X	X	1	1	3	ZNF619
123	Q9UK13	Zinc finger protein 221	2.4	X	X	1	1	3	ZNF221
124	Q9UK12-1	Zinc finger protein 222	2.4	X	X	2	1	3	ZNF222
125	O75373	Zinc finger protein 737	2.4	X	X	2	1	3	ZNF737
126	Q8NEM1-1	Zinc finger protein 680	2.4	X	X	2	1	3	ZNF680
127	P52434-5	Isoform 5 of DNA-directed RNA polymerases I, II, and III subunit RPABC3	1000.0	X	X	19	1	3	POLR2H
128	P51522-2	Isoform 2 of Zinc finger protein 83	2.4	X	X	2	1	3	ZNF83
129	P52737	Zinc finger protein 136	2.4	X	X	2	1	3	ZNF136
130	P17097	Zinc finger protein 7	2.4	X	X	1	1	3	ZNF7
131	Q3ZCX4-2	Isoform 2 of Zinc finger protein 568	2.4	X	X	2	1	3	ZNF568
132	Q6ZMS4	Zinc finger protein 852	2.4	X	X	2	1	3	ZNF852
133	Q8NB42-1	Zinc finger protein 527	2.4	X	X	1	1	3	ZNF527
134	Q13469-3	Isoform 3 of Nuclear factor of activated T-cells, cytoplasmic 2	1000.0	X	X	2	1	3	NFATC2
135	Q9NVN8	Guanine nucleotide-binding protein-like 3-like protein	1000.0	101	X	8	3	3	GNL3L
136	Q14588	zinc finger protein 234	2.4	X	X	1	1	3	ZNF234
137	Q13469-4	Isoform 4 of Nuclear factor of activated T-cells, cytoplasmic 2	1000.0	X	X	2	1	3	NFATC2
138	P17097-2	Isoform 2 of Zinc finger protein 7	2.4	X	X	1	1	3	ZNF7
139	Q9H9L3	Interferon-stimulated 20 kDa exonuclease-like 2	1000.0	159	X	8	2	3	ISG20L2
140	P17027	Zinc finger protein 23	2.4	X	X	1	1	3	ZNF23
141	Q3ZCX4	Zinc finger protein 568	2.4	X	X	1	1	3	ZNF568
142	O60765	Zinc finger protein 354A	2.4	X	X	3	1	3	ZNF354A
143	Q04837	Single-stranded DNA-binding protein, mitochondrial	2.1	X	X	10	1	3	SSBP1
144	Q9UK10	Zinc finger protein 225	2.4	X	X	3	1	3	ZNF225
145	Q96LW1	Zinc finger protein 354B	2.4	X	X	3	1	3	ZNF354B
146	P51815	Zinc finger protein 75D	2.4	X	X	2	1	3	ZNF75D
147	P48729-1	Casein kinase I isoform alpha	1000.0	8	X	6	1	3	CSNK1A1
148	Q8NB42-2	Isoform 2 of Zinc finger protein 527	2.4	X	X	2	1	3	ZNF527
149	Q92670	Putative zinc finger protein 75C	2.4	X	X	2	1	3	ZNF75CP
150	Q9Y383	Putative RNA-binding protein Luc7-like 2	197.1	42	X	5	2	3	LUC7L2

Table 3.2 continued.

Table Index	Accession	Description	Abundance Ratio Sample/Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
152	Q8N7M2	Zinc finger protein 283	2.4	X	X	1	1	3	ZNF283
153	Q9UK12-2	Isoform 2 of Zinc finger protein 222	2.4	X	X	2	1	3	ZNF222
154	Q16670-1	Zinc finger and SCAN domain-containing protein 26	2.4	X	X	2	1	3	ZNF187; ZSCAN26
155	Q6ZN06	zinc finger protein 813	2.4	X	X	1	1	3	ZNF813
156	Q12901	Zinc finger protein 155	2.4	X	X	2	1	3	ZNF155
157	Q13123	Protein Red	1000.0	X	X	2	1	3	IK
158	Q8TF32	Zinc finger protein 431	2.4	X	X	2	1	3	ZNF431
159	P51815-2	Isoform 2 of Zinc finger protein 75D	2.4	X	X	2	1	3	ZNF75D
160	P17027-2	Isoform 2 of Zinc finger protein 23	2.4	X	X	2	1	3	ZNF23
161	Q8N2I2-3	Isoform 3 of Zinc finger protein 619	2.4	X	X	2	1	3	ZNF619
162	Q8N2I2-2	Isoform 2 of Zinc finger protein 619	2.4	X	X	2	1	3	ZNF619
163	Q9NRF9	DNA polymerase epsilon subunit 3	1000.0	X	X	10	1	3	POLE3
164	Q7Z3B4	Nucleoporin p54	1000.0	X	X	3	1	3	NUP54
165	Q8TF20-2	Isoform 2 of Zinc finger protein 721	2.4	X	X	2	1	3	ZNF721
166	Q12962	transcription initiation factor TFIID subunit 10	31.6	41	X	22	3	3	TAF10
167	Q8N2I2	Zinc finger protein 619	2.4	X	X	2	1	3	ZNF619
168	Q13469-2	Isoform 2 of Nuclear factor of activated T-cells, cytoplasmic 2	1000.0	X	X	2	1	3	NFATC2
169	Q8N9F8	zinc finger protein 454	2.4	X	X	2	1	3	ZNF454
170	Q6PG37	Zinc finger protein 790	2.4	X	X	1	1	3	ZNF790
171	Q13469-5	Isoform 5 of Nuclear factor of activated T-cells, cytoplasmic 2	1000.0	X	X	2	1	3	NFATC2
172	Q9NVV9-1	THAP domain-containing protein 1	2.3	X	X	13	2	3	THAP1
173	Q9UK11	Zinc finger protein 223	2.4	X	X	2	1	3	ZNF223
174	Q9Y2H8	Zinc finger protein 510	2.4	X	X	1	1	3	ZNF510
175	P51522	zinc finger protein 83	2.4	X	X	2	1	3	ZNF83
176	Q9C0G0	zinc finger protein 407	2.4	X	X	0	1	3	ZNF407
177	Q99676	Zinc finger protein 184	2.4	X	X	1	1	3	ZNF184
178	Q8TB69	Zinc finger protein 519	2.4	X	X	3	1	3	ZNF519
179	Q9C0G0-2	Isoform 2 of Zinc finger protein 407	2.4	X	X	0	1	3	ZNF407
180	Q9NYT6	Zinc finger protein 226	2.4	X	X	1	1	3	ZNF226

Table 3.2 continued.

Table Index	Accession	Description	Abundance Ratio Sample/Control	CV [%] Sample	CV [%] Control	Coverage [%]	#Peptides	#PSMs	Gene Symbol
182	Q13469	Nuclear factor of activated T-cells, cytoplasmic 2	1000.0	X	X	2	1	3	NFATC2
183	O00425	Insulin-like growth factor 2 mRNA-binding protein 3	1000.0	X	X	5	2	3	IGF2BP3
184	Q9NZL3	Zinc finger protein 224	2.4	X	X	1	1	3	ZNF224
185	Q92900	Regulator of nonsense transcripts 1	1000.0	X	X	3	2	2	UPF1
186	P52701	DNA mismatch repair protein MSH6	38.6	89	X	2	1	2	MSH6
187	Q9NXV6	CDKN2A-interacting protein	1000.0	X	X	5	2	2	CDKN2AIP
188	Q4VC44-3	Isoform 3 of FLYWCH-type zinc finger-containing protein 1	6.5	X	X	4	2	2	FLYWCH1
189	Q9NV31	U3 small nucleolar ribonucleoprotein protein IMP3	1000.0	78	X	8	1	2	IMP3
190	Q5JTH9-1	RRP12-like protein	1000.0	X	X	3	2	2	RRP12
191	Q9Y4A0	Jerky protein homolog-like	29.8	94	X	5	2	2	JRKL
192	Q8N488	RING1 and YY1-binding protein	35.2	316	X	19	2	2	RYBP
193	Q92900-2	Isoform 2 of Regulator of nonsense transcripts 1	1000.0	X	X	3	2	2	UPF1
194	Q4VC44	FLYWCH-type zinc finger-containing protein 1	6.5	X	X	4	2	2	FLYWCH1
195	Q9H7Z6	Histone acetyltransferase KAT8	75.7	61	X	4	2	2	KAT8
196	Q9H7Z6-2	Isoform 2 of Histone acetyltransferase KAT8	75.7	61	X	4	2	2	KAT8
197	Q15723-4	Isoform 4 of ETS-related transcription factor Elf-2	1000.0	X	X	5	1	2	ELF2
198	Q8NBZ0	INO80 complex subunit E	1000.0	X	X	16	2	2	INO80E
199	P49756-1	RNA-binding protein 25	1000.0	X	X	2	1	2	RBM25

Table 3.2 continued.

