Analytical Platform Developments For Multiple Therapeutic Prognoses Initiatives

by KyOnese P. Taylor

B.S. in Chemistry, University of Houston
Minor in Mathematics University of Houston

A dissertation submitted to

The Faculty of
the College of Science at
Northeastern University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

August 10, 2015

Dissertation directed by

William S. Hancock
Professor of Chemistry and Chemical Biology
Dedication

“the two most important days in your life are the day you were born and the day you find out why” – Mark Twain

Love to my family and friends and God, whom have supported me in this manuscript completion. Thank you for all you have done.
Acknowledgements

To Dr. Jared Auclair and Dr. Francisca Gbormittah.

Jared, I want to thank you for your guidance during this process, you have been more than a guidance on this manuscript, you have been a true friend.

Francisca, no words can describe how much I care for you and thank you for your mentoring during my graduate career. Love and blessings to you.

To all of my fellow NEU Husky friends, thanks for being there for me with your laughter, kindness and support. Stephanie, Noreen, Lisseth, Jonathan, Angela, Fateme, Fan, Carolina, and the Jennifer’s.
Abstract

Scientific analytical techniques, have been a pivotal anchor in the progression of modern medicine and its therapeutics. In the bioanalytical field, we focus on the study and characterization of therapeutics to treat and cure different forms of illness or disease, with core principles of accuracy, precision and analytical reproducibility. Characterization studies are performed to reveal differential protein and genetic information pertaining to regulatory trends, fiscal material and allows one to optimize practiced analytical applications. For this thesis project, I developed an analytical platform to characterize a targeted low abundant protein and a yeast secretome. I also characterized chromosome 22 as part of the chromosome initiative.

Chapter 1 will provide a brief background on relevant topics to each chapter and appendix section. These topics will include the relevance of therapeutics and their contributions. It will also briefly describe the importance of analytical science, including validation, advancements in instrumentation, protein separation and understanding of different sample constituents. I will also discuss bioinformatics as a tool for a genetic approach and other characterization methodologies. Thus, chapter 1 will conclude with a robust outlook that comprises the resulting thesis chapters.

Chapter 2 describes a project, which shows a developed analytical platform to address target peptides for Carbonic Anhydrase Xii. This Clear Cell Renal Cell Carcinoma(ccRCC) low abundant protein, is a potential biomarker target for future drug therapies. Through linear protein isolation, validation and acknowledged analytical reproducibility, peptides were discovered as targets of Carbonic Anhydrase Xii for future ccRCC studies.
Chapter 3 encompasses the details of characterizing the *Komagataella Phaffii Pastoris* secretome. As part of the DARPA-MIT on-site drug delivery program, they set out to create a microfluidic bioreactor chamber to produce readily available drug therapies. Prior to having the stable drug products INFα2B and hGH, I was tasked with unveiling all host cell proteins in the *p.pastoris* 11430 yeast strain. The previously derived analytical platform was also applied in this chapter as well as the approaches for analytical reproducibility for secreted proteins. I discovered seventy-four host cell proteins in this *p.pastoris* strain and was able to compare and contrast proteins in fractionated and unfractionated sample sets. A secretomes full identity will help in the regulation of proliferated gene expression and thus aid in the development of on-site drug therapeutics.

Chapter 4 details a bioinformatics project, which shows a combination of proteomic and genomic databases to derive a ‘parts list’ for chromosome 22. The importance to understand all protein coding genes and their justification along the chromosome, can better the advancements of prognosis, diagnosis and treatment of resultant diseases and abnormalities. Here, I focused on uncharacterized gene products, focal oncogene significance of NF2 and other and demonstrate the conclusion of Single Nucleotide Polymorphisms(SNPs) or Alternative Splice Variants(ASV) of gene construction.

Appendix i, is a brief introductory summary of the acetylation of two cell lines in conjunction with Von Hippel-Lindau disease. The cell lines VHL(+) and pBABE(-), were analyzed through LC-MS/MS technology to access the acetylated proteins within each cell species. I thus, preliminarily identified acetylation modified proteins from the two cell lines. Unveiling such acetylome, would give medicinal advantages is targeted protein therapies.
To conclude this thesis a conclusion summary was produced to show future directions from the contribution of this published work. We are showing the application of our linear nanoLC-MS/MS platform for low abundant protein identification as well as the post translational modified proteins. This platform also help reveal a collection of host yeast cell protein for on-site drug therapeutics, which will used as protein modules in multi-varying product gene inserted through vector plasmids. Our continuation to fulfill the Chromosome-Centric Projects chromosome goal continues to give genetic details on all 24 chromosomes. Overall, this manuscript brings forth information to enlighten studies of genomic, proteomic and transcriptomics to better lives afflicted with illnesses and disease.
# Table of Contents

Dedication ii

Abstract iv

Table of contents vii

List of Figures xiii

List of Tables xv

Chapter 1: Introduction: Brief Overview Of Therapeutic And Analytical Science Implications 1

1.1 Therapeutic Drug Background 2

1.2 Therapy Application 3

1.3 Target Source For Therapeutics 4

1.4 Analytical Characterization Studies 5

  1.4.1 Protein Isolation 5

  1.4.2 Enzymatic Digestion 7

  1.4.3 Identification Through Mass Spectrometry 7

  1.4.4 Quantitation 8

1.5 Sample Constituents 10

1.6 Analytical Detection Instrumentation 11

1.6 Conclusion 14

1.7 References 15

Chapter 2: Development Of Bioanalytical Platform For Peptide Characterization Of Targeted Clear Cell Renal Cell Carcinoma Protein Carbonic Anhydrase Xii 21

2.1 Abstract 22
2.2 Introduction 

2.2.1 Cancer Background 

2.2.2 Renal Cell Carcinoma Background 

2.2.3 Analytical Techniques Background 

2.2.4 Carbonic Anhydrase Xii Analytical Development 

2.3 Materials and Methods 

2.3.1 Materials 

2.3.2 Methods 

2.3.2.1 rCAXii Solution 

2.3.2.2 SDS PAGE of rCAXii 

2.3.2.3 SDS PAGE Human Plasma 

2.3.2.4 SDS PAGE rCAXii Into Human Plasma 

2.3.2.5 InGel Digestion 

2.3.2.6 NanoLC-MS/MS Analysis And Peptide Sequencing 

2.4 Experimental Procedure 

2.4.1 Experimental Flow Diagram 

2.5 Results 

2.5.1 Tryptic Map Of Recombinant Carbonic Anhydrase Xii_{human} 

2.5.2 Human Plasma Tryptic Map Of Carbonic Anhydrase Xii 

2.5.3 ELISA Protein Assay Detection Of Carbonic Anhydrase Xii 

2.5.4 Establish 3 Caxii Peptides For Biomarker Candidacy By Spiking In rCAXii In Human Plasma 

2.6 Discussion 

2.6.1 CAXii Sequence Coverage
3.4.1 Experimental Flow diagram A and B

3.5 Results

3.5.1 Establish Protein Concentration: Null

3.5.2 Establish Protein Concentration: AEX

3.5.3 Host Cell Protein Secretome List

3.5.4 Comparative Description of Null Strain and Product Strain

3.6 Discussion

3.6.1 Host Cell Protein Analytical Characterization for CBI-InSCyT Platform

3.6.2 Komagataella Phaffii Protein Concentration of Unfractionated and Anion Exchange PreFractionated Wildtype-NRRL 11430

3.6.3 Elucidating Mass spectrometry Protein Identification

3.7 Conclusion

3.8 References

3.9 Supplementary Evidence

Chapter 4: Chromosome 22 Proteo-Genomic Analysis For Bioinformatics Contribution To The Centric Human Proteomic Project

4.1 Abstract

4.2 Introduction

4.2.1 Human Proteome Background

4.2.2 Background On The Chromosome

4.2.3 The Chromosome-Centric Human Proteome Project Assignment

4.2.4 Chromosome 22 Background

4.3 Materials and Methods

4.3.1 Experimental Flow Diagram
4.5 Results
  4.5.1 Genomic and Proteomic Bioinformatics Heat Map of Chromosome 22 121
  4.5.2 Ideogram Of Gene Families And Syndromes Identification 126
  4.5.3 NF2 Oncogene Focus to Assimilate Posttranslational Modifications of NF2 And Surrounding Genes 128

4.6 Discussion
  4.6.1 Missing Genomic Evidence Used To Illustrate Missing Proteins 132
  4.6.2 Implications Of Locating Families And Syndromes On Chromosome 22 133
  4.6.3 Focal Oncogene Nf2 Location And Its Variants In Multiple Post Translational Modifications 135

4.7 Conclusion 136
4.8 References 138
4.9 Supplementary Evidence 142
4.10 Acknowledgements 143

APPENDIX i: Acetylome Protein Characterization Of Von Hippel-Lindau(+) AND pBABE(-) Mammalian Cell Lines 144
  i.1 Abstract 145
  i.2 introduction 145
    i.2.1 Von Hippel-Lindau Syndrome Background 145
    i.2.2 Gene Manipulation Vector pBABE 146
    i.2.3 Acetylation Importance For Study In Science 147
    i.2.4 Review Of Analytical Testing Used To Study Diseases 148
  i.3 Materials and Methods 148
    i.3.1 Materials 148
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>CAXii Experimental Flow Diagram</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>SDS-PAGE CAXii Protein Localization</td>
<td>35</td>
</tr>
<tr>
<td>2.3</td>
<td>eXtraction Ion Chromatographs</td>
<td>37</td>
</tr>
<tr>
<td>2.4</td>
<td>CAXii Tryptic Map</td>
<td>40</td>
</tr>
<tr>
<td>2.5</td>
<td>MS/MS LNLPSDMHIQGLQSR</td>
<td>58</td>
</tr>
<tr>
<td>2.6</td>
<td>MS/MS QFLLTNNGHSVK</td>
<td>59</td>
</tr>
<tr>
<td>2.7</td>
<td>MS/MS TAEYYR</td>
<td>60</td>
</tr>
<tr>
<td>3.1</td>
<td>unFractionated HCP Experimental Flow Diagram</td>
<td>74</td>
</tr>
<tr>
<td>3.2</td>
<td>preFractionated HCP Experimental Flow Diagram</td>
<td>75</td>
</tr>
<tr>
<td>3.3</td>
<td>unFractionated HCP Elution Chromatographs</td>
<td>77</td>
</tr>
<tr>
<td>3.4</td>
<td>preFractionated HCP Elution Chromatographs</td>
<td>79</td>
</tr>
<tr>
<td>3.5</td>
<td>AEX Protein Count</td>
<td>81</td>
</tr>
<tr>
<td>3.6</td>
<td>unFractionated SDS-PAGE Protein Localization</td>
<td>101</td>
</tr>
<tr>
<td>3.7</td>
<td>AEX Elution Gradient Chromatograph</td>
<td>102</td>
</tr>
<tr>
<td>3.8</td>
<td>R1 clean-up Gradient Profile</td>
<td>103</td>
</tr>
<tr>
<td>3.9</td>
<td>Stacked R1 clean-up Whole Null blanks 1,4,5,7</td>
<td>104</td>
</tr>
<tr>
<td>3.10</td>
<td>Null 1 Chromatograph</td>
<td>105</td>
</tr>
<tr>
<td>3.11</td>
<td>Null 2 Chromatograph</td>
<td>105</td>
</tr>
<tr>
<td>3.12</td>
<td>Stacked R1 clean-up Whole Null 1 and 2</td>
<td>106</td>
</tr>
<tr>
<td>3.13</td>
<td>Stacked R1 clean-up AEX Blanks 1,2,3</td>
<td>107</td>
</tr>
<tr>
<td>3.14</td>
<td>Wild-Type 8.2 Chromatograph</td>
<td>108</td>
</tr>
<tr>
<td>3.15</td>
<td>Wild-Type 9.2 Chromatograph</td>
<td>108</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>3.16</td>
<td>Wild-Type 10.2 Chromatograph</td>
<td></td>
</tr>
<tr>
<td>3.17</td>
<td>Wild-Type 11.2 Chromatograph</td>
<td></td>
</tr>
<tr>
<td>3.18</td>
<td>Wild-Type 12.2 Chromatograph</td>
<td></td>
</tr>
<tr>
<td>3.19</td>
<td>Wild-Type 13.2 Chromatograph</td>
<td></td>
</tr>
<tr>
<td>3.20</td>
<td>Wild-Type 14.2 Chromatograph</td>
<td></td>
</tr>
<tr>
<td>3.21</td>
<td>Wild-Type 15.2 Chromatograph</td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Building A Chromatid</td>
<td></td>
</tr>
<tr>
<td>4.2</td>
<td>Chromosome 22 Experimental Flow Diagram</td>
<td></td>
</tr>
<tr>
<td>4.3</td>
<td>Proteo-genomic Heat Map</td>
<td></td>
</tr>
<tr>
<td>4.4</td>
<td>Legend of Genomic Color Coding</td>
<td></td>
</tr>
<tr>
<td>4.5</td>
<td>Chromosome 22 Ideogram</td>
<td></td>
</tr>
<tr>
<td>4.6</td>
<td>NF2 Heat Map</td>
<td></td>
</tr>
<tr>
<td>4.7</td>
<td>NF2 and Genes Post-Translation Modifications</td>
<td></td>
</tr>
<tr>
<td>4.8</td>
<td>NF2 STRING Map</td>
<td></td>
</tr>
<tr>
<td>i.1</td>
<td>Acetylome Experimental Flow Diagram</td>
<td></td>
</tr>
<tr>
<td>i.2</td>
<td>VHL and pBABE Gel</td>
<td></td>
</tr>
<tr>
<td>i.3</td>
<td>VHL and pBABE MS Chromatograms</td>
<td></td>
</tr>
<tr>
<td>i.4</td>
<td>Stacked VHL and pBABE MS Chromatograms</td>
<td></td>
</tr>
<tr>
<td>i.5</td>
<td>MS² Chromatograms in VHL A and pBABE A</td>
<td></td>
</tr>
</tbody>
</table>
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Fragmentation Ion Chart</td>
<td>37</td>
</tr>
<tr>
<td>2.2</td>
<td>CAXii Percent Recovery</td>
<td>46</td>
</tr>
<tr>
<td>3.1</td>
<td>Anion Exchange Concentrations</td>
<td>78</td>
</tr>
<tr>
<td>3.2</td>
<td>Host Cell Protein List</td>
<td>84-86</td>
</tr>
<tr>
<td>3.3</td>
<td>unFractionated and preFractionated Protein Identity Count</td>
<td>94</td>
</tr>
<tr>
<td>4.1</td>
<td>Cancer Associated Genes</td>
<td>128</td>
</tr>
<tr>
<td>i.1</td>
<td>Acetylation Sample Log</td>
<td>163</td>
</tr>
</tbody>
</table>
CHAPTER 1

Introduction:
Brief Overview Of Therapeutic And Analytical Science
Implications
1.1 Therapeutic Drug Background

Therapies come in an array of different forms, from gene therapy expression systems through antigen receptors or targeted protein species\textsuperscript{[1-3]}. The protein species can be with accession of small molecules or other proteins; including antibodies or a combination of both, as seen in the growth of ADCs, Antibody Drug Conjugates\textsuperscript{[4, 5]}. Regardless of the therapeutic approach, the goal is the same; to treat or cure a disease. We can help those suffering from ‘minor’ diseases as they carry-on everyday life such as asthma. As well as a ‘major’ disease, seen in someone living with leukemia. Both need treatment(s) to eradicate or stabilize the imbedded illness and with the advancements from modern scientific practices, we can do such tasks. The implication of an analytical approach was a pivotal turn for traditionally studied medicine, which was normally studied by a physician. With analytics, the focus is to essentially investigate the functions of molecular and cellular movement in the body\textsuperscript{[6, 7]}. Thus, analytical application instrumentation was developed which allowed scientists to quantify and qualify collected data through similar analytical experimentation. As such, it has lead to novel treatments and cures, exemplified thorough vaccines and it continues to push forward with revolutionized protein, gene, and blood component therapeutics.
1.2 Therapy Application

The transition from small molecule therapeutics (< 900 Daltons) to protein (18 kDa >) happened as the need for these biologics (protein drugs) proved to differ in ways of cost, administration, manufacturing and clinical efficacy\(^8\). The body can have a greater response, both positive and negative with biologics, thus the need for analytical characterization has proven to be of great importance\(^9, 10\). These studies ensure that intended gene, protein or other blood and blood components do not disturb other human biological functions. Therapy applications are continued to be explored to yield different factors concerning administration site, dosage, and effectiveness to the patient\(^11\). The purposefulness of therapeutics continue aid those in need of medicinal assistance. We want to address diseases from a genetic and proteomic vantage point in the development, efficacy and administration of any therapeutic molecule, both small and proteins\(^12, 13\). A necessity for full characterization studies is applied in order to produce biotherapeutics. Challenges arise in the biopharmaceutical industry in drug production and manufacturing\(^14, 15\). Making more technological advances in on-site drug therapeutics or optimized analytical platforms, a key to creating more specialized medicines to better mankind.
1.3 Target Source For Therapeutics

Analytical sciences focus on studying the molecular and cellular content of human illnesses, through core principle of accuracy and precision for the target substance. The study of biological systems followed the same pathway coined by Dr. Francis Crick known as central dogma. The central dogma states that DNA transcribes RNA, which translate to large molecular masses called proteins\[16\]. Over the course of time, each area; DNA, RNA and proteins, were chosen as a focal point for drug therapies as each has quantifiable advances in modern medicine\[17-20\]. In current times, it is no longer an understanding of not knowing what lies beyond DNA, but rather which is beneficial in areas of specific illness and a development for treatment\[21\]. Protein characterization is the study of proteins and their function is referred to as proteomics, termed by geneticist Dr. Marc Wilkins\[22\]. Of the three areas studied for therapeutics(DNA,RNA, proteins), protein therapeutics is the most recent entity to study described as the most complex\[23\]. Since proteins are derived from two preceding molecules (DNA and RNA), varying products are produced including those from standard and aberrant molecules added with Alternative Splice Variants and Post Translational Modifications\[24-26\]. Therefore the pool of proteins produced from a single gene can be a heterogeneous mixture. It is difficult to decipher which protein form is most advantages to target this chosen form when designing a therapeutic drug. However, this natural complex detail can too be advantageous, due to having a protein species with select ligand identification. Many proteomic studies can thus target proteins based on their change in PTM, abundance and macromolecule formation\[27, 28\]. In contrast to proteomics, genomic and transcriptomic studies, lack the tangible appeal of mass conglomerates of a protein(s), but they are a source for downstream protein modulation\[29, 30\]. Thereby addressing the idea to control the source will control the outcome. Because of this many organizations such as the Human
Genome Project, continue to identify gene regulation of diseases\textsuperscript{[31]}. All three methods of study prove to be beneficial by ways of mapping out all protein coding genes on a chromosome, finding ASV or Single Nucleotide Polymorphisms which contribute to protein modification outside PTMs and the dominant method of protein target therapeutics\textsuperscript{[32, 33]}.

1.4 Analytical Characterization Studies

Analytical characterization is an experimental process in which the analyst establishes the performance and structural traits of single molecules (proteins, DNA, RNA) cells or drug products\textsuperscript{[34]}. The goal is to produce quantifiable information to an ascribed target to aid in the betterment of mankind\textsuperscript{[35]}. The analytical practice compiles biological and technological sciences, such as molecular biology and computational mathematics, and engineering\textsuperscript{[36]}. This can been seen in databases like ExPASy and in instruments such as fluorescence detectors and mass spectrometers\textsuperscript{[37, 38]}. Today, analytical scientists play an important role in development of therapeutic treatments for disease\textsuperscript{[39]}. In these studies, many tests are performed to identify proteins through top-down and bottom-up proteomics\textsuperscript{[40]}. “Shot-gun” or bottom-up proteomics is more commonly used, based on instrumentation capability and the linear path of the following distinct areas to fully characterize a protein\textsuperscript{[41, 42]}.

1.4.1 Protein Isolation

In bottom-up, one works with piecing the peptide to the globular protein. We often do an isolate of target protein(s) through separation techniques such as sodium Dodecyl Sulfate-PolyAcrylamide Gel Electrophoresis (SDS-PAGE). SDS-PAGE allows, one to separate a mixture of proteins based on their molecular weight(MW) size\textsuperscript{[43]}. Prior to detection by a mass
spectrometer, we can purify the protein(s) through common separation techniques such as SDS-PAGE and liquid chromatography. In SDS-PAGE, the electric current, gel pore size and solvent for protein flow are important characteristics that help separate the protein groups in the mixture. The voltage current can be set at 50 volts(v) to 300v, to retain the principle for migratory protein flow. The voltage is proportional to the migration time. Therefore, for a faster separation one can use a high voltage. This technique is great, in that knowing a protein’s molecular weight can tell you if it will migrate at the top (high MW) or the bottom (low MW) helps you determine your set voltage and gel run time. In addition to electrical current being a factor in a neat SDS-PAGE separation, the gel pore size is also important. The pore size, referred to as a gel percentage, allows for your molecules to flow in a set voltage. This is the same as the voltage is set according to pore size. The percentage tells one what the molecular weight range of the pores can separate. At 12%, proteins between 10 kDa – 200 kDa will separate, while at 15% the range decreases to 3 kDa – 100 kDa. The percentage is chosen that will best separate an unknown or known protein mixture. The running buffer is used to migrate the molecules to the negative end of the electrode through its pH and ionic composition. Common buffers are Tris-Glycine, Bis-Tris and Tris-HCl. They are individually complementary to gel percentage, thus giving the best protein migration and separation. Understanding the points of gel buffer, percentage and voltage current effect capturing the protein through SDS-PAGE[44]. High Performance Liquid Chromatography(HPLC) is a multi versatile separation approach because is allows for multiple capture affinity particles, such as antibodies(immunoprecipitation), chiral moieties, proteins, porous and nonporous beads, gel-filtration and so forth[45, 46]. It is a system where you can target your molecule by chemical composition as well as elute and capture by chemical properties(ion-
exchange, buffer constituents\textsuperscript{[47]}. With the capture and now enriched protein fluid, proteins are broken into peptides through enzymatic digestion.

1.4.2 Enzymatic Digestion

Enzymatic digestion is what separates the two common proteomic approaches bottom-up and top-down. In top-down, we flow from protein separation directly to mass analyzation. With digestion we can follow a more direct path to get protein functional information from the large amount of mass analyzer. You can also have enzymatic digestion to release the post-translation modification, glycosylation. Digestion on the intact protein breaks the peptide backbone using many different enzymes. Trypsin breaks at the sight of Arginine and Lysine, Ly-C(Lysine) and Pepsin cleaves 11 of the 20 common amino acids. The benefit of using enzymatic digestion shows one can get explicit peptide sequence to further analyze, by the revolution of mass spectrometry\textsuperscript{[48, 49]}.

1.4.3 Identification Through Mass Spectrometry

Mass spectrometry(MS), is a young scientific application that allows one to analyze the mass of given protein species\textsuperscript{[38, 50]}. By measuring the mass-to-charge ratio (m/z), we can identify our protein and thus actively confirm with other precursor isolation analyzation\textsuperscript{[51]}. There are many MS instruments that suit to help identify targeted species. From the robust Linear Trap Quadrupole, to advanced traps such as OrbiTrap, or Triple Quadrupoles, Quadruple Time-Of-Flight, MALDI, and the Fourier Transform-ICR mass analyzer. All offer different specifications to trapping ability, resolution and resolving power, and detecting species\textsuperscript{[52]}. Having these analyzers help identify PTMs such as phosphorylation, glycosylation, and even methylation as well as non-PTMs like
deamidation and even metabolomics\textsuperscript{[53]}. This technological advancement has been the crucial player for therapeutic efficacy, stability and other forms for manufacturer and patient well being\textsuperscript{[54]}.

