LC-MS determination of glycosylation pattern on glycoproteins as critical quality attribute for biopharmaceuticals and potential markers for diseases

by Anna (Fan) Zhang
B.Sc in Biological medicine, Peking University

A dissertation submitted to

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

May 15, 2015

Dissertation directed by
William S. Hancock
Professor of Chemistry and Chemical Biology
Acknowledgements

It is my great honor to thank all of the people who helped, supported and encouraged me during my PhD life.

My foremost acknowledgement is for my advisor, Professor Bill Hancock for accepting me as his student and giving me several interesting projects with great trust. He worked with me patiently and nicely, offering me plenty of space, support, encouragement in my pursuit of career goal. The words “focus on science, well trained yourself” he said to me will significantly impact my career. I am very grateful to him for his mentoring during my doctor studies.

I also want to express my deepest gratitude to Professor Shiaw-lin Wu, for his patient supervision and nice instruction for my research. He worked with me closely on the data analysis and gave me a lot of valuable suggestions, not only in my career, but also in my life.

To my dissertation committee members: Professor Jeffrey Agar, Professor George O'Doherty and Professor Jordan Kreidberg. I truly thank them for reviewing my thesis and bringing up useful suggestions.

I would like to thank Professor Barry Karger for providing an excellent research environment at the Barnett Institute. I would like to acknowledge all the members of the Barnett Institute with whom I had opportunities to work with. In Professor Hancock’s lab, I appreciate the help of Dr. Qiaozhen Lu, Dr. Janet Zeng, Dr. Yi Wang, Dr. Suli Liu, Dr. Yue Zhang, Dr. Fangfei Yan, Dr. Francisca Gbormittah, KyOnese, Yu Wang, Yanjun Liu and Di Wu. In Karger’s lab, I appreciate the support of Dr. Wenqin Ni, Dr. Chen Li, Dr. Zhenke Liu, Dr. Siyang Li, Jason Wang, Siyuan Liu, Yuanwei Gao.
Thanks are extended to my collaborators: Professor Jordan Kreidberg, Dr. Yunjoon Jung, and Dr. Shan Qin from Boston Children’s Hospital; Professor J. Christopher Love and his labs from Massachusetts Institute of Technology; and Dr. Erno Pungor from BioMarin Pharmaceutical Inc. They are generous to provide us samples, share data and give useful discussion.

I also want to acknowledge the help from the staff in the department of chemistry and chemical biology and Barnett institute: Nancy Carbone, Andrew Bean, Cara Shockley and Richard Pumphrey.

Finally and most importantly, I wish to dedicate this dissertation to my family members: my husband, parents and parents-in-law. I appreciate their selfless support and generous understanding to me during my Ph.D. study.
Abstract

Protein glycosylation is known to be one of the most complicated and biologically significant post-translational modifications and glycoproteins play critical roles in both the pharmaceutical market and disease biomarker discovery. For glycoproteins as biopharmaceuticals, glycosylation influences the efficacy of protein drugs through the biological activity, stability, solubility, serum half-life and immunogenicity and for glycoproteins as potential disease biomarkers, the glycan-based alteration can serve as biomarkers for detecting disease in the early stage, monitoring disease progression, measuring the therapy response and observing recurrence. At present, liquid chromatography coupled with mass spectrometry (LC-MS) has been demonstrated to be a valuable technology in the characterization of protein glycosylation, from either glycopeptide approach or glycan approach. This thesis focus on the methods for the LC-MS determination of glycosylation pattern on glycoproteins as critical quality attribute for biopharmaceuticals and potential markers for disease.

Chapter 1 gives an overview of glycosylation pattern as critical quality attributes for biopharmaceuticals and potential disease markers. Basic concepts of oligosaccharides, biosynthesis pathways, and types of glycosylation are introduced to provide the background knowledge. In addition, the influence of glycan patterns on the efficacy of therapeutic glycoproteins and significant function of glycoproteins in disease biomarker discovery are explained to establish the importance of glycoprotein characterization. Finally, LC-MS based techniques for determination of protein glycosylation are summarized. Strategies for glycopeptide and glycan analysis are reviewed in details, respectively.
In chapter 2, we adopted two different LC-MS approaches to characterize the glycosylation patterns of biosimilar I-TNK and reference product G-TNK. Glycan heterogeneity is extensively analyzed at all three N-linked glycosylation sites (Asn-103, Asn-184 and Asn-448) and specifically glycosylation site Asn-103, generated from two mutations, only contains the desired complex-type glycans instead of original high-mannose glycans. In addition, biosimilar I-TNK and reference product G-TNK showed different distributions on sialic acid-containing glycans. The quantitative comparison of major sialylated glycans was performed by our two independent approaches and consistent results were obtained. Finally, mono-sialylated positional isomers were also identified and quantitated for two products by these two approaches in a complementary manner. Two isomers can be completely resolved by reversed phase chromatography and then relatively quantitated based on peak area in the glycopeptide analysis. On the other hand, the D ions generated in MS negative-ion mode in the glycan analysis could facilitate the determination of glycan positional isomers.

In Chapter 3, we have utilized a powerful HPLC-mass spectrometric approach to characterize the α3 integrin from mouse kidney cells with the N-linked glycosylation on α3 integrin light chain. We demonstrated that larger and different glycan structures were observed at Asn-925 and Asn-928 sites in Pkd1+/+ cells (normal cells), compared to Pkd1+/− cells (cells with polycystic kidney disease), which can explain the changes in mobility of α3 integrin light chain on SDS-PAGE electrophoresis. The unusual di-sialic acid glycan structures were observed solely in Pkd1+/− cells and this observation could also indicate the presence of altered glycosyltransferase expression in the disease state. We assumed that the study of glycan structures and resulting insights into biosynthetic mechanisms has the potential of increasing our understanding of molecular mechanisms involved in PKD, and thus the resulting knowledge of which glycotransferase are
involved in the disease process can be used to design potential therapeutic agents. The glycan structure identified specifically at each individual site with associated unique glycan structures can also enable us to develop potential disease glycoprotein markers derived from site-specific mass spectrometric approaches.

In chapter 4, we have characterized the recombinant human growth hormone produced from *Pichia pastoris* by MIT. The sequences with N-terminal peptide variants as well as post-translational modifications (i.e. oxidation and deamidation) were analyzed by the state-of-the-art LC-MS approach. The extent of N-terminal end sequence variants as well as oxidation and deamidation were measured by the ratio of peak areas of the non-modified peptide vs. the sum of peak area of the non-modified and modified peptide in the same LC-MS analysis. We successfully identified two additional N-terminal variants with amino acids from leader sequence attached to N-terminus in MIT produced rhGH. Also, in the relative comparison with Genentech produced rhGH, a higher amount of oxidation but an equivalent amount of deamidation was found in MIT produced rhGH. The concentration of MIT produced rhGH was also measured by external standard Genentech rhGH with its calibration curve. The technique we applied can also confidently and sensitively pinpoint the subtle but distinct difference due to a different manufacturing process and our analytical results provide critical and immediate feedbacks to the upstream manufacture groups and instruct them to improve the fermentation process.

In the supplemental section of chapter 4, we have characterized the recombinant Adeno-Associated Virus protein (type 5) by our HPLC-mass spectrometric approach, the protein sequence has been characterized and unexpected N-terminal processing was discovered and identified with N-terminal methionine cleavage and serine acetylation as the major variant, as well as only N-terminal methionine cleavage for the minor variant. The extent of N-terminal
variants was relatively quantitated by the ratio of peak area of specific variant on the peak area of all related peptides. Also, to enhance the sequence coverage by single-enzyme digestion, the trypsin autolytic digestion was considered and matched against the database by no-enzyme search and two large peptides were identified by using this strategy. The analytical approach we developed in this study will provide an effective method to study the unexpected sequence variance during the production process and our analytical results can provide immediate and critical feedbacks to our collaborators, assisting them in optimizing the production process with correct processing of the N-terminal sequence.
# Table of Contents

Acknowledgement .............................................................................................................. ii

Abstract ......................................................................................................................... vi

Table of Contents ............................................................................................................. viii

List of Figures .................................................................................................................. xiv

List of Tables ................................................................................................................... xviii

List of Abbreviations ....................................................................................................... xix

Chapter 1: Overview of LC-MS determination of glycosylation pattern on glycoproteins as critical quality attribute for biopharmaceuticals and potential markers for diseases ............. 1

  1.1. Abstract ..................................................................................................................... 1

  1.2. Glycosylation patterns as critical quality attribute for biopharmaceuticals and potential disease markers ................................................................................................. 1

    1.2.1. Basic concepts of glycosylation ............................................................................. 1

    1.2.2. Biosynthesis of N-glycans and O-glycans in human cells .............................. 4

    1.2.3. Glycoproteins as biopharmaceuticals ................................................................. 9

        1.2.3.1. Effect of glycosylation in therapeutic protein efficacy ................................. 9

        1.2.3.2. Factors influencing glycosylation in manufacture process ....................... 13

    1.2.4. Glycoproteins in disease biomarker discovery ............................................... 17

  1.3. Determination of protein glycosylation by LC-MS based techniques ................. 20

    1.3.1. Strategy of glycopeptides analysis by LC-MS .................................................... 20

        1.3.1.1. Isolating glycoproteins .............................................................................. 20
1.3.1.2. Reduction and alkylation ................................................. 21 
1.3.1.3. Enzymatic strategies ....................................................... 22 
1.3.1.4. Separation methods ......................................................... 26 
1.3.1.5. Fragmentation methods ................................................... 28 
1.3.1.6. Conclusion ................................................................. 35

1.3.2. Strategy of glycan analysis by LC-MS ................................. 36 
1.3.2.1. Enzymatic strategies ....................................................... 37 
1.3.2.2. Derivatization methods .................................................. 40 
1.3.2.3. Separation methods ....................................................... 42 
1.3.2.4. Ion activation strategies in MS analysis ............................ 46 
1.3.2.5. Conclusion ................................................................. 49

1.4 References ............................................................................ 49

Chapter 2: Identification of critical glycosylation differences between Biosimilar I-TNK and Reference product G-TNK ................................................................. 58

2.1 Abstract .............................................................................. 58

2.2 Introduction .......................................................................... 58

2.3 Experimental ......................................................................... 61 
2.3.1 Materials and reagents ....................................................... 61 
2.3.2 In-solution digestion .......................................................... 61 
2.3.3 Fluorescent labeling with 2-aminobenzamide ........................ 62
2.3.4 LC-MS analysis........................................................................................................62

2.3.5 Data analysis............................................................................................................64

2.4 Results and discussion................................................................................................64

2.4.1 Identification of complex-type glycan structures at all three N-linked glycosylation sites for both biosimilar I-TNK and reference product G-TNK........................................65

2.4.2 Sialic acid difference at particular site by glycopeptides approach.........................71

2.4.3 Determination of branch differences of sialic acid by glycan approach...................72

2.5 Conclusion.................................................................................................................76

2.6 References................................................................................................................78

Chapter 3: Identification of upregulation of sialic acid with unique di-sialylated structure in alpha3 integrin from Mouse kidney cells with polycystic kidney disease......................83

3.1 Abstract......................................................................................................................83

3.2 Introduction...............................................................................................................84

3.3 Materials and methods..............................................................................................85

3.3.1 Reagents.................................................................................................................85

3.3.2 Cells.......................................................................................................................85

3.3.3 Immunoprecipitation, SDS-PAGE and western blot............................................86

3.3.4 In-gel digestion......................................................................................................86
3.3.5 LC-MS………………………………………………………………………………87

3.3.6 Data analysis……………………………………………………………………………87

3.4 Results…………………………………………………………………………………………88

3.4.1 Glycosylation patterns of α3 integrin………………………………………………..88

3.4.2 Identification of α3 integrin from kidney cell……………………………………..90

3.4.3 Glycopeptide analysis of α3 integrin light chain…………………………………92

3.5 Discussion………………………………………………………………………………………102

3.6 Conclusion………………………………………………………………………………………105

3.7 References……………………………………………………………………………………110

Chapter 4: Mass spectrometric analysis of leader sequence attached to N-terminus of recombinant human growth hormone produced from a yeast strain of, pichia. Pastoris……….112

4.1 Abstract………………………………………………………………………………………112

4.2 Introduction……………………………………………………………………………………113

4.3 Materials and methods……………………………………………………………………….115

4.3.1 Materials………………………………………………………………………………………115

4.3.2 SDS-PAGE and in-gel digestion………………………………………………………115

4.3.3 LC-MS…………………………………………………………………………………………116

4.3.4 Protein identification…………………………………………………………………………117
4.4 Results .............................................................................................................................. 117

4.4.1 Protein and primary structure identification ............................................................... 117

4.4.2 Identification of N-terminal sequence variance ......................................................... 124

4.4.3 Oxidation ..................................................................................................................... 129

4.4.4 Deamidation ............................................................................................................... 132

4.4.5 Label-free quantitation by external standard with its calibration curve .................... 134

4.5 Conclusion .................................................................................................................... 135

4.6 References .................................................................................................................... 137

Supplemental: Mass spectrometric analysis of leader sequence attached to N-terminus of
recombinant viral protein produced from a CHO cell strain line ........................................ 139

S4.1 Abstract ....................................................................................................................... 139

S4.2 Introduction .................................................................................................................. 140

S4.3 Experimental Section ................................................................................................ 143

S4.3.1 Reagent and materials ............................................................................................. 143

S4.3.2 SDS-PAGE and in-gel digestion .............................................................................. 143

S4.3.3 LC-MS ..................................................................................................................... 144

S4.3.4 Protein identification ............................................................................................... 145

S4.4 Results and discussion ............................................................................................... 145
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4.4.1 Identification of Primary structure</td>
<td>146</td>
</tr>
<tr>
<td>S4.4.2 Identification of N-terminal processing</td>
<td>149</td>
</tr>
<tr>
<td>S4.4.3 Autolytic digestion of Trypsin Protease</td>
<td>153</td>
</tr>
<tr>
<td>S4.5 Conclusion</td>
<td>157</td>
</tr>
<tr>
<td>S4.6 References</td>
<td>159</td>
</tr>
<tr>
<td>Thesis conclusion</td>
<td>161</td>
</tr>
<tr>
<td>Additional research projects &amp; publications</td>
<td>162</td>
</tr>
</tbody>
</table>
List of Figures

Fig. 1-1 Pathway for biosynthesis of Dolichol-P-P-glycan precursor for protein N-glycosylation and transfer of the precursor to Asparagine residues (Asn-X-Thr/Ser) on nascent translated proteins by the oligosaccharyltransferase (OST) complex.................................................6

Fig. 1-2 Processing of the initial high mannose N-glycan in the endoplasmic reticulum and cis-Golgi to generate the core N-glycan substrate used for further diversification in the Golgi............7

Fig. 1-3 The types of N-linked glycans.................................................................8

Fig. 1-4 Core structures of O-linked glycans......................................................8

Fig. 1-5 Therapeutic antibody-induced ADCC in human blood...........................12

Fig. 1-6 Representation of N-glycan structures commonly found in human cells and in different expression systems available for recombinant protein production.............................16

Fig. 1-7 Two unusual glycan structures identified in sEGFR...............................20

Fig. 1-8 Formation and interpretation of a, b, c, x, y and z fragments in peptides........30

Fig. 1-9 CID fragmentation pathway...............................................................30

Fig. 1-10 MS2 and MS3 spectrums of glycosylated tryptic peptide of tissue plasminogen activator, using the hybrid LTQ-ETD instrument.........................................................31

Fig. 1-11 CID-MS2 and ETD-MS2 spectrums of glycosylated tryptic peptide of monoclonal antibody, using the hybrid LTQ-ETD instrument.........................................................33
Fig. 1-12 Nomenclature for describing fragment ions from carbohydrates according to Domon, and Costello………………………………………………………………………………...48

Fig. 1-13 Proposed mechanism for the formation of the $^{2,4}$A$_{R}$ ion from N-linked glycans……..48

Fig. 2-1 LC-MS analysis of the high-mannose glycan structure at Asn-117 from tryptic digests of t-PA was changed to the complex-type glycan structure at Asn-103 from tryptic digests of G-TNK and I-TNK…………………………………………………………………………………………67

Fig. 2-2 Summary of major identified glycans at Asn-103 of G-TNK and I-TNK with their corresponding molecular mass and relative abundance……………………………………………………68

Fig. 2-3 Summary of major identified glycans with their corresponding molecular mass and relative abundance in G-TNK and I-TNK…………………………………………………………………………..70

Fig. 2-4 Differentiation and Identification of positional isomers for 211 Glycan at Asn-448 site74

Fig. 2-5 Relative quantitation of the abundance of positional isomers for 211 glycan at Asn-448 site…………………………………………………………………………………………………………………………………………75

Fig.S2-1 Summary of major identified glycans at Asn-184 site of G-TNK and I-TNK with their corresponding molecular mass and relative abundance 81

Fig.S2-2 Summary of major identified glycans at Asn-448 site of G-TNK and I-TNK with their corresponding molecular mass and relative abundance 82

Fig. 3-1 Glycosylation patterns of $^{\alpha}$3 integrin…………………………………………………………90

Fig. 3-2 Identification of $^{\alpha}$3 integrin from kidney cells………………………………………………………92
Fig. 3-3 Identification of high-mannose glycan structure at Asn-937 of tryptic digested α3 integrin light chain from pkdl^{+/+} cells pkdl^{-/-} cells.................................................................95