Table Index	Accession	Description	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
1	Q13422	DNA-binding protein Ikaros	3.9	23	3	46	20	167	IKZF1
2	Q13422-7	Isoform Ik7 of DNA-binding protein Ikaros	3.6	39	9	44	18	159	IKZF1
3	Q13422-2	Isoform Ik2 of DNA-binding protein Ikaros	4.4	7	16	45	16	158	IKZF1
4	Q13422-3	Isoform Ik3 of DNA-binding protein Ikaros	3.6	40	9	44	16	155	IKZF1
5	Q9UKT9-1	Zinc finger protein Aiolos	2.3	15	10	43	19	129	IKZF3
6	Q13422-5	Isoform Ik5 of DNA-binding protein Ikaros	4.1	28	47	41	13	128	IKZF1
7	Q9UKT9-7	Isoform 7 of Zinc finger protein Aiolos	2.2	18	12	41	18	118	IKZF3
8	Q13422-6	Isoform Ik6 of DNA-binding protein Ikaros	4.3	50	61	39	9	114	IKZF1
9	Q9UKT9-4	Isoform 4 of Zinc finger protein Aiolos	3.0	1	5	43	18	113	IKZF3
10	Q9UKT9-3	Isoform 3 of Zinc finger protein Aiolos	2.4	6	1	39	16	102	IKZF3
11	Q9UKT9-6	Isoform 6 of Zinc finger protein Aiolos	2.4	6	1	42	16	102	IKZF3
12	Q9UKT9-9	Isoform 9 of Zinc finger protein Aiolos	3.3	13	15	41	15	101	IKZF3
13	Q02447-3	Isoform 3 of Transcription factor Sp3	2.1	13	1	16	8	98	SP3
14	Q02447-1	transcription factor Sp3	2.1	13	1	10	8	98	SP3
15	Q02447-5	Isoform 5 of Transcription factor Sp3	2.1	13	1	11	8	98	SP3
16	Q9UKT9-2	Isoform 2 of Zinc finger protein Aiolos	2.4	8	6	36	14	95	IKZF3
17	Q9UKT9-8	Isoform 8 of Zinc finger protein Aiolos	2.4	4	2	37	15	91	IKZF3
18	Q9UKT9-13	Isoform 13 of Zinc finger protein Aiolos	2.4	3	14	36	12	75	IKZF3
19	Q9UKT9-5	Isoform 5 of Zinc finger protein Aiolos	2.4	3	14	33	12	75	IKZF3
20	Q9UKT9-10	Isoform 10 of Zinc finger protein Aiolos	3.2	18	1	37	11	73	IKZF3
21	Q9UKT9-14	Isoform 14 of Zinc finger protein Aiolos	2.8	7	17	43	10	71	IKZF3
22	Q01664	Transcription factor AP-4	4.9	2	51	33	9	61	TFAP4
23	Q9UKT9-11	Isoform 11 of Zinc finger protein Aiolos	2.5	7	18	33	9	53	IKZF3
24	Q9UKT9-12	Isoform 12 of Zinc finger protein Aiolos	2.5	7	18	38	9	53	IKZF3
25	Q9UKT9-16	Isoform 16 of Zinc finger protein Aiolos	2.1	12	10	33	8	42	IKZF3
26	Q13422-8	Isoform Ikx of DNA-binding protein Ikaros	3.5	26	4	41	8	42	IKZF1
27	P32519-1	ETS-related transcription factor Elf-1	9.1	50	7	28	10	35	ELF1
28	P32519-2	Isoform 2 of ETS-related transcription factor Elf-1	9.1	50	7	29	10	35	ELF1
29	P48382	DNA-binding protein RFX5	2.5	58	118	21	9	29	RFX5
30	P48382-2	Isoform 2 of DNA-binding protein RFX5	2.5	58	118	22	9	29	RFX5

Table 3.3: Proteins with at least a 2x greater abundance in Ramos/chicken enhancer samples vs control. Proteins highlighted in grey are common between the Ramos/human and Ramos/chicken samples.

Table Index	Accession	Description	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
31	Q8IWI9-1	MAX gene-associated protein	2.3	20	33	6	10	28	MGA
32	Q8IWI9-4	Isoform 4 of MAX gene-associated protein	2.3	20	33	6	10	28	MGA
33	O15391	transcription factor YY2	2.2	6	49	11	4	27	YY2
34	Q8IWI9-3	Isoform 3 of MAX gene-associated protein	2.2	21	23	6	9	27	MGA
35	Q8N9N5-1	Protein BANP	2.6	22	18	15	7	26	BANP
36	P15923-2	Isoform E47 of Transcription factor E2-alpha	2.5	102	91	25	8	23	TCF3
37	Q5T5X7	BEN domain-containing protein 3	6.3	46	101	27	14	21	BEND3
38	P15923	Transcription factor E2-alpha	4.6	37	102	20	6	17	TCF3
39	P15923-3	Isoform 3 of Transcription factor E2-alpha	4.6	37	102	14	5	16	TCF3
40	Q13951	Core-binding factor subunit beta	2.0	35	13	29	3	15	CBFB
41	Q13951-2	Isoform 2 of Core-binding factor subunit beta	2.7	48	24	32	3	15	CBFB
42	Q13950-3	Isoform 3 of Runt-related transcription factor 2	2.9	65	105	10	4	15	RUNX2
43	Q15910-2	Isoform 2 of Histone-lysine N-methyltransferase EZH2	34.1	54	555	17	7	14	EZH2
44	Q15910-3	Isoform 3 of Histone-lysine N-methyltransferase EZH2	34.1	54	555	18	7	14	EZH2
45	Q15910	Histone-lysine N-methyltransferase EZH2	34.1	54	555	17	7	14	EZH2
46	P08047-2	Isoform 2 of Transcription factor Sp1	2.4	88	51	10	5	13	SP1
47	Q13950-2	Isoform 2 of Runt-related transcription factor 2	2.2	65	105	7	3	13	RUNX2
48	Q13950-1	runt-related transcription factor 2	2.2	65	105	7	3	13	RUNX2
49	Q15910-4	Isoform 4 of Histone-lysine N-methyltransferase EZH2	31.7	54	498	16	6	13	EZH2
50	P08047	Transcription factor Sp1	2.4	88	51	10	5	13	SP1
51	Q6DD87	Zinc finger protein 787	16.4	9	102	21	4	12	ZNF787
52	P25912-1	Protein max	2.5	39	48	31	4	12	MAX
53	P08047-3	Isoform 3 of Transcription factor Sp1	2.4	88	51	6	4	12	SP1
54	O00287	Regulatory factor X-associated protein	2.1	43	85	23	4	12	RFXAP
55	Q8NHQ9	ATP-dependent RNA helicase DDX55	1000.0	20	X	14	5	12	DDX55
56	P61244-2	Isoform 2 of Protein max	2.5	39	48	33	4	12	MAX
57	Q01081	Splicing factor U2AF 35 kDa subunit	2.4	27	77	13	2	12	U2AF1; LOC102724594; U2AF1L5
58	Q9UKS7-2	Isoform 2 of Zinc finger protein Helios	3.7	71	31	3	2	11	IKZF2
59	Q9UKS7-1	Zinc finger protein Helios	3.7	71	31	3	2	11	IKZF2
60	Q15910-5	Isoform 5 of Histone-lysine N-methyltransferase EZH2	1000.0	54	X	15	5	10	EZH2

Table 3.3 continued.