1.4.4 Quantitation

With mass analyzation, we can get complete primary sequence confirmation, of molecules in mass spectrometry characterization studies. An important attribute to this is measuring the quantity of the protein material, through quantitation assays. Assays can measure protein abundance by protein capture or the values in ion and peptide abundance produced through mass analyzers. The same is true from mass spectrometric measurements. Through MS, the bioinformatics output can also tell us the spectral count for peptides seen of a protein, which can be mapped to intact protein studies or protein expression from its protein coding gene\textsuperscript{[55]}. Having a quantitative measurement through mass spectrometry can give one molecular validation, analyte development, monolithic and transformed protein structures\textsuperscript{[56]}. Examples of such quantitation can be seen in ligand binding assays, such as the formation of antibody drug conjugates(ADCs)\textsuperscript{[57]}. Where the ligand antibody binds to linker peptide, fatty acid or other organic chain, which is also bound to therapeutic drug molecule. In addition to protein conjugates, which include other tags(immuno., chemical, etc.) the MS quantitation can also be viewed on spectrophotometry instruments as would immunofluorescence proteins in UV/UV-VIS fluorescence detectors. There are also assays to use for nucleic acid capture, which is used to capture and measure DNA genetic material. For thorough measurements, it is to have coupled quantitation values in your protein isolation
assays (MS, spectrophotometry, western blot)\textsuperscript{[58]}. Completing a characterization study with quantitation assays, signify for full methodical confirmation of protein(s) of interest. Overall, all aspects of protein separation, mass analyzation and quantitation focal experimentation are essential tools for full proteomic characterization studies.
1.5 Sample Constituents

Performing a detailed characterization includes protein(s) purification, mass spectrometric detection and quantitation. As one performs all parts for structural and molecular function, it is an important task to choose the right starting material, such as plasma, tissue, cell or blood. Cells can be used as a visible variation between conditions such as malignant and benign, and even whole cell production\textsuperscript{[59]}. Tissue is a challenge for proteomic analysis because we would need to dissect to the tissue and isolate proteins or use Matrix Assisted Laser Desorption Ionization (MALDI) on thinly sliced tissue. Thus for intact protein studies; i.e. top-down proteomics, MALDI is best for a characterized profile\textsuperscript{[60]}. But, new advancements such as mass spectrometry imaging are adding to the benefits of molecular composition on tissue samples.

Two highly common samples which are analyzed optically through binding assays, protein concentration and even capture affinity are blood serum and blood plasma. Plasma is the whole blood component treated with an anticoagulant, while serum contains clotting factors such as fibrinogen. One prime reason plasma and serum are used is the large amount of collectable sample quantity. Typically blood collected around the organ of disease contains the same affected proteins that our housed inside the organ, thereby eliminating painful biopsies to the patient. With the difference being the presence of leukocytes in the sample, one is choosing white blood cell use for experimentation. Plasma is more commonly used than serum\textsuperscript{[61]}. Although there is large sample quantity, a major disadvantage in analyzing plasma is that targeted proteins are overwhelmed by common blood proteins and their abundance\textsuperscript{[62]}. Such as albumin, TFA, and immunoglobulins\textsuperscript{[63]}. Most extract these abundant proteins and do a more focused protein isolation assays through HPLC affinity chromatography and SDS-PAGE\textsuperscript{[62, 64]}. 

With the correct sample constituent, science can reveal many optimal detailed and obscure information on their designated characterization study.

1.6 Analytical Detection Instrumentation

With a chosen sample makeup, appropriate mass detection analysis is determined by two distinct criteria. The criteria are complementary by protein purification through SDS-PAGE and HPLC. The collection of the isolated protein material can now be analyzed for its atomic mass through spectrometric instrumentation. Including spectrophotometry (fluorescence) or the popular mass spectrometry (MS). In MS full proteins or peptides are determined based on their mass-to-charge (m/z) ratio on specific spectrometers. Choosing the right mass spectrometer is dependent on sample constituent and instrument capability. These premises are key factors in proteomic research. From previous discussions, a full scale proteomics project is identifying a group of proteins, individual molecules or classification groups; as enzymes, glycoproteins and so forth. This helps us to point potential trends while focusing on a disease or illness. Imaging mass spectrometry is useful for visualizing molecular deposits in tissue samples\textsuperscript{[65]}. By seeing the protein or peptide localization in breast, pancreas or muscle tissues, is beneficial for the following reasons. It eliminates the tedious sample preparation of lysed cells and collection fluid, such as plasma or blood\textsuperscript{[66]}. The visualization also allows one to see drug mobility in the sample, as one way to measure pharmacokinetics and pharmacodynamics. This tissue matrix desorption ionization technique is a great addition to research, but it lacks the in-depth resolving and resolution power with and without chemical tags that other instruments through electrospray ionization have achieved. Electrospray ionization is a technique that changes the liquid sample material into a volatile spray through an applied voltage current. This gives better detection
because the molecules have become highly ionized species. Research has caused for the expansion from simple Linear Triple Quadrupoles to Ion-Trap and Time-Of-Flight Mass Spectrometers. Each has both advantages and disadvantages to the sample make-up. For instance, an LTQ (Linear Trap Quadrupole) is useful for highly abundant, easily separate protein species and to simply acquire compounds in the sample mix. This can be manipulated for targeting by having influences of multiple reaction monitoring coupled with Electron Transfer Dissociation (ETD); to find the disulfide bonds or Collision Induced Dissociation (CID). Chapter 2 describes the use of this Multiple Reaction Monitoring (MRM) LTQ-CID platform. LTQ can be used for many studies being a robust instrument, but for finding classifications such as glycans, glycopeptides a more focused instrument is required. Many have turn to use Ion Trap, to ‘trap’ the molecule(s) of interest, which separates it from unwanted material. Its capability to trap and thus enrich the protein population gives the analyst the select filtering system not used in the LTQ. The Ion Trap was based of Fourier Transfer-Ion Cyclotron Resonance MS and has been updated to an OrbiTrap; where the ions cycle in an orbital motion as they are trapped\[67\]. A similar highly selective filtration is the use of TOF-MS; a Time-Of-Flight Mass Spectrometer\[68\]. In TOF, the ions are separated on the principle of velocity. Keeping in mind that velocity is a function of distance and time, the ions will separate such that the given the same initial acceleration, higher m/z (heavy) will reach the detector after the lower (m/z) molecules\[69\]. All of the enrichment style mass spectrometers have high resolving and resolution power. This means that they are able to distinguish one mass from another most accurately and can resolve similar peak mass differences; lowest = LTQ, high = OrbiTrap/TOF, highest = FT-ICR\[70\]. This is so that each peak is correctly assigned the most correct m/z value\[71\]. This is a great additive in that being able to distinguish whether an acetyl group has been added or removed or whether a
lysine has been trimethylated. This is discussed in Appendix i of the thesis. The high resolution and resolving power instruments have been coupled to mass analyzers\textsuperscript{[72]}. Such as MALDI-TOF, LTQ-TOF, LTQ-OrbiTrap and so forth\textsuperscript{[73]}. The coupling allows for select filtration as well as a fuller collection of injected proteins. By having many options for collecting proteins from sample sources of tissue, blood or even plasma, one only needs to decide which instrument works best for the intended outcome. The instrument capabilities continue to grow to get better, faster and stronger detection of a protein.
1.7 Conclusion

This thesis embodies the robust aptitude for full scale proteomic and genomic characterization studies. I have combined three research projects, which show analytical science practices in their completion. Briefly the projects consist of analytical platform development for potential biomarker Candidate Carbonic Anhydrase Xii in Clear Cell Renal Cell Carcinoma. Along with analytical characterization continued for host cell protein secretome of *Pichia Pastoris* and bioinformatics genomic mapping of chromosome 22 in accordance with its protein coding genes as part of the Chromosome-Centric Human Proteome Project. To complete this manuscript, an appendix of an acetylome of VHL(+/−) cell lines was addressed. All of these projects have given the contribution of knowledge to increase the assessments for therapeutic drug development, both genomically and proteomically. They have shown the application of my optimized analytical nanoLC-MRM-MS/MS platform for peptide targets as well as extract a large quantity of proteins in a secretome. We also show the crossover of a proteo-genomic platform, which is continued guideline for other chromosome explorations for both coding and non-coding genes. With these analytically developed techniques my thesis can assist future scientific explorations in the world.
1.7 References


CHAPTER 2

Development Of Bioanalytical Platform For Peptide Characterization Of Targeted Clear Cell Renal Cell Carcinoma Protein Carbonic Anhydrase Xii
2.1 Abstract

Recent statistics show that one out of every sixty three people will be diagnosed with kidney (renal) cancer\textsuperscript{[1]}, the most common being Clear cell renal cell carcinoma(ccRCC). CCRTCC is caused by a genetic mutation of \textit{vhi}\textsuperscript{[2, 3]}, allowing for isomers of the carbonic anhydrase family to promote tumor growth\textsuperscript{[2]}. There are fifteen members of the carbonic anhydrase family, fourteen in mammalian cells\textsuperscript{[4]}. Currently, caix is the protein of choice in the carbonic anhydrase family for renal therapies and diagnosis\textsuperscript{[4-6]}. However, there has been increased interest in characterizing other members of the carbonic anhydrase family for renal therapies\textsuperscript{[4, 7]}. One carbonic anhydrase in particular is CAXii, because of its unique localization in varying regions of the nephron\textsuperscript{[4, 8]}. There is not much known about CAXii’s role in renal cell carcinoma, therefore, I set out to develop a novel analytical platform using liquid chromatography mass spectrometry. This technique can be used to address CAXii’s involvement of in pre-nephrectomy as well as post-nephrectomy patient procedures. The low levels of CAXii in patient samples and sample complexity made it necessary to develop this novel analytical platform, in which I was able to identify three peptide chains that can used for targeted ccRCC therapeutic approaches. Thus, I was able to develop a novel analytical platform that can detect CAXii in complex patient samples, and which can be used for potential diagnosis and future therapeutic development.
2.2 Introduction

2.2.1 Cancer Background

For centuries physicians have been diagnosing patients, who display similar phenotype characteristics of a disease\cite{9,10}. This disease is called cancer, which affects almost all parts of the human body. Each cancer is unique for its location in the body as well as the location of the cells in an organ or system\cite{10,11}. The display of tumors in various locations, lead researchers to discover more than 100 cancerous diseases\cite{10}. Such as, intraocular melanoma, where melanocytes of the eye form malignant tumors\cite{12,13}. Or, Hodgkin lymphoma blood cancer, developed from abnormal lymphocytes\cite{14}. Through research, carcinoma, which is the abnormal growth of epithelial cells, has been proven to be the most common type of cancer\cite{11,15,16}. Being that epithelial cells are located on the inside and outside of surfaces of the body, one can see its dominance in tumor growth, such as in the lungs, spleen and kidneys\cite{11,15,16}. Even though there are established sets of cancers, scientific research is dedicated to eradicating abnormalities, thereby finding a cure and ending cancer. There are a few challenging task to overcome, the obvious being halting the cellular growth of the tumor\cite{17,18}. To do this, research has observed removal of malignant or benign tumors are insufficient, due to return of tumors or depleting too much organ for survival. Modern science has turned to an analytical approach of targeting the cancer cells, through protein therapeutics\cite{19,20}. Turning away from the invasive chemotherapy, we focus on site-specific targeting of drug-to-protein interaction. For a successful targeted drug therapeutic, there are obstacles of difficult target-to-drug interaction sites and the more prevalent problem of obtaining a protein target that will eradicate tumor death and perhaps complementary to no new tumor formation\cite{21}. This project focused on developing an analytical technique that could be used in the development of targeted therapeutics.
2.2.2 Renal Cell Carcinoma Background

Renal Cell Carcinoma has been investigated, since the first public case by Dr. Miriel in 1810, on a mid-thirties woman who was thought to be pregnant\[22\]. However, her abdominal enlargement was due to a grand tumor on her right kidney\[23, 24\]. As more cases were developed on mistakenly pregnant women, physicians presented this detrimental phenomenon for scientific investigation\[22\]. Even during the early nineteenth century, therapeutics were provided to cease the carcinomatous growth\[23\]. There were only two practiced methods for a renal cell carcinoma(RCC) cure was to remove the malignant carcinomata from such affected kidney or full nephrectomy of the kidney\[25\]. In both cases, it was very well assumed that the patient would not survive during the recovery phase, because of reoccurrence of new carcinomas\[23\]. This led physicians to work with pathologists in finding a treatment to eradicate this disease. Currently, there have been significant improvements for all subsets of RCC cures, with focus on clear cell renal cell carcinoma(ccRCC), the most dominant subset\[2, 26\]. In the late 1980’s, it was established the cause of ccRCC is brought on by the Von-Hippel Lindau\((VHL)\) gene mutation on Chromosome 3\[27, 28\]. The genetic mutation of the tumor suppressor \(VHL\) gene causes the \(HIF1-\alpha\) protein to induce hypoxia survival conditions of the damaged cells and allowing for tumor growth\[28-30\]. With this sequence of events, protein cofactors such as growth factors and catalyzed enzymes were revealed in the regulatory functions of the kidneys, including the catalytic enzyme family of carbonic anhydrase\[4, 7\]. This family of enzymes is readily produced in the body, with particular density in the tubular nephrons of a kidney. The carbonic anhydrase family has been chosen as a focal study in ccRCC from their direct proportional correlation of regulation due to the \(VHL\) gene and proteins\[6\]. This is seen in previous research which shows upregulation of CAiX is due to \(mVHL\)[6]. Major contributors are CAiV, CAii, CAXii and the
current therapeutic target CA\textsubscript{i}X. However, attention has been drawn to therapeutic prognoses for CAX\textsubscript{ii}\cite{31}. Through a collaboration with the Dr. Othon Iliopoulos at Harvard Medical School, the study of the regulation trends of CAX\textsubscript{ii} in ccRCC was developed. There are selective therapies for RCC. Including tyrosine kinase inhibitions of such molecules as SCD1 for apoptosis of the tumor. There is also antiangiogenic drug targeting with FDA approved small molecules such as axitinib and sorafenib for advanced RCC\cite{32,33}. Both methods stay true to progression-free survival (PFS) While these methods have shown success for RCC treatment, there is a major need for targeting for early detection and less invasive treatments in the form of antibody-antigen interaction\cite{18,34}. There is also a need for a more strategic approach for targets outside the time consuming biochemistry and molecular biology assays. That is why I decided to apply analytical techniques to qualify targets in clear cell RCC. Being that carbonic anhydrase xii is a component for onset ccRCC, it is important to discover respectable targets for future drug therapies. This project showed the approach of a novel analytical platform to reveal three peptides with CAX\textsubscript{ii}.

2.2.3 Analytical Techniques Background

Modern medicine uses analytical methodologies for conclusive medicinal diagnosis, prognosis and treatment. With clear cell renal cell carcinoma; a subset of RCC, approved treatments have been with small molecule drug compounds and the monoclonal antibody, Bevacizumab\cite{28}. These molecules use the targeted angiogenesis approach whereby signal transduction cytokines and tryrosine kinases are targeted to inhibit tumors from recruiting and maintaining blood vessels for survival\cite{33,35}. However, a need for additional targets in eradicating RCC is clear. Common practices for protein targeting include biological assays for localization and identification
followed by detection mechanisms as confirmation. These techniques include immunoassays, immunoaffinity chromatography and SDS-PAGE, to name a few\textsuperscript{[36]}. Using any of these techniques would allow one to isolate one or multiple proteins of interest for further analysis. In this project, I isolated carbonic anhydrase xii from the biological fluid plasma, prior to chromatographic separation of the enzymatically digested peptides. With enrichment of the set protein, our chromatographic separation was used to separate peptides based on their hydrophobicity. This was achieved through reverse phase chromatography, entailing a hydrophobic stationary phase column of saturated alkyl carbon chains bound to silica\textsuperscript{[37]}. After peptide separation, the next step is to characterize this protein. We chose mass spectrometry detection to characterize CAXii. It allows for accurate identity of proteins and peptides, through its capability to decipher mass-over-charge difference through resolving and resolution power of the instrument\textsuperscript{[38, 39]}. By laying out a designated analytical platform, we were then able to accurately characterize carbonic anhydrase xii.

2.2.4 Carbonic Anhydrase Xii Analytical Platform development

Applying analytical applications for characterization of cancer molecules has lead researchers to significant information for creating opportunistic drug therapies in many cancers. Due to its ease of sample collection, plasma has been used in these studies\textsuperscript{[36]}. My platform of interfaced liquid chromatography-mass spectrometry (nanoLC-ESI-LTQ-MS) is ideal for plasma because of the robust capacities of the nanoLC-ESI-LTQ-MS instrumentation. One of the most advantageous qualities of nanoLC instrumentation is the ability to conserve the sample. With low nanoliter flow rates, I am able to utilize low sample volumes (1 uL), thereby allowing biological, analytical and technical reproducibility. In addition, my analytical platform is
optimized to avoid sample degradation and does not use crystallization in tissue matrix, thus preventing sample content loss. Complex protein mixtures required a novel analytical method to be developed to separate and capture a vast range of abundant proteins that are plausibly linked to our protein of interest such as IγG\textsuperscript{[40, 41]}. These abundant proteins have the ability to aid in CAXii’s transport through tissue membrane and into the surrounding blood stream through oncotic pressure\textsuperscript{[42]}. Carbonic Anhydrase Xii; Uniprot ID O43570, is a 39kDa small protein of low abundance. Therefore it was important to determine a limit of detection in mass spectrometry instrumentation. Our analysis consisted of studying human recombinant carbonic anhydrase xii(rCAXii\textsubscript{h}), human reference plasma(hP) and varying spiked rCAXii\textsubscript{h} into reference plasma. From these experiments including SRM (selected reaction monitoring) and MRM (multiple reaction monitoring), we have analytically derived a tryptic map of human recombinant CAXii and naturally occurring CAXii in human plasma\textsuperscript{[43, 44]}. Protein localization was obtained by 1D gel analysis from plasma. Thus, I concluded, that because of low protein detection(100ng), future endeavors, will lead to CAXii protein isolation for true regulation levels of clear cell renal cell carcinoma of pre-nephrectomy(+) and post-nephrectomy(-).
2.3 Materials and Methods

2.3.1 Materials

Recombinant Carrier-Free Human Carbonic Anhydrase XII was purchased from R&D Systems (Minneapolis, MN). Female human plasma from Bioreclamation LLC (Westbury, NY) was used as the human reference plasma for this project. Bis-Tris 12% mini-PROTEAN TGX Stain-free™ 10well 30µL comb gels (Biorad, Hercules, CA) were used for sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE). With running buffer of 1X Tris-Glycine from (Thermo) Fisher Scientific. HPLC-MS grade acetonitrile, water, formic acid, and all other reagents and buffers were purchased from Thermo Fisher Scientific (Waltham, MA). The enzyme-linked immuno assay (ELISA) kit was purchased from Cloud Cone® distributed by Cedarlane laboratories (Burlington, NC).

2.3.2 Methods

2.3.2.1 rCAXII Solution

Upon arrival, lyophilized recombinant human carbonic anhydrase xii (rCAXII) was stored at -20 °C, until needed. The 10µg rCAXII material was dissolved in 50µL or 20µL of 1X PBS (phosphate buffer saline) and mixed using a vortex (Vortex Genie 2 Scientific industries, Bohemia, NY) for 20 seconds. After vortexing the sample was incubated for 30 seconds incubation at 37°C (Dry Bath Incubator, Fisher Scientific, Waltham, MA). This cycle of brief vortex and incubation was performed for five cycles. Once mixing was complete, dilutions were made of 0.1 µg, 0.5 µg, 1 µg, 2 µg, 4 µg and 5 µg. These dilutions were from the initial concentration of 0.2 $\mu g/\mu L$ or 0.5 $\mu g/\mu L$. 
2.3.2.2 SDS PAGE of rCAXii

For gel electrophoresis 0.1 µg, 0.5 µg, 1 µg, 2 µg, 4 µg and 5 µg of rCAXii were transferred into 1.5 mL centrifuge eppendorf® tubes. After addition of Laemmli Sample Buffer to denature and solubilize (Biorad, Hercules, CA), 1 µL 1 M DTT was added and the samples were incubated at 68°C for 10 minutes to promote the disruption of disulfide bond linkages. The samples were then incubated with 2 µL iodoacetamide in the dark for 1 hr to induce alkylation of the broken cystine bond. Once loaded on the Bis-Tris 12% mini-PROTEAN TGX Stain-free™, electrophoresis was achieved in 30 minutes at 200V (Power Pac 3000, Biorad, Hercules, CA) in daily prepared 1X Tris-Glycine running buffer. The gels were then rinsed with DI H2O and Coomassie stained with Simple Blue™ SafeStain (Invitrogen, Carlsbad, CA) for 1hr or overnight. To remove the Comassie stain, gels were rinsed with DI H2O and rocked on low speed overnight to visualize all bands of the protein on the gel.

2.3.2.3 SDS PAGE of Human Plasma

Before protein visualization of human reference plasma(hP), overall protein concentration needed to be determined for accurate multi-protein concentration separation. A Thermo Fisher BCA kit was used to determine hP concentration. Along with provided kit standards, the reference plasma was tested with 1:2, 1:10, 1:50 and 1:100 dilutions. In the completion of the protein concentration assay, 66.42 µg/µL was deemed the final concentration. A 1:10 dilution was performed to give 6.642 µg/µL concentration to ensure accurate separation during gel electrophoresis. The same protocol of rCAXii SDS-PAGE was also performed on the reference plasma.
2.3.2.4 SDS PAGE rCAXii Into Human Plasma

Original rCAXii concentrated solutions were prepared and dosed into the 1:10 dilution of human reference plasma. The gels were then performed using the above describe SDS-PAGE method.

2.3.2.5 In-gel Digestion

In-gel Digestion(InGD) was performed as described below. 1mm x 1mm gel pieces were excised from the gel at the 50 kDa molecular marker indication(Promega, Madison, WI). The bands were then finely dissected before placing in 1.5mL eppendorf® tubes. The bands were then washed to remove the Coomaisse blue stain with 1mL additions of 100% Acetonitrile(ACN), followed by a brief vortex spin to dry out the stain. This volume was aspirated and 1mL of prepared 50 mM Ammonium Bicarbonate(ABC) pH.9 was added and again a brief vortex cycle before this volume was aspirated. This continued to ensure that a total of three ACN and two ABC cycles were performed. After the final ACN aspiration, samples were centrifuged at to collect any excess wash solution. Samples were then placed on ice for enzymatic digestion by Trypsin(Promega, Madison, WI). A 1:50 protein:enzyme ration resulted in 20 ng of trypsin being added to each sample tube, by mixing 100 µL of trypsin reconstitution buffer(Promega, Madison, WI) into 20 µg of sequence grade Trypsin. Followed by 10 µL of this solutions diluted with 500 µL of 50 mM ABC. 50 µL of the diluted trypsin was added into each tube, with additional 50mM ABC to cover gel pieces. Before overnight incubation at 37°C, samples were imbibed on ice for 1hr. After incubation, peptides were extracted from the gel pieces through 30% v/v, 60% v/v and 99% v/v of Acetonitrile/Formic Acid washes. With initial aspiration of trypsin solution into new centrifuged tubes, 200 µL of 30% v/v was added to samples tubes, followed by a quick vortex cycle and this collection was placed into the new centrifuged tube.
The same procedure was performed with both 60% v/v and 99% v/v solutions. The gel pieces of the samples were discarded and peptide extracts were lyophilized to dryness, before reconstitution with 20 µL of 1% aomic Acid (F.A.). The reconstituted samples were then centrifuged for 15 mins at 10,000 r.p.m(eypendorf® 5424R, Hauppauge, NY). Once samples were properly centrifuged, 10 µL of sample was placed in mass spectrometry vials(Thermo Scientific, Waltham, MA) for MS and MS/MS analysis.