Fig.3-4 Identification of complex glycan structure at Asn-971 of tryptic digested α3 integrin light chain from pkdl^{-/-} cells.................................................................97

Fig.3-5 Identification of different glycan at Asn-925 and/or Asn-928 of tryptic digested α3 integrin light chain from pkdl^{+/+} cells and pkdl^{-/-} cells.........................................................99

Fig.3-6 Identification of glycan structure at Asn-925 and/or Asn-928 of tryptic digested α3 integrin light chain from Pkd1^{+/+} and Pkd1^{-/-} cells.......................................................100

Supplemental Fig.S3-1 Immunopurification of α3 integrin.........................................................106

Supplemental Fig.S3-2 Identification of glycosylated peptide backbone by using deglycosylation method (PNGase F treatment) .................................................................107

Fig. 4-1 SDS-PAGE separation of recombinant human growth hormone produced from Pichia pastoris by MIT and GNE rhGH.................................................................120

Fig. 4-2 LC-MS Identification of T21 peptide (C-terminal) from the tryptic digests of rhGH121

Fig. 4-3 Summary of the identification of rhGH peptide sequence...........................................123

Fig. 4-4 Comparison of LC-MS analysis of the tryptic map of rhGH from two different manufacturers.................................................................124

Fig. 4-5 LC-MS analysis of the T1 and two additional peaks from the tryptic digests of MIT rhGH.................................................................126
Fig. 4-6 Extracted ion chromatography (XIC) of three N-terminal variants from tryptic digest of MIT rhGH and GNE rhGH……………………………………………………………………128

Fig. 4-7 Extracted ion chromatography (XIC) of T2 peptide with its oxidation counterparts from tryptic digests of MTI rhGH and GNE rhGH………………………………………………….131

Fig. 4-8 Extracted ion chromatography (XIC) of T15 peptide with its deamidation counterparts from tryptic digests of MTI rhGH and GNE rhGH……………………………………………133

Fig. 4-9 Linearity of quantitation of MIT rhGH by using external standards (T4 peptide and T21 peptide from GNE rhGH) with four different concentrations…………………………………….135

Supplemental Fig. S4-1 LC-MS analysis of the C-terminal (T36) peptide from the tryptic digests of recombinant viral protein AAV5…………………………………………………………………147

Supplemental Fig. S4-2 LC-MS analysis of the N-terminal peptide variant 1 (T1 peptide without Methionine) from the tryptic digest of recombinant viral protein AAV5…………………..150

Supplemental Fig. S4-3 LC-MS analysis of the N-terminal peptide variant 2 (T1 peptide without Methionine but with Serine acetylation) from the tryptic digest of recombinant viral protein AAV5…………………………………………………………………………………………152

Supplemental Fig. S4-4 LC-MS analysis of T13-part1 from the tryptic digests of recombinant viral protein AAV5…………………………………………………………………………..155

Supplemental Fig. S4-5 LC-MS analysis of T13-part2 from the tryptic digests of recombinant viral protein AAV5……………………………………………………………………………156
List of Tables

Table 1-1 Seven basic common monosaccharides in human…………………………………..….3

Table 1-2 Structural difference in the N-glycans produced difference expression systems, used as hosts for glycoprotein production, as compared to human glycosylation………………….15

Table 1-3 Commonly used enzymes in peptide/glycopeptides digestions……………………………..25

Table 1-4 Bovine α1-acid glycoprotein glycopeptides detected by nano LC-MS/MS………………..........................27

Table 1-5 Enzymes applied for glycan release……………………………………………………………………………..39

Table 3-1 Tryptic digested α3 integrin light chain with corresponding glycopeptides……………93

Table 3-2 Major glycans at each site of α3 integrin light chain in Pkd1+/+ cells and Pkd1−/− cells…………………………………………………………………………………………………………………104

Supplemental Table S3-1 Summary of the identified peptide sequences for α3 integrin heavy chain and light chain with precursor mass, charge, identified m/z and X correction…………107

Table 4-1 Summary of tryptic mapping of MIT produced rhGH (peptides identified without trypsin miscleavages)……………………………………………………………………………………………122

Table 4-2 Summary of tryptic mapping of MIT produced rhGH (peptides identified with 1 or 2 trypsin miscleavages)……………………………………………………………………………………123

Supplemental Table 4-1 Summary of tryptic peptides of recombinant viral protein, AAV5…148
List of Abbreviations

Glc: Glucose

Gal: Galactose

Fuc: Fucose

Man: Mannose

GlcNAc: N-acetylglucosamine

GalNAc: N-Acetylgalactosamine

Neu5Ac or NANA: N-Acetylneuraminic acid

Neu5Gc: N-Glycolylneuraminic acid

Ser: Serine

Thr: Threonine

OST: Oligosaccharyltransferase

FcR: Fc receptors

ADCC: Antibody-dependent cellular cytotoxicity

CHO: Chinese hamster ovary

BHK: Baby hamster kidney

DO: dissolved oxygen
CO2: Carbon dioxide:

sEGFR: The secreted form of EGFR

M-LAC: The multi-lectin affinity chromatography

Con A: concanavalin A

WGA: wheat germ agglutinin

GnHCl: Guanidine hydrochloride

SDS: sodium dodecyl sulfate

DTT: dithiothreitol

IAA: iodoacetamide

PNGase F: peptide N-glycosidase F

PNGase A: Peptide-N-glycosidase A, glycoamidase A

RPLC: reversed phase chromatography

PGC: Porous graphite carbon chromatography

HILIC: Hydrophilic interaction chromatography

CID: Collision-Induced Dissociation

PTM: post-translational modifications

ETD: Electron transfer dissociation

2-AB: 2-aminobenzamide
2-AA: 2-aminobenzoic acid

PA: 2-aminopyridine

APTS: 1-aminopyrene-3, 6, 8-trisulfonic acid

CGE: capillary gel electrophoresis

HPAEC-PAD: high-performance anion-exchange chromatography with pulsed amperometric detection

CE-LIF: capillary electrophoresis-laser induced fluorescence

DEA: diethylamine

MALDI: Matrix-assisted laser desorption ionization

ESI: electrospray ionization

LC-MS: liquid chromatography-mass spectrometry

ADPKD: Autosomal Dominant Polycystic Kidney Disease

RTK: receptor tyrosine kinases

XIC: extracted ion chromatogram

NSCL: non-small cell lung cancer

α-MF: α-mating factor leader sequence

Kex2p: Kex2 protease

Ste13p: Ste13 dipeptidyl aminopeptidase
MetO: Methionine Sulfoxide

NMC: N-terminal Methionine Cleavage

NTA: N-terminal Acetylation

MetAPs: Methionine Amino-peptidases

NER: N-end rule

AAV: Adeno-associated virus
Chapter 1: Overview of LC-MS determination of glycosylation pattern on glycoproteins as critical quality attribute for biopharmaceuticals and potential markers for diseases.

1.1 Abstract

In chapter 1, an overview of glycosylation pattern as critical quality attributes for biopharmaceuticals and potential disease markers. Basic concepts of oligosaccharides, biosynthesis pathway, and types of glycosylation are introduced to provide the background knowledge. In addition, influence of glycan patterns on glycoprotein in therapeutic protein efficacy and significant function of glycoproteins in disease biomarker discovery are explained to establish the importance of glycoprotein characterization. Finally, LC-MS based techniques for determination of protein glycosylation are summarized. Strategies of glycopeptides analysis and glycan analysis are reviewed in details, respectively.

1.2 Glycosylation patterns as critical quality attributes for biopharmaceuticals and potential disease markers

1.2.1 Basic concepts of glycosylation

The complexity of glycosylation lies in the fact that not all potential glycosylation sites are actually glycosylated (macro-heterogeneity) and variable glycan structures attached to the individual site (micro-heterogeneity) [1]. In addition, enantiomers and anomers of monosaccharide structures [1], different linkages of glycosidic bonds [2], and substitutions of glycans, such as sulfation, acetylation and phosphorylation, are also important factors leading to the complexity of glycosylation [2, 3].
In human, the glycan structures are only composed of seven basic monosaccharides under normal physiological conditions: glucose (Glc), galactose (Gal), fucose (Fuc), mannose (Man), N-Acetylglucosamine (GlcNAc), N-Acetylgalactosamine (GalNAc), N-Acetyleneuraminic acid (Neu5Ac or NANA)[1]. Of which, Neu5Ac is of great significance due to its highly variable chemical structures and inherent negatively charged properties. N-Acetyleneuraminic acid (Neu5Ac) belongs to sialic acids family, which comprises of about 40 derivatives of nine-carbon sugar neuraminic acid [5-7]. Table 1-1 listed seven basic common monosaccharides in human with structures and symbols. Other derivatives of sialic acids, such as N-Glycolyneuraminic acid (Neu5Gc), have been found in human only in particular cancers [6]. Except for the variability of structures, linkages of sialic acids also have several possibilities. Sialic acids usually seated at the terminal of sugar chain [8] and they are connected via C2 to position 3 or 6 of the penultimate sugar, which are Gal and GalNAc in most cases. Alternatively, they are linked via C2 with position 8 of another sialic acid molecule [9].
Table 1-1. Seven basic common monosaccharides in human [6]

<table>
<thead>
<tr>
<th>Monosaccharide</th>
<th>Structure</th>
<th>Symbol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucose (Glc)</td>
<td><img src="image1" alt="Glucose Structure" /></td>
<td><img src="image2" alt="Blue Circle" /></td>
</tr>
<tr>
<td>Galactose (Gal)</td>
<td><img src="image3" alt="Galactose Structure" /></td>
<td><img src="image4" alt="Yellow Circle" /></td>
</tr>
<tr>
<td>Fucose (Fuc)</td>
<td><img src="image5" alt="Fucose Structure" /></td>
<td><img src="image6" alt="Red Triangle" /></td>
</tr>
<tr>
<td>Mannose (Man)</td>
<td><img src="image7" alt="Mannose Structure" /></td>
<td><img src="image8" alt="Green Circle" /></td>
</tr>
<tr>
<td>N-acetylglucosamine (GlcNAc)</td>
<td><img src="image9" alt="GlcNAc Structure" /></td>
<td><img src="image10" alt="Blue Square" /></td>
</tr>
<tr>
<td>N-acetylgalactosamine (GalNAc)</td>
<td><img src="image11" alt="GalNAc Structure" /></td>
<td><img src="image12" alt="Yellow Square" /></td>
</tr>
<tr>
<td>N-Acetyleneuraminic acid (Neu5A; sialic acid in human;)</td>
<td><img src="image13" alt="Neu5A Structure" /></td>
<td><img src="image14" alt="Purple Diamond" /></td>
</tr>
</tbody>
</table>

This type of glycans are highly special because they behave as strong organic acids with PKa equal to 2.2 and introduce negative charges to glycoproteins under physiological conditions due to amino group at position 5 and the carboxyl group at position 1[5]. Thus, negatively charged
sialic acids can exert a simple but profound effect on its environment. They are involved in the binding and transporting of positively charged molecules, such as Ca$^{2+}$. They are also of significant function in cell attractions and repulsions [9]. Additionally, sialic acids act as protective shield for glycoproteins from being degraded and recognized by immune system, which are explained in the section 1.2.3.1[10].

1.2.2 Biosynthesis of N-linked glycans and O-linked glycans in human cells

Typically, the glycans associated with glycoproteins are divided into two types based on the linkages: N-glycans, attached to the Asparagine (Asn) [11] side chains; and O-glycans, which bind to the oxygen atom of Serine (Ser) or Threonine (Thr) side chains. N-glycosylation sites can be predicted by a conserved epitope of Asn-Xaa-Ser/Thr (where Xaa is any amino acid except proline) [12]. The biosynthesis of N-linked glycans involves a series of activities, with assistance of a multitude of glycosyltransferases and glycosidases [13]. It starts on the cytosolic side of the ER membrane, and then flips across the membrane bilayer into the lumen of the ER, finishing the assembly of Dolichol-P-P-glycan precursor Glc3Man9GlcNAc2. Then the Dolichol-P-P-glycan precursor is transferred to the Asparagine residue with consensus epitope [14]. The pathway for biosynthesis of N-linked glycans is summarized in the Fig.1-1 and Fig.1-2. [17].

Further processing reactions then occur after the attachment of glycan precursor to Asparagine residue. The initial trimming step is to remove the three Glc residues, which is associated with protein folding mechanisms. It also controls the retention time of glycoprotein in the ER.

Correctly folded glycoproteins are further transferred to cis-Golgi, where sequential activities are responsible for glycan diversification and maturation [15, 16]. Galactose, sialic acid, and fucose are added and extended in the final process. The mature N-linked glycans contain three types:
high-mannose, complex and hybrid types, but all type glycans preserve the same core structures with Man3GlcNAc2. The high-mannose type N-glycans may have two to six additional mannose residues linked to the core. Complex types have two or more branches GlcNAc attached to core, instead of mannose residues. The hybrid types contain one outer branch of the high-mannose type and another of complex type [15, 16]. The types of N-linked glycans with core structures circled are summarized in Fig.1-3.
Fig.1-1. Pathway for biosynthesis of Dolichol-P-P-glycan precursor for protein N-glycosylation and transfer of the precursor to Asparagine residues (Asn-X-Thr/Ser) on nascent translated proteins by the oligosaccharyltransferase (OST) complex [17].
Fig. 1-2. Processing of the initial high mannose N-glycan in the endoplasmic reticulum and cis-Golgi to generate the core N-glycan substrate used for further diversification in the Golgi [15].

1. Oligosaccharyltransferase
2. Glucosidase I; Glucosidase II; Endomannosidase
3. Mannosidase IA; Mannosidase IB
Fig. 1-3. The types of N-linked glycans and the core structure with Man3GlcNAc2 are circled in red [15].

Fig. 1-4. Core structures of O-linked glycans.

Core 1 $\beta_3$

Core 2 $\beta_6$

Core 3 $\beta_3$

Core 4 $\beta_6$

Core 5 $\alpha_3$

Core 6 $\beta_6$

Core 7 $\beta_6$

Core 8 $\alpha_3$
O-linked glycosylation process happens after the protein N-glycosylation, folding and oligomerization [1]. The biosynthesis of O-glycans is simpler than N-glycans, which is initiated by the addition of the GalNAc to Serine and Threonine residues, then the O-glycans are extended into different structures by adding GalNAc, GlcNAc, Xylose, Mannose, Fucose and Sialic acids [18].

1.2.3 Glycoproteins as biopharmaceuticals

Therapeutic proteins have been dramatically dominating the pharmaceutical market over the past years, and the majority of these approved therapeutic proteins are glycoproteins. Glycosylation is one of most important posttranslational modifications, influencing the efficacy of protein drugs through the biological activity, stability, solubility, serum half-life and immunogenicity. Consequently, monitoring and manipulating the glycosylation alterations are critical demands to current biopharmaceutical manufacturing and processing.

1.2.3.1 Effect of glycosylation on therapeutic protein efficacy

The therapeutic protein efficacy can be influenced by inappropriate glycosylation, which display on aspects of stability, circulatory half-life, solubility and immunogenicity [17]. For example, the attachment of glycans to some proteins (erythropoietin) assists in protein folding. Also, the size, branching and charges of glycans attached, are all related to proteolytic stabilization, especially for glycoproteins with sialic acid-containing glycans. Sialic acids exert an extraordinary impact on the quality and stability of any therapeutic glycoproteins for several reasons.

First, terminal sialic acid glycans are traditionally connected with galactose residues, which are the major determinants for serum half-life because exposed galactose residues can be bound with liver Gal-specific receptors, then removed from serum by endocytosis. Expression of extended
sialic acids protects galactose residues to be recognized and further prevents serum glycoproteins from degradation. Second, the receptors in immune systems are highly sensitive to the nonsialylated structures. Therefore, the immune systems will generate neutralizing antibodies against the foreign therapeutic glycoproteins with exposed terminal galactose. The presence of sialic acids extended from galactose residues is an important mask for those antigenic epitopes. Finally, negatively charged sialic acids improve protein thermal stability, solubility and resistance to proteolytic degradation due to the hydrophilic glycans. Thus, sialylation of glycoproteins is crucial for therapeutic glycoproteins production. Also, the consistency and homogeneity of standardized sialylation (identical degree of sialylation from batch to batch) are also major challenges for manufactures to ensure in the production process [19].

In order to enhance the therapeutic efficacy, a multiple of approaches have been developed to engineer glycan structures of protein therapeutics. Strategies of site-directed mutagenesis were mainly employed and achieved in many scientific accomplishments in the past decades. For example, the biosynthesis pathway of glycosylation can be modified specifically to obtain the desirable product and elevated therapeutic efficacy. Specific genes, encoding functional enzymes along the biosynthesis pathway, are knocked out and led the host cell to produce fitting glycan structures on glycoproteins, such as eliminating unwanted glycans (i.e. high mannose) and incorporate desired structure (i.e. complex type). This strategy can lead to extended serum half-life and improved in vivo biological activities [20, 21]. One of the major targets by this approach is Fut8, which encodes for the α1, 6-fucosyltransferase enzyme. Knockout of Fut8 results in the absence of α1, 6-fucosyltransferase enzyme and this is responsible for attachment of core fucose residue to the glycan [22, 23]. For mAb-based therapeutics, the clearance rate of mAb depends on the binding affinity of Fc receptors (FcR) to its attached glycans. It has been
reported that the binding affinity is also influenced by the attachment of fucose in the structure. Studies have shown that the binding affinity of non-fucosylated glycoforms to FcR receptors is up to 50-fold more than the corresponding fucosylated counterparts [24, 25], which triggers a dramatically elevated antibody-dependent cellular cytotoxicity (ADCC) up to 100 folds [24-27]. Thus, after knock-out of Fut8, the host cell can only produce defucosylated mAbs, which displayed highly activated ADCC[28, 29] at low doses[23, 30], achieving enhancement of therapeutics efficiency. Fig.1-5 illustrated the therapeutic antibody-induced ADCC activation in human blood. Additionally, RNA interferences are also engineered to hinder specific enzymes, which function in the biosynthetic pathways [31, 32]. Fut8 can also be silenced, by externally introduced interfering RNA, to obtain defucosylated mAb with enhanced ADCC and therapeutic efficacy [33, 34].
Fig. 1-5. Therapeutic antibody-induced ADCC in human blood [27].