Table Index	Accession	Description	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
61	Q7Z2E3-2	Isoform 2 of Aprataxin	16.2	17	280	18	2	10	APTX
62	Q15853-4	Isoform USF2c of Upstream stimulatory factor 2	3.6	23	4	11	3	8	USF2
63	Q15853	Upstream stimulatory factor 2	3.6	23	4	7	3	8	USF2
64	Q15853-3	Isoform USF2B of Upstream stimulatory factor 2	3.6	23	4	8	3	8	USF2
65	O95232-1	Luc7-like protein 3	27.2	369	X	8	3	8	LUC7L3
66	P22415	Upstream stimulatory factor 1	3.6	44	3	10	3	7	USF1
67	Q9BXK1	krueppel-like factor 16	2.4	45	76	25	3	7	KLF16
68	P46063	ATP-dependent DNA helicase Q1	30.5	36	X	8	3	7	RECQL
69	P22415-2	Isoform 2 of Upstream stimulatory factor 1	3.6	44	3	12	3	7	USF1
70	Q96C00	zinc finger and BTB domain-containing protein 9	2.7	44	103	9	2	6	ZBTB9
71	P48634-1	Protein PRRC2A	5.4	13	12	3	4	6	PRRC2A
72	Q99496	E3 ubiquitin-protein ligase RING2	34.9	95	X	7	2	6	RNF2
73	Q2VY69	zinc finger protein 284	46.9	82	X	5	1	6	ZNF284
74	P51610-1	Host cell factor 1	3.9	37	104	2	3	5	HCFC1
75	P51610-4	Isoform 4 of Host cell factor 1	3.9	37	104	2	3	5	HCFC1
76	Q8TAQ2	SWI/SNF complex subunit SMARCC2	45.0	31	240	3	2	4	SMARCC2
77	Q15723-1	Isoform 1 of ETS-related transcription factor Elf-2	1000.0	17	X	7	2	4	ELF2
78	Q15723-3	Isoform 3 of ETS-related transcription factor Elf-2	1000.0	17	X	8	2	4	ELF2
79	Q15723-5	ETS-related transcription factor Elf-2	1000.0	17	X	7	2	4	ELF2
80	Q9BVL2	Nucleoporin p58/p45	73.4	62	X	3	1	4	NUPL1; NUP58
81	Q15723-2	Isoform 2 of ETS-related transcription factor Elf-2	1000.0	17	X	8	2	4	ELF2
82	Q6ZMS4	Zinc finger protein 852	46.9	82	X	2	1	3	ZNF852
83	Q9Y2H8	Zinc finger protein 510	46.9	82	X	1	1	3	ZNF510
84	P17097	Zinc finger protein 7	46.9	82	X	1	1	3	ZNF7
85	Q9UK13	Zinc finger protein 221	46.9	82	X	1	1	3	ZNF221
86	Q12901-2	Isoform 2 of Zinc finger protein 155	46.9	82	X	2	1	3	ZNF155
87	Q8NEM1-1	Zinc finger protein 680	46.9	82	X	2	1	3	ZNF680
88	O14776-2	Isoform 2 of Transcription elongation regulator 1	1000.0	X	X	3	2	3	TCERG1
89	O60765	Zinc finger protein 354A	46.9	82	X	3	1	3	ZNF354A
90	Q8N9F8	zinc finger protein 454	46.9	82	X	2	1	3	ZNF454

Table 3.3 continued.

Table Index	Accession	Description	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
91	O75373	Zinc finger protein 737	46.9	82	X	2	1	3	ZNF737
92	Q7Z6E9	E3 ubiquitin-protein ligase RBBP6	1000.0	55	X	1	2	3	RBBP6
93	Q3ZCX4-2	Isoform 2 of Zinc finger protein 568	46.9	82	X	2	1	3	ZNF568
94	P41223	Protein BUD31 homolog	7.6	46	86	26	3	3	BUD31
95	Q13469-5	Isoform 5 of Nuclear factor of activated T-cells, cytoplasmic 2	48.1	50	X	2	1	3	NFATC2
96	Q13469-3	Isoform 3 of Nuclear factor of activated T-cells, cytoplasmic 2	48.1	50	X	2	1	3	NFATC2
97	Q3ZCX4	Zinc finger protein 568	46.9	82	X	1	1	3	ZNF568
98	P17097-2	Isoform 2 of Zinc finger protein 7	46.9	82	X	1	1	3	ZNF7
99	Q13469-4	Isoform 4 of Nuclear factor of activated T-cells, cytoplasmic 2	48.1	50	X	2	1	3	NFATC2
100	P51522-2	Isoform 2 of Zinc finger protein 83	46.9	82	X	2	1	3	ZNF83
101	Q96LW1	Zinc finger protein 354B	46.9	82	X	3	1	3	ZNF354B
102	Q8NB42-1	Zinc finger protein 527	46.9	82	X	1	1	3	ZNF527
103	O14776-1	Transcription elongation regulator 1	1000.0	X	X	3	2	3	TCERG1
104	Q8N2I2-4	Isoform 4 of Zinc finger protein 619	46.9	82	X	1	1	3	ZNF619
105	Q14588	zinc finger protein 234	46.9	82	X	1	1	3	ZNF234
106	Q9UK11	Zinc finger protein 223	46.9	82	X	2	1	3	ZNF223
107	Q8TF20-1	Zinc finger protein 721	46.9	82	X	2	1	3	ZNF721
108	Q9C0G0-2	Isoform 2 of Zinc finger protein 407	46.9	82	X	0	1	3	ZNF407
109	Q9NYT6	Zinc finger protein 226	46.9	82	X	1	1	3	ZNF226
110	Q8TF32	Zinc finger protein 431	46.9	82	X	2	1	3	ZNF431
111	P51815-2	Isoform 2 of Zinc finger protein 75D	46.9	82	X	2	1	3	ZNF75D
112	P17027-2	Isoform 2 of Zinc finger protein 23	46.9	82	X	2	1	3	ZNF23
113	Q7Z6E9-2	Isoform 2 of E3 ubiquitin-protein ligase RBBP6	1000.0	55	X	1	2	3	RBBP6
114	Q8N2I2-3	Isoform 3 of Zinc finger protein 619	46.9	82	X	2	1	3	ZNF619
115	Q8N2I2-2	Isoform 2 of Zinc finger protein 619	46.9	82	X	2	1	3	ZNF619
116	Q6PG37	Zinc finger protein 790	46.9	82	X	1	1	3	ZNF790
117	Q7Z3B4	Nucleoporin p54	55.5	141	X	3	1	3	NUP54
118	Q8TB69	Zinc finger protein 519	46.9	82	X	3	1	3	ZNF519
119	Q8TF20-2	Isoform 2 of Zinc finger protein 721	46.9	82	X	2	1	3	ZNF721
120	P51815	Zinc finger protein 75D	46.9	82	X	2	1	3	ZNF75D

Table 3.3 continued.