### 2.3.2.6 NanoLC-MS/MS Analysis and Peptide Sequencing

With the peptide extracts, the samples would undergo liquid chromatography prior to mass spectrometry detection. Reverse Phase prepared 150mm x 50mm i.d. capillary columns(New Objective, Woburn, MA) were packed in-house with a 2:1 Acetonitrile:isopropanol slurry mixture of 5 µm particle 300-Å pore size stationary phase Magic C_{18} (Michrom Bioresources, Auburn, CA) beads. The columns were equilibrated on a Eskigent 2D-nanoLC system (Dublin, CA) for 1hr of stabilizing 2000 psi with a solvent of 0.1% v/v formic acid in HPLC grade water, serving as buffer A and 1 hr of 0.1% v/v formic acid in HPLC grade Acetonitrile, serving as buffer B. 2 µL of peptide extractions from the sample mixture were loaded onto the C_{18} capillary column and desalted for 30 mins at a flow rate of 300nL/min using 98% buffer A. Separation of the mixture took place by the following linear gradient; Stationary at 2% buffer B for 20 mins; 5% Mobile phase B to 70% Buffer B for 65 mins; 70% buffer B to 90% buffer B for 15 mins; Stationary at 90% buffer B for 10 mins; 90% buffer B to 2% buffer B for 1 min; stationary at 2% buffer B for 15 mins. The Finnigan LTQ(ThermoFisher Scientific, Waltham, MA) mass spectrometer was operated in data dependent mode on full MS scan from range 220 m/z through 2000 m/z. With MS/MS fragmentation for the 8 most intense precursor ions from
the MS spectrum. For MRM scans, there were 4 scan events. They scanned as follows, full MS scan 220 m/z through 2000 m/z; scan 2 precursor ion set for 854 m/z; scan 3 set for 679.36 m/z; scan 4 set for 804.37 m/z. While SRM experimentation had one MS scan of targeted m/z per 2ul nanoLC injection. Dynamic exclusion was set with 2 repeating counts, repeat duration of 30s, exclusion list of 200, and exclusion duration of 40s, exclusion mass width 1.50 m/z low and 1.50 m/z high. The chromatographic MS/MS spectra were then searched using Thermo Fisher Proteome Discoverer 1.3 software with SEQUEST(Thermo Electron) as the database search algorithm. The parameters were as follows two missed cleavages; precursor ion mass tolerance of 2.0 Da; fragment ion mass tolerance 1.0Da; full trypsin as enzyme; carboxyamidomethylation(C) as a fixed modification. False Discovery Rate(FDR) was targeted at 1% to give one the best confidence score for peptide matching. SEQUEST (Thermo Electron) algorithm present in Thermo Fisher Proteome Discoverer 1.3 software suite for proteins identification.
2.4 Experimental Procedure

2.4.1 Experiment Flow Diagram
In figure 1, the flow for experimentation proceeded as indicated below.

Figure 1: Analytical Platform Flow Diagram
2.5 Results

2.5.1 Tryptic Map Of Recombinant Carbonic Anhydrase Xii\textsubscript{human}

Little is known about carbonic anhydrase xii (CAXii), a 39-42 kDa protein, in clear cell renal cell carcinoma, however its location is conserved along the nephron\cite{4, 7, 8}. In order to characterize CAXii in RCC, we set out to develop a novel analytical platform that is able to detect the low levels of CAXii in a complex matrix, human plasma. First, I used recombinant human CAXii(rCAXii\textsubscript{h}), to generate a tryptic map and to optimize our analytical platform. I used a linear range of CAXii concentrations spanning .02\mu g/\mu L to 0.5\mu g/\mu L in order to perform limit of detection studies. Briefly, 10 \mu g of commercially available rCAXii\textsubscript{h} was diluted to a stock concentration of 0.2 \mu g/\mu L, which was further used in serial dilutions of 0.02 \mu g/\mu L, 0.1 \mu g/\mu L and 0.5 \mu g/\mu L. Initial sample detection of 0.1 \mu g, 0.2 \mu g, 0.5 \mu g, 1 \mu g, 2 \mu g, 4 \mu g and 5 \mu g rCAXii\textsubscript{h} was done using SDS PAGE gel(Figure 2).
After SDS PAGE separation I was able to extract the CAXii commaisse stained bands for enzymatic digestion. Based on the protein sequence, I choose to use trypsin as my enzyme of choice, which cleaves at arginine (R) and lysine (K) residues. After trypsin digestion, the samples were prepared for full scan range mass spectrometry. The rCAXii is free of three possible N-acetylglucosamine glycosylation sites on Asn28, Asn80 and Asn162. The peptide database search using Sequest determined a limit of detection of 100ng of rCAXii. With a percent recovery of 19.24%, a total of eight
peptides were identified. In addition, searching the CAXii peptide data against the entire human genome revealed that six of the eight peptides were unique to CAXii. Of the six peptides, three were chosen as potential biomarker ccRCC targets based on their consistent reproducibility in the differential microgram load fractions. These peptides are shown in Figure 3A, with high fragmentation ion intensity and percent peptide recovery. After this initial discovery of peptides in full scan MS, I next used selected reaction monitoring (SRM) and multiple reaction monitoring (MRM) mass spectrometry to specifically target the three identified peptides in Figure 3A and 3B. In both SRM and MRM, I obtained similar results. Thus, I successful identified rCAXiih using the novel analytical method with a limit of detection of 100ng, by accurately identifying known polypeptides of carbonic anhydrase xii. Next, I used my analytical platform to characterize CAXii from a complex sample, human plasma.
<table>
<thead>
<tr>
<th>Amino Acid Sequence</th>
<th>MW(Da)</th>
<th>MH(^2)</th>
<th>MH(^3)</th>
<th>MRM Target</th>
<th>XIC fragment interest ion</th>
<th>Retention Time Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNLPSDMHIQGLQSR</td>
<td>1708.87</td>
<td>854.94</td>
<td>570.29</td>
<td>MH(^2)</td>
<td>741 m/z</td>
<td>1(^{st})</td>
</tr>
<tr>
<td>QFLTNGHSVK</td>
<td>1357.72</td>
<td>679.36</td>
<td>415.19</td>
<td>MH(^2)</td>
<td>615 m/z</td>
<td>2(^{nd})</td>
</tr>
<tr>
<td>TAEYYR</td>
<td>802.37</td>
<td>451.19</td>
<td>225.09</td>
<td>MH(^1)</td>
<td>501 m/z</td>
<td>3(^{rd})</td>
</tr>
</tbody>
</table>

**Figure 3A:** Fragment Ion detail chart of appointed qualified peptide biomarkers for clear cell renal cell carcinoma.

**Figure 3B:** eXtracted Ion Chromatograph (XIC) of 3 CAXii peptides.
2.5.2 Human Plasma Tryptic Map Of Carbonic Anhydrase Xii

CAXii has been implicated in renal cell carcinoma, thus the development of an analytical technique capable of characterizing CAXii in human plasma samples from both control and diseased patients. After validating my platform with rCAXii\textsubscript{h}, I next characterized CAXii from human plasma. The tryptic map of recombinant human CAXii protein identified the three peptides to be looked for and characterized in the patient samples. First, I spiked the same serial dilutions as above (.1µg, .5µg, 1µg, 2ug, 4ug, 5ug) of recombinant CAXii into human reference plasma(hP), in order to determine the limit of detection of CAXii in a complex matrix. Prior to spiking hP with rCAXii\textsubscript{h}, I determined the protein concentration of reference plasma to be 66.42µg/µL, which aided in using the size exclusion chromatography method, SDS-PAGE. In Figure 2, SDS PAGE gels show both the accurate location of normalized carbonic anhydrase in human plasma (hP), as well as solo rCAXii\textsubscript{h} and spiked rCAXii\textsubscript{h} into hP. Whole protein gel band extraction of CAXii took place within the multivariate matrices. I next performed enzymatic trypsin digestion, to develop a tryptic map for comparison with the tryptic map of pure rCAXii\textsubscript{h}. Although plasma has approximately 10 orders of magnitude more proteins compared to recombinant CAXii, the amount of trypsin injected was the same\cite{40}. In order to allow direct comparison of rCAXii\textsubscript{h} with naturally occurring hPCAXii, I used exactly the same analytical platform, synchronized nano-liquid chromatography with LTQ mass detection. Initial full scan mode MS was implemented for map characterization. This was to unveil the proteins in this complex mixture. It also showed proteins that are co-localized with CAXii, including carrier proteins, which could impact the signal intensity of CAXii. When doing my targeted analysis, it was key to apply different methodology strategies of SRM and MRM. This allowed me to rule out outlier protein, making the analytics more accurate. With selected reaction monitoring, I
extrapolated the individual masses of 854.94 m/z, 679.36 m/z and 802.37 m/z per sample injection. Multiple reaction monitoring scanning, allowed us to target these masses simultaneously. SRM allowed me to have selected mass focusing but with inefficient targeting. We determined the best analytical reproducibility was achieved using the MRM targeting platform. Targeting multiple peptides within the protein will ensue important aspects for future drug therapies. One major influence of synchronous MRM-MS/MS is that it shows the natural intensity levels of the peptides, by segregating the multi-complex proteins in the plasma matrix. With proper indication of peptide levels, our analytical platform shows evidence of carbonic anhydrase xii. As we now have identified CAXii in both unadulterated CAXii solution and naturally expressed hPCAXii, we were able to establish a hP tryptic map. The map in Figure 4A and 4B, shows similar peptides expressed in our sequence coverage depiction. Through nanoLC experimentation, we were able to draw conclusive evidence of CAXii’s expression in plasma. Having knowledge that displaces CAXii in matrices, will allow us to compare our targeted mass spectrometry detection method and other targeted analytical assays for protein validation.
rCAXii_Tryptic Map  

Human Plasma CAXii Tryptic Map

MPRRSLHAAAVLLLVILKEQPSSPAPVNGSK**WTYFG**

**PDGENSWSK**KYPSCGGLLQSPIDLHSDILQYDASLTP
LEFQGYNLSANK**QFLLTNNGHSVKLNLPSDMHIOQGL**
**OSR**YSATQLHLHWGNPNDPHGSEHTVSGQHFAAEL
HIVHYNSDLYPDASTASNKSEGGLAVLAVLIEMGSFN
PSYDK**IFSHLQHVKYKGQEAFVPGF**NIEELLPERTAE
**YYR**YRGSLTTPPCNPTVLWTVFRNPVQISQEQLLAL
TALYCTHMDDPSPREMINNFRQVQKFDERLVYTSFS
**QVQVCTAAGL**
SLGIILSLALAGILGICIVVVVSIWLFRKS**IKKGDNK**
GVIYKPATK**METEAHA**

MPRRSLHAAAVLLLVILKEQPSSPAPVNGSK**WTYFG**

**PDGENSWSK**KYPSCGGLLQSPIDLHSDILQYDASLTP
LEFQGYNLSANK**QFLLTNNGHSVKLNLPSDMHIOQGL**
**OSR**YSATQLHLHWGNPNDPHGSEHTVSGQHFAAEL
HIVHYNSDLYPDASTASNKSEGGLAVLAVLIEMGSFN
PSYDK**IFSHLQHVKYKGQEAFVPGF**NIEELLPERTAE
**YYR**YRGSLTTPPCNPTVLWTVFRNPVQISQEQLLAL
TALYCTHMDDPSPREMINNFRQVQKFDERLVYTSFS
**QVQVCTAAGL**
SLGIILSLALAGILGICIVVVVSIWLFRKS**IKKGDNK**
GVIYKPATK**METEAHA**

**Figure 4A and 4B:** Tryptic map of Carbonic Anhydrase Xii in recombinant and extracted occurrence in Human plasma(hP).
2.5.3 ELISA Protein Assay Detection Of Carbonic Anhydrase Xii

Due to the complexity of human plasma, I tested for identification of whole protein display of CAXii in immunocapture assays. Being able to observe a protein at lower levels than cytokines, which have been declared low abundant proteins (LAPts), would confirm the low signal intensities of plasma matrixed CAXii observed in my analytical platform. Immunocapture assays are advantageous not only for protein quantification, but also for the specificity to targeted antigens. With polyclonal and monoclonal epitope approaches, one can hope for molecular validation, given that the synthesized antibody is compatible with the antigen epitope. A commonly used immunoassy is an enzyme-linked immunosorbent assay, or ELISA. As an antibody dependent assay, the Cloud Clone Corp ELISA kit, uses a monoclonal antibody technique, to bind to a site specific epitope on carbonic anhydrase xii. With a singular epitope antigen site in the ELISA, the detection limit was 7.46 ng/mL within reference plasma. It was conferred that matrix complexity interferes with the capture signal due to the accurate standard protocol detections. My samples consisted of reference human plasma, a complementary dilution of 1:10 and 1:100 of rCAXii, preliminary experimentation with pre(+) and post (-) nephrectomy samples and kit standard ranging from 0.312ng/ml to 20ng/ml. It is clear through experimentation, that the extra activity of multiple protein complexes in plasma is one limiting factor in inhibiting carbonic anhydrase xii binding to the monoclonal antibody. An accurate standardized linear curve, using provided standards, indicated the robustness of the binding assay. This shows us that there is no analytical reproducibility being that the antibodies can bind to ELISA standards and hP samples, but only at seven-fold higher than none plasma matrix samples. Investigation of the results for the (+/-)
nephrectomy samples revealed unreliable connectivity of the CAXii antigen and the plated antibody. Thus, the imaging results suggest isolation of the target protein within the complex mixture impacts the detection of the detection system. Reverse phase liquid chromatography was better suited for the detection of CAXii based on the hydrophobic interaction of the stationary phase of the pre-qualified peptides. Next, I tested for the predetermined three peptides of Figure 3A in a series of spiked in solutions of our human reference plasma as the matrix.

### 2.5.4 Establish 3 CAXii Peptides For Biomarker Candidacy By Spiking In rCAXii In Human Plasma

The ELISA assay confirmed the detection of of carbonic anhydrase xii in standardized recombinant CAXii solutions. The data also showed that even in derived experimenting, detecting CAXii within human plasma, yielded results that proved CAXii is indeed a low abundant protein(LAPt). Because of this, I went on to optimize my current analytical platform to detect LAPt in complex matrices. My results from the analysis of rCAXiiₕ and CAXii’s base level in hP, highlighted three unique peptides to focus on: LNLPSDMHIQGLQSR, QFLLTNNGHSVK and TAEYYR. In order to use CAXii in ccRCC diagnostics, I needed to determine the limit of detection(L.O.D.), by our chosen method of MS/MS spectrometry. I applied the same multiple reaction scanning(MRM) methodology through a linear dose dependent spike-back study. In this study I used my rCAXiiₕ concentrations to spike .1µg, .2µg, .5µg and 1µg deliverables, into the reference plasma solution. I was accurately able to identify the targeted peptides in the nanoLC-LTQ-MRM-MS/MS platform, through extracted ion chromatograms as seen in Figure
3B. In alignment with the three specified peptides, six other peptides were also intermittently present, as exposed in the tryptic map of Figure 4A and 4B. Examining all of these polypeptides, gave us detailed chemical specificities of the molecule including hydrophobicity. The majority the peptides are thirteen or more amino acids long, indicating good surface area during reverse phase chromatography. They are also predominately in the core of the molecule. This suggested that the amino acid sequence shows the tertiary folding structure of the disulfide bond connectivity. The positions of the cysteine residues are at C50 and C230, which are located on the initial tail end leader sequence and the core end. There are three N-linked glycosylation sites of N-acetylglucosamine; GlcNAc, at N28, N80, N162; only one is present in the three polypeptides analyzed. Thus it would appear that glycosylation did not effect the signal intensity of my analytical platform. The production of a linear curve, led to congruent results of 100ng as a limit of detection of added CAXii into human plasma. In the L.O.D. curve, the standard deviation of the most prevalent peptide QFLLTNNGHSVK was $1.23 \times 10^{-4}$. Therefore, I showed the power of my newly developed analytical platform in detecting low abundance proteins, like CAXii, in complex mixtures like human plasma.
2.6 Discussion

2.6.1 CAXii Sequence Coverage

In Table 1 I demonstrate, the resulting uppermost peptide sequences of CAXii through our developed nanoLC-MRM-MS analytical platform. This was initiated by understanding that validation of the protein is the key step toward becoming a biomarker. Thus, we designed our experiment to reveal peptides as the features for future studies of CAXii in Clear Cell Renal Cell Carcinoma. By initially performing localization studies through SDS-PAGE, we saw the effects of concentration through visualization patterns. In the gels, the weakest coomassie stain was obtained at 0.1 µg, while there were equal depths in color for 2 µg, 4 µg and 5 µg. After we optimized both the linear gradient for primary chromatographic separation and tailored mass spectrometry detection for full scan MS or MRM-MS the MS/MS data was searched using the SEQUEST algorithm of Proteome Discoverer 1.3. It was confirmed, that through LC-MS/MS, our limit of detection(L.O.D.), was 100 ng. For a clear fragmentation pattern of ‘R’ group on the CAXii peptide backbone, we analyzed 5 µg of the InGD protein. All analysis of the samples showed technical replication and analytical replication, by proving reproducibility. The bioinformatics search was set to my acquired MS/MS spectra against UniProt protein entries. One condition was to search against the completed CAXii protein sequence, O43570. As seen in Table 1A, percent coverage increased from 9% to 28%, entailing that our protein was indeed identified in both MRM-MS and full scan range MS. Another step was to rule out the possible contaminant peptide sequences that would identify coincidentally with the CAXii sequence. That is why the spectra were searched against the full human proteome. Followed by the top hits in the
human protein; being Protein disulfide-isomerase A3(thioredoxin), Brain acid soluble protein 1, Stomatin-like protein 2, mitochondrial, ATP synthase subunit alpha, mitochondrial, ATP synthase subunit beta of the 68,500 proteins(Uniprot™) added into the CAXii sequence; Caxii C&H. As assessed, no outlier proteins were discovered. Thereby proving the sequence coverage was accurate. The trend of increased recovery is seen in sample sources of 0.1 µg MRM through 5 µg full scan MS. Within the experimentation, we also proved the complex plasma matrix effects the percent recovery. Under searching conditions Caxii C and Caxii C&H, reference plasma full scan MS and target(MRM) scan, the sequence coverage lessened compared to 5 µg full scan MS. Seeing zero recovery in target MS does not mean no peptides were present. It could mean that with a targeted method in human reference plasma (hP), the peptides are too low in intensity to be identified in the bioinformatics algorithm. To prove this we extracted target hP peptides in XIC; Supplementary Figure i, ii, iii. This assessment proved peptides; LNLPSDMHIQGLQSR, QFLLTNNGHSVK and TAEYYR are thoroughly extracted in rCAXiih targeted and full scan as well as targeted and full scan hP. Therefore, based on the percent recovery in Table 1B we can relate carbonic anhydrase xii in reference plasma to be comparable to L.O.D. of 0.1 µg.
### Table 1: CAXii Percent Recovery

A) The percentage of CAXii sequence coverage, that matches with Sequest™ Uniprot™ rCaXii only; C, protein sequence or rCaXii; and Human proteome sequence; C&H.

B) Eight CAXii peptides found in the complete sample set.

C) Peptide sequence legend.
2.6.2 CAXii Tryptic Map

As the targeted peptides were being discovered, the sequence coverage revealed a tryptic map depiction of rCAXiiₙ and human reference plasma CAXii. The initial phase of the analytical platform was the physical localization of CAXii protein content for further downstream analysis. A common way to achieve this was through SDS-PAGE. When we took a look at the standardize molecular weight markers, we see that carbonic anhydrase lies on the 36 kDa under Tris-Glycine run buffer conditions. This indicates relative anhydrases to be observed in any protein sample. However, the rCAXiiₙ was spotted 10-15 kDa above indicated carbonic anhydrase. It is not uncommon for proteins to be displaced outside of expected locations. Gel mobility is an important factor of protein separation in PAGE assays. As mentioned, we used a 12% Bis-Tris TGX gel. The monolithic gel will give better separation of proteins, due to the concept of the higher the percentage of the gel, the better separation of proteins in the mixture.⁴⁸ Since we dealt with a singular protein, the gel cast TGX or Bis-HCL gives a different migration pattern to the proteins. Thus, giving reasonable explanations as to why the observed CAXii was located at a different molecular weight. Gel bands of rCAXiiₙ, hP and rCAXiiₙ spiked into hP, were therefore extracted from the whole gel, based on the representation of rCAXiiₙ at 50 kDa. This confirmed authentication of CAXii protein. Before we chose targeted peptides, the sequence recovery unveiled multiple peptides ranging from the C-terminus, N-terminus and the core of the protein. Table 1A, depicts eight highly expressed CAXii peptide sequences, with most predominately in the core of the molecule. I then mapped these highly expressed peptides onto the CAXii protein sequence; shown in Figure 4, onto rCAXiiₙ and human reference plasma. This was
executed to see what balances and conflicts may arise during (mass spectrometry) detection. With that knowledge, it would help in future aspects of drug targets of CAXii in ccRCC. The peptides match very well against rCAXii, and human reference plasma. I then went back through all experimental samples to see if there were any other peptides that may not have shown in percent recovery, but were present through protein authentication of SEQUEST. Circled in blue in Figure 4, are amino acids that were essentially tagged on to the common dominant peptides. This represents the cleavage of the peptide. Particularly of peptides (R)KSIKKGDNK and (K)METEAHA(Figure 4B) of human plasma CAXii tryptic map. With the preceding Arginine(R) and Lysine(K) residues, with the proceeding Methionine(M) or multi-Lysine sites, trypsin cleavage[49-52] would dissect and release these peptides to be minimally seen in data algorithm acquisitions. The same is true for residues Tyrosine(Y) and again Lysine(K) being present in hP, but missing in rCAXii. In conceiving the tryptic maps, the developed analytical platform was proved accurate for justifying the chosen targeted peptides as biomarker candidacy for clear cell renal cell carcinoma.

2.6.3 LAPt (Low Abundant Protein) Approval Through ELISA

In order to prove our analytical platform was accurately giving low intensity signal for CAXii, we needed to ascertain carbonic anhydrase xii’s low abundance profile. This was performed by evaluating the binding capability through ELISA. In humans there are a significant number of proteins that are well known to contribute to the function of organs, cellular tissue and the excretion of waste by product. There are even such proteins that are necessary in the cytosol cellular compartment, and those that are regulatory factors in
human plasma. The plasma proteome has classic proteins such as albumin, fibrinogen, Immunoglobulin’s α/γ/μ (alpha/gamma/mu) with abundance of 6-12 orders of magnitude. With high levels of secretion in the blood stream, these proteins have been termed high abundant proteins (HAPts) \(^{[40,53]}\). Tissue leakage labeled proteins of TNF Binding, carcinoembryonic and well know TPA have been classified as average abundance in plasma, with 2-5 orders of magnitude \(^{[40,53]}\). This leaves predominately interleukins and tissue factors, less than one order of magnitude in abundance. Knowing the MW of CAXii, literature confirmed that the family of CA’s could potentially be carried with HAPts, thereby potentially decreasing the signal detection in analytical applications. That is why this antibody capture assay was devised to test the ability of CAXii presence in plasma. The CAXii ELISA profile was set up, by following the standard operating protocol of the Cloud Clone® CAXii ELISA kit. This kit used a monoclonal antibody to capture CAXii. The experiment samples were as follows; seven triplicate ELISA prepared standards, triplicate (10 ng/mL) rCAXii\(_h\), triplet hP, triplet hP + 1:10 rCAXii\(_h\) and triplet hP 1:100 rCAXii\(_h\). For preliminary testing purposes I tested a doublet of one (+/-) nephrectomy sample. The standards were prepared by diluting the ELISA stock solution 80ng/mL to 20ng/mL, 10ng/mL, 5ng/mL, 2.5 ng/mL, 1.25 ng/mL, 0.625 ng/mL and 0.312 ng/mL. The blank was also included in this range at 0 ng/mL and was composed of ELISA standard diluent. Detection was achieved through fluorescence detection(SpectraMesh 340 PC, Molecular Devices, Sunnyvale, CA) microwell plate absorbed at 450nm. The limit of detection for this assay was assigned at 1.25 ng/mL. For the two lower concentrations of 0.625 ng/mL and 0.312 ng/mL, the L.O.D. values were quite analogous to the blank of 0.0 ng/mL. Thereby yielding a parallel linear curve.
of Optical Density versus Concentration (ng/mL) disclosure in the ELISA kit literature. The preliminary (+/-) nephrectomy sample yielded (-) 8.36 ng/mL and (+) 6.92 ng/mL. The dosed in and pure rCAXii were excluded from consideration to the irreproducibility of rCAXii to dosed in human reference plasma. The testing of reference plasma yielded 7.46 ng/mL putting the sensitivity ten|fifteen fold higher than mass spectrometry detection of rCAXii. Based on the results of ELISA testing, we concluded that for quantitation measurements of CAXii, it is best to do a protein capture assay based on immunoaffinity or any protein enrichment procedure\[54-56\]. Other points such as multiple (+/-) nephrectomy samples would give a better understanding of the trend of CAXii in ccRCC. Overall, this ELISA was a tool to show the low abundant protein level of CAXii as confirmation for low intensities in the developed analytical platform.

2.6.4 Analytical Reproducibility Of Clinical Unsuccessful Samples

As seen throughout the experimental procedures of this project, carbonic anhydrase xii has been difficult to characterize as the single source for ccRCC. We have declared and proven analytical reproducibility through the analytical platform as well as show how the (+/-) nephrectomy sample are clinically unsuccessful. They are unsuccessful in that prior to CAXii peptide detection we could not detected CAXii in clear cell renal cell carcinoma. The same analytical platform to reveal targeted peptides, LNLPSDMHIQGLQSR, QFLLTNNGHSVK and TAEYYR was used in the (+/-) nephrectomy sample set. With linked SDS-PAGE-nanoLC-MS or nanoLC-MRM-MS, they were unable to be extracted in clinical samples. One clear observance is due to CAXii low abundance levels. Thereby enforcing the need as described by ELISA to have
an immunoaffinity or protein enrichment procedures to deplete competing proteins of CAXii signal detection. However, HAPts are carrier proteins for LAPts, thus the obvious trepidation required for analysis. My platform used both Chameleon® (Dionex, Sunnyvale, CA) and Xcalibur® (Thermo Fisher Scientific) for autosampler preparation of chromatographically separated protein mixture. Indicating critical sequence triggers were enabled for mass spectrometry detection, on the Dionex Ultimate 3000 UPLC LTQ (Thermo Fisher) or Eskigent 2D-nanoLC LTQ (Thermo Fisher). On both instrumentation validation of chosen targeted peptides were acquired. With approved scientific method, the next step was to test (+/-) nephrectomy samples. To the surprise in small batches; six clinical samples, equilibrating blanks, and complimentary hP per gel, LC-SRM-MS or LC-MS analysis was not suitable. With illegitimate results, initiative was taken to describe proper (+/-) nephrectomy sample validation for future project directions[44, 56].