Therapeutic antibodies show the same antigen binding activity irrespective of core fucosylation of the Fc. (a) Non-fucosylated antibodies overcome the competition with serum IgG to bind to the effectors cells through much higher binding affinity to FcRIIIa than serum IgG, and thereby induce high ADCC. (b) Fucosylated antibodies fail to recruit effector cells effectively due to low binding affinity to the FcRIIIa. (c) The high ADCC of non-fucosylated antibodies is inhibited by the fucosylated counterparts through the competition for binding to the antigen on target cells [27].
1.2.3.2 External factors influencing glycosylation

The consistency of glycosylation heterogeneity and minimization of undesirable glycoforms produced during manufacturing process are required by regulatory authorities, but the control and maintenance of homogeneity of glycosylation profile during production and storage is still a considerable challenge [13], due to the great variability inherent to the biosynthesis of glycosylation procedure and to the variability raised by environmental factors [35-38]. Therefore, to avoid the unexpected alterations during therapeutic protein production, it is crucial to understand how and at what extend these parameters affect glycosylation. First, the selection of an expression system is highly critical as the correct expression system can reach maximum product yields, ensure product quality and avoid cell line specific contaminants [39]. Many expression systems have been considered in the manufacture systems, but not all of them are suitable for productions of therapeutic proteins. Some therapeutic proteins, such as insulin or human growth hormone, are not glycosylated, so the expression system is easier to choose. However, the majorities of approved therapeutics, or those under development, are glycoproteins, so the expression system has to be seriously considered as proper glycosylation of therapeutic proteins has to be guaranteed. Improper glycosylation and non-human glycoforms can raise strong side effects on pharmacokinetic parameters and lethal immunogenicity issues [40-43]. The Chinese hamster ovary (CHO) cells, producing nearly 70% of recombinant protein therapeutics [39, 44], are the predominant choice in the current pharmaceutical market. Compared with other cell lines, the human-like post translational modifications produced, especially human-like glycosylation, are the most attractive advantages [39, 45-47]. Also, easy of transfection, extensive knowledge, as well as high product yield, are all benefits contributing
to the extensive use of CHO cells [39, 46-48]. However, an inherent disadvantage is that CHO cells lack the enzyme α2, 6-sialyltransferase, so it can only produce α2, 3- linkages between sialic acid and Gal in the glycans [48, 49]. In addition, GnT-III enzyme, which is responsible for addition of bisecting GlcNAc, is also absent in the CHO cells, so the therapeutic proteins produced from CHO cells, lack the bisecting GlcNAc glycan structures. These divergences can influence circulatory lifetime and biological activities [42, 48, 50, 51].

Other expression systems are also used in the production of therapeutic proteins, including bacteria, yeast, insect, plant, and other mammalian cells, for example baby hamster kidney (BHK) and murine myeloma (NS0, Sp2/0). However, they can produce different glycan structures, which are unfavorable in human. The different glycan structures, produced by various hosts for glycoprotein production, are summarized in table 1-2 and representation of N-glycan structures commonly found in human cells and in different expression systems are summarized in the Fig.1-6 [17].
Table 1-2. Structural difference in the N-glycans produced difference expression systems, used as hosts for glycoprotein production, as compared to human glycosylation [17].

<table>
<thead>
<tr>
<th>Expression systems</th>
<th>Difference to human</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td>absence of the glycosylation machinery</td>
<td>recombinant proteins cannot be glycosylated</td>
</tr>
<tr>
<td>Yeast</td>
<td>presence of yeast-specific high-mannose type glycans</td>
<td>fast clearance rate immunogenic</td>
</tr>
<tr>
<td>Insect cells</td>
<td>presence of α1,3-Fucose presence of β1,2-xylose absence of Gal/sialic acid</td>
<td>fast clearance rate immunogenic</td>
</tr>
<tr>
<td>Plant</td>
<td>presence of α1,3-Fucose presence of β1,2-xylose absence of Gal/sialic acid</td>
<td>immunogenic</td>
</tr>
<tr>
<td>Baby hamster kidney (BHK)</td>
<td>absence of α2,6-linked sialic acids NeuGc produced at minor fractions</td>
<td>Undersialylation, reduced serum half-life Not immunogenic at the levels produced</td>
</tr>
<tr>
<td>Murine myeloma</td>
<td>Presence of α1,3-Gal Predominance of NeuGc sialic acids</td>
<td>immunogenic</td>
</tr>
<tr>
<td>(NS0, Sp2/0)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fig. 1-6. Representation of N-glycan structures commonly found in human cells and in different expression systems available for recombinant protein production [17].

![Diagram showing N-glycan structures in various organisms](image-url)
Second, environmental factors and culture conditions are also important contributing factors for aberrant glycosylation of therapeutic glycoproteins. Optimal parameters of cell culture environment can promote cell growth and increase product yield [36, 45, 52, 53]. Therefore, control and manipulation of cell culture conditions are essential to manufactures to ensure the product quality and glycoform profile consistency [13].

Among a mixture of nutrients in the medium, glucose is often utilized as energy source in the cell culture [54, 55]. Limited glucose feeding can result in reduced site occupancy and absent glycosylation as glucose is a major component that comprises the glycan precursor. Galactose can also be employed as a substituent to serve energy for cell growth. However, galactose feeding is easy to produce hyper-galactosylated N-glycan profile [36, 56]. Also, Glutamine, essential component for nucleotide synthesis, can also be applied as an alternative energy source. But ammonia generated tends to accumulate in the medium to inhibit cell growth and increased pH in Golgi apparatus can influence the activities of some glycotransferase, and thereby alter the glycan structures [57-59]. The known alteration resulted from ammonia accumulation, is the decreased terminal sialylation, reported by many groups [52, 57, 60].

In addition, other cell environment factors, such as dissolved oxygen (DO), pH, carbon dioxide (CO₂) and temperature, are all affecting protein glycosylation to some extent in different ways, which will not be introduced in details here.

1.2.4 Glycoproteins in disease biomarker discovery

In human cells, it is estimated that 70% of proteins are glycosylated [61]. Aberrant glycosylation patterns were linked with specific diseases from more than 40 years ago [62] and glycoproteins play crucial roles in the clinical discovery of glycan-based biomarker for detecting disease in the
early stage, monitoring disease progression, measuring the therapy response and observing recurrence. Glycoproteins have higher chance to serve as disease markers, especially for cancer biomarkers for the reason that unlike protein synthesis, glycosylation is a non-template based synthetic process, undergoing a series of enzymatic pathways. This process can pass on a rich source of information, such as the expression of protein-coding genes, activities of involved enzymes, availabilities of monosaccharide donors as well as multiple upstream and downstream effectors [16].

The oligosaccharide units on glycoproteins have been established to affect the function of whole protein through many theories, so the functions of oligosaccharides are truly worthwhile to discuss. Oligosaccharides influence glycoprotein 3-D conformation, which is closely associated with protein inherent stability, solubility and protease resistance in human cells [16]. Specifically, the oligosaccharides on glycoproteins provide the function of “coating”, preventing the polypeptide chains from recognition and digestion by protease or antibody. For example, abnormal oligosaccharides on fibronectin make it more susceptible to protease digestion [63]. Lack or alteration of oligosaccharides on tissue-type plasminogen activator can increase proteins enzymatic activity and also alter conversion to two-chain form. In addition, oligosaccharides are commonly accepted on initiating the correct folding of glycoproteins. Many improper glycosylated proteins fold incorrectly and are inhibited to exit ER, then consequently degraded [16].

Glycoproteins also play crucial roles in biological activities. Oligosaccharides on glycoproteins, are often located on cell membranes or in extracellular matrix [64, 65], allowing them to be recognized as receptors, and significantly function in cell trafficking, cell adhesion and biological half-life. They mediate biochemical interactions such as binding with lectins and
ligands [66]. For insulin and insulin-like growth factor-I receptors, complete deglycosylation correlates with totally loss of binding affinity while partial glycosylation alterations change binding specificity of the protein to some extent. Some studies also demonstrate the loss of tyrosine kinases activity with abnormal glycosylation pattern of insulin [63].

Furthermore, one of the major changes in glycan structures in many cancer types are the increased expression levels of fucosylation and sialylation. Thus, the abnormal glycan structures can be used as potential glycan markers in the disease state. Epidermal growth factor receptor (EGFR) is an important oncoprotein and has been implicated with lung, ovarian and breast cancers. The secreted form of EGFR (sEGFR) from A431 cell line, exhibited two abnormal glycan structures, including di-sialic acid motif and HexNAc-[Fuc]Gal-GlcNAc motif. The presence of such unusual glycan structures at given sites with either more negative charges or bulky branched structures, may be developed into critical glycan markers for secretion or cancer metastasis [67]. Two unusual glycan structures identified in sEGFR are summarized in Fig.1-7.
1.3 Determination of protein glycosylation by liquid chromatography and Mass spectrometry (LC-MS) based techniques

1.3.1 Strategy of glycopeptide analysis by LC-MS

1.3.1.1 Isolating glycoproteins

Sample enrichment and purification is very necessary in glycopeptides analysis, in order to obtain useful MS data. Recombinant expressed glycoproteins are always available in high concentration and high purity, so there is no need for extra purification methods to be performed. But for endogenous glycoproteins, which are from biological samples, such as serum, human urine and cell lysates, etc, isolation of interesting glycoproteins is a key step prior to LC-MS analysis [68].

In protein-specific analysis, immunoprecipitation can be applied by utilizing corresponding antibodies, followed by subsequent gel electrophoresis for further separation [69].
analysis of broad types of glycoproteins, affinity chromatography is the main technique to use [70]. Lectins are known to bind with different types of carbohydrates [71].

The multi-lectin affinity chromatography (M-LAC) was developed in our laboratory and used for enrichment of plasma or serum glycoproteins [72]. The M-LAC affinity column consists of three types of lectins, concanavalin A (Con A), wheat germ agglutinin (WGA) and jacalin, which have affinity for glycoproteins with N-linked branched mannose, sialic acid as well as O-linked glycan structures, respectively [72]. In the proteomics and glycoproteomics analysis, the M-LAC integrated with upstream immunodepletion of abundant plasma proteins have become a useful strategy in the proteomics and glycoproteomics analysis [73], and the platform was implemented to study the glycosylation analysis of biologically significant proteins to breast cancer [74].

1.3.1.2 Reduction and alkylation

Glycoproteins have to be unfolded to achieve an efficient enzymatic digestion, so before proteolytic digestion, glycoproteins need to be denatured first. There are two ways to denature a protein. One is by adding denaturing agents, such as surfactants to the initial buffer solution. For example, Guanidine hydrochloride (GnHCl), urea, or sodium dodecyl sulfate (SDS). The other method is reduction by dithiothreitol (DTT) and alkylation by iodoacetamide (IAA) [75, 76]. DTT is usually used for denaturation of glycoprotein and cleavage of disulfide bond, which is responsible for holding the glycoprotein 3-dimentional structure, to two sulfhydryl groups. The sulfhydryl groups are subsequently capped by alkylating agents, to prevent unwanted reformation of disulfide bonds [77].
1.3.1.3 Enzymatic strategies

The selection of enzymes must be designed with considerations of the sequence of specific protein to be characterized. In term of specificity of enzymes, it is generally divided into two types: specific enzymes (i.e. Trypsin, GluC, LysC, etc) and non-specific enzymes (i.e. Protease K or Pepsin). Enzymes with defined cleavage sites are often taken into account in the first place as predictions of peptide or glycopeptide backbone are more straightforward.

Trypsin is the most widespread enzyme to use in the protein or glycoprotein characterization. It has well defined cleavage sites at C-terminal side of two charged amino acid residues Lys and Arg (except for the next amino acid residue is Proline). Peptides or glycopeptides digested by trypsin have appropriate peptide lengths and charges in LC-MS detection compared with those digested by GluC and LysC, so trypsin is always considered as the first choice in enzymatic digestions [78]. However, in many different cases, other enzymes are also utilized solely to obtain desired peptide lengths. For example, in ETD fragmentations, LysC, which only cleaves K, is the most favorable enzyme for digestion in order to generate larger peptides or glycopeptides, because ETD fragmentations work best for highly charged species and further lead to more informative MS/MS spectrums [79].

Multienzyme strategies are of great significance in glycopeptide analysis. Particularly for N-linked glycopeptides analysis, peptide N-glycosidase F (PNGase F) is the critical enzyme to be utilized because PNGase F can remove N-linked glycans, accompanied with the glycosylation site - Asparagine (N) converted to Aspartic acid (D) and resulted in 0.984 Da mass shifts on peptide precursor mass. Digestions by selected enzymes with and without PNGase F have to be performed in parallel as samples treated with PNGase F can be used as control to identify the
sequence of glycopeptides backbone and provide hints for site occupancy as well. Peptides are confirmed to be originally glycosylated if Asn changed to Asp after PNGase F treatment, while peptides are originally non-glycosylated if Asn existed without any change. Even though deamidation can also cause Asn changed to Asp, it is not considered in this situation because for Asn without glycosylation, samples without treating PNGase F, should have the same observations (Asn\(\rightarrow\)Asp) if the 1Da difference is introduced by deamidation, but for Asn with glycosylation, deamidation cannot occur to the same site.

In addition, multiple glycosylation sites, including O-linked and N-linked glycosylation, are possible to locate on the same glycopeptide if glycoprotein is only digested by a single enzyme, and this can complicate the data interpretation. Large and non-modified peptides from single enzymatic digestion can certainly suppress the signal of glycopeptides, which have much poorer ionization efficiency. To overcome these limitations, various enzymes can be selected to pair with conventional trypsin digestion in terms of the potential cleavage sites included in the glycopeptides backbone. For example, GluC cleaves C-terminus of D and E while AspN cleaves N-terminus of D. If proline (P) is the next amino acid after Lysine (K), Lys-C and trypsin can function together to cleave between K and P. Chymotrypsin, with preferred cleavage sites at C-terminal of large hydrophobic amino acids, Tryptophan(W), Tyrosine (Y), Phenylalanine (F), is another candidate to choose for enzymatic digestion, in order to obtain desired glycopeptides backbone.

If glycopeptides don’t contain any interesting cleavage sites by enzymes with specificity, an alternative approach for proteolysis is using the non-specific enzymes together with trypsin, such as proteinase K[78, 80], which digests glycoproteins in an undefined manner, but to produce shorter glycopeptides and non-modified peptides. Smaller glycopeptides always lead to higher
mass accuracy measurement and higher signal intensities compared with large glycopeptides backbone. Even though the predictions of glycopeptide backbone are more challenging, with PNGase F treated samples as control (N-linked glycopeptides), it can simplify the data analysis to some extent because after the removal of N-linked glycans, deglycopeptide backbone can be searched against the software and automatically identified. Therefore, Glycopeptides with glycan structures on the corresponding peptide backbone can be easily predicted. For O-linked glycopeptides, instead of using a specific enzyme to remove glycans (such as N-linked glycans), chemical reactions (i.e. β-elimination) need to be conducted to remove O-linked glycans. Then the deglycosylated peptides can be determined by software with the corresponding modifications on the sites (dehydration on Serine, Threonine and Tyrosine). Pepsin is another commonly used non-specific enzyme, but cannot be used concurrently with trypsin or other enzymes introduced above, as the well-suited environment of pepsin digestion is acidic and the pH is about 2. Pepsin is also not recommended to use in a step-wise manner with other enzymes, because the digestion buffer has to be adjusted to proper pH during proteolysis, which more or less, introduces external variances generated by sample preparations. Commonly used enzymes in peptides/glycopeptides digestions are summarized in Table 1-3.
Table 1-3. Commonly used enzymes in the peptide/glycopeptides digestions.