Table Index	Accession	Description	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
121	Q6ZN06	zinc finger protein 813	46.9	82	X	1	1	3	ZNF813
122	Q96ST3	Paired amphipathic helix protein Sin3a	4.2	61	52	2	1	3	SIN3A
123	Q9Y2L8	Zinc finger protein with KRAB and SCAN domains 5	46.9	82	X	1	1	3	ZKSCAN5
124	Q9Y383	Putative RNA-binding protein Luc7-like 2	1000.0	97	X	5	2	3	LUC7L2
125	Q8N201	integrator complex subunit 1	1000.0	X	X	2	2	3	INTS1
126	Q8NB42-2	Isoform 2 of Zinc finger protein 527	46.9	82	X	2	1	3	ZNF527
127	Q12901	Zinc finger protein 155	46.9	82	X	2	1	3	ZNF155
128	P51522	zinc finger protein 83	46.9	82	X	2	1	3	ZNF83
129	O15156	Zinc finger and BTB domain-containing protein 7B	25.8	303	X	5	1	3	ZBTB7B
130	Q99676	Zinc finger protein 184	46.9	82	X	1	1	3	ZNF184
131	P60842-2	Isoform 2 of Eukaryotic initiation factor 4A-I	136.0	59	X	5	1	3	EIF4A1
132	Q16670-1	Zinc finger and SCAN domain-containing protein 26	46.9	82	X	2	1	3	ZNF187; ZSCAN26
133	Q63HK3	Zinc finger protein with KRAB and SCAN domains 2	46.9	82	X	1	1	3	ZKSCAN2
134	O15156-2	Isoform 2 of Zinc finger and BTB domain-containing protein 7B	25.8	303	X	5	1	3	ZBTB7B
135	Q13469-2	Isoform 2 of Nuclear factor of activated T-cells, cytoplasmic 2	48.1	50	X	2	1	3	NFATC2
136	Q92670	Putative zinc finger protein 75C	46.9	82	X	2	1	3	ZNF75CP
137	Q9NZL3	Zinc finger protein 224	46.9	82	X	1	1	3	ZNF224
138	Q9H9L3	Interferon-stimulated 20 kDa exonuclease-like 2	201.0	17	419	8	2	3	ISG20L2
139	Q9UK12-1	Zinc finger protein 222	46.9	82	X	2	1	3	ZNF222
140	O43172-1	U4/U6 small nuclear ribonucleoprotein Prp4	1000.0	18	X	5	2	3	PRPF4
141	Q8N2I2	Zinc finger protein 619	46.9	82	X	2	1	3	ZNF619
142	Q9C0G0	zinc finger protein 407	46.9	82	X	0	1	3	ZNF407
143	O43172-2	Isoform 2 of U4/U6 small nuclear ribonucleoprotein Prp4	1000.0	18	X	5	2	3	PRPF4
144	P17027	Zinc finger protein 23	46.9	82	X	1	1	3	ZNF23
145	Q13469	Nuclear factor of activated T-cells, cytoplasmic 2	48.1	50	X	2	1	3	NFATC2
146	P60842	Eukaryotic initiation factor 4A-I	136.0	59	X	4	1	3	EIF4A1
147	Q8N7M2	Zinc finger protein 283	46.9	82	X	1	1	3	ZNF283
148	Q9UK12-2	Isoform 2 of Zinc finger protein 222	46.9	82	X	2	1	3	ZNF222
149	P52737	Zinc finger protein 136	46.9	82	X	2	1	3	ZNF136
150	Q9UK10	Zinc finger protein 225	46.9	82	X	3	1	3	ZNF225

Table 3.3 continued.

Table Index	Accession	Description	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
151	Q4VC44	FLYWCH-type zinc finger-containing protein 1	1000.0	3	X	4	2	2	FLYWCH1
152	B2RPK0	Putative high mobility group protein B1-like 1	17.6	40	X	10	2	2	CTCFL
153	Q7Z6E9-4	Isoform 4 of E3 ubiquitin-protein ligase RBBP6	1000.0	193	X	1	1	2	RBBP6
154	P49454	Centromere protein F	1000.0	132	X	1	2	2	CENPF
155	Q9H981-3	Isoform 3 of Actin-related protein 8	1000.0	107	X	4	1	2	ACTR8
156	P35680-1	Hepatocyte nuclear factor 1-beta	25.4	23	X	1	1	2	HNF1B
157	P20823-3	Isoform C of Hepatocyte nuclear factor 1-alpha	25.4	23	X	2	1	2	HNF1A
158	O75528-2	Isoform 2 of Transcriptional adapter 3	1000.0	13	X	7	1	2	TADA3
159	Q9NZI7-4	Isoform 2 of Upstream-binding protein 1	1000.0	X	X	5	2	2	UBP1
160	P56270-4	Isoform 4 of Myc-associated zinc finger protein	4.3	X	X	30	1	2	MAZ
161	Q92900-2	Isoform 2 of Regulator of nonsense transcripts 1	1000.0	X	X	3	2	2	UPF1
162	P35680-2	Isoform B of Hepatocyte nuclear factor 1-beta	25.4	23	X	2	1	2	HNF1B
163	Q8N0Y2-1	zinc finger protein 444	3.1	X	X	11	2	2	ZNF444
164	Q9HCG8	Pre-mRNA-splicing factor CWC22 homolog	1000.0	144	X	3	2	2	CWC22
165	P20823-1	Hepatocyte nuclear factor 1-alpha	25.4	23	X	1	1	2	HNF1A
166	P20823-6	Isoform 6 of Hepatocyte nuclear factor 1-alpha	25.4	23	X	2	1	2	HNF1A
167	P41223-2	Isoform 2 of Protein BUD31 homolog	7.6	46	86	19	2	2	BUD31
168	Q8N0Y2-2	Isoform 2 of Zinc finger protein 444	3.1	X	X	11	2	2	ZNF444
169	P78346-1	ribonuclease P protein subunit p30	1000.0	9	X	6	1	2	RPP30
170	Q9NZI7	Upstream-binding protein 1	1000.0	X	X	5	2	2	UBP1
171	Q15723-4	Isoform 4 of ETS-related transcription factor Elf-2	1000.0	X	X	5	1	2	ELF2
172	Q92900	Regulator of nonsense transcripts 1	1000.0	X	X	3	2	2	UPF1
173	Q13129	Zinc finger protein Rlf	1000.0	X	X	1	1	2	RLF
174	P35680-4	Isoform 4 of Hepatocyte nuclear factor 1-beta	25.4	23	X	2	1	2	HNF1B
175	Q9NPA8-1	Transcription and mRNA export factor ENY2	1000.0	454	X	17	1	2	ENY2
176	Q9C0K0-1	B-cell lymphoma/leukemia 11B	104.6	X	X	1	1	2	BCL11B
177	P20823-2	Isoform B of Hepatocyte nuclear factor 1-alpha	25.4	23	X	1	1	2	HNF1A
178	Q4VC44-3	Isoform 3 of FLYWCH-type zinc finger-containing protein 1	1000.0	3	X	4	2	2	FLYWCH1
179	Q9NP66	High mobility group protein 20A	32.7	12	X	11	2	2	HMG20A
180	P20823-7	Isoform 7 of Hepatocyte nuclear factor 1-alpha	25.4	23	X	1	1	2	HNF1A

Table 3.3 continued.

Table Index	Accession	Description	Abundance Ratio Sample / Control	CV [%] Sample	CV [%] Control	Coverage [%]	# Peptides	# PSMs	Gene Symbol
181	O75528	Transcriptional adapter 3	1000.0	13	X	6	1	2	TADA3
182	P78346-2	Isoform 2 of Ribonuclease P protein subunit p30	1000.0	9	X	5	1	2	RPP30
183	Q9C0K0-2	Isoform 2 of B-cell lymphoma/leukemia 11B	104.6	X	X	1	1	2	BCL11B
184	P35680-3	Isoform C of Hepatocyte nuclear factor 1-beta	25.4	23	X	2	1	2	HNF1B