This project developed and analytical platform that can be pertained to multiple characterization studies. Such as a full scale proteomic project to reveal a proteome or the proteomic approach to represent post translational modification in varying diseases. Therefore, by optimizing this method, we can continue to apply analytical scientific practices to improve the world.
2.7 Conclusion

Here, I have successfully developed an analytical platform to detect and characterize Carbonic Anhydrase xii, which can be used for future therapeutic biomarker studies. This enzyme is a regulator in neoplasms\cite{22} of the kidney, thereby making it a great target in therapeutic medicine for clear cell renal cell carcinoma. In summary, multivariant Caxii solutions, consisting of rCAXii\textsubscript{h}, human reference plasma(hP) and spiked in rCAXii\textsubscript{h} into hP, allowed for CAXii protein detection in this platform. My analytical platform includes SDS-PAGE, LC-MS/MS and LC-MRM-MS/MS, where I found three peptides; LNLPSDMHIQGLQSR, QFLLTNNGHSVK and TAEYYR that show analytical reproducibility in plasma and nonplasma matrices. I indicated, that protein quantification would be the next step for accurate therapeutic studies. Through ELISA, I showed that immunoaffinity protein capture, shows the precision antibodies have for the enzyme. It also showed, that other protein enrichments, such a depletion of High Abundant Proteins(HAPts), may increase MS signal intensity. However, removing HAPts, may remover carrier proteins of CAXii, thereby decreasing MS and preceding assays quantities. By having this platform, it can be applied as biomarker target characterization of clear cell renal cell carcinoma.
2.8 References


2.9 Supplementary Evidence

i) XIC MS/MS LNLPSDMHIQGLQSR

5ugB #2297  RT: 21.65  AV: 1  NL: 5.80E5
T: ITMS + c NSId Full ms2 855.28@cid28.00 [225.00-2000.00]

854.94 m/z
ii) XIC MS/MS QFLLTNNGHSVK

5ugB #1865  RT: 17.47  AV: 1  NL: 3.07E6
T: ITMS + c NSI d Full ms2 680.28@cid28.00 [175.00-2000.00]

QFLLTNNGHSVK
679.36 m/z
iii) XIC MS/MS TAEYYR

5ugB #1401  RT: 13.15  AV: 1  NL: 4.63E5
T: ITMS + c NSI d Full ms2 802.39@cid35.00 [210.00-1615.00]

TAEYYR

802.37 m/z
CHAPTER 3

*Komagataella Phaffii Pastoris* Secretome Host Cell Protein Characterization By Analytical Methodology And Bioinformatics Application
3.1 Abstract

Current medicinal therapeutics use varying host cell organisms for protein product production. The MIT Center for Biomedical Innovation (CBI) and the federal agency DARPA, have joined together to develop the Integrated and Scalable Cyto-Technology (InSCyT) platform, which will use yeast as the host cell organism for protein production. The InSCyT platform is being developed for the production of on-site protein therapeutics to be used around the world. Subsequent protein product from Human Growth Hormone and Interferon-alpha 2b genes, have be chosen to be the first protein therapeutics in the InSCyT system by transgenic yeast. Before the initial product genes of hGH and IFNα2b can be produced, the secreted proteins of the Saccharomyces Cerevisiae host organism strain, *Komagataella phaffii pastoris*(*k.phaffii*), must be characterized to illustrate all Host Cell Proteins (HCPs) and elucidate any contaminating proteins. The goal of my project, was to characterize the secreted fraction of the *k.phaffii Null* strain NRRL-11430 in order to identify potential contaminants during protein therapeutic production. I used the bioanalytical application of SDS-PAGE, in-solution digestion (InSD), RP-HPLC-MS/MS(Q-Tof), to monitor the secreted HCPs. To increase the dynamic range of the characterization study, we identified HCPs in both unfractionated *Null* and eight anion exchange(AEX) fractions. These proteins are being identified by different bioinformatics database searches within ProteinScape™. Once the lists of HCPs are obtained, we will then have the secreted proteome for *k.phaffii*, which can be compared with the secretome of product producing strains. Thus, the contribution of characterized *k.phaffii Null* strain NRRL-11430 will give researchers in the InSYcT platform development a scaffold of proteins during protein therapeutic production. As well as serve as a resource for biotechnology applications of many yeast production systems once the resulting data are deposited in online databases.
3.2 Introduction

3.2.1 Yeast Role In Science

Yeast has been a powerful tool as a model organism for fifteen years\(^1\). Being that they are simple eukaryotes that share many complementary genes to humans, it has lead science in the direction of developing protein therapeutics\(^2, 3\). This suggested to researchers that diseases arise from disruption on basic cellular process of DNA damage and repair, cell division and gene expression, which was confirmed with the completion of the human genome \(^4-7\). Science continues to use other model organism, such as mammalian Chinese hamster ovary cells and bacteria microbes but has taken a keen interest in yeast for its overall advantages in therapeutic medicine \(^8, 9\) \(^10, 11\). One major advantage to using a yeast microbe, is that proteins are readily expressed with human structural and functional characteristics, such as tertiary protein folding and simple glycan moieties \(^12\). Glycoengineered yeast lines have been shown to secrete highly complex needed glycoproteins for therapeutic drugs. Whereas bacteria do not, due to the lack of endoplasmic reticulum or golgi apparatus, which are essential in glycan additions to proteins \(^12-14\). Other advantageous topics are yeasts’ high protein titers production, making it easy to scale up genetic manipulation for biological production and time and cost efficiency of gene to product turn around cycle \(^12, 15-18\). One disadvantage is the glycoprotein-engineered pathway or adding glycans post secretion, is it makes it challenging for yeast to compete with mammalian cells for biopharmaceutical products, by adding non naturally occurring proteins \(^19\). However, yeast continues to make progressive strides in producing therapeutic medicines that rival with other host cell organisms.
3.2.2 Pichia Pastoris General Background

Modern day therapeutic production begins with protein production in a particular organism cell, which is compatible in all aspects of biological safety. Science has broadened the selection of host from mammals (animal) and microbes, both bacteria and yeast\cite{20, 21}. Out of 1500 species of yeast, Saccharomyces Cerevisiae (S. Cerevisiae) and Schizosaccharomyces pombe (S. Pombe) are the top two studied species\cite{22-25}. Science has been looking at Saccharomyces Cerevisiae as a model organism, since its complete DNA was sequenced about fifteen years ago\cite{1}. S. Cerevisiae has been continued to be studied as it has many homologous proteins to humans, thus making it a valuable tool in biomedical research\cite{26}. A strain of S. Cerevisiae, known as Komagataella phaffii pastoris (k.phaffii); sometimes referred to as Pichia Pastoris is currently the most widely used in such research\cite{27}. Currently, we see Pichia Pastoris being used for the high level production of foreign proteins, monitoring transformants and its ability to be optimized for secretion by minimalizing proteolysis. These advances make the organism an ideal candidate for more protein therapeutics. The Massachusetts Institute of Technology (Cambridge, MA) Center for Biomedical Innovation (CBI), generated optimized Pichia Pastoris strains to produce initial protein products; recombinant human growth hormone (hGH) and interferon-alpha 2b (IFNa2b). Their substrain NRRL-11430 was used to generate a secretome of Pichia Pastoris. With this yeast, the CBI has teamed up with DARPA (Defense Advanced Research Projects Agency) to create the InSYcT microfluidic protein therapeutic device for on-site drug delivery.
3.2.3 On-Site Protein Therapeutic Delivery

The development of the Integrated and Scalable Cyto-Technology (InSCyT) platform from the Center for Biomedical Innovation (CBI) was authenticated as the need for therapeutic drug deliverables has increased in demand. The demand has increased due to concerns with cost efficiency, rapid biological production and most importantly accessibility to the public \([12]\). With the world population averaging around seven billion inhabitants the need for therapeutic treatments will need to accommodate people in remote locations without access to modern medicinal facilities \([28]\). On-site therapeutics have already had an impact in modern medicine, with advances in areas such as those diagnosed with Type I diabetes or asthma \([29, 30]\). Both of these diseases have targeted biologics or small molecule drug therapeutics delivered on-site to the patient. There are even slow release and stable therapies as is the case for ethinyl estradiol, the main component in multiple female contraception applications \([31, 32]\). Although such described medication is available to most consumers in highly metropolitan populated countries, these standard medicines can be difficult to obtain on a regular basis and even more so when natural disaster takes place. Such was the case in Haiti of 2010, when a 7.0 Richter scale earthquake cutoff the supply of insulin to those with Type I diabetes, according to the international diabetes foundation. That is why this InSCyT platform will have a broad impact in therapeutic medicine. Its capability to be derived and delivered to patients in any location in the world, will make it desirable and a beneficial product in medicine.
3.2.4 InSCyT Program Contribution

Having yeast as a host organism has munificent benefits to fulfill desired public demands. The demands include multiple vantage points of being cost effective, quick production turnaround time, high yield titer levels and differential engineering aspects for varying protein production and efficacy \(^{[12, 33]}\). The InSCyT platform has been constructed to fulfill all of these public demands. With their set up, CBI has outlined benefits in this microfluidic system using yeast in the DARPA Biologically Derived Medicines on Demand Proposal. For example, financial expenses would decreases when using yeast, because the freeze-dry capability eliminates refrigeration pre and post biologic formation and automatic secretion reduces purification steps.

There are also significant benefits in this system of biomanufactoring because of the elevated quantitation levels of natural yeast functionality. This is in reference to its rapid growth ability and high titer levels of therapeutic protein produced. One highly desirable trait to using yeast is the ability for it to be genetically engineered to excrete product with complex post-translation modifications; especially those of complex glycan moieties. As one is able to engineer the cell, it gives a plethora of distinct product cell lines to be created and easily stored until needed for desired use at the on site location. This platform shows copious potential for public use of therapeutics. The initial intent of the microfluidic system is for patients who need a supply of hGH or interferon-alpha 2b. But, it is clear to see this system can be used for a range of therapeutic drugs, thereby making broad impact in the biopharmaceutical industry as well as to the worlds 7 billion inhabitants\(^{[33, 34]}\).
3.2.5 Analytical Application For Yeast

With the development of any biomanufactured compound, it is important to verify the quality of the product, to ensure it meets efficacy and safety therapeutic drug standards. These two principles are used in association with all modern therapeutic practices and is used as quality assurances for the CBI-InSCyT program. For my project, the yeast species Saccharomyces Cerevisiae (S. Cerevisiae) secretome was characterized, through enzymatic digestion high performance liquid chromatography-mass spectrometry. The secretome from the microbioreactor was delivered from a CBI-in-house optimized P. Pastoris strain NRRL-11430. This strain had no product gene insertion and was therefore named NULL-NRRL-11430. By deciphering the secretome, we would know all host cell proteins (HCPs) associated in this k.phaffii (p.pastoris) strain. This would be relevant in understanding the role these HCPs partake in the regulatory processes with and without product gene insertion. Thereby knowing the role the proteins play on the product modifications and cleavages. Our analytical platform of protein quantitation, enzymatic digestion and LC-MS/MS was performed on non-fractionated and AEX fractionated samples of NULL-NRRL-11430, thereby giving a full spectrum of HCPs through an added protein enrichment application. A key aspect of this project was using bioinformatics tools to define yeast host proteins. With the assistance of Proteome Discoverer® and Protein ProteinScape™, we were able to create the de novo proteome for P. Pastoris NRRL-11430. For robustness of the newly found secretome we compared HCPs in the hGH product gene yeast cell line. Having this protein list will make a great impact to biotherapeutics because it will allow scientist to monitor the affects of HCPs on all downstream production of the chosen product protein.
3.3 Materials and Methods

3.3.1 Materials

Optimized *Komagataella Phaffii*-NRRL 11430 shake-flask and bioreactor yeast secretion solution were obtained from Massachusetts Institute of Technology (Dr. Luv, Cambridge, MA) CBI-InSCyT collaboration. PreFractionation was performed on the Ultimate 3000 nanoLC (Dionex, California, USA) system, through a Thermo Fisher Scientific (Waltham, MA) 75 μm, 50 cm Acclaim® PepMap RSLC column. SDS-PAGE experimentation was performed on a Bis-Tris 12% mini-PROTEAN TGX Stain-free™ 10well 30μL comb gels (Biorad, Hercules, CA). POROS® R1 50μm Bulk Media and HPLC self-packing device (Applied Biosystems, Framingham, MA) was used to pack a 4.6mm × 30mm PEEK column (Life Technology, Milford, MA) for de-salting the protein mixture. In-Solution digestion reagents Urea and Dithiothreitol were purchased from Sigma-Aldrich (St. Louis, MO). While Iodoacetamide along with Pierce™ BCA Assay kit were purchased from Life Technologies (Carlsbad, CA). HPLC-MS grade acetonitrile, water, formic acid, and all other reagents and buffers were purchased from Thermo Fisher Scientific (Waltham, MA).

3.3.2 Methods

3.3.2.1 Protein Concentration

UnFractionated and preFractionated NRRL-11430 Null yeast solution were transferred into two-unF and three replicates per eight AEX-preF (performed through personal communication of Dr. Jared Auclair) 1.5 mL eppendorf tubes. All twenty-six samples were lyophilized and reconstituted in 200 μL 1X Phosphate Buffer Saline (PBS) and stored at -20°C until needed for experimentation. A Thermo Fisher BCA kit was used to determine the concentration of all
samples following the manufacturer’s protocol. Dilutions were made with DI H₂O as follows; 1:2 and 1:10 on unF and 0:0, 1:2, 1:10 on AEX fractions #8 and #11. The resulting concentrations were achieved at 280 nm fluorescence detection (SpectraMesh 340 PC, Molecular Devices, Sunnyvale, CA) on both unF and AEX-preF.

### 3.3.2.2 SDS-PAGE Protein Visualization

Yeast secretion of unF-Null was used for gel electrophoresis and dilutions of 1:2, 1:5 and 1:10 with DI H₂O were used; 15 µL of Laemmli Sample Buffer (Biorad, Hercules, CA) was then added to each sample containing a final concentration of 0.021 mM with the addition of 1M DTT (Dithiothreitol). This was done to promote the disruption of disulfide bond linkages; the sample mixture was incubated at 68 °C for 10 minutes. The samples were then incubated at a final concentration of 8.33mM with 2 uL iodoacetamide in the dark for 1 hour, prompting alkylation of the broken cystine bonds. All three samples were loaded on a Bis-Tris 12% mini-PROTEAN TGX Stain-free™, at 200V on the Power Pac 3000 (Biorad, Hercules, CA) for 30 minutes with freshly prepared 1X Tris-Glycine as the running buffer. The gels were rinsed with DI H₂O and Coomassie stained with Simple Blue™ SafeStain (Invitrogen, Carlsbad, CA) for 1hr or overnight. The Comomassie stain was removed by rinsing the gels with DI H₂O and rocked on low speed overnight to visualize the protein bands on the gel.
3.3.2.3 In-Solution Digestion

In-Solution digestion (InSD) was performed on the nine samples, which composed of the unF-Null and the eight AEX-preF-Null reconstituted solutions. With a concentration of 15.090 µg/µL, 10 µL of unF-Null was extracted and placed in a 1.5 µL eppendorf tube in order to digest 150 µg of protein. Next, 200 µL of 6 M Urea was added to denature and solubilize the proteins, followed by 1 µL of 1 M DTT. The sample was then incubated for 10 minutes at 68°C to break the cystine bonds. 21 µL of 100 mM IAA (Iodoacetamide) was added to the sample, before incubation at room temperature (25°C), for 45 minutes to 1 hour in the dark. After alkylation of the free cysteines with IAA, R1-reverse phase sample clean up was performed. The POROS® R1 (4.6 mm × 30 mm, PEEK) column was operated on one of the three valves on the 2-dimensional Shimadzu HPLC (Columbia, MD). The sample was then ‘cleaned-up’; de-salted by the EZstart software (Shimadzu, Columbia, MD), which programmed the valve on-line and off-line, sample flow rate and gradient. This seven minute HPLC method is described as follows. First, 250 µL of sample was injected into the HPLC. The R1 column was cleaned and equilibrated with 5 column volumes; CV of Mobile Phase A (0.1% Trifluoroacetic acid/milli-q-H2O) v/v at a flow rate of 1.00 mL/min for about 3 minutes. These conditions also eluted out the salts from the protein mixture. Next, the flow rate was increased to 3.00 mL/min with 2% Mobile Phase B (0.1% Trifluoroacetic acid/Acetonitrile) v/v. This condition was kept for 1 minute to equilibrate the proteins for elution and to ensure all salts have been removed. With the steady flow-rate, proteins were eluted off the R1 column over the course of 2 minutes with a gradient of 2% to 70% Mobile Phase B (0.1% Trifluoroacetic acid/Acetonitrile) v/v. These bound, then eluted proteins were collected into eppendorf tubes and placed on ice. To end the HPLC program sequence for the column, the column was then washed and equilibrated at a flow rate of 3.00
mL/min with 70% to 2% Mobile Phase B. Supplementary Table Q, shows the Null R1 clean-up elution timed profile and chromatograph. The eluted proteins on ice, were then enzymatically digested with 0.2 μg/µL Trypsin in 100mM Ammonium Bicarbonate, to keep the digestion ratio 25 μg sample : 1 μg enzyme. After being incubated at 37°C over night, the reaction was stopped by adding 100 µL of 0.1% Formic Acid (F.A.). The sample was lyophilized to dryness, before reconstitution with 20 µL of 1% F.A. The reconstituted samples were then centrifuged for 15 mins at 10,000 r.p.m(eppendorf 5424R, Hauppauge, NY). Once samples were properly centrifuged, 10 µL of sample was placed in autosampler vials(Thermo Scientific, Waltham, MA). For InSD of the eight AEX-preF-Null reconstituted solutions, 200 µL of each fraction was to be digested. Thus, 20 µg, 33 µg, 50µg and 100 µg of protein would to be digested based on each fractions concentration, refer to Table 1. In the InSD-mixture eppendorf tube, 100 µL of 6 M Urea, 1.5 mL of 1 M DTT and 31 µL of 100 mM IAA was added into each sample. All samples then followed the sample protocol of reduction and alkylation. As well as de-salting through R1 HPLC clean-up as prescribed on unF-Null. All samples were now ready to be analyzed with Q-TOF-MS.
3.3.2.4 Liquid Chromatography/Mass Spectrometry

With the collection of the tryptically digested peptides, they underwent UPLC (Dionex, California, USA) separation prior to being detected by Q-TOF-MS (Bruker, Billerica, MA). For unF-Null, a 1:10 dilution (sample : DI H₂O), was necessary to be in capacity of the maximum 3 µg load on the 75 µm, 50 cm Acclaim® PepMap RSLC column (Thermo Fisher). 2 µL of the 1.5 µg/µL unF-Null was loaded onto the equilibrated column and the gradient for maximum peptide separation was as follows. 5 % mobile phase B (0.1% Formic Acid/ Acetonitrile) at time points 0-5 mins, steadily increase from 5% buffer B to 40% buffer ending at 245 mins, stationary 40% mobile phase B from 245 mins to 248 mins, increase to 80% solvent B at time points 248 mins to 262 mins, decrease to 5% mobile phase B until final end point at 265 mins. All AEX preF-Null final sample concentration after InSD were; 0.38 µg/µL (#8 & #15), 0.632 µg/µL (#9 & #14), 0.976 µg/µL (#10 & #13), 2.074 µg/µL (# 11 & #12). Appropriate volumes were injected to load 2 µg-3 µg were loaded onto the column as well as follow the same UPLC separation gradient. Both set of samples underwent Q-TOF mass spectrometry analysis with intensity-dependent rate of 4-20Hz between precursor scans (Northeastern University, Dr. Jared Auclair). The resulting MS/MS spectra were characterized by ProteinScape (Bruker).
3.3.2.5 Characterize Proteins Through Peptide Sequencing Identification

The chromatographic data produced from Q-TOF-MS instrumentation came from unF-Null and the eight AEX-preF samples. Each sample was searched against all taxonomies (species) and Saccharomyces Cerevisiae s288c strain. The parameters for protein identification is described below. Two missed cleavages; precursor ion mass tolerance of 2.0 Da; fragment ion mass tolerance 1.0Da; full trypsin as enzyme; carboxyamidomethylation(C) as a fixed modification. False Discovery Rate(FDR) was targeted at 1% to give one the best confidence score for peptide matching. SEQUEST (Thermo Electron) algorithm present in Thermo Fisher Proteome Discoverer 1.3 software suite for proteins identification.
3.4 Experimental Procedure Flow diagrams

3.4.1 Experimental Flow diagram: A

Figure 1A: 
*Komagataella Phaffii Pastoris* unFractionated Null NRRL-11430 analytical platform
3.4.1 Experimental Flow diagram: B

Figure 1B: *Komagataella Phaffii Pastoris AEX-preFractionated* Null NRRL-11430 analytical platform
3.5 Results

3.5.1 Establish Protein Concentration: Null

The goal for the CBI-InSCyT program is to produce protein therapeutic deliverables on-site in the host organism yeast. Therefore, I set out to validate the yeast strain, *pichia pastoris* NRRL-11430 strain, for the use in on-site protein production. As part of the validation process, we set out to characterize the host cell proteins (HCPs) of this strain that are excreted. After receiving the small batch NRRL-11430 (Null), SHAKE-FLASK yeast secretion, the immediate step was to determine the protein concentration of the mixture. The concentration was determined to be 15.090 \( \mu g/\mu L \), using the BCA kit (Fisher scientific) and a 1:2 dilution and 1:10 dilution of the Null host cell protein mixture. I determined the concentration of our sample mixture in order to understand its contents for accurate enzymatic digestion, thus yielding quantifiable mass spectrometry detection of the proteins. Along with protein concentration, I ran an SDS-PAGE of the sample as well. The gel is shown in Supplementary Figure R. We see that, even with high protein yield, the proteins do not localize well on the gel, indicating the better protein digestion can be achieved through in-solution digestion. In solution protein digestion (InSD), is a method that was advantageous due to our low sample volume and high protein concentration. The protocol is briefly described as follows; solubilizing and denaturing our proteins in Urea reduction and alkylation with DTT and IAA, de-salting the mixture for optimal ionization through a reverse phase-R1 liquid chromatography column, and proteolysis with the addition of Trypsin. Figure 2 A-H, shows the R1-column successfully eluted *k.phaffii*-Null (unF) proteins for Q-TOF mass spectrometry detection. This analytical process contributed to my platform development for maiden characterization of the Null secretome.
Figure 2A: (R1) Reverse Phase protein elution concentration profile.

Figure 2B: Blanks to clean R1 column prior to protein sample injection.

Figure 2C: Reference blank (refBlk) to protein samples; Area indicated.

Figure 2D: refBlk and null 1 aligned.

Figure 2E: refBlk and replicate null; null 2 aligned.

Figure 2F: refBlk and null 1 and null 2; Area indicated. Insert shows all samples aligned with one another.

Figure 2 G: refBlks before and after each null sample injection.

**Figure 2 A-H:**
*Komagataella Phaffii*
Unfractionated(unF) NULL-11430 protein elution chromatographs

Figure 2H: refBlks after each null run comparison.
3.5.2 Establish Protein Concentration: AEX

In order to try to increase the number of HCPs identified, we preFractionated the sample using anion exchange chromatography. In Supplementary Figure S, we see the complete gradient and elution profile of AEX *k.phaffii*-Null fractions (with collaboration of Dr. Jared Auclair, senior scientist Barnett Institute). Of the sixteen automated collected fractions, numbers eight through fifteen were used as the fractionated set for AEX *k.phaffii*-Null. As with unfractionated (unF) Null, it was imperative to know the protein concentrations within each fraction/collectant. The concentrations were obtained through a standard BCA assay kit (Fisher Scientific). BCA experimentation was performed on samples #8 and #11, with 0:0, 1:2 and 1:10 dilutions for a range of testing. Based on the concentration of fraction #11 of 0.488 µg/µL, I was able to give approximate concentrations of all other fractions, by comparing mAU amounts. Table 1 below displays concentration of the fractions.