<table>
<thead>
<tr>
<th>Types</th>
<th>Enzyme</th>
<th>Cleavage sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specific</td>
<td>Trypsin</td>
<td>C-terminal of K and R*</td>
</tr>
<tr>
<td></td>
<td>Glu-C</td>
<td>C-terminal of D and E</td>
</tr>
<tr>
<td></td>
<td>Lys-C</td>
<td>C-terminal of K#</td>
</tr>
<tr>
<td></td>
<td>Asp-N</td>
<td>N-terminal of D</td>
</tr>
<tr>
<td></td>
<td>Chymotrypsin</td>
<td>C-terminal of W, Y and F</td>
</tr>
<tr>
<td>Nonspecific</td>
<td>Pepsin</td>
<td>Undefined manner</td>
</tr>
<tr>
<td></td>
<td>Protease K</td>
<td>Undefined manner</td>
</tr>
</tbody>
</table>

*Trypsin cannot cleave K and R if the next amino acid residue is Proline (P)  
#Lys-C with trypsin digestion can cleave between K and P.
1.3.1.4 Separation methods

**Reversed phase chromatography (RPLC)**

The reversed phase chromatography (RPLC) is widely used in the glycopeptide analysis. The elution times of glycopeptides on RPLC heavily depend on the hydrophobic interaction of peptide portion with the stationary phase, which are determined by the peptide lengths and composition of amino acids. Relative to the deglycosylated counterparts, the glycopeptides are eluted slightly earlier due to the polarity of glycan molecules in most cases, except for the glycopeptides with very short peptide backbones [81]. Also, different glycan structures associated with same glycosylation sites are co-eluted or eluted within a short retention time window on RPLC due to the identical peptide backbone. And this attribute can facilitate us to identify more glycan structures even those with relative low intensity. In addition, the glycan structures with different number of sialic acid can be nicely separated on RPLC and the glycan isomers with branches differences can also be resolved in some cases, such as A1G2F and A1G1F. However, the separation of isomers may be affected by either length or amino acid compositions of peptide backbones, which still need further study [82-83].

**Porous graphite carbon chromatography (PGC)**

Porous graphite carbon chromatography adapts porous graphite carbon as a stationary phase, which is made of intertwined carbon ribbons. The separation principle is determined by the disperse interactions between the solute and the stationary phase. Compared to C18 columns, PGC columns are well-suited for glycopeptides with short peptide backbones, so glycoproteins are always subjected to pronase digestion before using PGC columns [84]. Pronase is a generic enzyme which digests glycoproteins in a harsh way. Pronase hydrolyzed each peptide bond of
the glycoprotein, except for those neighboring to the glycosylation sites due to the steric hindrance caused by the glycans. Thus, the remaining glycopeptides only contain a few amino acids in the peptide backbone, which are well suited for the PGC-MS analysis. In addition, PGC column can also resolve some glycan isomers with peptide backbone, such as G1F and G1F’[86].

A study of glycopeptide analysis using PGC has been implemented by Thermo Fisher Scientific (San Jose) and the results were compared with the same glycopeptide analysis using traditional C18 column. The results are illustrated as Table 1-4. From the table 1-4, it is obvious to see that the glycopeptides with short peptide backbone have better retentivity than C18 column but glycopeptides with long peptide backbone cannot be analyzed on both PGC and C18 columns as the long peptides are two hydrophobic to elute out. Instead, C8 or C12 column may achieve the analysis.

Table 1-4: Bovine α1-acid glycoprotein glycopeptides detected by nano LC-MS/MS [85].

<table>
<thead>
<tr>
<th>Peptides</th>
<th>Type of LC column</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PGC</td>
</tr>
<tr>
<td>103 QNGTLSK109</td>
<td>√</td>
</tr>
<tr>
<td>53 NPEYNK58</td>
<td>√</td>
</tr>
<tr>
<td>91 CVYNCSFIK99</td>
<td>√</td>
</tr>
<tr>
<td>128 TFMLAASWNGTK139</td>
<td>√</td>
</tr>
<tr>
<td>19 QSPECANLMTVAPITNATMDLLSGK14</td>
<td>×</td>
</tr>
</tbody>
</table>

27
**Hydrophilic interaction chromatography**

Hydrophilic interaction chromatography (HILIC) is a part of normal phase liquid chromatography and separation principle is dependent on liquid-liquid partition of analytes. HILIC utilized the hydrophilic materials (i.e. non-polar bonded silica) as stationary phases and high organic solvent (i.e. acetonitrile) as eluents. As the high organic solvent composition decrease, the analytes elute in order of their polarity. Specifically for glycopeptides, peptide backbone with large glycans (i.e. A1G2F) are always polar than that with small glycans (i.e. G0), so glycopeptides with small glycan structures are eluted earlier than those with large glycans. Also, pronase treatment is also preferable for HILIC column as short peptide length can enhance the retentivity of glycopeptides on HILIC column and further achieve a better separation [87].

1.3.1.5 **Fragmentation methods**

**Collision-Induced Dissociation (CID)**

The CID process basically involves two sequential steps. The first step is very fast and only takes $10^{14}$ to $10^{16}$ s. The precursor ions of analytes collide with the helium or nitrogen gas present in the collision cell. During the collision, part of ion energy is translated into internal energy, leading the ions into an excited state. The second step is the activated ions decomposed into fragment ions, which is referred as unimolecular decomposition. CID can be performed with high collision energies, by using tandem sector and time-of-flight instruments, or with low energies (between 1 to 100eV), in triple quadruple, ion trap, ICR or hybrid instruments [88].

CID remains the most useful fragmentation method applied in analytical laboratories today. It widely employed in the proteomics analysis and characterization of therapeutic proteins. The peptides fragmented by CID, induce dissociation of amide bond, producing b and y ions [88].
Fig.1-8 indicated the formation and interpretation of b and y ions, together with a, c, x and z ions generated from other fragmentations. Fig.1-9 illustrated the CID fragmentation pathway.

However, most posttranslational modifications (PTM) of proteins are labile under CID, such as glycosylation and phosphorylation, etc, because the bonds linked with those modifications are weaker than peptide bonds, which are dissociated first [89].

CID fragmentation method is applied extensively for glycopeptides analysis. In the low energy CID, glycosidic bonds are easier to dissociate, generating b- and y- type glycosidic ions on sugar chain, but glycopeptide backbone keep intact and unfragmented. Major peaks in the CID spectra, resulted from consecutive losses of single oligosaccharide residues from precursor mass, are dominating in the CID-MS2 spectrum. At low mass regions, oxonium ions of oligosaccharides, such as 366, 527, 657Da can be observed and used as marker ions to identify the presence of glycopeptides. Critical ions correspond to peptide with one GlcNAc can also be observed in the CID-MS2 spectrum in most cases, giving hints on glycopeptides precursor mass. Fig.1-10 showed MS2 and MS3 spectrums of glycosylated tryptic peptide of tissue plasminogen activator, using the hybrid LTQ-ETD instrument.
Fig.1-8. Formation and interpretation of a, b, c, x, y and z fragments in peptides

a, b or c ions represent N-terminal fragments with charges. x, y or z ions represent C-terminal fragments with charges. The subscript number indicates the number of amino acid residues in the fragment.

Fig.1-9. CID fragmentation pathway [91]
Fig.1-10. MS2 and MS3 spectrums of glycosylated tryptic peptide of tissue plasminogen activator, using the hybrid LTQ-ETD instrument.

Panel A: MS2 targeting of the m/z 1059.45 ion. Signals from consecutive losses of Mannose (from 1 to 5) are detected in CID-MS2. Triplet charged peptide with GlcNAc ions (1074.12, 3+) are identified and isolated for CID-MS3. Panel B: MS3 scan of the m/z 1074.12 ion. B/Y ions resulted from peptide backbone cleavage are annotated in the figure. Some y ions are in the form of peptide with the first attached GlcNAc.
In the high energy CID spectra, peptide backbone cleavages of b and y ions generate and dominate the spectral while the glycan structures are lost. However, the b and y ions generated are not traditional fragment ions from the deglycosylated peptides, instead, the b and y ions are observed by retaining the first attached GlcNAc. Alternatively, targeting analysis of CID-MS3 on the species of peptide with GlcNAc isolated from CID-MS2 can generate not only peptide backbone information by CID-MS3, but also glycan structural information by CID-MS2, which was shown in the Fig.1-10. The restriction of this method is that the analysis can only be implemented in the linear ion trap with capability of targeting MS3, instead of mass spectrometry only equipped with Orbitrap. Meanwhile, the MRM method with CID fragmentation on triple quadruple analyzer can also be used in the glycopeptide quantitation. Briefly, the instrument is firstly operated in the precursor ion scan mode for detecting all possible glycosylated precursor ions by using the diagnostic ions such as m/z 204, 274, 292, and 366. After the identification of glycopeptide precursor ions, the MRM method is subsequently employed to quantitate these glycopeptides by monitoring m/z at 204, 274, 292, 366 and specific fragmented glycan ions from particular peptides [90].

**Electron Transfer Dissociation (ETD)**

The principle of ETD reactions is based on electron transfer, involving three basic steps. 1) Production of analyte cations and reagent anions. 2) Delivery of the cations and anions into the same space within a short time. 3) Mass analysis of generated ions [92]. In ETD reactions, generation of radical anions is a key step, as radical anions prefer electron transfer. Coon et al.
tested several potential reagents for ETD reaction and finally found out that fluoranthene is the

![CID-MS2 spectrum](image1)

**Fig. 1-11.** CID-MS2 and ETD-MS2 spectrums of glycosylated tryptic peptide of monoclonal antibody, using the hybrid LTQ-ETD instrument.

Panel A: CID-MS2 spectrum of the precursor ion m/z 729.38 (z=2) ion. Consecutive losses of mannose residues (from 1 to 2 in this case) are detected and peptide backbone cleavages of b and y ions are generated in the CID-MS2 spectrum.

Panel B: ETD-MS2 spectrum of the precursor ion m/z 729.38 (z=2) ion. Peptide backbone cleavages (with mannose residues) of c and z ions are generated. Neutral losses of mannose residues are not observed in the ETD-MS2 spectrum.
favorite reagent [93]. The hybrid instrument, such as LTQ-Orbitrap ETD, introduced a back to front configuration. Anions transferred from back to front efficiently and fluoranthene radical anions generated from the rear were injected to the LTQ in the front in a stable and robust manner, which only took about 4-8 msec. ETD for analysis of various PTMs, are widely employed in the area of Serine/Threonine Phosphorylation [94], N or O-linked Glycosylation [93, 95], Glycation [96], Isoaspartic acid differentiation [96] and etc.

Electron transfer dissociation (ETD) is another widely used fragmentation approach in the glycopeptides analysis, because in ETD fragmentation, extensive peptide backbone cleavages can be obtained, but keep glycan moieties intact. ETD dissociates the N-Cα bond and produces fragment C/Z ions. One of the major advantages of ETD in proteomics and protein characterization is the ability to localize the exact site of PTMs that can be missed by CID, therefore, a combination of CID and ETD fragmentation strategies can produce a rich of information including peptide sequence and glycan structure [93].

ETD bears obvious superiority on fragmentations of large peptides, which often have higher charges (z≥3) [97-99]. In the data-dependent acquisition, if the precursor ions of relative short peptides with lower charge states are automatically selected for fragmentation, the ETD-MS2 spectrum generated is often of limited information, instead, charge-reduced (odd electron) species are dominated in the spectrum. In this case, an additional CID activation step on the charge-reduced species can be further targeted and extensive information from product ions (c and z ions) can be generated with more amplified intensity compared to those in the ETD spectrum. ETD fragmentation is very useful in the glycosylation analysis. In the previous published paper [79], an on-line LC-MS approach (CID MS2→ETD MS2→CRCID) was implemented to identify both the glycosylation sites and peptide backbone information. This
approach can generate glycan structural information in CID-MS2 step and peptide backbone information in the ETD and CRCID steps, without deglycosylation steps in the sample preparation (PNGase F treatment). Especially in the identification of O-linked glycopeptides, this LC-MS approach is more helpful since no efficient enzymes available so far to remove O-glycans. The paper illustrated two impressive examples of identification of N-linked glycopeptides in EGFR and O-linked glycopeptides in t-PA. This approach was demonstrated to provide maximal information on the glycostructure without the knowledge of the attached glycan molecular weight and without de-glycosylation step [79].

1.3.1.6 Conclusion

Even though data interpretation of glycopeptides is still challenging and time-consuming, especially for glycoproteins with multiple glycosylation sites, this method allows for the determination of glycan structures related to specific site and this is the major strength of this approach, relative to the released glycan approach. In addition, glycopeptide approach consumes low amount of materials (20µg glycoproteins) by using RPLC-MS, compared with glycan analysis (>100ug) by HILIC-MS because glycan analysis depends on how heavily the protein is glycosylated, while the glycopeptides analysis can avoid this issue. Researchers have already shown that glycopeptides analysis can satisfy diverse and extensive needs in recently years and still keeping a fast pace in developing solutions to problems found in sample preparation, MS analysis and especially data interpretation with the aid of software. Thus, there is no doubt that glycopeptides analysis will continue to blossom into an essential component of basic biological and biochemical research.
1.3.2 Strategy of glycan analysis by LC-MS

Glycan analysis plays an increasingly crucial role in biological research, clinical analysis and pharmaceutical production [101, 102]. Many glycosylation patterns have been discovered to associate with disease or therapeutic efficacy in the past years. There are various approaches developed for glycan analysis. The released glycans (Native glycans, unlabeled) are hydrophilic molecules without chromophores, so native glycans cannot be detected by UV or fluorescence detection, instead, native glycans can be determined by mass spectrometric analysis, based on precursor mass and MS/MS fragments. In order to facilitate other routinely used detection methods, such as fluorescence detection, glycans are usually subjected to derivatization. Derivatized glycans, containing a chromophore or fluorophore in the structure, are more hydrophobic than native, unlabeled glycans, which increase or decrease the interactions between the glycans and the stationary phase of particular chromatography. Separations of glycans by chromatography are widely applied in the glycan profiling, which are introduced in details in sections 1.3.2.3. Traditionally, sequential treatments of exoglycosidases have been widely used for glycan structure determination. After serial exoglycosidases digestions, the glycan structure can be deduced by detecting the shifts of glycan elution in LC or m/z difference in MS measurement [100]. Today, glycan composition, sequence, branching or linkage information can be directly obtained by mass spectrometry using CID fragmentation in both negative mode and positive mode, which made the glycan analysis easier than before. However, lost of site information is the intrinsic disadvantage in the glycan analysis. In addition, the derivatization process is tedious, especially for the cleanup step, proficient sample preparation skills are highly demanding to minimize inevitable sample loss [100].
In this part, enzymatic strategies for glycan release and various chemical derivatization methods are summarized first, followed by current analytical techniques used for glycan analysis.

### 1.3.2.1 Enzyme strategy

Various enzymes (glycoamidases, endoglycosidases and exoglycosidases) used for removal of glycans from glycoproteins have been commercially available (Table 1-5) and the most popular enzyme is peptide N-glycosidase F (PNGase F), which is the most effective enzymatic method to release all protein-bound N-linked carbohydrates except for an α1-3 Fucose residue linked to the innermost GlcNAc residue. This type of glycan is most commonly found in plant and some insect glycoproteins, rather than human glycoproteins. Instead of PNGase F, glycans with α1-3 Fucose is found to be sensitive to PNGase A [100].

Sialidase (α2-3, 6, 8 Neuraminidase) is another important exoglycosidases in the glycan analysis of glycoprotein characterization, which is used for removal of terminal sialic acids with all types of linkages (α2-3, α2-6, and α2-8). Sialidase treatment plays a crucial function in converting all negatively charged glycans to neutral glycans, while high-mannose structures are not affected. Then neutral and desialylated glycans (uncharged) are detected in MS positive mode. In addition, to separate glycans by antennary structures or fucosylation level with hydrophilic interaction chromatography, these glycans have to be desialylated by sialidase digestions. Thus, enzymatic strategies of PNGase F digestion with or without Sialidase are often applied for different purpose in the released glycan analysis. There are also many exoglycosidases available with substrate specificity. For example, α2-3 Neuraminidase prefers catalyzing the hydrolysis of α2-3 linked sialic acid residues at terminal from oligosaccharides than sialic acid with other linkages, such as
α2-6, α2-8 linkages. Those enzymes with specific activity provide a fast distinction of isomers, with different linkages [100].

Endoglycosidases are usually more specific than glycoamidases. Endo H is the popular endoglycosidases used for releasing N-linked glycans, but limitations of this enzyme is that it can only hydrolyze the glycosidic bond between the two GlcNAc residues with high-mannose and hybrid structures. Instead, Endo F is another alternative candidate, which also cleaves between the two GlcNAc residues of the chitobiose core, but for biantennary complex N-glycans and high-mannose glycans as well. In addition, the information associated with the core fucose is lost by using endoglycosidases, therefore, this type of enzymes are considered to be invaluable for a complete characterization of glycan structures [100].

Unlike N-glycans, there is an absence of a generic enzyme, which cleaves all O-linked oligosaccharides from any protein. Enzymatic release of O-glycans is restricted by a few enzymes available and their narrow substrate specificities. Therefore, O-glycans of glycoproteins are commonly removed by chemical cleavage methods, such as β-elimination.
Table 1-5. Enzymes applied for glycan release [100].