Table 3.3 continued.

```
Ikaros ----------------------------MDADEGQDMSQVSGKESPPVSDTPDEGDEPMP 32
Aiolos MEDIQTNAELKSTQEQSVPAESAAVLNDYSLTKSHEMENVDSGEG-PANEDEDIGDDSMK 59
Helios ---------------------------------METEAIDGYITCDNELSPEREHSNMA 26

Ikaros IPEDLSTTSGGQQSSKSDRVVASNVKVETQSDEENGRA-----CEMNGEECAEDLRMLDA 87
Aiolos VKDEYSERDE--------NVLK---------------------SEPMGNA--EEPEIPYS 88
Helios IDL-TSSTPNGQHASPSHMTSTNSVKLEMQSDEECDRKPLSREDEIRGHDEGSSLEEPLI 85

Ikaros SGEKMNGSHRDQGSSALSGVGGIRLPNGKLKCDICGIICIGPNVLMVHKRSHTGERPFQC 147
Aiolos YSREYNEYENIKLERHVVSFDSSRPTSGKMNCDVCGLSCISFNVLMVHKRSHTGERPFQC 148
Helios ES---SEVADNRKVQELQGEGGIRLPNGKLKCDVCGMVCIGPNVLMVHKRSHTGERPFHC 142

Ikaros NQCGASFTQKGNLLRHIKLHSGEKPFKCHLCNYACRRRDALTGHLRTHSVGKPHKCGYCG 207
Aiolos NQCGASFTQKGNLLRHIKLHTGEKPFKCHLCNYACQRRDALTGHLRTHSVEKPYKCEFCG 208
Helios NQCGASFTQKGNLLRHIKLHSGEKPFKCPFCSYACRRRDALTGHLRTHSVGKPHKCNYCG 202

Ikaros RSYKQRSSLEEHKERCHNYLESMGLPGTLYPV-------IKEETNHSEMAEDLCKIGSER 260
Aiolos RSYKQRSSLEEHKERCRTFLQSTDPGDTA----------------SAEARHIKAEMGSER 252
Helios RSYKQRSSLEEHKERCHNYLQNVSMEAAGQVMSHHVPPMEDCKEQEPIMDNNISLVPFER 262

Ikaros SLVLDRLASNVAKRKSSMPQKFLGDKGL-S---DTPYDSSASYEKENEMMKSHVMDQAIN 316
Aiolos ALVLDRLASNVAKRKSSMPQKFIGEKRHCF---DVNYNSSYMYEKESELIQTRMMDQAIN 309
Helios PAVIEKLTGNMGKRKSSTPQKFVGEKLMRFSYPDIHFDMNLTYEKEAELMQSHMMDQAIN 322

Ikaros NAINYLGAESLRPLVQTPPGG-SEVVPVISPMYQLHKPLAEGTP---RSNHSAQDSAVEN 372
Aiolos NAISYLGAEALRPLVQTPPAPTSEMVPVISSMYPIALTRAEMSN---GA-----PQELEK 361
Helios NAITYLGAEALHPLMQHPPSTIAEVAPVISSAYSQVYHPNRIERPISRETADSHENNMDG 382

Ikaros LLLLSKAKLVPSEREASPSNSCQDSTDTESNNEEQRSGLIYLTNHIAPHARNGLSL-KEE 431
Aiolos KSIHLPEKSVPSERGLSPNNSGHDSTDTDSNHEERQNHIYQQNHMVLSRARNGMPLLKEV 421
Helios PISLIRPKSRPQEREASPSNSCLDSTDSESSHDDHQS--YQGHPALNPKRKQSPAYMKED 440

Ikaros HRAYDLLRAASENSQDALRVVSTSGEQMKVYKCEHCRVLFLDHVMYTIHMGCHGFRDPFE 491
Aiolos PRSYELLKPPPICPRDSVKVINKEGEVMDVYRCDHCRVLFLDYVMFTIHMGCHGFRDPFE 481
Helios VKALDTTKAPKGSLKDIYKVFNGEGEQIRAFKCEHCRVLFLDHVMYTIHMGCHGYRDPLE 500

Ikaros CNMCGYHSQDRYEFSSHITRGEHRFHMS 519
Aiolos CNMCGYRSHDRYEFSSHIARGEHRALLK 509
Helios CNICGYRSQDRYEFSSHIVRGEHTFH-- 526
```

Figure 3.10: Sequence alignment of the proteins Ikaros, Aiolos, and Helios. Regions highlighted in yellow were sequenced by peptides unique to that protein. Regions in pink were sequenced from peptides common to more than one of the three proteins.

```
1 TQSDEENGRACEMNGEECAEDLRMLDASGEKMNGSHRDQGSSALSGVGGIRLPNGKLKCD 120
2 ------------------------------------------------------------
3 TQSDEENGRACEMNGEECAEDLRMLDASGEKMNGSHRDQGSSALSGVGGIRLPNGKLKCD 120
4 TQSDEENGRACEMNGEECAEDLRMLDASGEKMNGSHRDQGSSALSGVGGIRLPNGKLKCD 76
5 TQSDEENGRACEMNGEECAEDLRMLDASGEKMNGSHRDQGSSALSGVGGIRLPNGKLKCD 120
6 ------------------------------------------------------------
7 TQSDEENGRACEMNGEECAEDLRMLDASGEKMNGSHRDQGSSALSGVGGIRLPNGKLKCD 120
8 TQSDEENGRACEMNGEECAEDLRMLDASGEKMNGSHRDQGSSALSGVGGIRLPNGKLKCD 120

1 ICGIICIGPNVLMVHKRSHTGERPFQCNQCGASFTQKGNLLRHIKLHSGEKPFKCHLCNY 180
2 -------------------VGERPFQCNQCGASFTQKGNLLRHIKLHSGEKPFKCHLCNY 93
3 ICGIICIGPNVLMVHKRSHTGERPFQCNQCGASFTQKGNLLRHIKLHSGEKPFKCHLCNY 180
4 ICGIICIGPNVLMVHKRSHTGERPFQCNQCGASFTQKGNLLRHIKLHSGEKPFKCHLCNY 136
5 ICGIICIGPNVLMVHKRSHT---------------------------------------- 140
6 ------------------------------------------------------------
7 ICGIICIGPNVLMVHKRSHTGERPFQCNQCGASFTQKGNLLRHIKLHSGEKPFKCHLCNY 180
8 ICGIICIGPNVLMVHKRSHTGERPFQCNQCGASFTQKGNLLRHIKLHSGEKPFKCHLCNY 180

1 ACRRRDALTGHLRTHSVGKPHKCGYCGRSYKQRSSLEEHKERCHNYLESMGLPGTLYPVI 240
2 ACRRRDALTGHLRTHSVGKPHKCGYCGRSYKQRSSLEEHKERCHNYLESMGLPGTLYPVI 153
3 ACRRRDALTGHLRTHS------------------------------------------- 196
4 ACRRRDALTGHLRTHS------------------------------------------- 152
5 ------------------------------------------------------------
6 ------------------------------------------------------------
7 ACRRRDALTGHLRTHS-------------------------------------------VI 198
8 ACRRRDALTGHLRTHS-------------------------------------------VI 198

1 KEETNHSEMAEDLCKIGSERSLVLDRLASNVAKRKSSMPQKFLGDKGLSDTPYDSSASYE 300
2 KEETNHSEMAEDLCKIGSERSLVLDRLASNVAKRKSSMPQKFLGDKGLSDTPYDSSASYE 213
3 -------------------------------------------GDKGLSDTPYDSSASYE 213
4 -------------------------------------------GDKGLSDTPYDSSASYE 169
5 -------------------------------------------GDKGLSDTPYDSSASYE 157
6 -----------------------------------------VGDKGLSDTPYDSSASYE 70
7 KEETNHSEMAEDLCKIGSERSLVLDRLASNVAKRKSSMPQKFLGDKGLSDTPYDSSASYE 258
8 KEETNHSEMAEDLCKIGSEISRAGQTSK------------------------------- 226
```