<table>
<thead>
<tr>
<th>Fraction Number</th>
<th>mAU</th>
<th>Concentration [µg/µL]</th>
</tr>
</thead>
<tbody>
<tr>
<td>#8</td>
<td>~150</td>
<td>0.0976</td>
</tr>
<tr>
<td>#9</td>
<td>~200</td>
<td>0.162</td>
</tr>
<tr>
<td>#10</td>
<td>~400</td>
<td>0.244</td>
</tr>
<tr>
<td>#11</td>
<td>~700</td>
<td>0.488</td>
</tr>
<tr>
<td>#12</td>
<td>~650</td>
<td>0.488</td>
</tr>
<tr>
<td>#13</td>
<td>~400</td>
<td>0.244</td>
</tr>
<tr>
<td>#14</td>
<td>~200</td>
<td>0.162</td>
</tr>
<tr>
<td>#15</td>
<td>~150</td>
<td>0.0976</td>
</tr>
</tbody>
</table>

**Table 1:** Anion Exchange Fraction Collection Concentrations
SDS-PAGE was not performed since the fractions are part of the complete unF-Null. Instead the InSD protocol for unF-Null was also applied to these eight AEX wild-type Null samples.

Multiple calculated steps were procured to assure sample volume concentration of solubilizing with Urea was $\geq 0.1 \mu g/\mu L$ and Trypsin : sample ratio was $1 \mu g : 25 \mu g$. In Figure 3 A-F, we again see the R1-column ‘clean-up’ sample elution profile, for maximum mass spectrometry ionization detection. With the AEX wild-type Null samples analyzed with Q-TOF-MS, an adjunct host cell protein list was created. This list was then compared, contrasted and compiled to produce the classified *Komagataella Phaffii* Null-NRRL 11430 secretome list.

**Figure 2A:** (R1)Reverse Phase protein elution concentration profile.

**Figure 2B:** Blanks to clean R1 column prior to protein sample injection.

**Figure 2C:** Reference blank(refBlk) to protein samples; Area indicated.

**Figure 2D:** refBlk and null 1 aligned.

**Figure 2E:** refBlk and replicate null; null 2 aligned.

**Figure 2F:** refBlks before and after each null sample injection.

**Figure 2G:** refBlks after each null run comparison.

**Figure 2H:** refBlks after each null run comparison.

**Figure 3 A-F:** *Komagataella Phaffii* Anion Exchange preFractionated NULL-11430 protein elution chromatographs
3.5.3 Host Cell Protein Secretome List

The host cell protein secretome, was determined through peptide identification using LC-Q-TOF mass spectrometry. Table 2 below shows a detailed list of seventy-four; 74, *pichia pastoris* NRRL-11430 host cell proteins. These proteins were obtained by analysis of unfractionated-Null and in-house anion exchange-Null yeast protein secretions. All samples were searched using ProteinScape™ bioinformatics protein identification software. With search criteria databases of all taxonomies and Saccharomyces Cerevisiae_s288C strain. From the variable databases and sample constituents, Table 3 was created to show different facets in the multiple sample pool. It was proven that all taxonomies search would yield more protein hits, but would lack authenticity of *pichia pastoris* protein members. Thereby, allowing us to rule out false-positive identification of the 74 HCPs. Figure 4, shows the total counts of proteins identify in unF-Null and AEX-preF-Null.
Figure 4:
Extracted count of Anion Exchange proteins in classified HCP *k.phaffii* Null Secretome
3.5.4 Comparative Description of Null Strain and Product Strain

Continuous experimentation of *Komagataella Phaffii* Null-NRRL 11430 secretome, led to profiling proteins within Massachusetts Institute of Technology in-house hGH and INFα2b yeast secretion solutions. There were also recombinant manufactured standards; rINFα2b(Sigma Aldrich, Ray Biotech) and rhGH(Novo Nordisk) that were used as comparative controls. The proteomics core facility used a compiled database of all Uniprot S. Cerevisiae, *Pichia Pastoris*, hGH and INFα2b from the Thermo Fisher Proteome Discover to reveal proteins within MIT-in-house and manufactured standards. I was then able to search and extract proteins found in my focus of bioreactor produce samples. The bioreactor analysis with and without an osmotic stabilizer; in this case sorbitol. Then the gel bands were excised at 22 kDa, thereby decreasing HCP count significantly. In searching both INFα2b (-/+)-sorbitol and hGH (-/+)-sorbitol, there was only one protein matching from the *k.phaffii*-Null secretome. This protein is Phosphoglycerate kinase, a functionally known enzyme that catalyzes the reversible reaction of ADP to become ATP \[^{[37,38]}\]. It can now be concluded that there are no additional HCPs outside of the secretome characterized list. Thereby, improving future possibilities of less adverse affects of up-regulated or down-regulated HCPs of hGH and INFα2b on-site conjured therapeutics.
3.6 Discussion

Table 2:
Descriptive table of classified *k.phaffii* host cell proteins. Column heading describes column content; Proteinscape™ identification (ID); Uniprot alternative ID; Spectral Count = total spectra identified for protein; Evidence, Function and Enzyme collected from Uniprot % coverage = percent of protein sequence covered in chromatograph spectrum.
Pages 84-86.
Table 2: page 1
Descriptive table of classified *k.phaffii* host cell proteins.

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Uniprot name</th>
<th>Spectral Count (peptide)</th>
<th>Evidence Observance Level</th>
<th>Function</th>
<th>Enzyme</th>
<th>% coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vacular protein sorting-</td>
<td>Vacular protein sorting-</td>
<td>2</td>
<td>Protein</td>
<td>Protein</td>
<td></td>
<td>2.00</td>
</tr>
<tr>
<td>associated protein 16</td>
<td>associated protein 16</td>
<td></td>
<td></td>
<td>regulates transcription</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vacular morphogenesis protein</td>
<td>9</td>
<td></td>
<td></td>
<td>TATA binding protein</td>
<td>yes</td>
<td>1.10</td>
</tr>
<tr>
<td>SWI/SNF chromatin-remodeling</td>
<td>SWI/SNF chromatin-remodeling</td>
<td>1</td>
<td>Protein</td>
<td>Transcriptional activation</td>
<td></td>
<td>1.20</td>
</tr>
<tr>
<td>complex subunit SWH1</td>
<td>complex subunit SWH1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anaphase spindle elongation</td>
<td>Anaphase spindle elongation</td>
<td>1</td>
<td>Protein</td>
<td>DNA directed DNA polymerase</td>
<td></td>
<td>0.00</td>
</tr>
<tr>
<td>Protein RAD61</td>
<td>DNA polymerase V</td>
<td>1</td>
<td>Protein</td>
<td>DNA repair cellulose</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>DNA polymerase epsilon</td>
<td>DNA polymerase epsilon</td>
<td>1</td>
<td>Protein</td>
<td>DNA repair yeast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Catalyses the synthesis of</td>
<td>Catalyses the synthesis of</td>
<td>1</td>
<td></td>
<td>DNA repair yeast</td>
<td>2.60</td>
<td></td>
</tr>
<tr>
<td>dihydouridine(20) synthase</td>
<td>dihydouridine(20) synthase</td>
<td>1</td>
<td></td>
<td>DNA synthesis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphoglycerate kinase</td>
<td>Phosphoglycerate kinase</td>
<td>1</td>
<td>Protein</td>
<td>Phosphoglycerate kinase activity</td>
<td></td>
<td>2.40</td>
</tr>
<tr>
<td>Protein MIF2</td>
<td>Protein MIF2</td>
<td>1</td>
<td>Protein</td>
<td>Required for nuclear protein import and mediates docking of import substrate to distinct nucleoporins</td>
<td></td>
<td>0.80</td>
</tr>
<tr>
<td>Vacular amino acid transporter</td>
<td>Vacular amino acid transporter</td>
<td>1</td>
<td>Protein</td>
<td>Amino-acid transport</td>
<td></td>
<td>1.20</td>
</tr>
<tr>
<td>Involved in resistance to</td>
<td>Involved in resistance to</td>
<td>1</td>
<td>Protein</td>
<td>DNA replication</td>
<td></td>
<td>0.40</td>
</tr>
<tr>
<td>ionizing radiation</td>
<td>ionizing radiation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probable E3 ubiquitin-protein</td>
<td>Probable E3 ubiquitin-protein</td>
<td>1</td>
<td>Protein</td>
<td>ubiquitin/proteasome pathway</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>ligase HUL5</td>
<td>ligase HUL5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probable formate dehydrogenase</td>
<td>Probable formate dehydrogenase</td>
<td>2</td>
<td>Protein</td>
<td>NONE</td>
<td></td>
<td>2.00</td>
</tr>
<tr>
<td>FORMATE DEHYDROGENASE 2</td>
<td>FORMATE DEHYDROGENASE 2</td>
<td>1</td>
<td>Protein</td>
<td>formate catabolic process</td>
<td></td>
<td>7.20</td>
</tr>
<tr>
<td>Probable regulatory protein</td>
<td>Regulatory protein SIR4</td>
<td>1</td>
<td>Protein</td>
<td>Transcription regulation</td>
<td></td>
<td>1.20</td>
</tr>
<tr>
<td>Silent information regulator</td>
<td>Silent information regulator</td>
<td>1</td>
<td>Protein</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncharacterized protein</td>
<td>Uncharacterized protein</td>
<td>1</td>
<td>Protein</td>
<td></td>
<td></td>
<td>4.00</td>
</tr>
</tbody>
</table>
## Table 2: page 2
Descriptive table of classified *k.phaffii* host cell proteins.

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Protein scape</th>
<th>Uniprot name</th>
<th>Spectral Count(peptide)</th>
<th>Evidence Observance Level</th>
<th>Function</th>
<th>Enzyme</th>
<th>% coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aminomethyltransferase, mitochondrial</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>degradation of GLY</td>
<td>yes</td>
<td>2.00</td>
</tr>
<tr>
<td>DNA mismatch repair protein MSH6</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>DNA repair</td>
<td></td>
<td>0.60</td>
</tr>
<tr>
<td>DNA repair and recombination protein RAD54</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>DNA repair</td>
<td>yes</td>
<td>1.20</td>
</tr>
<tr>
<td>Protein NNF2</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>protein binding</td>
<td></td>
<td>2.20</td>
</tr>
<tr>
<td>Delta-1-pyrrrole-5-carboxylate dehydrogenase, mitochondrial</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>glutamate and proline biosynthesis processes</td>
<td>yes</td>
<td>2.30</td>
</tr>
<tr>
<td>Putative uncharacterized protein YBR174C</td>
<td></td>
<td></td>
<td>1</td>
<td>uncertain</td>
<td>NONE</td>
<td></td>
<td>15.40</td>
</tr>
<tr>
<td>T-complex protein 1 subunit delta</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>Molecular chaperone</td>
<td></td>
<td>2.70</td>
</tr>
<tr>
<td>Protein DIA2</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>ubiquitination and subsequent proteasomal degradation of target proteins</td>
<td></td>
<td>2.50</td>
</tr>
<tr>
<td>Pleiotropic ABC efflux transporter of multiple drugs</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>Confers resistance to numerous chemicals</td>
<td></td>
<td>1.10</td>
</tr>
<tr>
<td>Serine/threonine-protein kinase SKM1</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>cytoskeletal functions</td>
<td>yes</td>
<td>1.40</td>
</tr>
<tr>
<td>Uncharacterized protein YLL007C</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>NONE</td>
<td></td>
<td>2.00</td>
</tr>
<tr>
<td>Inheritance of peroxisomess protein 2</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>peroxisome-specific receptor</td>
<td></td>
<td>1.30</td>
</tr>
<tr>
<td>Protein transport protein DSL1</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>ER-Golgi transport</td>
<td></td>
<td>1.20</td>
</tr>
<tr>
<td>U3 small nuclear RNA-associated protein 22</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>Ribosome biogenesis</td>
<td>yes</td>
<td>1.60</td>
</tr>
<tr>
<td>Uncharacterized protein YGR042W</td>
<td></td>
<td></td>
<td>1</td>
<td>predicted</td>
<td>NONE</td>
<td></td>
<td>5.90</td>
</tr>
<tr>
<td>Uncharacterized protein YGR153W</td>
<td></td>
<td></td>
<td>1</td>
<td>predicted</td>
<td>NONE</td>
<td></td>
<td>3.70</td>
</tr>
<tr>
<td>Cleavage factor two protein 2</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>mRNA processing</td>
<td>yes</td>
<td>1.60</td>
</tr>
<tr>
<td>DNA repair protein RAD1</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>DNA repair</td>
<td></td>
<td>2.30</td>
</tr>
<tr>
<td>E3 SUMO-protein ligase SIIZ2</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>DNA</td>
<td>yes</td>
<td>1.40</td>
</tr>
<tr>
<td>Mitochondrial chaperone TCM62</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>stress response</td>
<td></td>
<td>2.40</td>
</tr>
<tr>
<td>Condensin complex subunit 2</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>Regulatory subunit of the condensin complex</td>
<td></td>
<td>1.20</td>
</tr>
<tr>
<td>Mitochondrial Rho GTPase 1</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>Mitochondrial trafficking</td>
<td>yes</td>
<td>2.10</td>
</tr>
<tr>
<td>Protein HRB1</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>RNA and nucleotide binding</td>
<td></td>
<td>2.60</td>
</tr>
<tr>
<td>Uncharacterized protein YIR024C</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>NONE</td>
<td></td>
<td>3.20</td>
</tr>
<tr>
<td>RNA polymerase I-specific transcription initiation factor RRN3</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td>Transcription regulation</td>
<td></td>
<td>3.00</td>
</tr>
<tr>
<td>Mitotic check point protein BFA1</td>
<td></td>
<td></td>
<td>1</td>
<td>Protein</td>
<td><em>cell cycle mitosis</em></td>
<td></td>
<td>2.80</td>
</tr>
</tbody>
</table>
### Table 2: Descriptive table of classified \textit{k.phaffii} host cell proteins.

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Protein scape</th>
<th>Uniprot name</th>
<th>Spectral Count (peptide)</th>
<th>Evidence Observance Level</th>
<th>Function</th>
<th>Enzyme</th>
<th>% coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein PAM1</td>
<td>Protein PAM1</td>
<td>1</td>
<td>Protein</td>
<td>protein binding</td>
<td>Ribosome biogenesis</td>
<td></td>
<td>1.30</td>
</tr>
<tr>
<td>Ribosome biogenesis protein SLX9</td>
<td>Ribosome biogenesis protein SLX9</td>
<td>1</td>
<td>Protein</td>
<td>Ribosome biogenesis</td>
<td>rRNA processing</td>
<td></td>
<td>5.70</td>
</tr>
<tr>
<td>Factor arrest protein 10</td>
<td>Factor arrest protein 10</td>
<td>1</td>
<td>Protein</td>
<td>cell cycle</td>
<td></td>
<td></td>
<td>2.50</td>
</tr>
<tr>
<td>Protein NUB1</td>
<td>Protein NUB1</td>
<td>1</td>
<td>Protein</td>
<td>cell cycle</td>
<td></td>
<td></td>
<td>2.70</td>
</tr>
<tr>
<td>Hexose transporter HXT16</td>
<td>Hexose transporter HXT16</td>
<td>1</td>
<td>Inferred from homology</td>
<td>sugar transport</td>
<td></td>
<td></td>
<td>1.10</td>
</tr>
<tr>
<td>Pre-mRNA-splicing factor SYF2</td>
<td>Pre-mRNA-splicing factor SYF3, PRP19 complex protein 31, Synthetic lethal with CDC40 protein 2</td>
<td>1</td>
<td>Protein</td>
<td>mRNA processing</td>
<td>Golgi apparatus</td>
<td>yes</td>
<td>0.60</td>
</tr>
<tr>
<td>Probable phospholipid-transporting ATPase DRS2</td>
<td>Probable phospholipid-transporting ATPase DRS2</td>
<td>1</td>
<td>Protein</td>
<td>ATP-binding</td>
<td>Nucleotide-binding</td>
<td>unfolded protein binding</td>
<td>yes</td>
</tr>
<tr>
<td>Heat shock protein SSC3, mitochondrial</td>
<td>Heat shock protein SSC3, mitochondrial, Extracellular mutant protein 10</td>
<td>1</td>
<td>Protein</td>
<td>ATP-binding</td>
<td>Nucleotide-binding</td>
<td>unfolded protein binding</td>
<td>yes</td>
</tr>
<tr>
<td>Serine/threonine-protein kinase HAL4/SAT4</td>
<td>Serine/threonine-protein kinase HAL4/SAT4, Halotolerance protein 4</td>
<td>1</td>
<td>Protein</td>
<td>ATP-binding</td>
<td>Nucleotide-binding</td>
<td>unfolded protein binding</td>
<td>yes</td>
</tr>
<tr>
<td>Ribosome assembly 1 protein</td>
<td>Ribosome assembly 1 protein, EF-2-like GTPase, Elongation factor-like 1</td>
<td>1</td>
<td>Protein</td>
<td>Ribosome biogenesis</td>
<td></td>
<td>yes</td>
<td>3.40</td>
</tr>
<tr>
<td>Probable target of rapamycin complex 2 subunit BIT2</td>
<td>Probable target of rapamycin complex 2 subunit BIT2, Binding partner of TOR2</td>
<td>1</td>
<td>Protein</td>
<td>NONE</td>
<td></td>
<td></td>
<td>2.00</td>
</tr>
<tr>
<td>E3 ubiquitin-protein ligase DMA1</td>
<td>E3 ubiquitin-protein ligase, Checkpoint forhead associated with RING domains-containing protein 1, Defective in mitotic arrest protein 1</td>
<td>1</td>
<td>Protein</td>
<td>cell cycle retarding</td>
<td></td>
<td>yes</td>
<td>2.60</td>
</tr>
<tr>
<td>5'-3' exoribonuclease 2</td>
<td>5'-3' exoribonuclease 2, Ribonucleic acid-trafficking protein 1, p116</td>
<td>1</td>
<td>Protein</td>
<td>mRNA processing</td>
<td>rRNA processing</td>
<td>transcription processes</td>
<td>yes</td>
</tr>
<tr>
<td>Uridine nucleosidase</td>
<td>Uridine nucleosidase, Uridine ribohydrolase</td>
<td>1</td>
<td>Protein</td>
<td>catalytic activity</td>
<td></td>
<td>yes</td>
<td>2.60</td>
</tr>
<tr>
<td>Uncharacterized protein YCR101C</td>
<td>Uncharacterized protein YCR101C</td>
<td>1</td>
<td>Predicted</td>
<td>NONE</td>
<td></td>
<td></td>
<td>4.40</td>
</tr>
<tr>
<td>Bile pigment transporter 1</td>
<td>Bile pigment transporter 1</td>
<td>1</td>
<td>Protein</td>
<td>transport</td>
<td></td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>Nucleolar protein NET1</td>
<td>Nucleolar protein NET1</td>
<td>1</td>
<td>Protein</td>
<td>rDNA binding</td>
<td>protein binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Serine/threonine-protein kinase CHK1</td>
<td>Serine/threonine-protein kinase CHK1, Checkpoint kinase 1</td>
<td>1</td>
<td>Protein</td>
<td>DNA damage</td>
<td>catalytic activity</td>
<td></td>
<td>yes</td>
</tr>
<tr>
<td>Initiation-specific alpha-1,6-mannosyltransferase</td>
<td>Initiation-specific alpha-1,6-mannosyltransferase</td>
<td>1</td>
<td>Protein</td>
<td>Golgi apparatus</td>
<td></td>
<td>yes</td>
<td>2.90</td>
</tr>
<tr>
<td>N(+)K(+)H(+) antipporter</td>
<td>N(+)K(+)H(+) antipporter</td>
<td>1</td>
<td>Protein</td>
<td>antiporter activity</td>
<td></td>
<td></td>
<td>0.90</td>
</tr>
<tr>
<td>Target of rapamycin complex 1 subunit KOG1</td>
<td>Target of rapamycin complex 1 subunit KOG2, Kontroller of growth protein 1, Local anesthetic-sensitive protein 24</td>
<td>1</td>
<td>Protein</td>
<td>ubiquitin binding</td>
<td>protein binding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37S ribosomal protein S22, mitochondrial</td>
<td>37S ribosomal protein S22, mitochondrial</td>
<td>1</td>
<td>Protein</td>
<td>structural constituent of ribosome transferase activity</td>
<td></td>
<td>yes</td>
<td>1.30</td>
</tr>
<tr>
<td>Pyruvate dehydrogenase E1 component subunit alpha, mitochondrial</td>
<td>Pyruvate dehydrogenase E1 component subunit alpha, Pyruvate dehydrogenase complex component E1 alpha</td>
<td>1</td>
<td>Protein</td>
<td>glycolysis</td>
<td></td>
<td>yes</td>
<td>3.60</td>
</tr>
<tr>
<td>Spindle pole body-associated protein CJK1</td>
<td>Spindle pole body-associated protein CJK1, Chromosome instability and karyogamy protein 1</td>
<td>1</td>
<td>Protein</td>
<td>microtubule motor activity</td>
<td></td>
<td></td>
<td>2.90</td>
</tr>
</tbody>
</table>
3.6.1 Host Cell Protein Analytical Characterization for CBI-InSCyT Platform

In developing the *k.phaffii*-Null secretome shown in Table 2 we were able find descriptive evidence for all of the 74 Host Cell Proteins (HCPs). Knowing each proteomic function, would benefit in downstream isolation of the purified target protein. In Table 2, we see that about 11% of the proteins pertain to a form of RNA processing. Whether it be linked/directed to RNA polymerases, nucleotide binding or an objective of the RNA genome. Such claims can be found upon studying Transposon Ty3-G Gag polyprotein(#5), Transcription factor TFIIIB component B"(#13) and U3 small nucleolar RNA-associated protein 22(#38). Acquiring RNA descriptive proteins is as relevant as revealing DNA activity in our NRRL-11430 S. Cerevisiae strain. With 24% of HCPs related to genomic construction of both DNA and RNA, we have become aware of the impact incorrect plasmid vector design may have on target protein secretion. The known DNA functionalities of these proteins consist of DNA damage repair, DNA replication, and directed polymerases among other processes. This can be seen in DNA repair protein REV1(#11), DNA polymerase V(#12), and E3 SUMO-protein ligase SIZ2(#43). There are also protein classes of energy participants, but more directly referred with post translation modification aptitudes. Multiple proteins such as Protein transport protein DSL1(#37) or Initiation-specific alpha-1,6-mannosyltransferase(#69) exist for the endoplasmic reticulum and golgi apparatus organelles. As is the case for Hexose transporter HXT16(#55), Heat shock protein SSC3, mitochondrial(#58) and Serine/threonine-protein kinase HAL4/SAT4(#59) contribute in ATP binding. Other organelles such as the mitochondria and its ribosomes make up 12% of protein function of HCPs is this yeast secretome. Giving a total of 20% of proteomic activity is solely toward perfect protein compilation of S. Cerevisiae. To declare the proteins and their purpose in S. Cerevisiae, Uniprot protein database allowed me to describe cellular evidence.
Cellular evidence is significantly important as not only a way to rule out false-positive ProteinScape™ (Bruker, Billerica, MA) identification, but it gives scientist the opportunity to add onto the identity platform and fill in previously non-existing proteins. This can be seen in Table 2 Evidence Observance Level column, where I described the proteins to be inferred as noted within Uniprot. There are four levels of potential evidence observance levels for the seventy-four host cell proteins. These levels indicate that through validated scientific experiments, laboratory research groups have seen these proteins with protein product observance, transcript observance, genomic observance or inferred from homology observance. The homology is a reference to very preliminary homologous evidence of the protein, but nothing conclusive enough on the protein, gene or transcript levels. Of the proteins, 97% of the HCPs have been seen in experimental evidence at protein level, through varying scientific experimentation. However, two of the seventy-four proteins have been inferred by homology. I have added to the database architects that it has now been observed on a protein level. With the contribution of the k.phaffii-Null secretome, I have identified proteins that would participate in the InSCyT on-site protein therapeutic devices. This was an substantial role for the device, since it is important to understand all components for any analytical instrumentation. These proteins would be involved in an array of manufacturing techniques including, the miniscule scale platform, product purification and even solvent permeation of all molecules. My analytical platform was developed to explore all Host Cell Proteins of NRRL-11430 strain, and in achieving that, give a clear proteomic path for future product protein production in this strain.
3.6.2 *Komagataella Phaffii* Protein Concentration of Unfractionated and Anion Exchange PreFractionated Wildtype-NRRL 11430

The goal of this project was to descriptively identify all host cell proteins within the CBI-InSCyT optimized *Saccharomyces Cerevisiae* *Pichia Pastoris* Null strain. Prior to HCP analysis, I determined the shake-flask (unF) and bioreactor (pref) yeast secretion concentrations; both unF(unfractionated)-Null and AEX(Anion Exchange)-Null. The BCA assay of unF-Null, showed a concentration of 15 µg/µL. This indicated two important points. One point being, this manipulated strain did indeed secrete a high yield of proteins. Which was a conformation given to the microbiologist who optimized the strain. The concentration also suggested that there was adequate protein for proteomics experiments. I identified 74 HCPs from unF-Null. In order to gather a complete collection of HCPs, it was important to analyze separate sections of the unF-Null. These sections were created by injecting the unF-Null secretion for liquid chromatographic protein separation by an anion exchange column. Anion exchange chromatography separates proteins based on their elution off of a positive stationary phase resin as the ionic strength of the salt concentration mobile phase increases over time \[39, 40\]. Ion exchange separation is a great compliment for separating the protein mixtures, based on the proteins’ net charge in the presence of an acidic or basic pH solvent front. This will cause the protein to be negative (pH above the PI) or positive (pH below the PI). Thus it is complementary to anion (negative proteins) or cation (positive proteins) chromatography. In Supplementary Figure S, we see that the increase in ionic strength gradient; represented by the light green and brown lines, as increases protein content within the entire elution profile. Only eight of the total sixteen collected fractions were used as a part of the AEX-Null samples. The mAU(milli-absorbance unit) ranges from 150 through 700, in the chosen eight samples. With the identified samples, the BCA concentration
experiment, was designed to capture the concentration of the lowest mAU and the highest mAU. We therefore used samples #8 and #11, which have the described mAU thresholds. The concentration profile of the AEX-Null samples are depicted in Table 1. To determine other concentrations, I devised a ratio comparison of the true calculated #11 concentration. As #11 has 700 mAU with 0.488 µg/µL, #10 has ½ that at 400 mAU, therefore the concentration would be ½, making it 0.244 µg/µL. The other ratio concentrations are as follows; #9 at ⅓, #8 at ⅕, #12 = #11, #13 at ½, #14 at ½ and #15 at ⅓. The next step was performing in-solution digestion on both unF-Null and AEX-Null. The InSD is designed to give accurate results as long as the newly solvated and denatured sample-urea concentration is ≥ 0.1 µg/µL. For accurate digestion, there also needed to be a final solution concentration of 5 mM DTT and 10 mM IAA. Thus, reasons as to why different amounts of detergents was used in unF and AEX-preF. Having sample-urea mixture is important so that the proteins will still be able to denature fully with the amount of urea and other solvents added\textsuperscript{[41]}. The final step in the InSD procedure was to de-salt our mixture. That is because, salts are extremely disadvantageous to mass spectrometry detection. I used a R1-reversed phase POROS\textsuperscript{®} column to remove all salts from the mixture\textsuperscript{[42, 43]}. Looking in Figure 2 and Figure 3, R1 columns were thoroughly cleaned prior to initial sample injection, with mobile phases A and B solvent solutions. We can see in Figure 2G the comparison of the clean blank before each duplicate unF-Null, as well as the blanks in between Null1 and Null2 injections. Blanks between each injection were done to assure no sample was carried over into the following run. This sample principle R1-column chromatography methodology was carried over when analyzing AEX-Null samples. In looking at these profiles, I immediately noticed the slight decrease in area value. This is due to there no longer being a large conglomerate of proteins at one time. That would defeat the purpose of fractionation, which is to separate the
protein mixture. Examining blk_after_sample13.2 (Figure 3F), the spikes in the salt elution area, indicated just that. There seemed to be a build up of salts from all the samples and it needed to be flushed out. This had no impact on AEX-wt-null 14.2 sample injection. All experiments before mass spectrometry analysis are key parts in my analytical platform for host cell protein characterization of k.phaffii-null NRRL-11430.