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>EC number</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-Glycanase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PNGase F (Peptide-N-glycosidase F, glycoamidase F)</td>
<td>3.5.1.52</td>
<td>Cleaves between Asn of oligomannose, complex, or hybrid N-glycans (except for α 1-3 fucose is present); requires at least one amino acid at both the amino terminal and carboxyl terminal of Asn</td>
</tr>
<tr>
<td>PNGase A (Peptide-N-glycosidase A, glycoamidase A)</td>
<td>3.5.1.52</td>
<td>Cleaves between Asn of oligomannose, complex or hybrid N-glycans; will work if core is α 1-3 fucosylated, but ineffective when sialic acid is present; requires at least one amino acid at both the amino terminal and carboxyl terminal of Asn</td>
</tr>
<tr>
<td>Endo-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endoglycosidase H</td>
<td>3.2.1.96</td>
<td>Cleaves between the two N-acetylglucosamine residues in the core of high mannose N-glycans</td>
</tr>
<tr>
<td>Endoglycosidase F</td>
<td>3.2.1.96</td>
<td>Cleaves between the two N-acetylglucosamine residues in the core of high mannose or biantennary complex N-glycans</td>
</tr>
<tr>
<td>Exo-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sialidase (neuraminidase)</td>
<td>3.2.1.18</td>
<td>Nonreducing terminal sialic acids</td>
</tr>
<tr>
<td>Fucosidase</td>
<td>3.2.1.51</td>
<td>Nonreducing terminal fucose</td>
</tr>
<tr>
<td>β-galactosidase</td>
<td>3.2.1.23</td>
<td>Nonreducing terminal β-galactose</td>
</tr>
<tr>
<td>α-mannosidase</td>
<td>3.2.1.24</td>
<td>Nonreducing terminal α-mannose</td>
</tr>
<tr>
<td>β-N-Acetylhexosaminidase</td>
<td>3.2.1.52</td>
<td>Nonreducing terminal HexNAc</td>
</tr>
</tbody>
</table>
1.3.2.2 Derivatization methods

In contrast to the case of peptides and proteins, the MS detection of native oligosaccharides remains a difficult task due to the poor ionization efficiency and low response factor. Various derivatization strategies have been developed to enhance the ionization efficiency and response factors of oligosaccharides, leading to a more sensitive LC-MS detection. In addition, derivatized glycans containing a chromophore or fluorophore facilitates fluorescence detection after chromatographic separation, which is easy to monitor glycan profile in QC environment. The most common derivatization methods employed are the reductive amination and permethylation. Alternatively, Michael addition or hydrazide labeling may also be applied. In this thesis, only reductive amination and permethylation approaches are discussed.

Reductive amination

In the reaction of reductive amination, glycan species are required to have reducing ends, in aldehyde form, rather than amine form and then primary amine groups of the labeling reagent are reacted with aldehyde group in glycan structures by condensation reaction, yielding a secondary amine though an imine intermediate. Most prevalently used labeling reagents are 2-aminobenzamide (2-AB), 2-aminobenzoic acid (2-AA), 2-aminopyridine (PA), and 1-aminopyrene-3, 6, 8-trisulfonic acid (APTS). The major merit of this labeling approach is that each glycan structure can only be introduced with one single fluorescence tag, so a straightforward quantitation in terms of fluorescence absorbance is available [103].

2-AB is the labeling reagent without negative charges, which is commonly used for HILIC-MS analysis. Relative to unlabeled native glycans, 2-ab labeled glycans can achieve a more efficient chromatographic separation and more sensitive mass spectrometric detection. Also, as 2-ab
labeled glycans have equal mass difference (120Da) for all structures, an extensive well-developed database can support glycan structural assignments in mass spectrometric analysis.

With HILIC coupled to fluorescence detection, structural identifications are retrieved to standardized elution positions [103,104]. In contrast to 2-AB labeling, the 2-AA labeling reagent introduces one negative charge to each glycan structure, which can also be employed in the capillary electrophoresis separations with either fluorescence detection or negative-mode mass spectrometric detection, allowing detection of both neutral and sialylated glycan species [105].

APTS, with three negative charges on the sulfonic groups, is well suited in CE analysis, however, MS analysis of APTS-labeled glycans are very difficult [106].

**Permethyla**

In permethylation, methyl groups are transferred to hydroxyl groups, amine groups and carboxyl groups of oligosaccharides replacing the original hydrogen atoms. The significant advantage of permethylation is that sialic acids can be neutralized and stabilized by the methyl ester formation, preventing the loss of sialic acids in MALDI or electrospray analysis. Also, permethylated glycans have enhanced measurement sensitivity in positive ion mode and the generated spectrals are easier to interpret during off-line MSn analysis based on the sequential loss of mono- or oligosaccharides. A complete permethylation facilitates simultaneous quantitation of both acidic glycans and neutral glycans in the positive ion mode. However, partial permethylation always occurs in the experiment because the more branched and larger glycan structures are less effectively permethylated.

The permethylation condition has been optimized by several groups [107-109], and the best permethylation efficiency has been achieved with a solution, consisting of 60% DMSO, 37.2%
methyl iodide and 2.8% water for 2-AB labeled glycans and 70.8% DMSO, 26.4%, and 2.8% water for native glycans. DMSO and methyl iodide need to be taken from original bottle by syringe to prevent absorbing extra water from air. Compared to native glycans, 2-ab labeled glycans have four additional substituted hydrogens, so the percent of methyl iodide (the donor of methyl group) is higher. Also, the suspended 2-ab glycan solution is loaded to the spin column, packed with sodium hydroxide, centrifuge, take and re-load to the spin column again for 15 times to maximum yield the fully permethylated glycans. However, in the permethylation of native glycans, only 8 times recycles are needed and over-reaction will lead to an oxidative degradation and unwanted breakages of glycosidic bonds [108].

Another derivatization method, mediated by DMT-MM/methanol, in which sialic acid residues in the glycan structure are stabilized and derivatized in a linkage specific manner. A senior student in our group published a paper last year and introduced a nano-HILIC-Orbitrap-MS platform coupled with DMT-MM derivatization. After derivatization reaction, α 2,3-linked sialic acid produce cyclic lactones, resulted in loss of 18 Da, while α 2,6-linked sialic acid becomes methylated and lead to an addition of 14Da mass. Therefore, the linkage information can be directly expressed from precursor mass [113].

1.3.2.3 Separation methods

Typically employed separation methods in glycan analysis involve hydrophilic interaction chromatography, reversed phase chromatography and porous graphitized carbon chromatography. Each method is well suited to couple with mass spectrometry, which is the focus of the section. Other techniques, such as high-performance anion-exchange chromatography with pulsed amperometric detection (HPAEC-PAD), capillary electrophoresis-laser induced fluorescence
(CE-LIF) or CE-MS, are also broadly used for carbohydrate analysis, but will not be discussed in the thesis.

**Hydrophilic interaction chromatography**

Hydrophilic interaction chromatography is an effective separation mode for glycans, as glycans are of hydrophilic properties and can be retained very well. In contrast to reversed phase chromatography, water is the elution solvent. Labeled glycans, containing hydrophobic groups in the structure, are more hydrophobic compared with native glycans, so the retention time of labeled glycans shifts earlier than unlabeled, native glycans. In addition, Glycans with larger size are eluted later than smaller glycans, due to the relation between size and hydrophilicity. HILIC column is also reported to have certain selectivity for isomers, such as Waters BEH columns, with great capability of distinguishing branches or linkage isomerism. However, not all glycan isomers can be separated on Waters BEH column. For example, A1G2F (211 glycan) and A1G2F’ (211’glycan) can only be partially resolved, while A1G1F and A1G1F’ glycans can be completely separated based on different branches. Additionally, the major drawback of Waters BEH column is that heavily sialylated glycans cannot achieve good separations due to limitations of its inherent separation mechanism, compared with other commercially available columns.

Therefore, Water BEH column is well suited for monoclonal antibody with relatively simple glycan profile (i.e. human IgG), rather than recombinant glycoproteins with complicated glycan patterns (i.e. TPA & TNK).

Instead, complex glycan samples with highly sialylated glycans can achieve a remarkable resolution on a weak anion-exchange resin in HILIC mode, indicating a promising application of glycan separations in mix mode column [110]. GlycanPac AXH columns were newly developed
by Thermo Fisher in 2012, which have advanced superiority on separation of charged species containing multiple sialic acids. Mixed-mode stationary phase combined the separation mechanisms of weak anion exchange and HILIC. Basically, glycans are separated by the number of negatively charged sialic acids in the structure and within each charge group, glycans with different branches are also separated. With the unique separation pattern, this column is very useful in the glycan profile characterization and quantification. Both labeled and native glycans can have good separations, based on the charge, size and polarity, but labeled glycans are partially resolved with corresponding unlabeled glycans (resulted from incomplete labeling) on GlycanPac AXH column, which is different with that on Water BEH column. Labeled glycans are largely separated with the corresponding unlabeled glycans on Water BEH column, which complicate the base peak recognition of each glycan type.

Combined with fluorescence detection, only labeled glycans can be identified. However, with mass spectrometric determination, both labeled and unlabeled glycans can be detected and the separation pattern on GlycanPac AXH column will generate a simpler and more straightforward chromatogram, which facilitates manual interpretation of MS data.

**Reversed phase liquid chromatography**

Reversed phase liquid chromatography generally is not used in the glycan analysis, because hydrophilic glycans cannot be retained very well on the reversed phase column. However, in some cases, labeling strategies coupled with RPLC separation are applied. Labeling reagents can improve the hydrophobicity of glycan species, leading to the better retentivity.

In order to increase the separation efficiency of acidic glycans, ion paring reagents such as diethylamine (DEA) are added to mobile phase with formic acid. Positively charged ion pairing
reagents, will shield the negatively charged glycans, resulted in enhanced interactions between glycans and stationary phase. However, high concentration of salts makes this separation method difficult to couple with mass spectrometry detection due to signal suppression effect caused by ion paring reagents.

The GlycanPac AXR column is developed later than AXH columns. It is also silica-based column with weak anion exchange on the reversed phase. Compared with the previously introduced AXH column, the advantages of AXR column lie in the separation of isomers with different linkages.

**Porous graphitic carbon chromatography**

Chromatography with Porous graphitic carbon (PGC) as stationary phase has unique separation efficiency on polar compounds, such as oligosaccharides. PGC is composed of intertwined graphitic ribbons, which is much hydrophobic than C18 stationary phase. The surface of graphite is very flat, without any functional group. The separation mechanism is very complicated but mainly based on hydrophobic and/or dipole-dipole interactions. The retention time is determined by the shape of analytes when they are touching the graphite surface. Analytes with more planar shape, having more interaction points with flat graphite surface, thus, have longer retention time than those rigid molecules.

PGC stationary phase is available in two formats: chip and column for glycan analysis. Glycans are analyzed mostly in reduced form on PGC. The dominant advantage of PGC is that it allows separation of structural isomers, such as A1G2F (211 glycan) and A1G2F’ (211’glycan), which cannot fully separated on HILIC column. It also exhibit excellent reproducibility due to the stable physical and chemical properties. Agilent HPLC-PGC chip-Q-TOF is a robust platform
for glycan profiling of monoclonal antibody with only 6 minutes gradient. The glycan data is performed using the Agilent Mass Hunter software. The structures of glycans are identified by searching against Agilent mAb glycan database. However, PGC is not fit for glycoproteins with complex glycan pattern, such as TPA, as sialic acids containing glycans are not well separated on the column.

1.3.2.4 Ion activation strategies in MS analysis

To date, mass spectrometry has become the most useful analytical technique in the glycan analysis, with unique superiority not only on the mass measurement of the precursor ion, but also on the fragmentation spectral generated from the precursor ion. The fragmentation pattern generated can be interpreted with a rich of information regarding glycan structures and linkages.

Matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) are two commonly applied (soft) ionization methods, because other (hard) ionization methods always result in a complete destructive vaporization of glycan samples during ionization[111].

Glycan profiles are conventionally obtained by MALDI, which can give a quantitative response and negligible fragmentation for neutral glycans of varying structures. However, this technique is not able to couple with on-line liquid chromatography. Also, the sialylated glycans are easy to lose sialic acids in MALDI detection, which have to be neutralized and stabilized by permethylation derivatization for improved determination. Instead, electrospray ionization (ESI), with its capability of coupling with on-line LC, has been increasingly applied in the oligosaccharides analysis. As ESI is the ionization method we used in the lab, in this chapter, we focused on the oligosaccharides analysis of ESI ionization in positive and negative mode, with their particular fragmentation patterns.
In the positive electrospray ionization mode, oligosaccharides tend to form [M+H]^+ ions and [M+Na/K]^+ pseudo-molecular adducts. Fragmentations of those ions lead to glycosidic b- and y-type cleavages, which provide extensive information on glycan composition and sequence information, however, no linkage and positional isomer information can be obtained in the positive mode. Meanwhile, the signals of negatively charged sialylated glycans are very low in the positive mode compared with neutral glycans. Desialylated glycans, in order to study the antennary distributions, are always determined in positive mode. Permethylated glycans (negative charges have been neutralized) are detected in positive mode and applied for isomer structure elucidation.

In the negative electrospray ionization mode, mobile phase with elevated pH is used to promote glycan deprotonation or adduction with anions. 100mM Ammonium formate with pH 4.5 has been used in the glycan analysis and to reduce the formation of anions adduct, in-source collision energy (with 30%) has been optimized. Compared with positive ESI, cross-ring fragmentations of glycans are dominated in the negative ESI, which provide more structure and linkage information (e.g. D- and E-type ions) [112]. The useful information of the negative ion fragmentation spectra arises from abstraction of protons from individual and specific hydroxyl groups which promote specific electron movements leading to the diagnostic fragments as illustrated in Fig1-11 and Fig1-12.

In addition, even though the overall sensitivity of neutral glycans is lower in the negative mode, the negatively charged sialylated glycans are observed with intensely increased sensitivity. For example, 2-ab labeled highly-sialylated glycans (with 3~4 sialic acids), they have intense signals in the negative ion mode; however, they even cannot be observed in the positive ion mode.
Fig.1-12. Nomenclature for describing fragment ions from carbohydrates according to Domon and Costello. [112]

Fig.1-13. Proposed mechanism for the formation of the $^{2,4}_{A_R}$ ion from N-linked glycans. [112]
1.3.2.5 Conclusion

The analysis of glycan structures is unavoidable as the glycan structures can determine the efficacy of therapeutic drugs and in order to fully understand the structure-function relationships for glycoproteins, the glycan composition and linkage has to be elucidated with mass spectrometry, in either positive ion mode or negative ion mode, with different strengths. However, in the QC environment of pharmaceutical industry, HILIC separations with fluorescence detection can be performed in a robust manner, which is the most applied technique for routinely monitoring the glycan profile change.

1.4 References


63. Varki, A., Biological roles of oligosaccharides: all of the theories are correct. 

64. Dennis, J.W., M. Granovsky, and C.E. Warren, Glycoprotein glycosylation and cancer progression. 


p. 237-64.

67. Wu, S.L., et al., Identification of potential glycan cancer markers with sialic acid 


90. Wikipedia, Peptide fragmentation.


Chapter 2: Identification of critical glycosylation differences between Biosimilar I-TNK and Reference product G-TNK.

Industrial collaboration with Bioanalytix Inc. and Thermo Fisher Scientific Inc.

2.1 Abstract

In this work, we adopted two different LC-MS approaches to characterize the glycosylation patterns of biosimilar I-TNK and reference product G-TNK (Innovator). Glycan heterogeneity was extensively analyzed at all three N-linked glycosylation sites (Asn-103, Asn-184 and Asn-448) and specifically glycosylation site Asn-103, generated from two mutations, only contained the desired complex-type glycans instead of original high-mannose glycans. In addition, biosimilar I-TNK and reference product G-TNK showed different distributions on sialic acid-containing glycans. The quantitative comparison of major sialylated glycans was performed by our two independent approaches and consistent results were obtained. Finally, mono-sialylated positional isomers were also identified and quantitated for two products by these two approaches in a complementary manner. Two isomers can be completely resolved by reversed phase chromatography and then relatively quantitated based on peak area in the glycopeptide analysis. On the other hand, the D ions generated in MS negative-ion mode in the glycan analysis could facilitate the determination of antenarity of glycan positional isomers.

2.2 Introduction

Protein glycosylation is known to be one of the most complicated and biologically significant post-translational modifications (PTMs) [1]. The assessment of glycosylation on biological therapeutics is very critical as inappropriate glycosylation patterns can affect the efficacy of therapeutic proteins on aspects of stability, circulatory half-life, solubility and immunogenicity.
In the production of a biosimilar, which requires high similarity to the reference product, glycosylation presents the major challenge for the final approval due to its inherent complexity and heterogeneity \[1\]. Therefore, the development of an analytical platform to systematically and comprehensively characterize glycosylation patterns of biosimilar therapeutics is of great interest to the analytical community.

Recently, liquid chromatography- mass spectrometry (LC-MS) has been demonstrated to be a valuable technology in the characterization of protein glycosylation. Two approaches are commonly used to characterize glycans by using mass spectrometry \[1\]. In the first approach, glycoproteins are directly subjected to proteolysis and the generated glycopeptides are characterized \[1,2,3,4\]. By this approach, the glycan compositions are easily correlated to specific site, in terms of the peptide backbone to which they are attached \[1\]. Glycopeptides are traditionally separated by reversed phase liquid chromatography and glycoforms from the same glycosylation site are eluted at similar retention times as the elution is primarily determined by hydrophobic interactions between peptide backbones and the stationary phase \[1,9\]. Mass spectrometry, with collision-induced dissociation (CID) can be used to identify the glycan structures at individual sites because glycosidic bonds are more easily fragmented in comparison to the peptide backbone by CID \[1,2,3\].

Besides, another more convenient approach of analyzing glycans is to cleave glycans from the glycoproteins enzymatically or chemically \[1\], and then the released glycans are purified and subjected to LC-MS analysis, in either native or derivatized form \[1,2,14\]. However, this approach is limited in the analysis of a complex glycoprotein with multiple glycosylation sites because site information is lost and glycan compositions fail to correlate to their original site. Glycans are often separated by hydrophilic liquid chromatography and analyzed by mass
spectrometry in either positive- or negative-ion mode. In the positive-ion mode, composition and sequence information of glycans can be obtained in terms of generated glycosidic B-/Y- type ions; however, linkage information cannot be obtained. In the negative-ion mode, besides B-/Y-type ions, A-/X- type cross-ring fragments, glycosidic C- type ions as well as highly informative D-/E- type ion series are generated, which provides not only glycan composition information, but also information about the linkage and positional isomers.