Figure 3.11: Sequence alignment of the 8 Ikaros sequence variants. Highlighted regions were sequenced by mass spectrometry.

```
1 KENEMMKSHVMDQAINNAINYLGAESLRPLVQTPPGGSEVVPVISPMYQLHKPLAEGTPR 360
2 KENEMMKSHVMDQAINNAINYLGAESLRPLVQTPPGGSEVVPVISPMYQLHKPLAEGTPR 273
3 KENEMMKSHVMDQAINNAINYLGAESLRPLVQTPPGGSEVVPVISPMYQLHKPLAEGTPR 273
4 KENEMMKSHVMDQAINNAINYLGAESLRPLVQTPPGGSEVVPVISPMYQLHKPLAEGTPR 229
5 KENEMMKSHVMDQAINNAINYLGAESLRPLVQTPPGGSEVVPVISPMYQLHKPLAEGTPR 217
6 KENEMMKSHVMDQAINNAINYLGAESLRPLVQTPPGGSEVVPVISPMYQLHKPLAEGTPR 130
7 KENEMMKSHVMDQAINNAINYLGAESLRPLVQTPPGGSEVVPVISPMYQLHKPLAEGTPR 318
8 ------------------------------------------------------------

1 SNHSAQDSAVENLLLLSKAKLVPSEREASPSNSCQDSTDTESNNEEQRSGLIYLTNHIAP 420
2 SNHSAQDSAVENLLLLSKAKLVPSEREASPSNSCQDSTDTESNNEEQRSGLIYLTNHIAP 333
3 SNHSAQDSAVENLLLLSKAKLVPSEREASPSNSCQDSTDTESNNEEQRSGLIYLTNHIAP 333
4 SNHSAQDSAVENLLLLSKAKLVPSEREASPSNSCQDSTDTESNNEEQRSGLIYLTNHIAP 289
5 SNHSAQDSAVENLLLLSKAKLVPSEREASPSNSCQDSTDTESNNEEQRSGLIYLTNHIAP 277
6 SNHSAQDSAVENLLLLSKAKLVPSEREASPSNSCQDSTDTESNNEEQRSGLIYLTNHIAP 190
7 SNHSAQDSAVENLLLLSKAKLVPSEREASPSNSCQDSTDTESNNEEQRSGLIYLTNHIAP 378
8 ------------------------------------------------------------

1 HARNGLSLKEEHRAYDLLRAASENSQDALRVVSTSGEQMKVYKCEHCRVLFLDHVMYTIH 480
2 HARNGLSLKEEHRAYDLLRAASENSQDALRVVSTSGEQMKVYKCEHCRVLFLDHVMYTIH 393
3 HARNGLSLKEEHRAYDLLRAASENSQDALRVVSTSGEQMKVYKCEHCRVLFLDHVMYTIH 393
4 HARNGLSLKEEHRAYDLLRAASENSQDALRVVSTSGEQMKVYKCEHCRVLFLDHVMYTIH 349
5 HARNGLSLKEEHRAYDLLRAASENSQDALRVVSTSGEQMKVYKCEHCRVLFLDHVMYTIH 337
6 HARNGLSLKEEHRAYDLLRAASENSQDALRVVSTSGEQMKVYKCEHCRVLFLDHVMYTIH 250
7 HARNGLSLKEEHRAYDLLRAASENSQDALRVVSTSGEQMKVYKCEHCRVLFLDHVMYTIH 438
8 ------------------------------------------------------------

1 MGCHGFRDPFECNMCGYHSQDRYEFSSHITRGEHRFHMS 519
2 MGCHGFRDPFECNMCGYHSQDRYEFSSHITRGEHRFHMS 432
3 MGCHGFRDPFECNMCGYHSQDRYEFSSHITRGEHRFHMS 432
4 MGCHGFRDPFECNMCGYHSQDRYEFSSHITRGEHRFHMS 388
5 MGCHGFRDPFECNMCGYHSQDRYEFSSHITRGEHRFHMS 376
6 MGCHGFRDPFECNMCGYHSQDRYEFSSHITRGEHRFHMS 289
7 MGCHGFRDPFECNMCGYHSQDRYEFSSHITRGEHRFHMS 477
8 ---------------------------------------
```

Figure 3.11 continued.

3.5 Conclusions

The use of DNA to capture proteins for mass spectrometry analysis is limited. Our experiments show the technique can be successfully applied for the identification of transcription factors selectively binding to short regions of DNA by nanoflow LC and MS/MS. However, we believe there is substantial room for improvement to our protocols. The current sample preparation method includes a high number of wash steps designed to remove detergents and salt prior to MS analysis. Unfortunately these could also be removing additional proteins of interest, especially if these proteins are not interacting directly with the DNA, but instead part of a larger regulatory complex. The high number of sample replicates we performed provide a baseline data set that can be used to compare against as incremental improvements are made in the sample preparation scheme.

The label free quantitation method performed surprisingly well. Isotopic labeling and isobaric mass tag have become popular options for quantitative mass spec protein analysis, but they are expensive and require additional sample processing steps that can hurt sensitivity. Label free quantitation is largely regarded as unreliable by the protein mass spectrometry community, but our experiments suggest it can be a valuable option if applied correctly. As seen in **Figures 3.4 and 3.9**, normalized protein abundances are highly consistent across all sample replicates. The coefficients of variance listed in **Tables 3.1, 3.2, and 3.3** would be considered high by most analytical chemists, but the majority are <50%. As a result we can be confident in identifying proteins with a sample/control abundance ratio of 2x or higher as being enriched.

Figure 3.8 suggests a high degree of inconsistency in the proteins identified between the three biological replicates of the Ramos/human enhancer samples. However, the proteins not common among the samples are largely those of low abundance with few PSMs. Of the 100 most abundant proteins in the Ramos/Human biological replicates 65% are present in all three samples, and 95% in at least two. One sample in particular, the first biological replicate, is entirely responsible for the lack of protein identification overlap. Results from the other two replicates are highly consistent. The proteins identified from the Ramos/chicken, and DT 40/chicken samples also show a high degree of overlap between biological replicates.

In conclusion, we were able to meet the primary goal of our experiments and identify by mass spectrometry proteins that selectively bind to the Enhancer DNA of Chickens and Humans. These proteins, particularly Ikaros and Aiolos, are promising candidates for further biological study by our collaborators.

3.6 References

1. Teng, G. & Papavasiliou, F. N. Immunoglobulin Somatic Hypermutation. *Annu. Rev. Genet.* **41,** 107–120 (2007).
2. Papavasiliou, F. N. & Schatz, D. G. Somatic Hypermutation of Immunoglobulin Genes: Merging Mechanisms for Genetic Diversity. *Cell* **109,** S35–S44 (2002).
3. Oettinger, M. A., Schatz, D. G., Gorka, C. ; & Baltimore, D. RAG-1 and RAG-2, Adjacent Genes That Synergistically Activate V(D)J Recombination. *Science (80-.).* **248,** (1990).
4. De Silva, N. S. & Klein, U. Dynamics of B cells in germinal centres. *Nat Rev*

Immunol **15,** 137–148 (2015).

5. Shulman, Z. *et al.* Dynamic signaling by T follicular helper cells during germinal center B cell selection. *Science (80-.).* **345,** (2014).
6. Eisen, H. N. & Siskind, G. W. Variations in Affinities of Antibodies during the Immune Response. *Biochemistry* **3,** 996–1008 (1964).
7. Peled, J. U. *et al.* The Biochemistry of Somatic Hypermutation. *Annu. Rev. Immunol.* **26,** 481–511 (2008).
8. Liu, M. *et al.* Two levels of protection for the B cell genome during somatic hypermutation. *Nature* **451,** 841–845 (2008).
9. Delbos, F., Aoufouchi, S., Faili, A., Weill, J.-C. & Reynaud, C.-A. DNA polymerase η is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. *J. Exp. Med.* **204,** (2007).
10. Martomo, S. A. *et al.* Different mutation signatures in DNA polymerase eta- and MSH6-deficient mice suggest separate roles in antibody diversification. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 8656–61 (2005).
11. Mitchell, P. & Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science (80-.).* **245,** (1989).
12. Kohler, K. M. *et al.* Identification of Core DNA Elements That Target Somatic Hypermutation. *J. Immunol.* **189,** (2012).
13. McDonald, J. J., Alinikula, J., Buerstedde, J.-M. & Schatz, D. G. A Critical Context-Dependent Role for E Boxes in the Targeting of Somatic Hypermutation. *J. Immunol.* **191,** (2013).

14. Blagodatski, A. *et al.* A cis-Acting Diversification Activator Both Necessary and Sufficient for AID-Mediated Hypermutation. *PLoS Genet.* **5,** e1000332 (2009).

15. Buerstedde, J.-M., Alinikula, J., Arakawa, H., McDonald, J. J. & Schatz, D. G. Targeting Of Somatic Hypermutation By immunoglobulin Enhancer And Enhancer-Like Sequences. *PLoS Biol.* **12,** e1001831 (2014).

16. Spruijt, C. G. *et al.* Dynamic Readers for 5-(Hydroxy)Methylcytosine and Its Oxidized Derivatives. *Cell* **152,** 1146–1159 (2013).