3.6.3 Elucidating Mass spectrometry Protein Identification

Identifying all proteins found in the Null strain was the crucial focal point in order to evaluate product induced strains; hGH and IFNα2b. With the knowledge of all proteins, we would truly be able to access the downstream product productivity for patient treatment. A detection system with high resolution power was needed, because such instruments can capture peptide fragments that would otherwise be unnecessarily lost. We chose both unF-Null and AEX-Null samples to be analyzed by Quadrupole-Time-Of-Flight mass spectrometry\(^{[44, 45]}\). To enrich the secretome list, we optimize protein separation with HPLC application\(^{[46]}\). This included having a steadily increasing mobile phase B (0.1% Formic Acid/ACN v/v) gradient of 5% - 80% over a 265 minute time period, through the 75 µm, 50 cm Acclaim® PepMap RSLC column (ThermoFisher). Thus time, column capacity and gradient, were factors in optimization of detection of each peptide in our digested k.phaffii Null solutions. The next step after spectrometry detection was to characterize the peptides that have been produced in this analytical platform. Characterizing these peptides to corresponding yeast proteins, was an important asset for accurate validation to being the host organism for the InSCyT program\(^{[47]}\). For this, particular organism proteomes were using the ProteinScape™ (Bruker Daltonics). Two distinct categories were the grounds for our search criteria. With the intent to maximize protein
recovery, we were exclusive in choosing proteomes to match with resulting peptide spectra. The first category was to search the spectra again S. Cerevisiae s288c strain [48]. As we examined other yeast strains such as the multiple Pichia Pastoris, [49, 50] it did not produce acceptable results, based on factors such as extremely low to negative sequence coverage or numerically low protein list. Referring back to there being no complete proteome for Pichia Pastoris, we expected low protein identification. Therefore, we chose to go with surveyed S. Cerevisiae strains, picking s288c. Another approach to finding all proteins of k.phaffii-null was to search the spectra against all species taxonomies. This was performed as a quantitation method, to see how many other proteins match the same peptides analyzed from the Q-TOF. Looking in Table 3, the is an average of four to 5 fold difference in S. Cerevisiae s288c and all species. Taking a more precise focus, shows the same difference in S. Cerevisiae proteins in the all species search criteria. Proving, all proteins can be accounted for in my characterized list. An interesting note, is that by doing the AEX fractionation, there was an increase in spectral count for the number of peptides per protein observed. As for the unF-Null, three S. Cerevisiae protein were found in the all species search while zero were found in originated secretome. Of the 1,502, we expected to see more than described above. This is a reflection of yeast not being prevalent in databases. Here we address the issue of not having a completed genome and having varying candidates in one species, can alter its protein content, as is the case for choosing our particular s. cerevisiae strain. That is why it was important to identify HCPs in this organism, so that the information can be donated to improving yeast genome knowledge, thereby improving more protein identifications. Though initial isolation of InSD trypsin digestion and HPLC chromatographic separation to bioinformatics contribution through selected organism strains. The analytical
platform has now been completed for the identification of host cell proteins in Komagataella Phaffii-NRRL 11430.
Table 3: Compiled Characterization Accountants of unF-Null and preF AEX-Null varying protein identity detail.

<table>
<thead>
<tr>
<th>unF</th>
<th>Proteinscape search Parameter</th>
<th>Total Proteins [# one-hit pep.ID],[# 0-hit pep.ID]</th>
<th>Total classified sacc.cerev null proteins within s288c database proteins [# one-hit pep.ID],[# 0-hit pep.ID]</th>
<th>Total of s288c database proteins on in this fraction [# one-hit pep.ID],[# 0-hit pep.ID]</th>
<th>Total ALL_SPECIES database proteins within classified sacc.cerev null proteins [# one-hit pep.ID],[# 0-hit pep.ID]</th>
<th>Total (s. cere.) found in ALL_SPECIES database proteins on in this fraction [# one-hit pep.ID],[# 0-hit pep.ID]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>All taxonomies</td>
<td>1501</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AEX8</td>
<td>Database search Parameter</td>
<td>Number ID Proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#8</td>
<td>All taxonomies</td>
<td>2619</td>
<td>596</td>
<td>596[434],[96]</td>
<td>2[1],[1]</td>
<td>49[37],[12]</td>
</tr>
<tr>
<td></td>
<td>Saccharomyces cerevisiae s288c</td>
<td></td>
<td>18[10],[4]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#9</td>
<td>All taxonomies</td>
<td>2923</td>
<td>697</td>
<td>697[491],[132]</td>
<td>2[2],[0]</td>
<td>65[46],[18]</td>
</tr>
<tr>
<td></td>
<td>Saccharomyces cerevisiae s288c</td>
<td></td>
<td>18[10],[5]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#10</td>
<td>All taxonomies</td>
<td>2588</td>
<td>617</td>
<td>617[421],[124]</td>
<td>1[0],[1]</td>
<td>45[37],[8]</td>
</tr>
<tr>
<td></td>
<td>Saccharomyces cerevisiae s288c</td>
<td></td>
<td>22[13],[2]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#11</td>
<td>All taxonomies</td>
<td>2822</td>
<td>669</td>
<td>669[454],[134]</td>
<td>1[1],[0]</td>
<td>42[29],[12]</td>
</tr>
<tr>
<td></td>
<td>Saccharomyces cerevisiae s288c</td>
<td></td>
<td>21[15],[3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#12</td>
<td>All taxonomies</td>
<td>2629</td>
<td>606</td>
<td>606[420],[105]</td>
<td>2[2],[0]</td>
<td>44[34],[10]</td>
</tr>
<tr>
<td></td>
<td>Saccharomyces cerevisiae s288c</td>
<td></td>
<td>24[16],[1]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#13</td>
<td>All taxonomies</td>
<td>2886</td>
<td>615</td>
<td>615[418],[131]</td>
<td>2[1],[1]</td>
<td>53[38],[15]</td>
</tr>
<tr>
<td></td>
<td>Saccharomyces cerevisiae s288c</td>
<td></td>
<td>26[14],[6]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#14</td>
<td>All taxonomies</td>
<td>2642</td>
<td>637</td>
<td>637[456],[117]</td>
<td>1[1],[0]</td>
<td>43[29],[11]</td>
</tr>
<tr>
<td></td>
<td>Saccharomyces cerevisiae s288c</td>
<td></td>
<td>25[18],[3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>#15</td>
<td>All taxonomies</td>
<td>2715</td>
<td>620</td>
<td>620[468],[85]</td>
<td>1[1],[0]</td>
<td>42[32],[10]</td>
</tr>
<tr>
<td></td>
<td>Saccharomyces cerevisiae s288c</td>
<td></td>
<td>20[16],[1]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.7 Conclusion

For many years, the advancement of therapeutic treatments have led physicians, pathologists and scientists to target the source; proteins, that cause the illness or disease. Thus, multiple avenues, from manufacturing, cost efficiency, drug safety and efficacy and of course drug (protein) production are important facets in this new wave of protein therapeutics. One hurdle, especially in the developing world, is drug delivery. In collaboration with CBI Integrated and Scalable Cyto Technology (InSCyT), I set out to validate the production organism for their on-site drug therapeutics micro-fluidic chamber apparatus. I then developed an analytical platform, which characterized all host cell proteins, in the M.I.T(Cambridge, MA) optimized Saccharomyces Cerevisiae Komagataella Phaffii Pastoris NRRL-11430 strain. This secretome was produced through tailored in-solution digestion protocols from two approaches of the Null (NRRL-11430) strain. The primary approach was to unveil all HCPs from the shake-flask (M.I.T.) optimized yeast secretion. This gave me an overall look of the secretome. Whereas, my secondary approach was to analyze eight anion exchange preFractionated yeast secretion from the bioreactor experiment (M.I.T.). Thus, giving me a more detailed look, with the entire Null secretion being allocated into parts, based on their mAU quantity. Upon, Q-TOF mass spectrometry and congruent protein identification through ProteinScape™ database searches, 74 host cell proteins were identified and therefore a part of the official classified secretome of S. Cerevisia NRRL-11430. By identifying these proteins, the InSCyT program can monitor HCP involvement which leads to the overall safety of on-site drug therapeutics in their platform. To develop drugs, concerns in any change in manufacturing process, from initial protein product in cells to secretion
of product can be monitored with the help of these identified HCPs in the NRRL-11430 S. Cerevisiae strain. With this monitoring, we can therefore meet the initial goal of having hGH and IFNa2b be the first gene induced protein products for this on-site protein therapeutic delivery platform.
3.8 References


3.9 Supplementary Evidence

3.9.1 Figure R:
1D SDS-PAGE of unF-Null NRRL 11430. Molecular Marker lane 1; lane 2 = 1:2 dilution; lane 3 = 1:5; lane 4 = 1:10 dilution.
3.9.2 Anion Exchange Gradient and Elution Chromatograph

[Graph showing the chromatogram with various peaks and volumes marked on the x-axis and absorbance on the y-axis.]
3.9.3 Table Q: R1 clean-up whole Null Chromatographs

<table>
<thead>
<tr>
<th>Time</th>
<th>Module</th>
<th>Action</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>Pumps</td>
<td>Toatal Pump A Flow</td>
<td>1.000</td>
</tr>
<tr>
<td>0.02</td>
<td>SPD-10Avp(Det.A)</td>
<td>Zero</td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td>Controller</td>
<td>Event</td>
<td>23</td>
</tr>
<tr>
<td>0.04</td>
<td>Pumps</td>
<td>Pump B Conc.</td>
<td>0.0</td>
</tr>
<tr>
<td>2.80</td>
<td>Pumps</td>
<td>Toatal Pump A Flow</td>
<td>1.000</td>
</tr>
<tr>
<td>2.81</td>
<td>Pumps</td>
<td>Toatal Pump A Flow</td>
<td>3.000</td>
</tr>
<tr>
<td>2.82</td>
<td>Pumps</td>
<td>Pump B Conc.</td>
<td>2.0</td>
</tr>
<tr>
<td>4.00</td>
<td>Pumps</td>
<td>Pump B Conc.</td>
<td>2.0</td>
</tr>
<tr>
<td>4.01</td>
<td>Pumps</td>
<td>Pump B Conc.</td>
<td>70.0</td>
</tr>
<tr>
<td>6.00</td>
<td>Pumps</td>
<td>Pump B Conc.</td>
<td>70.0</td>
</tr>
<tr>
<td>6.01</td>
<td>Pumps</td>
<td>Pump B Conc.</td>
<td>2.0</td>
</tr>
<tr>
<td>6.30</td>
<td>Pumps</td>
<td>Pump B Conc.</td>
<td>2.0</td>
</tr>
<tr>
<td>7.00</td>
<td>Controller</td>
<td>Stop</td>
<td></td>
</tr>
</tbody>
</table>
3.9.4 R1 clean-up whole Null Chromatographs
Runs: blnks 1,4, 5,7 stacked. All run before null 1 & 2
3.9.4 R1 clean-up whole Null Chromatographs

Runs: Null 1

3.9.4 R1 clean-up whole Null Chromatographs

Runs: Null 2
3.9.4 R1 clean-up whole Null Chromatographs
Runs: blk, null 1, null 2 stacked
3.9.5 R1 clean-up AEX Null Chromatographs
Runs: blk1, 2, 3
3.9.4 R1 clean-up AEX Null Chromatographs
Runs: WT 8.2

3.9.4 R1 clean-up AEX Null Chromatographs
 Runs: WT 9.2
3.9.4 R1 clean-up AEX Null Chromatographs
Runs: WT 10.2

3.9.4 R1 clean-up AEX Null Chromatographs
Runs: WT 11.2
3.9.4 R1 clean-up AEX Null Chromatographs

Runs: WT 12.2

---

3.9.4 R1 clean-up AEX Null Chromatographs

Runs: WT 13.2
3.9.4 R1 clean-up AEX Null Chromatographs
Runs: WT 14.2

3.9.4 R1 clean-up AEX Null Chromatographs
Runs: WT 15.2
CHAPTER 4

Chromosome 22 Proteo-Genomic Analysis For
Bioinformatics Contribution To The Chromosome Centric
Human Proteomic Project
4.1 Abstract

In the proteomic and genomic researching fields, bioinformatics is an area of science that combines experimental data and computational approaches. The goal of this chapter is to integrate proteomic and genomic data, using data from the transcriptome (RNAseq). The completion of the human genome in 2003 was just the first step to understanding the blueprint of a human. The next step is to define protein coding genes and their functionality and responsibility in diseases such as cancer and fatally rare syndromes. This is to be accomplished by mapping such genes or gene set on all chromosomes. Dr. William Hancock’s lab has provided a landmark view of such data for chromosome 17, thus I have set out to use similar bioinformatics approaches to provide new proteo-genomic information on chromosome 22. Chromosome 22 is a part of the three assigned chromosomes (10, 17, 22) to be characterized in United States, while other world country collaborations are assigned the others. The first step was to produce an informative heat map. This heat map allows one to see the expression of genes at different genomic levels. Followed by ideogram regional identification of gene families, with a conclusion relevance on prominently known oncogene, Neurofibromin 2 (NF2). To do this, we will characterize the major protein form for each protein coding gene, as well as an *Alternative Splice Variant (ASV)* and a *single nucleotide polymorphisms (SNPs)*. We will contribute a parts list of chromosome 22, which shows the depiction of its protein coding genes. This is in alignment with *Chromosome-centric Human Proteome Project (HPP)* and all work has been deposited into the CHPP online depository.
4.2 Introduction

4.2.1 Human Proteome Background

In 2003, the Human Genome Project (HGP) revealed there are about 21,000 protein coding genes with in humans\textsuperscript{[1, 2]}. With its inception of ideas to study the human genome in 1984, the project was officially created though National Institute of Health and U.S. Department of Energy in 1990. In development of the human genome by the HGP, the expected 100,000 or more genes decreased as researchers identified that some genes are non-protein coding, but do code for other functions within a system, such as regulatory RNA, and telomeres, introns and pseudogenes\textsuperscript{[3-5]}. Although the goal was addressed to map the DNA sequence for \textit{homo sapiens}, scientist realized that in order to fulfill the project potential, new technologies needed to be developed in order to map the human genome (exome) as well as isolate genes that cause and are effected by a disease\textsuperscript{[5-9]}. The early potential was through initial mapped genetic diseases such as Huntington’s disease (1983, gene mutation) and novel gene variation such as BRCA1 (Breast Cancer 1) in 1994\textsuperscript{[10, 11]} With discoveries such as these, the need to focus on genes, prior to protein isolation for genetic disorders, was evident. Out of the 24 chromosomes; 1-22 and sex chromosomes x and y, chromosome 22 was the first appointed and mapped chromosome. It was chosen due to being the second smallest autosome and acrocentric, with its protein coding genes on the long q-arm. Outside of its genetic ease of study, the next step was to address the disease affiliation with the chromosome. Chromosome 22 has been widely known for rare congenital anomaly disorders, which are due to the translocation of genetic material, the absolute deletion of genes, repetitive replication or even transfusion on this chromosome unveils the desire to be studied. To address its disease affiliation to the genes, the Human Proteome Organization developed the Chromosome-Centric Human Proteome Project (C-HPP), to characterize all
protein and genomic information about all chromosomes. This project was designed to address the C-HPP of Chromosome 22.

**4.2.2 Background On The Chromosome**

In this 21st century, the world is well aware of Dr. Francis Crick’s idea of the central dogma of molecular biology: DNA $\rightarrow$ RNA $\rightarrow$ Protein$^{12,13}$. Scientists have focused on all three sets of human functional product in DNA, RNA and proteins. We have seen that protein studies are quite predominant in that they are an easier tangible target. However, many diseases have been shown to be from genetic abnormalities, such as the case in clear cell Renal Cell Carcinoma, where a mutated VHL tumor suppressor gene cascading to hypoxia induced conditions for tumor survival$^{14,15}$. Thus, proving the desire to target the genes, which are located in DNA. Genes are a set of genomic Deoxyribonucleic Acid (DNA) molecules, which are located on a chromosome$^{16,17}$. As researchers are trying to fully characterize diseases based on genetics, they are truly representing the chromosome(s) that is causing affliction in the body. The ‘Genetics 101: DNA, Chromosome and Genes’ of Figure 1 (Kintalk; University of California San Francisco), shows the physical makeup of a chromosome, as our nucleic acids are tightly packed and wound around histones so that all 24 pairs of the molecule can fit into every cell nucleus$^{18,19}$. From the early discovery in the late 1800s, we have learned chromosomes contribute, to our gender, eye color, height and overall health$^{20,21}$. With such a characteristic for overall human functionality, the contribution of chromosome studies can benefit efforts to eradicate disease, as it allows one to target disease prevention, prognosis and treatment.
Figure 1: DNA → Genes → Chromosome
A pictorial schematic building a chromatid. Permission Requested.
4.2.3 The Chromosome-Centric Human Proteome Project Assignment

The HGP (Human Genome Project), lead to the formation of new organizations such as the Human Proteome Organization (HUPO) and the Chromosome-Centric Human Proteome Project (C-HPP). My project focuses on the contributions presented to all three organizations, and was fashioned due to the goal of C-HPP. Their goal has been established so that one day the entire human proteome will be mapped by the relating protein coded genes on each chromosome\textsuperscript{[22, 23]}. Combining both proteomic data with a genomic scaffold leads to informative knowledge of our complex biological system and furnish access to protein level data. With these sets of goals the focus for genomic aspects has gained acceptance and encouragement from clinical diagnosis to preventative dispositions from parent to child. In creating this “parts list” on each chromosome, there were discoveries in this proteo-genomic crossover, that lead to the concept of missing proteins\textsuperscript{[24]}. We will thus identify these missing proteins and characterize why there genomic or proteomic entry is not fully defined\textsuperscript{[25]}. The C-HPP was also designated to study oncogenesis on each chromosome, which leads to common diseases, such as breast, gastric and colorectal cancers. With a focal oncogene, researchers can align not only gene proliferation or decline on aligned chromosome, but also cross regional alignment. My proteo-genomic analysis of chromosome 22 addresses the involvement in the both normal and cancer (breast, gastric, colorectal) cell lines, describe Neurofibromin 2 oncogene association and exposes rare syndromes and gene families\textsuperscript{[26-29]}. 
4.2.4 Chromosome 22 background

Through the collaboration with the United States, Germany, England, China, Japan and France, chromosome 22 was the first chromosome to have its DNA genetic sequence mapped\(^6,30\). With its completion in 1999, it showed the world how much of an impact displaying genetic code can have on the complete functionality of the human species\(^{30}\). The proper placement of human DNA, is to have all 24 pairs of chromosomes; 1-22 and X and Y, one from each parent, inside the nucleus of each cell\(^{31}\). As cells divide and replicate, the chromosome are also replicated\(^{32}\). However, any type of aberration initially on the chromosome or formed during replication, transfers to an abnormal chromosome and will thus cause effects both miniscule and severe\(^{33,34}\). Many deformative physical or functional effects as well as intellectual or behavioral developments are found on chromosome 22\(^{35,36}\). It is linked to multiple rare diseases and syndromes\(^{37,38}\). All do in part to a malfunction of this acrocentric; all protein coding genes on the q-arm, chromosome\(^{30}\). Being that it is acrocentric, any manipulation on its protein coding long q-arm will cause these effects\(^{30}\). One such manipulation is seen with translocation genetic material to cause Cat Eye Syndrome (CES)\(^{39,40}\). As the cells multiply, there is an extra copy of chromosome 22 in the cell, making a total of three chromosome 22s\(^{18}\). This extra copy is not a complete copy of chromosome 22, but a small copy of short p-arm translocated onto the long q-arm. There is evidence of 3 Mb of duplicated DNA onto the q-arm near the center of the chromosome; the centromere. This translocated material causes Coloboma of the iris and anal atresia with fistula, heart defects, pits in front of ear to name a few\(^{18}\). Opposing to duplication of DNA, chromosome 22 causes other severe effects from depletion of genetic material. This is seen in DiGeorge Syndrome and Velocardiofacial Syndrome, where a region of DNA at 22q11.2(chromosome 22, section 1-band 1-subband 2), the same genes translocated in CES\(^{41,42}\).
Because of this deletion the effects are seen in heart disease and a cleft palate, Graves disease, kidney abnormalities and more\textsuperscript{[43]}. Other human anomalies from chromosome 22, include partial trisomy (extra piece in each cell), partial monosomy (missing segment in each cell), and a ring formation of the chromosome\textsuperscript{[44, 45]}. With a plethora of genetic contribution diseases, it has proven its need to be matched with the abnormal genetic material. That is why we focus on oncogene involvement of Neurofibromin 2 (syndrome), missing proteomic material and putting together a complete parts list of chromosome 22, to be a road map for future therapeutic targeting\textsuperscript{[46, 47]}. 
4.3 Materials and Methods

4.3.1 Experimental Flow Diagram
All bench work experiments for database entries were performed with outside sources of the C-HPP, HGP and HUPO organizations.