TNK-tissue plasminogen activator (TNK-tPA) is a generic therapeutic drug, derived from nature tissue plasminogen activator (t-PA) and approved for treatment of acute myocardial infarction and ischemic stroke [1,2,3]. TNK-tPA is a glycoprotein consisting of 527 amino acids with three N-linked glycosylation sites at Asn-103, Asn-184 and Asn-448. The amino acid sequence of TNK-tPA has mutations at three locations compared to natural t-PA. Of which, two mutations contributed to the shift of glycosylation sites from Asn-117 to Asn-103, which lead to a replacement of original unwanted high-mannose structures at Asn-117 site by desired complex-type glycans at Asn-103 site [1,2]. In a previous study, a comprehensive characterization of two TNK-tPA products from innovator (G-TNK) and follow-on (I-TNK) manufactures has been performed and critical differences have been observed in glycosylation occupancy and chain cleavages at the activation sites [1]. However, the detailed glycan structures at individual sites were not defined and compared for these two products.

In this work, we adopted two different LC-MS approaches to characterize the glycosylation patterns of biosimilar I-TNK and reference product G-TNK. Glycan heterogeneity was extensively analyzed at all three N-linked glycosylation sites and critical differences in sialylated glycans were observed for two products. In addition, glycan isomers with sialic acid at different
branches were identified and quantitated, which also displayed important differences between the two products.

2.3 Experimental

2.3.1 Materials and reagents

Reference protein G-TNK (Genentech, So. San Francisco), a lyophilized product, consisted of 50 mg of recombinant tissue plasminogen activator. The follow-on product, I-TNK (Gennova Biopharmaceuticals Ltd, Hinjwadi, Pune, India), a lyophilized product, consisting of 52.5 mg of recombinant tissue plasminogen activator and similar recipients as for Genentech. Trypsin (Sequencing grade) was obtained from Promega (Madison, WI). The LC-MS grade water was purchased from JT baker (Phillipsburg, NJ) and HPLC grade acetonitrile was from Thermo Fisher Scientific (Fairlawn, NJ). Microcon YM-10 centrifugal filter unit was obtained from Millipore (Bedford, MA). PNGase F was obtained from New England Biolabs (Ipswich, MA). Guanidine hydrochloride, ammonium bicarbonate and formic acid (FA), 2-Aminobenzamide (2-AB), sodium cyanoborohydride (reagent grade, 95%), acetic acid, ammonia, dimethyl sulfoxide, were purchased from Sigma-Aldrich (St. Louis, MO).

2.3.2 In-solution digestion

For glycopeptide analysis, the protein solution (2.5µg/µL, for both G-TNK and I-TNK) was denatured by 6M guanidine hydrochloride, reduced by 10mM DTT at 56 °C for 30 min and alkylated by 50mM IAA for 60 min at room temperature in the dark. The reduced and alkylated protein solution was buffer exchanged to 100 mM ammonium bicarbonate using 10kDa molecular weight filter for three times. Trypsin was added into the protein solution at 1:25 ratio
(enzyme to protein) and incubated for 6 hours. The enzymatic reaction was stopped by 5% formic acid and the digest was stored at -80°C until analysis.

For glycan analysis, the protein solution (2.5µg/µL, for both G-TNK and I-TNK) was denatured by 6M guanidine hydrochloride, reduced by 10mM DTT at 56 °C for 30 min and alkylated by 50mM IAA for 60 min at room temperature in the dark. The reduced and alkylated protein solution was buffer exchanged to 100 mM ammonium bicarbonate using 10kDa molecular weight filter for three times. (Same as glycopeptide analysis). PNGase F was added into the protein solution at v/v=1:20 and incubated at 37°C overnight. The enzymatic reaction was stopped by 5% formic acid and the digest was stored at -80°C until analysis.

2.3.3 Fluorescent labeling with 2-aminobenzamide

Released glycans were labeled with 2-aminobenzamide (2-AB) fluorescent dye (50mg/mL) via reductive amination in the presence of sodium cyanoborohydride (60mg/mL) in 70:30 v/v DMSO/acetic acid solution at 65°C for 2-3 hours. Then 90µL water and 900µL acetonitrile was added to the reaction mixture. Excess fluorescent label was removed by PhyNexus Normal Phase PhyTips.

2.3.4 LC-MS analysis

For glycopeptide analysis, An Ultimate 3000 nanoLC pump (Dionex, Mountain View, CA) and a self-packed C18 column (Magic C8, 200Å pore and 5 µm particle size, 75 µm i.d. × 15 cm) (Michrom Bioresources, Auburn, CA) was coupled on-line to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source (New Objective, Woburn, MA). The mobile phases used were as follows: (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile. In the sample loading, the flow rate was
maintained at 250 nL/min and 2% B for 25 minutes and decreased to 200 nL/min at 2% B for 10 minutes. And in the elution and separation, the gradient was maintained at 200 nL/min flow rate and increase linearly from 2% to 40% B over 60 minutes and then increase from 60% to 80% over 10 minutes, and finally maintained isocratic at 80% for 10 minutes. The LTQ-Orbitrap MS was operated as follows: survey full-scan MS spectra (m/z 400 – 2000) were acquired in the Orbitrap cell with mass resolution of 100,000 at m/z 400 (with the ion target value of 2×10⁶ ions), followed by 8 sequential MS2 scans using LTQ portion. For an inadequate assignment, the desired ions were subjected to targeting analysis to obtain additional information. For targeting CID-MS2 analysis, the isolation width window was set at 5 m/z and the CID was set at 35% with an activation Q=0.25 for 10 ms.

In the glycan analysis, two types of commercial columns were used. For Thermo GlycanPac column (Thermo Fisher Scientific, Waltham, MA), the mobile phases used were as follows: (A) 80% Acetonitrile in water and (B) 80mM Ammonium formate (pH=4.4). The column was balanced at 0.2mL/min and 2.5% B for 15 minutes and increased to 12.5% for 30 minutes, then increase to 25% within 5 minutes, and further increase to 37.5% for another 5 minutes. The separation temperature was 30°C. For Waters BEH column (100mm x 2.1 mm, 1.7 µm BEH particles) (Waters, Milford, MA), 100 mM ammonium formate (pH=4.4) was used as solvent A and acetonitrile as solvent B. The separation gradient started from 75% acetonitrile and decrease to 62% acetonitrile at a flow rate of 0.4 mL/min in a 20 min analytical run. The separation temperature was 40°C. The FT-MS range was set to 380-2000 in negative ion mode. For the heated ESI source, the heater temperature was 350 °C and the capillary temperature is 285°C, the sheath gas flow rate setting was 35arb and the aux gas flow rate setting was 10arb. Glycan
samples were prepared in 80:20 acetonitrile/water and maintained at 5°C in the thermostatted autosampler prior to injection.

2.3.5 Data analysis

For the peptide mapping of both G-TNK and I-TNK, the raw data with acquired MS/MS scans were searched against the single protein database using the SEQUEST algorithm incorporated in the software Proteomic Discovers (version 1.3, Thermo Fisher Scientific). The peptide mass tolerance was set to 1.5 Da, and the fragment ion mass tolerance was 1.0 Da. Peptides are automatically identified if the peptide has an Xcorr score (+1 > 1.5; +2 > 2.0; +3 > 2.5) with trypsin cleavage at both ends. LTQ-Orbitrap provides accurate mass measurement (<5 ppm) for confirmation of peptide identification. Thus, the Xcorr scores (+1 > 1.5; +2 > 2.0; +3 > 2.5) were used for the initial screening. Further, manual inspections were also performed for the MS/MS spectral of each peptide, to exclude the false positive or false negative results generated by software. For glycan structure identification, theoretical masses of glycan structures were added to the enzymatic peptide backbone. The anticipated glycopeptide masses with different charges were extracted from the LC-MS chromatogram and the observed precursor mass from raw data should be less than 5ppm mass accuracy to the theoretical mass.

2.4 Results and Discussion

Biosimilar I-TNK and its reference product G-TNK have three identical N-linked glycosylation sites (Asn-103, -184, and -448). As mentioned in the introduction, glycosylation occupancy has been studied, but the full glycan structures were not characterized. However, a detailed understanding of glycosylation patterns is critical to studying the efficacy of therapeutic drugs.
Thus, to establish the similarity of biosimilar I-TNK to reference product G-TNK, the glycosylation patterns of the two products are analyzed and compared. Glycan structures described below are presented in “three-number rule”, the first Arabic number represents branching, the second Arabic number represents the number of fucose, and the third number represents the number of sialic acids.

2.4.1 **Identification of complex-type glycan structures at three N-linked glycosylation sites for both biosimilar I-TNK and reference product G-TNK.**

In the analysis of glycopeptides, the glycans identified at three sites (Asn-103, -184, and -448) were all predominantly complex biantennary structures with lower amounts of complex tri- and tetra-antennary structures, for both I-TNK and G-TNK. Particularly at Asn-103 site, generated from mutation, only complex-type glycans were identified instead of original high-mannose glycans, which were found in t-PA. As shown in the Fig.2-1, the high-mannose type glycan at Asn-117 in t-PA with accurate precursor mass (m/z=1059.4492) was measured by FTMS (As insert in Fig.2-1A) and further fragmented by CID-MS2 (Fig.2-1A), the characteristic fragmentation pattern was observed with consecutive losses of 1 to 5 mannose residues from the mass of the precursor ion. The high-mannose type glycans was completely changed to the complex-type glycans at Asn-103 site, 211 glycan was shown as an example in Fig.2-1B for G-TNK and Fig.2-1C for I-TNK. The accurate precursor ion (m/z=1277.0297) was also shown as inserts and the fragmentation by CID-MS2 was illustrated as labels. The typical ion observed at m/z 657 in the spectrum indicated the presence of sialic acid in the structure. The ion m/z 1483.25 was also indicated the neutral loss of NGlc-Gal-SA.
Similarly, various complex-type glycan structures extended from bi-antennary to tri-antennary, with 0~3 terminal sialic acids were identified at Asn-103 site for two products. Fig.2-2 listed major identified glycans at Asn-103 site of G-TNK and I-TNK with their corresponding molecular mass and relative abundance. Identified glycans at Asn-184 and Asn-448 sites were also summarized in the Supplemental Fig.S2-1 and Fig.S2-2.
Fig. 2-1. LC-MS analysis of the high-mannose glycan structure at Asn-117 from tryptic digests of t-PA was changed to the complex-type glycan structure at Asn-103 from tryptic digests of G-TNK and I-TNK. A: CID-MS2 spectrum of the tryptic glycopeptide with Man5 at Asn-117 in t-PA, m/z 1058.95 (4+), with the FT-MS measurement of the precursor ion in the insert. B: CID-MS2 spectrum of the tryptic glycopeptide with 211 Glycan at Asn-103 in G-TNK, m/z 1058.95 (4+), with the FT-MS measurement of the precursor ion in the insert. C: CID-MS2 spectrum of the tryptic glycopeptide with 211 Glycan at Asn-103 in I-TNK, m/z 1058.95 (4+), with the FT-MS measurement of the precursor ion in the insert. Key to symbols: green circle, mannose residue; yellow circle, galactose; blue square, N-acetylglucosamine; the purple diamond, sialic acid.
Fig.2-2. Summary of major identified glycans at Asn-103 of G-TNK and I-TNK with their corresponding molecular mass and relative abundance.

A. Major glycans identified at glycosylation site Asn-103 in G-TNK and I-TNK with their corresponding molecular mass and relative abundance. B. Glycoform distribution at Asn-103 in G-TNK. C. Glycoform distribution at Asn-103 in I-TNK.
In order to further confirm the diversity of complex-type glycan structures, N-linked glycans of two products were released by PNGase F and then subjected to LC-MS analysis respectively, without considering the specific site of glycosylation. In this approach, released glycans were derivatized by 2-Aminobenzamide and profiled on Thermo GlycanPAc AXH column. Fig.2-3 lists all major identified glycans with their corresponding molecular mass and relative abundance. In this approach, high-mannose glycan structures were not identified as well, which lead to a consistent result that the elimination of high-mannose structure on Asn-103 site is complete. Thus, the glycan profiling by this approach can double confirm both products have complex-type glycans from biantennary to tetra-antennary with up to four sialic acids.
Fig. 2-3. Summary of major identified glycans with their corresponding molecular mass and relative abundance in G-TNK and I-TNK.

A. Major glycans identified at glycosylation site Asn-103 in G-TNK and I-TNK with their corresponding molecular mass and relative abundance. B. Glycoform distribution at Asn-103 in G-TNK. C. Glycoform distribution at Asn-103 in I-TNK.
2.4.2 Sialic acid difference at particular site

In the positive ESI-MS analysis, the positive charges mainly attach to the glycopeptide backbone, so the intensity of each glycoform should primarily depend on the peptide portion of the molecule. It is then quite rational to assume that the response factors for different glycoforms with same peptide backbone are similar. Also, these various glycoforms were identified simultaneously in the same LC-MS run. Thus, the relative quantitation for glycoform distributions at individual sites can be easily performed. As shown in Fig.2-2, the percent of glycoform at Asn-103 is estimated by each individual glycoform divided by the sum of all glycoforms peak area, utilizing a similar method as described before [21]. As a result, we observed the most abundant glycoform at Asn-103 site for biosimilar I-TNK is 212 glycan (~41%), followed by 211 glycan (~25%) and 210 glycan (~11%). However, for the reference product G-TNK at this site, the most abundant glycoform is 211 glycan (~38%), followed by 210 glycan (~28%) and 212 glycan (~15%). The reported percentages were the average of 3 measurements with %CV less than ~10% in every measurement.

Similarly, using the method depicted above, we then calculated the relative percent of various glycoforms on the other two sites for both products. As a result, even though a similar trend of glycan structures were observed at Asn-448 site, a different distribution for major sialic acid-containing glycans was still observed at the Asn-184 site (Fig.S2-1 and Fig.S2-2). That is, for biosimilar I-TNK, 212 glycan is the most abundant glycoform (~24%) in I-TNK, followed by 211 glycan (~20%) and 312 glycan (~16%), but for reference protein G-TNK, 211 glycan is the predominant glycoform (~36%), and followed by 312 glycan (~21%) and 212 glycan (~14%). Therefore, critical differences in sialylated glycans were observed in I-TNK at both Asn-103 and Asn-184 sites, relative to the reference product G-TNK. The difference we observed should
raise a concern as to whether the products are biosimilars, because different distributions of sialylated glycans will definitely affect the efficacy of therapeutic proteins.

In addition, rather than comparing the major sialylated glycoforms in the glycopeptide approach, the major glycan structures could also be quantitated and compared in the released glycan approach. As shown in Fig.2-3, we still observed the different glycan distribution for biosimilar I-TNK and reference protein G-TNK. The 211 glycan was measured as 35% for G-TNK and 31% for I-TNK, while the 212 glycan was measured as 24% for G-TNK and 35% for I-TNK, which also demonstrated that the most abundant glycoforms in G-TNK is 211 glycan, but in I-TNK is 212 glycan. Therefore, the critical differences of major sialylated glycans for biosimilar and reference product were confirmed by two independent approaches, in either glycopeptide analysis or glycan analysis.

2.4.3 Determination of branch differences of sialic acid

In addition to the variation observed for sialylated glycans, subtle but distinct difference of positional isomers was also observed for 211 glycans at Asn-448 site between biosimilar I-TNK and reference product G-TNK. In the analysis of Asn-448 site, the glycoform with the 211 glycan structure has an m/z of 1063.44 (z=3). Two peaks were extracted from the LC chromatogram based on the accurate precursor mass as indicated in Fig.2-5A for G-TNK and Fig.2-5B for I-TNK. These two resolved peaks represent two isomeric glycan structures with different branch of sialylation, i.e., sialic acid on a Man (α1-6) or Man (α1-3) antennae. However, there are no practical MS techniques so far that can differentiate branch-specific glycan isomers with peptide backbone. As a result, we differentiated the isomeric glycan structures by a glycan-based approach. In MS negative ion mode, CID can generate branch diagnostic D ion series.
along with other cross-ring cleavages in the MS/MS spectrum, but the prerequisite for differentiation of two isomers is that the base peaks have to be resolved or partially resolved on HILIC column, which can be further targeted for detailed MS/MS analysis. Unfortunately, the previously used HILIC column (GlycanPac AXH) cannot achieve baseline resolution for the 211 glycan isomers. Thus, we used another HILIC column from Waters instead, which is well-known for the capability of separating glycan isomers. As shown in the Fig.2-4A, 2-ab labeled 211 glycan has a m/z of 1097.91 (2-), which was extracted from the LC-chromatogram. Even though two isomeric glycan structures were only partially resolved by the HILIC column, we could still select two skim regions at pre-peak and post-peak, comparing the subtle difference based on different D ion series in the MS/MS spectrums. The D ions of 979(1-) and 961(1-) (Fig.2-4B) are diagnostic ions for 6-arm antenna, suggesting that the sialic acid was on the 6-arm branch. Also, different D ions: 689(1-) and the D-18 ion 671(1-) are shown (Fig.2-4C), suggesting that the sialic acid was on the 3-arm branch. Thus, the 211 glycan with sialic acid on Man (α1-6) branch was eluted earlier than the glycan with sialic acid on Man (α1-3) branch by the HILIC phase. Based on a reversed phase separation, the elution order of glycopeptides with isomeric glycan structures should be reversed, which indicates that the first resolved peak in the glycopeptide analysis is for 211 glycan with sialic acid on Man (α1-3) branch and the second peak is for sialic acid on Man (α1-6) branch.