17. Spruijt, C. G., Baymaz, H. I. & Vermeulen, M. in 137–157 (2013). doi:10.1007/978-1-62703-284-1_11

18. Baymaz, H. I., Spruijt, C. G. & Vermeulen, M. in 207–226 (2014). doi:10.1007/978-1-4939-1142-4_15

Appendix A

Nuclear Extract Preparation

1. Grow DT40 or Ramos cells to approximately 1 x 10^6 cells/mL at a 100 mL volume in T-175 flasks. Collect 300-600 x 10^6 cells in 50 mL centrifuge tubes.
2. Centrifuge the cells at 300 x g for 8 min and aspirate the supernatant.
3. Wash cells with 50 mL of ice cold PBS and centrifuge the cells at 300 x g for 8 min. Aspirate the supernatant.
4. Resuspend cells in each 50 mL centrifuge tube in 1 mL of ice cold PBS. Transfer to 1.5 (or 1.7) mL eppendorf tubes.
5. Centrifuge the cells at 300 x g for 8 min. Aspirate the supernatant.
6. During centrifugation pipet 5-15 mL of Buffer A into a 15 mL conical tube. Add **Aprotinin, Pepstatin A, AEBSF** to make a 1X solution.
7. Estimate the volume of the cell pellet. Add 5 volumes of complete **Buffer A** and resuspend the cells by gently pipetting up and down until homogenous.
8. Allow the cells to incubate on ice for 10 min. Centrifuge the cells at 300 x g for 8 min. Aspirate the supernatant.
9. Due to the osmotic uptake caused by the hypotonicity of Buffer A, the cells should have swelled. Determine the new volume of the pellet.
    a. If the cells are DT40 cells, add 3 volumes of complete Buffer A and resuspend the pellet.
    b. If the cells are Ramos cells, add 8-10 volumes of complete Buffer A and resuspend the pellet.
10. Gather an appropriate size type B (tight) dounce based on the volume of cell solution you will be douncing. Rinse the dounce with Buffer A and remove any remaining buffer with a 1 mL pipette.
11. Transfer the cell solution to the dounce.
12. Dounce the cells.
    a. For DT40 cells apply 30-40 slow and firm dounces (see step 14), being careful not to introduce bubbles into the solution. Stop every ten dounces and take a 45-60 sec break and leave the dounce on ice. This will prevent the cells from overheating.
    b. For Ramos cells apply 20-30 slow and firm dounces. Stop every ten dounces for a break to prevent overheating.
13. After 20 dounces of the DT40 cell solution or 12-15 dounces of Ramos cell solution, remove the pestle and take a 5 uL sample of the dounced solution. Dilute the sample at least 1:1 with hypotonic buffer. Make a 1:1 solution of the diluted sample and trypan blue. Using a hematocytometer and a light microscope, check for cell lysis. The nuclei of lysed cells will appear light blue (lysed nuclei will appear dark blue). About 90-95% cell lysis is ideal.
    a. Note: During the douncing steps if the cell solution starts to become viscous, stop douncing immediately and check the cells under the light microscope (as described above). Viscoscity is indicative of lysis of nuclei. Nuclear lysis during the dounces is more likely with Ramos than DT40. If it occurs, the best solution

is to dilute the solution with complete hypotonic solution further and use a bigger dounce and pestle.

14. Depending on the volume of the cell solution either transfer the suspension to a fresh 1.5 (or 1.7) mL eppendorf tube or a 15 mL conical tube and centrifuge at 500 x g for 15 min.
15. The supernatant is the cytoplasmic extract. Collect or discard the supernatant. If collecting, add glycerol to a 10% concentration, aliquot and snap freeze in a dry ice and ethanol bath.
16. The pellet is composed of the nuclei. Estimate the volume of the pellet. Add 10 volumes of hypotonic buffer. Gently resuspend and centrifuge at 500 x g for 15 min in a 1.5 (or 1.7) mL eppendorf tube.
17. While cells are spinning aliquot 5 mL of cold Buffer C in a 15 mL conical tube. Add Aprotinin, Leupeptin, Pepstatin A, and AEBSF to make a 1x solution.
18. Gently remove the supernatant. Determine the volume of the pellet and slowly add 2 volumes of complete Buffer C. Gently resuspend the pellet making sure to avoid nuclear lysis.
19. Incubate the suspension on a rotating wheel at 4C for 1 hr.
20. Centrifuge the solution at 20,800 x g for 20 min.
21. The supernatant is the nuclear extract. The pellet contains the chromatin fraction (composed of DNA and tightly bound proteins). Transfer the supernatant to eppendorf tubes in 100-200 uL aliquots. Make sure to save a 10 uL aliquot for protein concentration determination.
22. Snap freeze the nuclear extract in a dry ice and ethanol bath. Store at -80 C.

DNA Preparation

1. Design and order complimentary pairs of primers to amplify your sequence of interest from a plasmid. Either the forward or reverse primer should be biotinylated (not both).
    a. Note: All DIVAC sequences for ITA to date have been cloned in JMB's pIgL-GFP2 construct at the SpeI-NheI site. Primers used to amplify are RKD021 and RKD022, with RKD022 being 5' biotin triethylene glycol(BTEG)-labeled and purchased from IDT.
2. Using Phusion (or any other reasonably high fidelity polymerase), amplify the sequences of interest.
    a. Note: To get sufficient amounts of DNA, I usually run eight 100 uL reactions for 35 cycles.
    b. Note: Make sure to also amplify negative control sequences alongside your DIVAC sequence of interest. This can either be a mutated sequence or portions of the chicken IgL known to have no SHM stimulating activity.
3. Make agarose gel while PCR runs.
4. Pool PCR products. Take 10 uL of PCR reaction and run on agarose gel to check if reaction has run as expected and appropriate size product is produced without nonspecific bands or smearing.
5. Estimate the volume of pooled PCR reaction. Add two volumes of Buffer NTI from the **Macherny and Nagel Nucleospin Gel and PCR purification kit.**
6. Add Nucleospin column to holder. Mix the solution and add 700 uL to the column. Centrifuge at 11,000 x g for 30 sec. Discard flow through.
7. Repeat Step 6 until the entirety of the sample has passed through the column(s).

8. Add 700 uL of **Buffer NT3** to column. Centrifuge at 11,000 x g for 30 sec. Discard flow through.
9. Repeat Step 8.
10. Dry the membrane by placing the column back in the centrifuge. Centrifuge at 11,000 x g for 60 sec to remove any remaining excess buffer NT3.
11. Place the column on a fresh eppendorf tube. Add 50-100 uL of autoclaved ddH20. Incubate for 1 minute at RT.
12. Centrifuge at 11,000 x g for 60 sec.
13. Collect eluent and measure DNA concentration on Nanodrop.

Immobilized DNA Template Assay

1. Take Invitrogen M-280 beads and vortex briefly to resuspend beads until the solution is uniform. Aliquot 75 uL of beads into two separate eppendorf tubes (one for reaction and another for a negative control).
2. Add 1 mL of 1X Binding and Washing (BxW) Buffer and place the tubes in a magnetic Eppendorf holder.
3. Once the solution is clear, aspirate the supernatant.
4. Reconstitute the beads in each tube in 500 uL of 1X B&W Buffer.
5. Add 15 ug of either biotinylated DIVAC sequence or biotinylated control sequence to each tube. Invert the tubes until beads are completely resuspended.
    a. Note: According to the manufacturer the beads come in a concentration of 10 mg/mL. 1 mg (~100 uL) of beads is expected to have a binding capacity of approximately 10 ug of biotinylated DNA (with larger DNA fragments having lower capacity due to steric hindrance). It is expected that 15 ug of biotinylated DNA should saturate 75 uL of beads.
6. Allow the beads to incubate at RT on a rotation wheel for 1 hr.
7. During incubation, cast agarose gel.
8. Centrifuge briefly and place the tubes on a magnetic rack. After the solution has cleared, remove the supernatant and check coupling of the DNA to the beads by assessing the depletion of the DNA from the solution on an agarose gel.
9. Wash the beads two times with 0.5 mL of B&W Buffer using the magnetic rack.
10. Wash the beads two times with Protein Binding Buffer using the magnetic rack.
11. Add 400 ug of nuclear extract and 25 ug of polyDADT (or polyDIDC) in a total volume of 600 uL of protein binding buffer to each tube.
12. Incubate for 90 min at 4 C on a rotation wheel.
13. Wash the beads 2X with 0.5 mL of protein binding buffer using magnetic rack.
14. Wash the beads 2X with 0.5 protein washing buffer using magnetic rack.
15. Resuspend the beads in 0.5 mL protein washing buffer and transfer to a new tube.
16. Wash the beads 2X with 0.5 protein washing buffer using magnetic rack.
17. Centrifuge samples and aspirate supernatant.
18. Add 50 uL of digestion buffer (2M Urea in 100 mM Tris-Cl, pH7.5) to the beads.
19. Add .505 uL of 1M DTT to the beads and incubate on thermoshaker at RT and 1400 RPM for 20 minutes.
20. Add 5 uL of .55M IAA (in 50 nM ammonium bicarbonate) and incubate for RT, 1400 RPM for 20 min.