Figure 2: Flow Diagram of Experimental Procedure
4.5 Results

4.5.1 Genomic and Proteomic Bioinformatics Heat Map of Chromosome 22

To complete the C-HPP parts list for chromosome 22, I first identified all protein coding genes that contribute in human functionality. To do this, I located all genes and located potential gene obscurities by extracting characteristics from both proteomic and genomic data sets. I first assessed all 888 protein coded genes out of roughly 4,300 total genes from Mapviewer (http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?taxid=9606&chr=22). From this list we saw that the first coded gene on chromosome 22 is POTEH (POTE ankyrin domain family, member H) at 22q11.1 and the last at the telomere 22q13.33 is RABL2B (RAB, member of RAS oncogene family-like 2B). From other corroborated chromosome parts lists, the next step was to verify these proteins in the official protein coding gene database, neXtProt. It is the leader for ratified annotation of protein coding genes. As I combed through neXtProt for the 888 genes, a resulting 444 genes were observed and established as the official protein coding proteome for chromosome 22. In making this list, the most notable observation was that genes were found in the neXtProt list, but not Mapviewer, such as, zinc finger protein 73 (ZNF73). Both the Mapviewer and the neXtProt lists were characterized for specific proteomic and genomic data. In the Mapviewer list, protein evidence was defined by observance at protein level, transcript level or RNAseq level. While neXtProt used protein level, transcript level, homology and uncertain descriptions. To gain further protein and genomic descriptions of these genes they were cross referenced against proteome spectral count, Snyder mononuclear cell database and Global Proteome Machine Database. From these observances in the Mapviewer protein coding genes of chromosome 22, we noticed some genes having missing identification from either neXtProt, GPM or spectral count information. It is here, that we identified 59 ‘missing proteins’,
due to them having uncharacterized gene product. In Figure 3, is a detailed view of all ‘missing proteins’ on a color coded heat map affirming stages of observed to not observed. After determining what genes had missing information, these genes were again cross referenced against normal cells, cancer cells and the compilation of other evidence identification including tissue observance and western blot identification. In choosing the cancer cell lines of breast, colorectal and gastric, they are linked in that they all are ERRB2(+) a well known oncogene, expressing cell lines. By doing a comparison of cancer expressing cell lines and normal cells, it leads to future experimentation of gene expression causing a cascade of regulation effects, to knock out gene of expression. This has made a significant contribution to the C-HPP, in that we now have targeted genes for experimentation so that these ‘missing’ genes can fully be identified.
**Figure 3A: Genomic – Proteomic Crossover Bioinformatics ‘Heat Map’ of Uncharacterized Gene Product; missing proteins**

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Protein Evidence</th>
<th>RPKM-SKBR3</th>
<th>RPKM-A431</th>
<th>RPKM-SNU16</th>
<th>RPKM-KATOIII</th>
<th>RPKM-LIM1899</th>
<th>RPKM-LIM1215</th>
<th>RPKM-LIM2405</th>
<th>HPA Evidence</th>
<th>rna_tissue_evidence</th>
<th>log(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>XKR3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CECR1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>UFD1L</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>MED15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>AIFM3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>POM121L7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>YPEL1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>POM121L1P</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>RTDR1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>RGL4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>SLC2A11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>ADORA2A-AS1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>GGT1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>TMEM211</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>PITPNB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>RHBD2D3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>EWSR1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>HORMAD2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>LIF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>RNF215</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>SEC14L3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>GAL3ST1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>DUSP18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>SFI1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>DEPDC5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>RFLPL2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>RFLPL3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>RFLPL3-AS1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>BPIFC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>FBXO7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Protein Evidence</th>
<th>RPKM-SKBR3</th>
<th>RPKM-A431</th>
<th>RPKM-SNU16</th>
<th>RPKM-KATOIII</th>
<th>RPKM-LIM1899</th>
<th>RPKM-LIM1215</th>
<th>RPKM-LIM2405</th>
<th>HPA Evidence</th>
<th>rna_tissue_evidence</th>
<th>log(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>LARGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>ISX</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>APOL4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>TEX33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>IL2RB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>MFGNG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>LGALS2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>POLR2F</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>SLC16A8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>DMCI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>APOBEC3H</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>MGAT3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>TEF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>MEI1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>SHISA8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>WBP2NL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>SERHL2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>KIAA1644</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>PHF21B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>NUP50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>51</td>
<td>UPK3A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>FAM118A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>53</td>
<td>PPARA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>GRAMD4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>FAM19A5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56</td>
<td>IL17REL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>MOV10L.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>58</td>
<td>ADM2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>59</td>
<td>MAPK8IP2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 3B: Genomic – Proteomic Crossover Bioinformatics ‘Heat Map’ of Uncharacterized Gene Product; missing proteins

The uncharacterized gene products of chromosome 22. The gene names are listed by Ensembl and clarified by neXtprot. Protein Evidence is defined by neXtprot: (ftp://ftp.nextprot.org/pub/current_release/chr_repos/nextprot_chromosome_22.txt). The RPKM (reads per kilobase per million mapped reads) columns in the table are the RNAseq values of the following ERBB2 cancer expressing cell lines: SKBR3/SUM190 (breast cancer), KATO III/SNU16 (gastric cancer), LIM1889/ LIM1215/LIM 2405 (colorectal cancer), A431 (EGFR epithelial carcinoma) and H9 (stem cells). RNAseq evidence values were determined from transcriptomic data sets. HPA evidence was tabulated according to C-HPP development group stating it is calculated based on the manual execution of Western blot, tissue profiling (IHC) and subcellular location (IF). evidence_summary is a tabulated summary of protein evidence of each gene. rna_tissue_evidence is a division of the genes in reference to their RNA evidence. GPMDB evidence was derived from EC_GPMDB Human Proteome.

The colors coding used for the table are as follows: Green, RNAseq RPKM ≥ 15, Protein Evidence at Protein Level in neXtprot, GPMDB log(e) ≤ −10, HPA Evidence High, evidence summary High = protein level in Uniprot and high HPA, RNA tissue evidence = high third expression; Yellow, RNAseq RPKM 3~15, protein evidence at Transcript Level in neXtprot, GPMDB log(e) −5~−10; HPA Evidence Medium, evidence summary Medium = protein level in Uniprot or high HPA, RNA tissue evidence = medium third expression; Red, RNAseq RPKM 1~3, Protein Evidence Uncertain, GPMDB log(e) −1~−5, HPA Evidence Low, evidence summary = transcript/inferred from homology/predicted/uncertain Uniprot evidence and medium HPA, RNA tissue evidence = low third expression; Black, RNAseq RPKM < 1, Protein Evidence no information, GMPDB log(e) no information, HPA Evidence Very Low, evidence summary = transcript/inferred from homology/predicted/uncertain Uniprot evidence and very low or none HPA and/or High or Medium RNA evidence, RNA tissue evidence = not detected.
Figure 4: Color Code Legend of Genomic – Proteomic Crossover Bioinformatics

<table>
<thead>
<tr>
<th>Color Coding Legend</th>
<th>Protein evidence</th>
<th>UniProt evidence</th>
<th>HPA evidence</th>
<th>RNA tissue evidence</th>
<th>Evidence summary</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Protein level</td>
<td>Transcript level</td>
<td>Homology level</td>
<td>Uncertain NA</td>
<td>Uncertain NA</td>
</tr>
<tr>
<td>Protein evidence</td>
<td>Protein level</td>
<td>Transcript level</td>
<td>Homology level</td>
<td>Uncertain NA</td>
<td>Uncertain NA</td>
</tr>
<tr>
<td>UniProt evidence</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>Very Low</td>
<td>NA</td>
</tr>
<tr>
<td>HPA evidence</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>NA</td>
<td>Only RNA</td>
</tr>
<tr>
<td>RNA tissue evidence</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>NA</td>
<td>Only RNA</td>
</tr>
<tr>
<td>Evidence summary</td>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>Only RNA</td>
<td>Only RNA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Evidence summary</th>
<th>Protein evidence</th>
<th>UniProt evidence</th>
<th>HPA evidence</th>
<th>RNA tissue evidence</th>
<th>Evidence summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNAseq RPKM/FPKM</td>
<td>&gt;= 15</td>
<td>3~15</td>
<td>1~3</td>
<td>&lt; 1</td>
<td></td>
</tr>
<tr>
<td>Proteome Spectral count</td>
<td>&gt;= 10</td>
<td>3~10</td>
<td>1~3</td>
<td>NA</td>
<td></td>
</tr>
<tr>
<td>GPM log ( e )</td>
<td>&lt; -10</td>
<td>-5~10</td>
<td>-1~5</td>
<td>NA</td>
<td></td>
</tr>
</tbody>
</table>
4.5.2 Ideogram Of Gene Families And Syndromes Identification

After identifying the protein coding genes on chromosome 22, we looked deeper into the chromosomal data uncovering rare disease syndromes and gene families located on the chromosome. Figure 5 is an ideogram of chromosome 22 with regional locations of syndromes and gene families. I first set out to identify what gene families were associated with this chromosome. By looking into the neXtProt protein coding gene list, there were several outlined families through out this geneome including RIMS binding protein(q11.21) and Immunoglobulin lambda Chain C(q11.22) near the centromere. As well as those who are in varying regions like Ras-related protein Rab, which is located in both q11.22 and q13.1. And then there are the telemere quadrants which houses a family of DNA editing enzymes named APOBEC3. We too found syndromes such as the popular Cat eye syndrome(q11.1) and DiGeorge syndrome(q11.2). Looking at the figure, we see that there are disputed location of families and syndromes along the chromosome. The density of genes on the chromosome was mapped to illustrate what regions were densely populated compared to those that were sparse. The density was mapped by extracting the identified numbered genes per million bases interval as illustrated in Genecards(Figure E). The density was calculated as an average of 64 genes per interval. Those regions above the average were marked with a black bar underneath the interval on the ideogram. It gives researchers a perspective of either co-expression during proliferation or regression of one gene in the family, as well as do multiple families increase or decrease within a diseased or healthy patient. These principles also apply to the syndromes. By locating these conglomerates on chromosome 22, it gave researchers intuition on the location and unveils non-noted rare-syndromes and families.
Figure 5: Chromosome 22 Gene Families and Syndromes Ideogram

Ideogram of chromosome 22, showing selected gene family, rare syndrome and heavy gene density location. The protein families are shown in boxes with the matching color of the corresponding band location. Inside the box, denotes the shorthand name of the gene family, exact band location and comparison genes on chromosome 22 and the entire human genome. The comparisons are deciphered as follows: Genes in one band region / total genes in family on entire chromosome 22 / total genes in human genome; For IGLC and SEC14L3, 5 of 10 and 1 of 7 are pseudogenes in total human genome genes. The solid bars represent the regions with a gene density above the average on chromosome 22, namely 64 protein coding genes per Mb. (http://genecards.weizmann.ac.il/geneloc-bin/gene_densities.pl?chr=22&gc_id=GC22U900840)
4.5.3 NF2 Oncogene Focus to Assimilate Posttranslational Modifications of NF2 And Surrounding Genes

As the identification of protein coding genes on chromosome 22 were characterized in accordance with genomic (neXtProt, Mapviewer) and proteomic (GPMDB) databases, we set out to focus on its cancer genes, also known as oncogenes. Identifying gene assimilation was a purposeful start, to search for an oncogene. Thus all oncogenes were identified. A crude search for cancer affiliation by searching in Genecards.org, was performed. This identified cancers, such as leukemia, colon carcinomas and breast cancer have been identified with 24 genes found. The table below shows a list of the 24 identified genes.

<table>
<thead>
<tr>
<th>Band Location</th>
<th>22q11.21</th>
<th>22q11.23</th>
<th>22q12.1</th>
<th>22q12.2</th>
<th>22q12.3</th>
<th>22q13.1</th>
<th>22q13.2</th>
<th>22q13.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>BID&lt;sup&gt;C&lt;/sup&gt;</td>
<td>BCR&lt;sup&gt;W,C,S&lt;/sup&gt;</td>
<td>CHEK2&lt;sup&gt;C,S&lt;/sup&gt;</td>
<td>EWSR1&lt;sup&gt;C,S&lt;/sup&gt;</td>
<td>MYH9&lt;sup&gt;S&lt;/sup&gt;</td>
<td>MKL1&lt;sup&gt;5&lt;/sup&gt;</td>
<td>BIK&lt;sup&gt;C&lt;/sup&gt;</td>
<td>EP300&lt;sup&gt;S&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>CLTCL1&lt;sup&gt;S&lt;/sup&gt;</td>
<td>GSST1&lt;sup&gt;C&lt;/sup&gt;</td>
<td>MN1&lt;sup&gt;C,S&lt;/sup&gt;</td>
<td>LIP&lt;sup&gt;C&lt;/sup&gt;</td>
<td>TIMP3&lt;sup&gt;W&lt;/sup&gt;</td>
<td>PDGFB&lt;sup&gt;W,C,S&lt;/sup&gt;</td>
<td>CYP2D6&lt;sup&gt;W,C&lt;/sup&gt;</td>
<td>TYMP&lt;sup&gt;W&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td>COMT&lt;sup&gt;C&lt;/sup&gt;</td>
<td>MIF&lt;sup&gt;C&lt;/sup&gt;</td>
<td>NF2&lt;sup&gt;W,C,S&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRKL&lt;sup&gt;W,C&lt;/sup&gt;</td>
<td>MMP11&lt;sup&gt;C&lt;/sup&gt;</td>
<td>PATZ1&lt;sup&gt;5&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEPT5&lt;sup&gt;S&lt;/sup&gt;</td>
<td>SMARCB1&lt;sup&gt;C,S&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Sanger = s, Cancer Index = c, Waldman = w. Web sites that list oncogene information are: http://www.sanger.ac.uk/genetics/CGP/Census/, http://waldman.ucsf.edu/GENES/completechroms.html, and http://www.cancerindex.org/geneweb/genes_d.htm.

Within the 24 genes, we did further exploration, by assessing which genes are documented in novel cancer databases. These included Sanger ‘driver oncogenes’, Waldman and Cancer Index. All websites have been dedicated to revealing notable cancer genes. The most focus was on those who had accreditation in all three databases, resulting in BCR, PDGFB and NF2 as the upmost ‘driver’ oncogenes on chromosome 22. Also known as a gene, that drives and leads to cancer. However, only Neurofibromin 2; NF2 has direct physical association with chromosome 22. The others have translocation material with chromosome 9 and chromosome 17. With the
chosen focal oncogene, NF2, the related disease is neurofibromatosis type 2 (Merlins). These schwannomas; benign tumors allocated along the central nervous system greatly affect many with the genetic disorder. We started to study NF2 physical affiliation with other genes, to assess if there is a relationship with its expression. Figure 6 and Figure 7 show 10 genes located before and after NF2. In Figure 6, a similar ‘heat map’ to the protein coding genome was created. There is a comparison of gene expression in the ERBB2 positive cancer cell lines; SUM190 and SKBR3, to normal cells. Within this list we dictated the lowest (best) expectation value observed for that protein with the GPMDB log(e) values. This figure shows transcriptome versus proteome log(e) coverage in the cancerous and non-cancerous cell lines and affirmed proteome evidence about these genes. The overall goal was to express these genes with a more direct proteo-genome crossover in reference to the driver oncogene NF2. Only three genes are in the cancer index on Sanger oncogene database. Taking these same genes, the next step was to address what post translational modifications(PTMs) are associated on NF2 and its co-located genes. Figure 7 shows the effects of alternative spliced variants have on protein expression with differential PTMs. This also shows the databases Esembl and neXtProt show causing different variants will increase the band of PTMs per gene. However, not all variants have been identified, and thus delaying the dictated proliferation of PTM on a gene. We see that neXtProt identifies predominately more ASVs (alternative splice variants). Overall, this figure represents the association an oncogene has with its surrounding PTMs.
<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Protein Evidence</th>
<th>Transcriptome Avg.</th>
<th>Proteome Avg.</th>
<th>Transcriptome log(e)</th>
<th>Proteome log(e)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 RHBDD3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 EWSR1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 GAS2L1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 RASL10A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 AP1B1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 RFPL1-AS1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 RFPL1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 NEFH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 THOC5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 NIPSNAP1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 NF2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 CABP7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13 ZMAT5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 UQCR10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 ASCC2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 MTMR3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 HORMAD2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 LIF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 OSM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 GATSL3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 TBC1D10A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 6: 10 genes above and below focused oncogene Neurofibromin 2 (Merlin); NF2 Heat Map**

Pink Asterick = Sanger identified oncogenes. Purple circle = CancerIndex identified oncogene. Cancer = SKBR3 and SUM190 breast cancer cell lines with the average Fragments Per Kilobase of transcript per Million fragments mapped (FPKM) and average log(e) values. Normal = normal cells (cufflink cuffcom-tracking) FPKM values and log(e) Snyder mononuclear white blood cells. Protein Evidence are NexProt evidence levels.

The colors coding used in the table are as follows: Green, Protein Evidence at Protein Level in neXprot, Transcriptome Avg. ≥ 15, Proteome Avg. ≥ 10, Transcriptome ≥ 15, Proteome ≥10, GPM log(e) ≥ -10; Yellow, Protein Evidence at Transcript Level in neXprot, Transcriptome Avg. 3 ~ 15, Proteome Avg. 3 ~ 10, Transcriptome 3 ~ 15, Proteome 3 ~ 10, GPM log(e) -5 ~ -10; Red, Protein Evidence at Other Level in neXprot, Transcriptome Avg. 1 ~3, Proteome Avg. 1 ~ 3, Transcriptome 1 ~ 3, Proteome 1 ~ 3, GPM log(e) -1 ~ -5; Black Protein Evidence none in neXprot, Transcriptome Avg. < 1, Proteome Avg. NA, Transcriptome < 1, Proteome NA, GPM log(e) NA.
10 genes surrounding focal oncogene NF2. The following graphics are used to denote different types of information: Blue boxes have genes with PTM information; Red circles, number of alternative splice variants in Ensembl/neXtProt; Purple circles, number of single nucleotide variants in NeXtProt; Hexagon shape in yellow, Ser/Thr phosphorylation; Hexagon shape in dark blue, phosphorylation at Tyr; Pink Rectangular shape acetylation; Triangle, N-glycosylation; Gray square, O-glycosylation. The first number in each shape represents the number of PTMs for the major ASVs and a second number relates to the number of PTMs in secondary ASVs.
4.6 Discussion

4.6.1 Missing Genomic Evidence Used To Illustrate Missing Proteins

Careful examination of Figure 3, which highlights the uncharacterized gene products on Chromosome 22, identifies neXtProt as the most informative database. NeXtProt was chosen because it provides proteomic evidence for proteins, and is a compilation between the Swiss Institute of Bioinformation(SIB) and Geneva Bioinformatics(GeneBio). These manual annotation experiments, gave unique and greater leverage to other databases, in that it provides certainty for protein production. With a large availability of databases, a uniformity for selection should be set, thus, this is why the C-HPP set theirs to be neXtProt. It was noted that by doing an alignment with neXtProt genes with Mapviewer, there are twice as many in Mapviewer than neXtProt. We expected a change in number, since the manual annotation for protein, transcript or RNAseq evidence is based on experimentation. This leads to postulation that some gene product was at a lower abundance to the base level for a quantification or that there are co-expression of genes for the same product, thus eliminating the secondary gene and decreasing the total accountable genes. This ‘heat map’ of missing gene products was designed to show, genes, whose Global Proteome Machine DataBase expectation value, proteomic spectral count and neXtProt values were not present. This missing information played a large contribution for researchers to accurately confirm gene presence on the chromosome. The color coding system was a way to visualize gene or protein expression with the familiarity of a traffic light. Only a small percentage (13.28%) of the total neXtProt approved genes have missing information, thus giving hope that all genomic and proteomic information for chromosome 22 can be obtained. This in turn, will complete the genomic profile of protein coding genes and thus be shared to affiliated organizations such as C-HPP, HGP and all other scientific and medicinal fields.
4.6.2 Implications Of Locating Families And Syndromes On Chromosome 22

By building a schematic representation of chromosome 22 to display its condensed gene density, gene family location and congenital disease was done to illustrate those points and to give rise to other concerns of co-expression, co-location and general gene regulation. The ideogram itself was derived from locating the first and last protein coding genes, POTEH and RABL2B at positions 22q11.1 and 22q13.33. Careful analysis of both Mapviewer and neXtProt, showed a cluster of genes with similar nomenclature. This indicated to us, that the genes are a part of a family. It also explained another reason for a decrease in these two lists, could be due to each database not having the same indicated number of familiar members, as is the case for zinc finger protein family. It was noticed that more gene family members were present in neXtProt than in Mapviewer. Out of the original list acquired through both neXtProt and Mapviewer, several had to be removed, such as the POTEG family. This could have been classified with POTEH as well as the G2B2-associated binding protein family, member 4. It was also unveiled that our original gene family list, some only had as high as transcript level evidence or a combination of homology, transcript or protein evidence. We chose to only incorporate protein level evidence for the entire gene family, which left out members from the Apolipoprotein L family. By having set qualification used to depict the families, we were assuring to the C-HPP committee that this portion of the overall parts list was accurate in accordance to the manual observed evidence experiments, performed in neXtProt. Another criteria for ideogram location was that we wanted to express genes that were physically located by one another. Most displayed members that are thousands of kilobase to kilobase pairs next to each other. While others may skip to one-hundred thousand kilobase, as is the case for family RIMBP3. We chose to dictate this distant relationship for each family, by recollecting how many families are next to
one another, compared to how many total on chromosome 22 versus how many are in the entire human genome. This lead to a pathway for expression pathways to look at co-location of genes for chromosome 22. This entire principle was performed for the rare disease syndromes on the chromosome. One difference was that the notorious DiGeorge/Velocardiofacial syndrome was not present in neXtProt, but can be seen in Mapviewer. This absence is however, one we were expecting to see, due to the nature of the syndrome, which is also known as the deletion syndrome. These deleted 30-40 genes (3mb) are not seen at the original designated q11.2, instead the translocated bases are observed in Cat Eye Syndrome are seen in neXtProt and in Mapviewer. Lastly, it was important to show the essential gene displacement along the chromosome. The gene density, is dense as a whole but sparse on the length of chromosome 22. The dense areas do match up with the gene population chromosomal location in all databases. This was reassuring that this chromosome only makes up 1.6%–1.8%[^30] of the human genome. The completion of the pictorial description for novel gene landmarks on chromosome 22, gives the scientific community a visual tool to researching gene family contributions to overall human functionality and disease influences.
In completing this bioinformatics project, we chose to look at the influence of ERBB2 supporters, such as Neurofibromin 2 (NF2). Neurofibromin is a gene, that when mutated leads to Neurofibromatosis type 2. The gene regulates the production of the protein Merlin, which plays a role during embryonic development and can be found in nerve cells. This protein also plays a role in tumor suppression and we can see its relation to “driver” oncogene ERBB2 in Supplementary Figure A. With the mutation causing a shortening of the protein, there is no longer tumor suppression regulation, thus benign and malignant tumors in both the central and peripheral nervous system develop. The goal of this project was to document the physical protein production products and to depict NF2 location, which may be influenced by or influence other oncogenes. Looking in Figure 7 we see many varying alternative splice variants affecting PTMs of standard and isoform proteins. This documentation shows two key elements. One element is that not all PTMs have been accounted for in many of the databases. Esembl and neXtProt were used to determine both alternative spliced variants and single nucleotide variants and its isoforms(neXtProt). This identification was a tool in expressing how these DNA modifications change the resulting protein expression. And being that the goal is to isolate all protein products, it has been a major stride in addressing these C-HPP confirmed proteins. Being that NF2 is a well known oncogene, we set out to classify it as a ‘driver’ oncogene for Neurofibromin type 2 and other associated diseases. We conferred into three major oncogene databases, as a way to determine if there were any possible other oncogenes nearby NF2, that might cause co-expression or regulation patterns for diseases and overall health. Our figure 6 found only 3 ‘driver’ oncogenesis, including NF2. Future studies would need to be conducted, to
determine if these three genes do in fact, perform with or without one another. Having this knowledge, tells one that when studying diseases with chromosome 22, a good implication would be to also look for other oncogenes, which have been proven to be near NF2. Overall, having a focal concentration on an oncogene, gave us insight to the physical relationship and how variants play a role in gene proliferation.