After elucidation of two isomer structures with branch specificity, quantitation was performed in the glycopeptides analysis. The % of monosialylated glycan with sialic acid on Man (α1-6) branch could be determined by the ratio of Man (α1-6) branch divided by the sum of Man (α1-6) branch and Man (α1-3) branch peak area. As shown in the Fig.2-5C, for the reference product G-TNK, sialic acid on Man (α1-3) branch exhibited 3 fold differences than on Man (α1-6) branch.
However, sialic acid on Man (α1-3) branch was determined to have equivalent amount as Man (α1-6) branch in biosimilar I-TNK.

Fig. 2-4. Differentiation and Identification of positional isomers for 211 Glycan at Asn-448 site. A. extracted ion chromatogram (XIC) of the glycopeptides containing Asn-448 in G-TNK and I-TNK. B. CID-MS2 spectrum of the pre-peak of anticipated glycan precursor ion. C. CID-MS2 spectrum of the post-peak of anticipated glycan precursor ion.
Fig. 2-5. Relative quantitation of the abundance of positional isomers for 211 glycan at Asn-448 site. A: extracted ion chromatogram (XIC) of the glycopeptides with 211 glycan in G-TNK. B: extracted ion chromatogram (XIC) of the glycopeptides with 211 glycan in I-TNK. C: Determination of the abundance of positional isomers for 211 glycan at Asn-448 site.
However, except for the Asn-448 site, we only found one peak for the 211 glycan structure at Asn-103 and Asn-184 sites. There was also a consideration about the influence of peptide length on the separation of glycopeptides with isomeric glycan structures, so we employed multi-enzyme strategies to cut the peptide backbone with Asn-103 site to a similar length as that with the Asn-448 site. Still, the corresponding glycopeptide with 211 glycan was also showed as one peak (Data not shown). The same result was also observed for the glycopeptide with Asn-184 site. Thus, we conclude that only the Asn-448 site has isomeric glycan structures, rather than Asn-103 and Asn-184 sites.

2.5 Conclusion

The comprehensive characterization and comparison of glycosylation patterns between biosimilar and reference product were achieved by two independent LC-MS approaches, in analysis of glycopeptides as well as released glycans. Both products were identified with complex-type glycan structures at three N-linked glycosylation sites, including AsN-103 site, which is generated by mutation. In addition to the glycan structures, we also focused on the analysis of sialylation, which exerts significant influence on the efficacy of therapeutic drugs by mainly affecting serum half-life and immunogenicity. However, a concern was raised from the results, since biosimilar I-TNK and reference product G-TNK showed different degrees of sialylation. The quantitative comparison of major sialylated glycans was performed by our two independent approaches and consistent results were obtained. In the glycopeptides approach, the quantitation of two major glycoforms was performed at individual site, which makes the data analysis time-consuming. However, the sensitive glycopeptide approach bears the merits of lower sample consumption and easier sample preparation compared to glycan analysis. On the other hand, in glycan analysis, irrespective of site information, major sialylated glycans were
derivatized and identified by mass spectrometry, which makes the data analysis easier. However, more tedious sample preparation and larger sample consumption are required for the glycan approach.

In addition to the orthogonal comparison of sialylated glycans, monosialylated positional isomers are also identified and quantitated for two products by these two approaches in a complementary manner. Two isomers can be completely resolved on reversed phase and then relatively quantitated based on peak area in the glycopeptide analysis but they cannot be differentiated due to identical MS/MS spectra. On the contrary, even though the two isomers cannot be quantitated due to the partially resolved peaks, but in terms of D ions generated in MS negative-ion mode, the antenarity of two isomers can be determined.

In conclusion, we have demonstrated that the powerful analytical platform, combining two LC-MS approaches as described here, allows for the extensive characterization and quantitative comparison of the glycosylation pattern between two products. The two independent but complementary approaches, applied orthogonally can guarantee a high degree of confidence and sensitivity to pinpoint the subtle but distinct differences between biosimilar and reference products. Thus, the use of LC-MS as a proteomic technique is proven once again to be highly valuable in the characterization of glycosylation pattern and comparison between biosimilar and reference products.
2.6 References


A: Major glycans identified at glycosylation site Asn-184 in G-TNK and I-TNK with their corresponding molecular mass and relative abundance.

<table>
<thead>
<tr>
<th>Site</th>
<th>Glycoform</th>
<th>[M+H]+</th>
<th>G-TNK</th>
<th>I-TNK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn-184 (T17)</td>
<td>211</td>
<td>5132.92</td>
<td>35.6±4.2</td>
<td>20.4±1.2</td>
</tr>
<tr>
<td></td>
<td>212</td>
<td>5424.01</td>
<td>14.1±4.2</td>
<td>24.4±1.1</td>
</tr>
<tr>
<td></td>
<td>211-Gal</td>
<td>4970.86</td>
<td>5.9±1.2</td>
<td>0.6±0.1</td>
</tr>
<tr>
<td></td>
<td>311</td>
<td>5498.05</td>
<td>7.6±2.1</td>
<td>4.3±0.2</td>
</tr>
<tr>
<td></td>
<td>312</td>
<td>5789.14</td>
<td>21.2±2.1</td>
<td>16.1±1.0</td>
</tr>
<tr>
<td></td>
<td>313</td>
<td>6080.24</td>
<td>5.5±1.3</td>
<td>13.3±1.1</td>
</tr>
<tr>
<td></td>
<td>411</td>
<td>5863.12</td>
<td>1.5±0.3</td>
<td>0.9±0.1</td>
</tr>
<tr>
<td></td>
<td>412</td>
<td>6154.28</td>
<td>4.6±0.8</td>
<td>3.2±0.3</td>
</tr>
<tr>
<td></td>
<td>413</td>
<td>6445.37</td>
<td>1.7±0.5</td>
<td>5.2±0.5</td>
</tr>
<tr>
<td></td>
<td>414</td>
<td>6736.46</td>
<td>2.3±0.6</td>
<td>10.1±1.6</td>
</tr>
</tbody>
</table>

B: Glycoforms distribution at Asn-184 in G-TNK

C: Glycoforms distribution at Asn-184 in I-TNK

Fig.S2-1. Summary of major identified glycans at Asn-184 of G-TNK and I-TNK with their corresponding molecular mass and relative abundance.

A. Major glycans identified at glycosylation site Asn-184 in G-TNK and I-TNK with their corresponding molecular mass and relative abundance. B. Glycoform distribution at Asn-184 in G-TNK. C. Glycoform distribution at Asn-184 in I-TNK.
Fig. S2-2. Summary of major identified glycans at Asn-448 of G-TNK and I-TNK with their corresponding molecular mass and relative abundance.

A. Major glycans identified at glycosylation site Asn-448 in G-TNK and I-TNK with their corresponding molecular mass and relative abundance. B. Glycoform distribution at Asn-448 in G-TNK. C. Glycoform distribution at Asn-448 in I-TNK.

<table>
<thead>
<tr>
<th>Site</th>
<th>Glycoform</th>
<th>[M+H]+ G-TNK</th>
<th>G-TNK</th>
<th>I-TNK</th>
</tr>
</thead>
<tbody>
<tr>
<td>200-Gal</td>
<td>2589.10</td>
<td>1.03±0.16</td>
<td></td>
<td>2.55±0.15</td>
</tr>
<tr>
<td>210</td>
<td>2897.21</td>
<td>4.99±0.40</td>
<td>6.75±0.73</td>
<td></td>
</tr>
<tr>
<td>211</td>
<td>3188.31</td>
<td>17.07±1.99</td>
<td></td>
<td>22.89±1.42</td>
</tr>
<tr>
<td>212</td>
<td>3479.40</td>
<td>41.67±3.86</td>
<td>38.27±3.01</td>
<td></td>
</tr>
<tr>
<td>210-2xGal</td>
<td>2573.10</td>
<td>1.51±0.13</td>
<td>0.18±0.02</td>
<td></td>
</tr>
<tr>
<td>210-2xGal-GlcNAc</td>
<td>2370.03</td>
<td>0.67±0.07</td>
<td>0.12±0.02</td>
<td></td>
</tr>
<tr>
<td>210-Gal</td>
<td>2735.16</td>
<td>1.58±0.10</td>
<td>0.46±0.06</td>
<td></td>
</tr>
<tr>
<td>211-Gal</td>
<td>3026.25</td>
<td>3.70±0.36</td>
<td>0.26±0.03</td>
<td></td>
</tr>
<tr>
<td>301-Gal</td>
<td>3245.33</td>
<td>0.53±0.07</td>
<td>0.76±0.02</td>
<td></td>
</tr>
<tr>
<td>302-Gal</td>
<td>3536.42</td>
<td>1.33±0.27</td>
<td>1.67±0.12</td>
<td></td>
</tr>
<tr>
<td>310</td>
<td>3262.34</td>
<td>1.33±0.09</td>
<td>1.61±0.18</td>
<td></td>
</tr>
<tr>
<td>311</td>
<td>3553.44</td>
<td>2.58±0.37</td>
<td>3.03±0.56</td>
<td></td>
</tr>
<tr>
<td>312</td>
<td>3844.53</td>
<td>4.51±0.21</td>
<td>6.23±0.20</td>
<td></td>
</tr>
<tr>
<td>313</td>
<td>4135.63</td>
<td>7.21±0.69</td>
<td>6.76±0.18</td>
<td></td>
</tr>
<tr>
<td>411</td>
<td>3918.51</td>
<td>0.47±0.08</td>
<td>0.95±0.09</td>
<td></td>
</tr>
<tr>
<td>412</td>
<td>4209.67</td>
<td>0.65±0.01</td>
<td>1.60±0.04</td>
<td></td>
</tr>
<tr>
<td>413</td>
<td>4500.76</td>
<td>1.21±0.28</td>
<td>2.64±0.05</td>
<td></td>
</tr>
<tr>
<td>414</td>
<td>4791.85</td>
<td>0.46±0.10</td>
<td>0.93±0.20</td>
<td></td>
</tr>
</tbody>
</table>

B: Glycoforms distribution at Asn-448 in G-TNK

C: Glycoforms distribution at Asn-448 in I-TNK
Chapter 3: Identification of novel glycans with di-sialylated structures in α3 integrin from Mouse kidney cells with the phenotype of polycystic kidney disease.


Co-authors’ work in this chapter: Anna Fan Zhang: experimental design and perform data analysis, manuscript writing and revision; Shiaw-lin Wu: idea contribution, experimental design, data analysis; Yunjoon Jung: experimental design and perform (western blot experiments on lectin affinity precipitates), manuscript writing. Shan Qin: idea contribution, experimental design and perform (cell preparation and western blot experiments.) William S. Hancock: idea contribution, manuscript writing and revision. Jordan A. Kreidberg: idea contribution, manuscript writing and revision and grant support.

3.1 Abstract

Autosomal Dominant Polycystic Kidney Disease (ADPKD) is a common genetic disorder caused by mutations in the Pkd1 or Pkd2 genes, in which large cysts replace normal kidney tissue, leading to end stage kidney disease. In this study we have utilized a powerful nanoHPLC-mass spectrometric approach to characterize patterns of normal and abnormal N-linked glycosylation of α3 integrin subunit in Pkd1<sup>+/−</sup> cells derived from mouse kidneys. Higher molecular weight glycan structures with a different monosaccharide composition were observed at two sites, namely Asn-925 and Asn-928 sites in α3 integrin isolated from Pkd1<sup>+/−</sup> cells, compared to Pkd1<sup>+/+</sup> cells. In addition, an unusual and unique di-sialic acid glycan structure was observed solely in
Thus, these studies suggest that abnormal protein glycosylation may have a role on the pathogenesis of cyst formation in ADPKD.

3.2 Introduction

Autosomal Dominant Polycystic kidney disease (ADPKD) is a common genetic disease, initiated by mutation of either of two genes, PKD1 and PKD2. These two genes encode protein polycystin-1 and polycystin-2[1], respectively and absence of polycystins in the cells will result in cyst formation in the kidneys and gradually lead to renal failure. However, there is still no standard and effective treatment available to prevent cyst formation and conserve kidney function in individuals with ADPKD. α3β1 integrin receptor is a heterodimeric transmembrane protein that mediates cell-extracellular matrix and cell-cell interactions. Our previous study using western blots indicated differential glycosylation of the α3 integrin subunit between Pkd1+/+ and Pkd1−/− cells. Compared to other transmembrane receptors, integrin-mediated signaling has a bi-directionality, in that integrins convey signals to the cell based on interactions with the extracellular matrix [2], but signals inside the cell also affect an integrin’s ability to interact with the matrix (inside-out signaling) [2]. Additionally, integrins are able to signal coordinately with receptor tyrosine kinases (RTK) to influence cellular survival, apoptosis, proliferation and differentiation[3]. α3β1 integrin has previously been shown to signal coordinately with the receptor tyrosine kinase c-Met to regulate cell survival and the expression of Wnt7b in the developing kidney [4].

More recently, we have studied the role of α3β1 integrin and c-Met in ADPKD. It was observed that most α3β1 integrin was sequestered in the Golgi, along with the E3-ubiquitin ligase c-Cbl. As a consequence of c-Cbl being sequestered in the Golgi, c-Met the receptor tyrosine kinase
target for c-Cbl, was overly abundant on Pkd1\(^{-/-}\) cells, resulting in excessive signaling through mTOR and other signal transduction pathways. To examine the molecular basis for the sequestration of \(\alpha_3\beta_1\) integrin and c-Cbl in the Golgi, we studied the structure of abnormal glycosylation of the \(\alpha_3\) integrin subunit after developing a procedure for the isolation of \(\alpha_3\) integrin from Pkd1\(^{-/-}\) cells and Pkd1\(^{+/+}\) cells. Unlike most integrin subunits, the \(\alpha_3\) subunit is post-translationally cleaved to yield a larger N-terminal peptide and a smaller c-terminal peptide (hereafter referred to as the \(\alpha_3\) integrin light chain) that remain covalently linked through di-sulfide bonds. The \(\alpha_3\) integrin light chain includes the transmembrane domain and a juxta-membrane region that has previously been demonstrated to importantly affect interactions. Here we present an analysis of glycan structures of the \(\alpha_3\) integrin light chain in murine Pkd1\(^{+/+}\) and Pkd1\(^{-/-}\) cells. The glycan structures on each site of \(\alpha_3\) integrin light chain were then characterized using state-of-the-art LC-MS techniques.

### 3.3 Materials and methods

#### 3.3.1 Reagents

Rabbit polyclonal anti-mouse \(\alpha_3\) integrin was custom prepared by Invitrogen. Biotinylated Lectins including SNA (B-1305), MALII (B-1265), PHA-L (B-1115), and UEA (BA-0064) were purchased from Vector lab. Secondary Rabbit IgG, HRP was purchased from GE Healthcare.

#### 3.3.2 Cells

Pkd1\(^{+/+}\) and Pkd1\(^{-/-}\) cell lines (used at passage 8-13) were previously described [5]. All cells were cultured in DMEM/F12 containing 1% fetal bovine serum, 10 U/ml \(\gamma\)-interferon, 5 \(\mu\)g/ml Insulin,
5 μg/ml Transferin, 0.05 μM Sodium Selenite, 100 nM Hydrocortizone, 0.002 μM T3, 100 U/ml penicillin/streptomycin, 25 ng/ml PGE1 under 33 °C and 5% CO₂.

3.3.3 Immunoprecipitation, SDS-PAGE and western blot

Immunoprecipitation and western-blot analysis were performed using whole cell lysates. Confluent cells were collected, washed with PBS, lysed with lysis buffer (20 mM Tris/Cl, 1 mM EDTA, 150 mM NaCl, 1% Triton X-100) containing proteinase inhibitor cocktail tablet (Roche 1697498) at 4°C for 30 minutes. After centrifugation at 15,700 g for 15 minutes, supernatants were incubated with specific antibodies or lectins at 4°C for 4 hour, followed by incubation with Protein G–conjugated beads (Pierce Biotechnology) at 4°C for overnight, and then samples were washed in lysis buffer. Samples were reduced by 1M dithiothreitol (DTT) at 37°C for 30 min before running on 8% acrylamide gel and transferred to PVDF membranes, followed by immunoblotting with specified antibodies.

3.3.4 In-gel digestion

The gel bands with the correct molecular weight, which were shown on the western blot image, were cut for digestion. Briefly, the gel sliced, after removed coomassie blue stain, were reduced with dithiothreitol (DTT) by the addition of 200 μL of 10 mM DTT in 0.1 M ammonium bicarbonate and incubated for 35 min at 56 °C, and alkylated with 200 μL of 55 mM iodoacetamide (IAA) in 0.1 M ammonium bicarbonate at room temperature for 1 hour in the dark. Then, a trypsin digestion buffer (12.5 ng/μL trypsin in 50 mM ammonium bicarbonate, pH 8.0) was added for 30–35 min at 4 °C, followed by further incubation overnight at 37 °C. For multienzyme (Trypsin, GluC and PNGase F) digestion, a digestion buffer containing trypsin,
GluC and PNGase F were used and incubated the same way as trypsin digestion. The digested peptides were extracted from the gel with 25 mM ammonium bicarbonate, then acetonitrile (37 °C for 15 min), and further extracted with 5% formic acid at 37 °C for 5 min. All supernatants were collected and concentrated for the subsequent LC-MS analysis.