21. Add 2.5 uL of trypsin solution (0.1 mg/mL trypsin in 50 mM acetic acid) solution (0.04 mg/mL in 10 mM Tris-HCl pH 7.5) and incubate for 2 hr.
22. Spin down beads and remove supernatant.
23. Add 50 uL of fresh digestion buffer and incubate for 5 min at RT with shaking.
24. Spin down beads and remove supernatant.
25. Combine supernatant fractions and add 1 uL of trypsin and incubate O/N.
26. Flash Freeze on dry ice and place in -80 C

REAGENTS

DT40 Media

500 mL RPMI 1640

50 mL FBS

5 mL Chicken Serum

5 mL Penicillin/Streptomycin/Glutamine

3.5 uL 2-mercaptoethanol

Ramos Media

500 mL RPMI 1640

50 mL FBS

5 mL Penicillin/Streptomycin/Glutamine

Buffer A

10 mM HEPES, pH 7.9

1.5 mM $MgCl_2$

10 mM KCl

0.2 mM AEBSF

2 ug/mL Aprotinin

1 uM Leupepetin

1 uM Pepstatin

Buffer C

20 mM HEPES, pH 7.9

20% glycerol

0.42 M NaCl

2 mM $MgCl_2$

0.2 mM EDTA

0.1% NP-40

0.2 mM AEBSF

2 ug/mL Aprotinin

1 uM Leupepetin

1 uM Pepstatin

DNA Binding and Wash Buffer (2x)

10 mM Tris-HCl (pH 7.5)

1 mM EDTA

2 M NaCl

Protein Binding Buffer

150 mM NaCl

50 mM Tris–HCl pH 8.0

1 mM DTT

0.25%Igepal CA-630

0.2 mM AEBSF

2 ug/mL Aprotinin

1 uM Leupepetin

1 uM Pepstatin

Protein Wash Buffer

150 mM NaCl

50 mM Tris–HCl pH 8.0

1 mM DTT

0.2 mM AEBSF

2 ug/mL Aprotinin

1 uM Leupepetin

1 uM Pepstatin

Appendix B

Human Enhancer Sequence

AGCAGCCCTGAAACTGCCCCATGCTCCTCCTGGGCCACACCTGGGCCTGTTTGTCACTCATCCCATGCCCG
GGTGGCCATGAGCTCAGTTTCTCTTCCTCTTATTTTTCTCCTTTTGTCACTCTGAGTTCTGGTTTCAGCCAAC
TTGGGGTTAAATTTAGCCTGGGGATTTCCAGGGGTGGCCAGCTGCAGGCAGGGCCACCAGAGCTGGGA
AAGCGCATCCCCCAACCCCATCCGCTAGTTTTTCAGTTTCGGTCAGCCTCGCCTTATTTTAGAAACGCAAA
TTGTCCAGGTGTTGTTTTGCTCAGTAGAGCACTTTCAGATCTGGGCCTGGGCAAAACCACCTCTTCACAAC
CAGAAGTGATAAATTTACCAATTGTGTTTTTTTGCTTCCTAAAATAGACTCTCGCGGTGACCTGCTTCCTGC
CACCTGCTGTGGGTGCCGGAGACCCCCATGCAGCCATCTTGACTCTAATTCATCATCTGCTTCCA

Human Control Sequence

AGCAGCCCTGAAACTGCCCCATGCTCCTCCTGGGCCAGGCCTGGGCCTGTTTGTCACTCATCCCATGCCC
GGGTGGCCATGAGCTCAGTTTCTCTTCCTCTTATTTTTCTCCTTTTGTCACTCTGAGTTCTGGTTTCAGCCA
ACTTGGGGTTAAATTTAGCCTGGGGATTTCCAGGGGTGGCCAGCTGCAGGCAGGGCCACCAGAGCTGG
GAAAGCGCATCCCCCAACCCCATCCGCTAGTTTTTCAGTTTCGGTCAGCCTCGCCTTATTTTAGAAACGCA
AATTGTCCAGGTGTTGTTTTGCTCAGTAGAGCACTTTCAGATCTGGGCCTGGGCAAAACCACCTCTTCACA
ACCAGAAGTGATAAATTTACCAATTGTGTTTTTTTGCTTCCTAAAATAGACTCTCGCGGTGACCTGCTTCCT
GCGGCCTGCTGTGGGTGCCGGAGACCCCCATGCAGCCATCTTGACTCTAATTCATCATCTGCTTCCA

Chicken Enhancer Sequence

CCGTCGACGGATCCACTAGTAGCCATGGCCAAAATGGGGCCCAAATCAGCATTTCCTGCTCCTGGCTCTG
CTTGCTGCTGTGCCGGGCAGCGGCGCTCGTGGTGCCACATACTGGAATGTTTGTGCCGAGGAAGCAGAG
CGTGCCGTGGCCACGGCTATTCCTGTCCTCCAGCTGCTGTTTGGCCGAAGTTGCCACGTGTTTTCTTCCCG
GCTACAATTTTTAGCAAGAGAAACTGTTGCCCCCGTGGCTGCGGTCAGCACATCTGGGAGAGTGCAGGC
TCAGCTGTGGGGCTGGCGGCACGTGGCGGTACCCGTATGCCCGGGCCCATGCTGCCCCTCCAAACATGC
CCAAAAACATGGCAGAAACACTTAAACCTTGTGCTCAGCCCATGCAGAGGCTGCAGAGAGATGGGAAG
GCCGCTAGTGCGCGGCGCTCGCTCTGCTTCACACGTCAGTGCTTTCTGGAGACTGTTTCGTTTCCTTTTTTT
GGCCGGCGTGGGCCCCTCAGCTTCAGTTTCTGATTGGAGACTGAGAAGTAAATTTAGCTTGGGGATGCC
CCGGCCgCACAGCTGCATGGTGGGGCTGAGCGTGGCTAGCCTGCTCCCTGCTTG

Chicken Control Sequence

CCGTCGACGGATCCACTAGTAGCCATGGCCAAAATGGGGCCCAAATCAGCATTTACTGCTCCTGGCTCTG
CTTGCTGCTGTGCCGGGCAGCGGCGCTCGTGGTGCCACATACTTTAATGTTTGTGCCGATTAAGCAGAGC
GTGCCGTGGCCACGGCTATTAATGTCCTCTTGCTGCTGTTTGGCCGAAGTTGCTTCGTGTTTACTTAACGG
CTACAATTAAAAGCAAGATAAACTGTTGCCCCCGTGGCTAAAGTCAGCACATCTTGGAGAGTGCAGGCTT
TGCTGTGGGGCTGGCGGTTCGTGGCGGTACCCGTATGCCCGGGCCCATGCTGCCCCTCCAAACATGCCC
AAAAACATGGCATAAACACTTAAACCTTGTGCTCAGCCCATGCAGAGGCTGCAGAGAGATGTTAAGGCC
GCTAGTGCGCGGCGCTCGCTCTGCTTCACACGTCAGTGCTTACTGGAGACTGTTTAGTTTAATTTTTTTGG
CCGGCGTGGGCCCCTCAGCTTCAGTTACTGATTGGAGACTGAGAAGTAAATGTCCCTTGGGGATGCGGC
GGCCGCAGGGCTGCATGGTGGGGCTGAGCGTGGCTAGCCTGCTCCCTGCTTG