4.7 Conclusion

This chapter focused on contributing to the global proteome and genome advancements for human gene functionality. We achieved this by constructing proteo-genomic characterized compilations for chromosome 22, as outlined by the Chromosome-Centric Human Proteome Project. The organization has dedicated the study of all 24 chromosomes to different countries. Here in the United States, Dr. Hancock’s laboratory has completed the parts list for chromosome 17 and we were thus tasked with the same goals for chromosome 22. To begin the study, we characterized all 444 protein coding genes on the q-arm acrocentric chromosome. This was completed by dictating its manual annotations out of four distinct levels; protein, transcript, homology or uncertain, from the neXtProt database. We were then able to extract the termed ‘missing protein’ based on their absence of proteomic log(e) values and dual protein evidence(neXtProt and spectral counts). Identifying these missing genes, was a milestone in that it gives researchers an opportunity to try and characterize the gene product, which moves forward the detailed human proteome. Our final approach was to analyze physical attributes of the chromosome, by illustrating both gene families, disease syndromes and Sanger gene list affiliations. The ideogram allows one to see the cluster of gene families that reside on the chromosome. It tells you the role it plays, by having a lot of regulatory abundant protein
contributions. While the syndromes, open a window for more studies on deletion and translocation of genes, which is very prominent on chromosome 22. Having a collection of cancer databases (Sanger, Cancer Index, Waldman), gave us the opportunity to gather serious cancer genes. This produced a focal oncogene NF2, which has been previously known for the congenital disorder of Schwannomas from Neurofibromatosis type 2. We found that two other oncogenes are near this ‘driver’ gene NF2 and one of them (EWSR1) is also a driver oncogene. Having these discoveries from this project has led to intent for future studies. This would include, possibly co-location and co-expression for both normal and cancer cells, as well as continued manual annotation for protein level expression so that all genomic databases match. Overall, the parts list for chromosome 22, gives the Human Genome Project and its subsequent parties; HUPO and C-HPP, proteo-genomic information on the human genome through bioinformatics research.
4.8 References


4.9 Supplementary Evidence

**Figure A:** Confidence view of STRING interaction network of NF2 and surrounding proteins. The red stars indicate focal oncogenes NF2 and ERBB2. The thicker the line, the stronger the proteins are associated with each other.
4.10 Acknowledgements
I would like to signify special acknowledgments to those who help me construct and close this project. Thank you, Dr. Suli Liu, Dr. Fangfei Yan, Fan Zhang and Jennifer Poirier.
APPENDIX i

Acetylome Protein Characterization Of Von Hippel-Lindau(+) And pBABE(-) Mammalian Cell Lines
i.1 Abstract

Acetylation of proteins has been implicated in playing a role in multiple diseases. Currently, we are characterizing the acetylome (all acetylated proteins) of the well know human congenital genetic disorder, Von Hippel-Lindau syndrome. Von Hippel-Lindau (VHL) is a disease where benign and malignant tumors as well as cysts form in different parts of the body. In collaboration with Dr. Othon Iliopoulos of Harvard Medical School, we used analytical methodologies to characterize the acetylome of VHL positive and negative cell lines. The analysis of the cell lines; VHL and pBABE, were performed based on the extracted SDS-PAGE gel bands by visual appearance of bound acetylation antibody (Lysine-K antibody). In order to complete a more thorough analysis, we would need to determine the similar, distinct and unique protein trends from the different cell line contributors. It would also be beneficial to perform an antibody binding assay on a high contributing protein as an approach to plausible biomarker candidacy. By deducing the acetylome of Von Hippel-Lindau syndrome, the analytical advancements of our optimized HPLC-MS/MS platform on this post translation modification can contribute to potential drug therapeutic targets for this genetic disorder.

i.2 Introduction

i.2.1 Von Hippel-Lindau Syndrome Background

Von Hippel-Lindau (VHL) syndrome has been the predecessor for many subsequent diseases, such as cancer, blindness and fatalities\textsuperscript{[1]}. With no cure, others have begun to investigate potential therapeutics through attempting to understand the genetics and proteomics of VHL\textsuperscript{[2, 3]}. Due to direct linkage to clear cell renal cell carcinoma and hemanglobastomas of the retina and central nervous system, VHL is useful in developing analytical characterizations\textsuperscript{[4, 5]}. The
resulting illnesses are due to the mutation of the \textit{VHL} gene on chromosome 3\textsuperscript{[6]}. The mutation of this tumor suppressor, causes a cascade of events leading to hypoxia by HIF (Hypoxia-Inducible-Factor), thus sustaining and promoting tumor growth\textsuperscript{[7]}. The proliferation of upregulated proteins due to mutated \textit{VHL} (\textit{mVHL}) may have altered molecular modification, thus causing abnormalities from VHL syndrome\textsuperscript{[8]}. One such modification is the addition of an acetyl group (-OCCH\textsubscript{3}) onto an amino acid such as lysine\textsuperscript{[9]}. It can be seen as a reversible process on the core histone tail and protein p53 regulation, both through DNA damage. Due to a contributor of cell signaling, acetylation is involved in regulation of proteins and molecular functions outside of DNA damage\textsuperscript{[10, 11]}. We aim to study the acetylation of proteins in VHL syndrome and its contributions for drug therapies.

\textbf{i.2.2 Gene Manipulation Vector pBABE}

The understanding of Von Hippel-Lindau syndrome by characterizing its acetylation protein profile allows a better understanding of the genetic and proteomic functions in the disease\textsuperscript{[12]}. To expose the acetylation profile, mammalian derived renal cancer cells were used for experimentation. The Von Hippel-Lindau positive (+) cell line proliferates proteins as a result of \textit{mVHL}(mutated \textit{VHL}) gene, while pBABE expresses proteins as a result of HIF from non-mutated \textit{VHL} gene. In pBABE, gene manipulation technologies allow for controlled gene expressions of this plasmid vector\textsuperscript{[13]}. By controlling the expression of Hypoxia-Inducible-Factor, we are thus expressing natural healthy cell filaments. This gene manipulation technology is a great advancement in modern medicine as it enables the study one or multiple genes and the resulting proteins\textsuperscript{[14]}. As we study acetylated proteins from Von Hippel-Lindau disease, we will be unveiling possible therapeutic studies to eradicate the disease\textsuperscript{[15]}. 

146
i.2.3 Acetylation Importance For Study In Science

Acetylation is a common post translational modification, with relevant functionality to cell signaling processes. It also modulates protein functionality within the cell such as apoptosis and energy metabolisms\(^{[16]}\). For over a decade, science has used the influence of proteomic studies to focus on post translational modifications (PTMs) such as glycosylation, methylation, ubiquitination and phosphorylation\(^{[17]}\). The characteristic attributes can help to understand protein’s biological function in normal and diseased *Homo sapiens*. The sequential understanding of PTMs is that sectioning and collecting patterns, trends and relevance to systemic functionality, will allow science to alleviate obstacles for discerned drug targets\(^{[18]}\).

Choosing a modification which participates to cell signaling, gives the advantage of accessing disease targets on both the proteomic and genomic molecular levels\(^{[19, 20]}\). Acetylation is a good target for characterization studies in that defined modulation changes in primary protein structure become prevalent and accessible targets for complete understanding\(^{[21]}\). With Von-Hippel Lindau syndrome, it is relevant to determine the cause of the mutation on the *VHL* gene as well as the subsequent protein products. Choosing to study the acetylation profile is addressed from the well known acetylation and deacetylation on DNA histones to control gene expression\(^{[22]}\). In VHL syndrome we also seen the formation of hemangioblastomas; tumors made from newly formed blood vessels, found in the brain and spinal cord. One way to study these tumor masses is to understand the tumor developments, which research has shown to have association with acetylated proteins. Such is the case of defected cilium microtubules to produce renal cysts\(^{[23]}\). From this knowledge we can therefore use the acetylation profile to understand the impact these proteins have in Von-Hippel Lindau syndrome.
i.2.4 Review Of Analytical Testing Used To Study Diseases

Common analytical scientific applications used on non-PTM modified proteins are also applied to post translational modified proteins\(^{[24,25]}\). Using an integrated linear platform of protein separation followed by HPLC-MS/MS Q-TOF mass analyzation attunes to PTM proteins, while adding an acetyl search criteria to identified acetylated proteins\(^{[26]}\). Thus, the acetylome for Von Hippel-Lindau syndrome is disclosed. The mass analyzer for acetylated protein identification is an important element for this study\(^{[27]}\). All tandem mass spectrometers can locate site specific acetylation if they have high resolving power can to distinguish the +42.0106 Da mass shift on the ε-amino group on Lys or free amino group on the N-terminus protein, from the trimethylated (+42.0470 Da) of Lys, thus being termed to have mass accuracy, \(^{[28]}\). In choosing mass spectrometry methods for PTM identification we are also creating an analytically reproducible profile for this identification. Thus, allowing this application to be applied in acetylated targeted drug products. By having the idea to modulate disease through its PTMs, we are seeing all possible molecular and biological functions with Von Hippel-Lindau syndrome.

i.3 Materials And Methods

i.3.1 Materials

All SDS-PAGE gel materials, solutions and experimentation purchased and performed with Dr. Othon Iliopoulos of Harvard Medical School (Charlestown, MA). Subsequent trypsin in-gel digestion enzyme was purchased from Promega (Madison, WI). All other reagents and buffers including those used in HPLC-MS analysis were purchased from Thermo Fisher Scientific (Waltham, MA).
i.3.2 Methods

i.3.2.1 VHL(+) and pBABE(-) gel extraction sample collection

Collaborators performed SDS gel electrophoresis with 30 µg and 80 µg VHL(+) and pBABE(-) cell lysate material. A parallel gel of 30 µg cell lysate was executed and transferred to a PVDF membrane and incubated with K-acetylation antibody. This was then developed to visualize antibody attachment to proteins as a guide to extract non-antibody incubated experiments. Twenty-six gel band were extracted into 1.5 mL eppendorf® tubes, logged as seen in Supplementary Table 1 and placed into -20°C refrigeration until needed for experimentation.

i.3.2.2 In-Gel Digestion

Enzymatic In-Gel Digestion(InGD) was performed as described below. 1 mL of freshly prepared 50 mM Ammonium Bicarbonate(ABC) at pH.9 was added to the samples. Prior to this addition samples were briefly washed and vortex on a short spin cycle with 1 mL Acetonitrile(ACN) to remove any unnecessary contaminants from the pre-sliced extracted gel pieces. After removal of the 50 mM ABC by vortexing and aspiration, another 2.5 cycles of ACN and ABC were executed for thorough dryness and hydration, which exposes proteins for maximum enzymatic digestion. After the final ACN aspiration another vortex and aspiration technique was applied to the samples to remove all excess fluids. The samples were transferred to an ice chamber apparatus for the addition of the designated enzyme; Trypsin. A 1:50 protein:enzyme ratio of sequence grade Trypsin(Promega, Madison, WI) was added to the samples. This resulted in 20 µg of Trypsin diluted into 100 µL of trypsin reconstitution buffer(Promega). 10 µL of this fluid was diluted in 500 µL of 50 mM ABC, to give a final concentration of .0004 µg/µL. A total of 20 ng of Trypsin was added into each eppendorf® tube, with additional 50 mM ABC to cover the gel pieces. Samples were then imbibed over ice for 1hr.
and then placed on a heated incubator at 37°C overnight. After incubation, the peptides were extorted from the gel pieces through a series of Acetonitrile/Formic Acid(Formic Acid) solvent washes. Once the initial Trypsin-peptide solution was aspirated and transferred into a new sample tube, 200 µL of 30% v/v was added to the gel pieces, gently vortex and collected into the new sample tube. This same procedure was adapted when using both 60% v/v and 99% v/v ACN/F.A. solutions. The gel pieces of the samples were discarded and the peptide extracts were lyophilized to dryness, before reconstitution with 20 µL of 1% formic acid (F.A.). These reconstituted samples were placed for centrifugation for 15 mins at 10,000 r.p.m(eppendorf 5424R, Hauppauge, NY). 10 µL of sample was placed in mass spectrometry vials(Thermo Scientific, Waltham, MA) for MS and Q-TOF MS/MS analysis.

**i.3.2.3 LC-MS/MS Analysis And Protein I.D.**

For liquid chromatography and mass spectrometry linear analysis, the tryptically digested peptides followed the described procedure. Peptides were separated by Ultra Performance Liquid Chromatography(UPLC, Dionex, California, USA) by loading 2 µL of sample onto the 5 µm, 50 cm Acclaim® PepMap RSLC column (Thermo Fisher). The column was equilibrated before and after each sample with a blank under the same gradient conditions as the sample. The 300 minute gradient went as follows. 5 % mobile phase B (0.1% Formic Acid/ Acetonitrile) at time points 0-5 mins, steadily increase from 5% buffer B to 40% buffer ending at 245 mins, stationary 40% mobile phase B from 245 mins to 248 mins, increase to 80% solvent B at time points 248 mins to 262 mins, decrease to 5% mobile phase B until final end point at 300 mins. For mass detection samples were analyzed by a Quadrupole Time-of-Flight(Q-TOF, Bruker, Billerica, MA). It was set at an intensity-dependent rate of 4-20Hz between precursor
scans (Northeastern University, Dr. Jared Auclair). The resulting MS/MS spectra were characterized by ProteinScape (Bruker), set to search for modification oxidation, deamidation and the PTM characterization product acetylation. As well as two missed cleavages; precursor ion mass tolerance of 2.0 Da; fragment ion mass tolerance 1.0Da; full trypsin as enzyme; carboxyamidomethylation(C) as a fixed modification. False Discovery Rate (FDR) was targeted at 1% to give one the best confidence score for peptide matching.
\textit{i.4 Experimental Procedure}

\textit{i.4.1 Flow Diagram}

1. Preliminary Collaborator Experimentation
2. SDS-PAGE
3. Extracted gel bands
4. inGel Trypsin Digestion
5. HPLC QTOF-MS/MS
6. ProteinScape i.d.


**i.5 Discussion and Results**

**i.5.1 Protein Load For Von Hippel-Lindau Acetylome Analysis**

Our collaborators at Massachusetts General Hospital pre-determined cellular contents of 30µg and 80µg analyzed through SDS-PAGE experimentation. Figure 1, shows both the acetyl-k antibody chemioluminescence PVDF membrane and an SDS-PAGE without antibody acetyl-k addition. It was important to visualize and assist in quantity determination of these VHL acetylated proteins. The three sets of experiments from the microgram load were utilized to expressing the full acetylome in this disease. This showed the biological reproducibility from the samples. By having reproducibility from the sample set we can show molecular function, biological relevance and protein regulation trends, thus provide a characterized acetylome profile of mVHL(mutated VHL) in Von Hippel Lindau syndrome.

![Image of SDS-PAGE gel](image)

**Figure 1:** *A*Acetyl-K antibody applied to PVF membrane of 80 µg VHL and pBABE cell lysate transferred from an SDS-PAGE Gel. *B*SDS-PAGE Gel; 30 µg of VHL(+) and pBABE(-) cell lysate.
i.5.2 Acetylated Protein Sequencing of VHL(+) And pBABE (-)

A brief overview of the VHL(+) and pBABE(-) datasets show that each cell line produces a large quantity of acetylated proteins. As described previously, the extracted sections a, b, c were three symmetrical cuts within the protein stained SDS-PAGE gel, these identified proteins were from an 80µg VHL(+) and pBABE(-) load quantity. Looking at the two cell lines, we can see a trend of section b, increasing in both runs 1 and 2. From this assessment, we looked into Figure1 and noticed that the Lysine-antibody stained SDS-PAGE gel does indeed have a dark hue in the VHL lane in the center of the gel compared to pBABE. This shows that there are more acetylated proteins in the center section of VHL between 75 kDa and 50 kDa. Going through the list of proteins we see that there are similar identifications. This includes Keratin Type 2 Cytoskeletal 1 & 2, Heat Shock Protein 90 & 70A and Vimentin, while others are unique to VHL(+) such as Ribosome biogenesis methyltransferase WBSCR22, Reelin and Protein RIC-3. A continuation approach is to have a deeper comprehensive look into this set and other sets of extracted bands would reveal all comparable and unique proteins in VHL(+) and pBABE(-) cell lines. With these details, the researcher would get a full understanding of acetylation profile providing plausible therapeutic targets to cure or assist those living with the effects of Von Hippel-Lindau syndrome.

i.5.3 Mass Spectrometer Contributing Analysis

With the application of our mass spectrometry analytical platform, we were able to use Quadrupole Time-of-flight for mass detection. The crucial point in this project was the distinction of mass accuracy for the PTM acetylation. With a +42.0106 Da mass shift, high resolving power was a necessity. We need an instrument that would also have high
reproducibility of the technical and biological replication of the samples. The preliminary 18 samples show great reproducibility by examining the chromatographic spectra shown in Figure 2. In Figure 2 we see great separation, even in the low intensity peptides at 100 minutes to 300 minutes time range. By looking at Figure 3 we see the retention time shift on the stacked chromatographs. This is telling us that molecular weight is being added or cleaved as seen in extracted ion chromatograms of the VHL syndrome proteome. This gives a surface meaning that has naturally occurring acetylated proteins as opposed to degenerative mVHL gene producing acetylated PTM proteins. A further in depth analysis of tryptic peptides in the MS²(MS/MS) spectra would pin point the exact location of the acetyl group addition or removal on the peptide. This can be seen with preliminary data analysis of this acetylome. Figure 4A shows the MS/MS contributing spectra of a protein named Signal Transducer And Activator Of Transcription 1-Alpha/Beta from pBABE A and VHL A samples. The naturally occurring acetylation found in pBABE; the VHL negative cell line, shows the N-terminal end of the protein with acetylated Lysines at positions 4, 12 and 20. In contrast, Figure 4B; VHL positive cell line, shows the SDQ⁸KQEQLLLKKMYLMNDNK peptide with the deacetylation on the Lysines. To complete the acetylome of Von Hippel-Lindau the complete sample set would be characterized to identify all modification sites and reveal a collection of proteins by MS and MS² analysis.
Figure 2: Q-TOF base peak chromatographs of VHL and pBABE.
Figure 3: Q-TOF stacked base peak chromatographs of VHL and pBABE.
**Figure 4**: MS/MS Chromatograph of Signal Transducer And Activator Of Transcription 1-alpha/Beta Protein.

**Figure 4A**: pBabe A

**Figure 4B**: VHL A
i.6 Conclusion

The symptoms of Von Hippel-Lindau syndrome are a result of the formation of benign and malignant tumors in the body. Those associated with the central nervous system or kidneys have gained much attention, as seen in Renal Cell Clear Cell Carcinoma research studies. Thus, a detailed characterization of this disease was necessary. This project has successfully revealed acetylation proteins of Von Hippel-Lindau. We have begun to explore this acetylome, but future researchers will be needed to complete the project. This includes completing the allotted sample set for a firm robust analyzation of tryptically digested peptides. To reveal the acetylation trends, patterns and regulations in the VHL(+) and pBABE(-) cell lines, a bioinformatician would be advantageous in developing a program to extract acetylated proteins from the large dataset comprising of about 60,000 or more proteins. We have shown preliminary results of studying six samples and illustrate the linear UPLC-Q-TOF-MS/MS analytical platform shows analytical reproducibility and productivity in this acetylome. Overall, having an acetylome of VHL disease will greatly impact science in the development therapeutic drugs in eradication of the disease, as well as other targeted post translational modification analysis.
References


**Table 1:** Log of Acetylated Protein Dissected Gel Pieces.

<table>
<thead>
<tr>
<th>Number</th>
<th>Top of tube label: 2 cell lines: pBABE(PB) &amp; VHL</th>
<th>Gel Extraction Target</th>
<th>Experiment date</th>
<th>µG Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PB A</td>
<td></td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>2</td>
<td>PB B</td>
<td></td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>PB C</td>
<td></td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>4</td>
<td>VHL A</td>
<td></td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>VHL B</td>
<td></td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>6</td>
<td>VHL C</td>
<td></td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>7</td>
<td>PB A 50KDA</td>
<td>exact arnd 50kda band</td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>8</td>
<td>PB B 50KDA</td>
<td>exact arnd 50kda band</td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>9</td>
<td>PB 75KDA</td>
<td>exact arnd 75kda band</td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>10</td>
<td>PB 250KDA</td>
<td>exact arnd 250kda band</td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>11</td>
<td>VHL A 50KDA</td>
<td>exact arnd 50kda band, piece #1</td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>12</td>
<td>VHL B 50KDA</td>
<td>exact arnd 50kda band, piece #2</td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>13</td>
<td>VHL 75KDA</td>
<td>exact arnd 75kda band</td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>14</td>
<td>VHL 250KDA</td>
<td>exact arnd 250kda band</td>
<td>5/3/13</td>
<td>80</td>
</tr>
<tr>
<td>15</td>
<td>PBA C</td>
<td>everything less than 50kda</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>16</td>
<td>PBA A</td>
<td>100kda-150kda</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>17</td>
<td>PBA B</td>
<td>75kda-100kda &amp; 50kda-75kda</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>18</td>
<td>PBA 75KDA</td>
<td>exact arnd 75kda band</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>19</td>
<td>PBA 250KDA</td>
<td>exact arnd 250kda band</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>20</td>
<td>PBA A 50KDA</td>
<td>exact arnd 50kda band</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>21</td>
<td>VHL A 50KDA</td>
<td>exact arnd 50kda band</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>22</td>
<td>VHL C</td>
<td>everything less than 50kda</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>23</td>
<td>VHL A</td>
<td>100kda-150kda</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>24</td>
<td>VHL B</td>
<td>75kda-100kda &amp; 75kda-100kda</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>25</td>
<td>VHL 75KDA</td>
<td>exact arnd 75kda band</td>
<td>5/7/13</td>
<td>30</td>
</tr>
<tr>
<td>26</td>
<td>VHL 250KDA</td>
<td>exact arnd 250kda band</td>
<td>5/7/13</td>
<td>30</td>
</tr>
</tbody>
</table>
i.9 Acknowledgements
I would give a special acknowledgement to Dr. Paulo Gameiro who coordinated with me on accessing the samples and questions regarding preliminary experimentation outside of the Hancock laboratory.
CHAPTER 5

CONCLUSION SUMMARY
In this dissertation I have provided a detailed synopsis of the analytical approaches that may impact science for the progression of therapeutic medicine. We show the strides of procured knowledge to develop optimized treatment of many diseases and illnesses. To that end, I demonstrated the application of our in-house developed analytical platform. This platform consisted of a reverse phase nanoliquid chromatography system. It localized target peptides on the protein carbonic anhydrase xii(CAXii). This low abundant protein is being investigated as a biomarker, due to its regulation pattern in clear cell renal cell carcinoma. Clear Cell Renal Cell Carcinoma; commonly known as ccRCC, is a distinct kidney cancer in that there are transparent (“clear cell”) tumors formed in the kidney and have potential to metastasize. The three hydrophobic peptides that we discovered have been proposed to be studied as biomarkers in targeting CAXii in ccRCC. Being in the top ten cancers in the United States, it is still imperative to focus efforts on treating those with all forms of kidney cancer. This can be achieved by producing biologics to affect the body on a genomic and proteomic level including regulated proteins CAXii, CAiX and others in the carbonic anhydrase family. One can also address the mutated $VHL$ gene found on chromosome 3, thus repairing the damage from the malformed DNA. These efforts and implications hypothesized from others can contribute to treatment of those affected by kidney cancer and in turn be applied to other diseases. We also showed the reverse phase analytical platform applied in characterizing the CAXii was also applied for the study of host cell proteins of yeast. Yeast has been used as the model organism for drug therapeutics due to its ability to be genetically manipulated to tailor many biological needs, from simple oligosaccharides to full systematic protein production. As such, working with product development manufacturing has allowed for the Integrated and Scalable Cyto-Technology(InSCyT) program to create on-site protein therapeutics. The initial target
therapeutics, human growth hormone (hGH) and interferon-alpha 2b (IFNα2b) are to be used on the microfluidic bioreactor chamber to give to those who do not have direct access to standard medical facilities. This program has the most beneficial attribute, it allows for those in isolated parts in the world to have the drug produced and distributed directly to them. That is why it was important to characterize the NRRL-11430 Komagataella phaffii pastoris Null yeast strain, to shows its versatility for future drug therapies. The k.phaffii seventy-four host cell proteins were chosen as yeast regulators and thus maybe affected in product gene insertion. Comparative studies of fractionated and unfractionated samples were performed as well as study the relevance from product gene proteins produced. Overall, this scientific application of engineering and analytical science has increased hopes in affordable and accessible treatments for all. To conclude this document, a genomic approach has been brought forward for biotherapeutics treatments. In this thesis, I was aligned with the C-HPP goals to fully characterize chromosome 22 in the human proteome. In this I discovered all protein coding genes, missing proteins and labeled the prevalent syndromes and disorders associated with the chromosome. This was a monumental step for future genomic treatments. With the addressing of disorders and protein coding genes, one can therefore know which genes are responsible for disease afflictions in the body and thus create tailored therapeutics for patients. By revealing missing proteins we are aiming to search for their missing genomic information, through the continually expanding bioinformatics databases. The database has been filled with the proteo-genomic information generated in this thesis. The expansion of the databases helps scientists fill in the gaps of proteomic, genomic and transcriptomic data, which can be used in many scientific studies. In addition to these chapters, an appendix was brought forth to incorporate the preliminary findings of acetylated proteins in Von Hippel-Lindau syndrome. This syndrome causes the benign and
malignant tumor development in varying places of the body, including the central nervous system. More detailed data extraction is necessary for full protein regulation trends in the VHL positive and pBABE(VHL negative) cell lines. With acetylation being a highly attributed post translation modification(PTMs), understanding its proteomic function outside of cell signaling would increase the proteome profile of PTMs. We optimized our analytics to fully drive out all acetyl (-OCCH₃) peptides, by using high resolving instrumentation allowing for differentiation of other chemically or translationally modified peptides. A full profile of the acetylome in VHL syndrome would aid in the treatment for other diseases such as ccRCC, which is due to mVHL. All in all, this thesis was used to show results from analytically driven projects and to give more information in development of therapeutic products.