3.3.5 LC-MS

An Ultimate 3000 nanoLC pump (Dionex, Mountain View, CA) and a self-packed C18 column (Magic C8, 200Å pore and 5 μm particle size, 75 μm i.d. × 15 cm) (Michrom Bioresources, Auburn, CA) was coupled on-line to an LTQ-Orbitrap mass spectrometer (Thermo Fisher Scientific, San Jose, CA) through a nanospray ion source (New Objective, Woburn, MA). The mobile phases used were as follows: (A) 0.1% formic acid in water and (B) 0.1% formic acid in acetonitrile. In the sample loading, the flow rate was maintained at 250 nL/min and 2% B for 25 minutes and decreased to 200 nL/min at 2% B for 10 minute. And in the elution and separation, the gradient was maintained at 200 nL/min flow rate and increase linearly from 2% to 40% B over 60 minutes and then increase from 60% to 80% over 10 minutes, and finally maintained isocratic at 80% for 10 minutes. The LTQ-Orbitrap MS was operated as follows: survey full-scan MS spectra (m/z 400 – 2000) were acquired in the Orbitrap cell with mass resolution of 100,000 at m/z 400 (with the ion target value of 2×10^6 ions), followed by 8 sequential MS2 scans using LTQ portion. For an inadequate assignment, the desired ions were subjected to targeting analysis to obtain additional information. For targeting CID-MS2 analysis, the isolation width window was set at 5 m/z and the CID was set at 35% with an activation Q=0.25 for 10 ms.

3.3.6 Data analysis
For identification of α3 integrin, the raw data with acquired MS/MS scans were searched against mouse database using the SEQUEST algorithm incorporated in the software Proteomic Discovers (version 1.3, Thermo Fisher Scientific). For the peptide analysis, the raw data with acquired MS/MS scans were searched against α3 integrin sequence (light chain sequence or heavy chain sequence, respectively.) The peptide mass tolerance was set to 1.5 Da, and the fragment ion mass tolerance was 1.0 Da. Peptides are automatically identified if the peptide has an Xcorr score (+1 > 1.5; +2 > 2.0; +3> 2.5) with trypsin cleavage at both ends. LTQ-Orbitrap provides accurate mass measurement (<5 ppm) for confirmation of peptide identification. Thus, the Xcorr scores (+1 > 1.5; +2 > 2.0; +3> 2.5) were used for the initial screening. Further, manual inspections were also performed for the MS/MS spectral of each peptide, to exclude the false positive or false negative results generated by software. For glycan structure identification, theoretical masses of glycan structures were added to the enzymatic peptide backbone. The anticipated glycopeptide masses with different charges were extracted from the LC-MS chromatogram and the observed precursor mass from raw data should be less than 5ppm mass accuracy to the theoretical mass. Specifically in the analysis of α3 integrin light chain, the data of PNGase F treated samples were searched against the light chain sequence by the software Proteomic Discovers and the deglycopeptide counterparts were identified with Asn converted to Asp (1 Da mass difference). The glycopeptide with linked glycans were eluted within a similar retention time window and the observed precursor mass were further confirmed by the corresponding CID-MS2 fragmentation.

3.4 Results

3.4.1 Glycosylation patterns of α3 integrin
To determine the global glycosylation patterns of $\alpha3$ integrin in $Pkd1^{-/-}$ cells in comparison with $Pkd1^{+/+}$ cells, western blot for $\alpha3$ integrin antibody was performed on lectin affinity precipitates. The lectins used were SNA recognizing $\alpha(2-6)$ sialic acid; MalII recognizing $\alpha(2-3)$ sialic acid; PHA-L recognizing complex glycans; and UEA recognizing fucose-containing lectins (Fig.3-1). More $\alpha3\beta1$ integrin appeared to be bound in $Pkd1^{-/-}$ cells, indicating greater modification with $\alpha(2-6)$ sialic acid. PHA-L binding was similar and neither MALII or UES appeared to binding the $\alpha3$ integrin subunit.
**FIG3-1. Glycosylation patterns of α3 integrin.** Western blot for α3 integrin of lectins, affinity precipitates. For each lectin, from left to right, the bead only precipitate, the lectin precipitate and a loading control are shown. WT- Pkd1+/+ cells; KO: Pkd1-/- cells.

### 3.4.2 Identification of α3 integrin from kidney cell.

Cell lysates from *Pkd1*+/+ and *Pkd1*−/− cells were collected for subsequent IP by polyclonal anti-α3 integrin antibody. In the supplemental Fig.S3-1, panel A shows an image of a 1D SDS-PAGE gel stained with coomassie blue staining with the following samples: *Pkd1*+/+ and *Pkd1*−/− cell lysates, reduced α3 integrin IP extracts from *Pkd1*+/+ and *Pkd1*−/− cells and reduced polyclonal anti-α3 integrin antibody. The corresponding western blot for reduced α3 integrin IP extracts is also shown in Fig.S3-1, panel B. In the western blot image, the light chain from two-chain form (30 KDa band) and a small portion of full-chain form (150 KDa band) α3 integrin but not the heavy chain form were observed with fluorescence detection which is consistent with the specificity of the antibody binding site for the light chain of α3 integrin. The different mobility of α3 integrin light chain from *Pkd1*+/+ and *Pkd1*−/− cells was readily seen in the electrophoresis, which is consistent with our previous results [8]. In Fig. S3-1 the corresponding bands at the expected
molecular weight for 30 KDa (light chain from two-chain form), 130 KDa (heavy chain from two-chain form) and 150 KDa (full-chain form) from Pkd1+/+ and Pkd1−/− cells were marked with red squares. The indicated regions in Fig.S3-1A were cut for in-gel digestion and LC-MS proteomics analysis, and then searched against the entire mouse sequence database. As a result, the 30 KDa and 130 KDa bands were confirmed to contain α3 integrin light chain and heavy chain, respectively. In addition, 150 KDa band was proved to be the full single-chain form, without post-translation cleavage, which has not been reported before.

After identification of α3 integrin-containing bands, the sequence coverage of each section was studied by multi-enzyme cleavage strategies. α3 integrin was identified with 87.2% sequence coverage for heavy chain and 70.3% for light chain by using trypsin, Glu C and PNGase F. The identified peptide sequences for each region, along with precursor charges, and m/z are listed in the supplementary table S3-1. Due to the limited amount of single-chain form available for study, the sequence coverage for single-chain form is very low and was not studied further and the protein coverage observed for the various regions can be seen in Fig.3-2.
**Fig. 3-2. Identification of α3 integrin from kidney cells.** The sequence of α3 integrin (1053 amino acid) is illustrated by bars. Red bar indicates the α3 integrin heavy chain and yellow bar indicates the α3 integrin light chain. Identified tryptic peptides are denoted by shading. Purple circles indicate potential N-linked glycosylation sites (Asn-X-Ser/Thr where X: any amino acids except for proline). Glycosylation sites with solid purple circles have been reported [5,6] and those with light purple have never been reported.

### 3.4.3 Glycopeptide analysis of α3 integrin light chain

As reported in previous studies [9, 10], α3 integrin has 14 potential N-linked glycosylation sites, with 4 of these sites located on the light chain. We firstly focused on the glycosylation analysis for α3 light chain by using our LC-MS approach as the integrin light chain was the protein species with a significant difference in electrophoretic mobility. The corresponding enzymatic peptide fragments for these 4 glycosylation sites, along with precursor mass and glycosylation
occupancy, are listed in Table 3-1. Among the 4 glycosylation sites, the Asn-937 site, was shown to contain high-mannose glycans, while the other three sites were identified with complex-type and hybrid-type glycans. Although most glycan structures observed at an individual site are identical compared between *Pkd1*+/+ and *Pkd1*−/− cells, significant site-specific differences in glycan structures were observed. In the following section, we will describe the measurement by our LC-MS approach of the glycan structures at individual site of α3 integrin and especially the unusual glycan structures, which are unique in *Pkd1*−/− cells.

Table 3.1: **Tryptic digested α3 integrin light chain with corresponding glycopeptides.**

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Peptide sequence</th>
<th>[M+H]+</th>
<th>Glycosylation sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>913</td>
<td>932</td>
<td>CVWLECPLPDTSN925ITN928VTVK</td>
<td>2346.1468</td>
<td>One-site-p</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>The other site-f</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(complex and hybrid)</td>
</tr>
<tr>
<td>935</td>
<td>949</td>
<td>VWN937STFIEDYKDFDR</td>
<td>1934.8919</td>
<td>N937-f</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(High-mannose)</td>
</tr>
<tr>
<td>962</td>
<td>972</td>
<td>TSIPTINM*EN971K</td>
<td>1263.6249</td>
<td>N971-f</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(Complex and hybrid)</td>
</tr>
</tbody>
</table>

*M*: Met oxidation on N971 glycopeptide.

C: all Cys residues are modified by carbamidomethylation.

Monoisotopic [M+H]+ masses for the non-glycosylated peptides are listed.

Site occupancy are reported as f (fully glycosylated) and p(partial glycosylated) with glycan types (high-mannose, complex or hybrid type) annotated in parenthesis.
High Mannose glycans at Asn-937 site

In the analysis of Asn-937, high mannose structures were identified mainly as Man9, along with less abundant Man8 and Man7 structures in both $Pkd1^{+/+}$ and $Pkd1^{-/-}$ cells. As shown in the Fig.3-3, these glycopeptides are shown in an extracted ion chromatogram at 47.9 min in $Pkd1^{+/+}$ cells (Fig.3-3A) and 47.3 min in $Pkd1^{-/-}$ cells (Fig.3-3B), the accurate precursor ion (m/z 1267.1827, 3+) was measured by FTMS (Fig.3-3C), and the precursor was fragmented by CID-MS2 (Fig. 3-3D). The characteristic fragmentation pattern was observed with consecutive losses of 1 to 9 mannose residues from mass of the precursor ion, as indicated in the figure. Also, the signal observed at m/z 1069.8, was assigned to a doubly charged peptide backbone with a single GlcNAc residue and confirmed that the Man9 structure was attached at Asn-937 (VWNSTFIEDYKDFDR). The family of the related glycan structures with the same peptide backbone, such as Man8 and Man7 associated with this site, eluted at similar retention times and were also interpreted by MS/MS fragmentation in both $Pkd1^{+/+}$ and $Pkd1^{-/-}$ cells (Data not shown). Thus, we determined that identical high-mannose glycan structures with similar distribution of glycan forms were observed at this site in $Pkd1^{+/+}$ cells, with no significant differences at this site relative to $Pkd1^{-/-}$ cells.

Meanwhile, the glycopeptide backbone was also confirmed by deglycosylation methods. Samples were digested by trypsin with and without PNGase F in parallel. With PNGase F treatment, N-linked glycans were removed and Asparagine was converted to aspartic acid with a 0.984 Da mass shift on the precursor mass. Without the hydrophilic glycans, the deglycosylated peptide should be eluted a little later than the original glycosylated peptide. As shown in the supplemental Fig. S3-2, the deglycopeptide with Asn-937 was eluted at 51min (panel A) and the glycopeptide with N-linked glycosylation was eluted at 45min (panel B). In the analysis of CID-
MS2 of deglycopeptide (panel C), b and y ions were fragmented and Asn-937 was identified to convert to Aspartic acid. And the family of related glycan structures was identified to be Mannose 7, Mannose 8 and predominantly Mannose 9, which was shown in the panel D.

Fig. 3-3. Identification of high-mannose glycan structure at Asn-937 of tryptic digested α3 integrin light chain from *Pkd1*^+/+ and *Pkd1*^-/- cells.

A, extracted ion chromatogram (XIC) of the glycopeptide containing AsN-937 from *Pkd1*^+/+ cells; B, extracted ion chromatogram (XIC) of the glycopeptide containing AsN-937 from *Pkd1*^-/- cells; C, mass and charge of the trypsin digested peptides with the anticipated glycan structure; D, CID-MS2 spectrum of the precursor ion from B. Key to symbols: green circle, mannose residue; yellow circle, galactose; blue square, N-acetylglucosamine; the purple diamond, sialic acid.
Complex-type and hybrid-type glycans at Asn-971 site

In the analysis of the site Asn-971, both complex-type and hybrid-type glycan structures were identified in the α3 integrin light chain from both Pkd1<sup>+/+</sup> and Pkd1<sup>−/−</sup> cells (Fig.3-4). For complex-type glycans, as shown in the figure, a structure consisting of bi-antennary branches without core fucose and with one terminal sialic acid was determined. Fig.3-4, shows the location of the glycopeptides in the LC-MS map (Fig.3-4A for Pkd1<sup>+/+</sup> cells, Fig.3-4B for Pkd1<sup>−/−</sup> cells), the accurate precursor ion (m/z 1059.4178, 3+) (Fig.3-4C) and the fragmentation of the precursor ion by CID-MS2 (Fig.3-4D). The accurate precursor mass measured indicated that the complex glycan structure attached to the glycopeptide backbone contained a methionine oxidation. This result was also confirmed by the ion at 733.73 (2+) from the MS/MS spectra, which shows a single GlcNac residue attached to this glycopeptide backbone. In addition, consecutive loss of single oligosaccharide residues from precursor mass was observed in the fragmentation pattern. The typical ion observed at m/z 657 in the spectrum indicated the presence of sialic acid in the structure. Another glycoform in complex type at Asn-971 was also determined at the same retention time, which was confirmed as a bi-antennary glycan structure with a terminal sialic acid on one branch and Gal-Gal connections on the other branch. (Data not shown)

Besides the complex-type glycan structures, we identified another series of hybrid glycan structures at Asn-971 site the α3 integrin light chain isolated from Pkd1<sup>−/−</sup> cells. The major hybrid glycan structure is composed of three unsubstituted mannose residues and one GlcNAc-Gal-GlcNAc branch linked to the core (Data not shown). Compared the glycopeptides at Asn-971 site between Pkd1<sup>+/+</sup> and Pkd1<sup>−/−</sup> cells, similar to Asn-937 site, identical glycan structures and similar glycan distributions at this site were found in Pkd1<sup>+/+</sup> and Pkd1<sup>−/−</sup> cells.
FIG. 3-4. Identification of a complex glycan structure at Asn-971 of tryptic digested α3 integrin light chain from Pkd1−/− cells.

A, extracted ion chromatogram (XIC) of the glycopeptide containing AsN-937 from Pkd1+/+ cells; B, extracted ion chromatogram (XIC) of the glycopeptide containing AsN-937 from Pkd1−/− cells; C, mass and charge of the trypsin digested peptides with the anticipated glycan structure; D, CID-MS2 spectrum of the precursor ion from B. Key to symbols: green circle, mannose residue; yellow circle, galactose; blue square, N-acetylglucosamine; the purple diamond, sialic acid.
Di-sialylated glycans at N925 & N928 sites in Pkd1<sup>−/−</sup> cells

More complicated was the analysis of the Asn-925 and Asn-928 sites as these two sites are very close in the amino acid sequence and thus difficult to find an enzymatic strategy that generates a single-site glycopeptide. In the analysis of glycopeptide containing two sites, as shown in the Fig.5, different glycopeptides were identified in extracted ion chromatograms at 48 min and 46 min, in Pkd1<sup>+/+</sup> and Pkd1<sup>−/−</sup> cells, respectively (Fig.3-5A and Fig.3-5B). The glycopeptides containing the N925 and N928 sites have a mass 4 Da larger for the 4+ charged species in Pkd1<sup>+/+</sup> vs. Pkd1<sup>−/−</sup> cells(Fig.3-5C and Fig.3-5D, approximately 20 KDa for an individual glycan structure). The aberrant glycopeptide precursor mass in Pkd1<sup>−/−</sup> cells was observed in biological replicates from different cell cultures as well as analytical replicates (n=5). The significant differences in glycan structures can also explain the difference in mobility of α3 integrin light chain on SDS-PAGE electrophoresis.
FIG.3-5. Identification of different glycans at Asn-925 and/or Asn-928 of tryptic digested α3 integrin light chain from Pkd1+/+ and Pkd1−/− cells cells.

A and B, extracted ion chromatogram (XIC) of the glycopeptides containing Asn-925 and Asn-928 sites in Pkd1−/− and Pkd1+/+ cells, respectively; C and D, mass and charges of the tryptic peptides observed in Pkd1−/− and Pkd1+/+ cells respectively;
FIG. 3-6. Identification of glycan structure at Asn-925 and/or Asn-928 of tryptic digested α3 integrin light chain from *Pkd1*+/+ and *Pkd1*−/− cells.

A, CID-MS² spectrum of the anticipated glycopeptide precursor ion 1465.34 (4+) at Asn-925 and/or Asn-928 in *Pkd1*+/+ cells; B, CID-MS² spectrum of the anticipated glycopeptide precursor ion 1461.09 (4+) at Asn-925 and/or Asn-928 in *Pkd1*−/− cells;
However, it was very difficult to interpret the glycan structures on each site, as the product ions generating from this glycopeptides are a mix of cleavages simultaneously occurring at both glycosylation sites. But from typical diagnostic ions and characteristic mass loss present in the CID-MS/MS spectrum, we still can obtain partial but useful information for oligosaccharide compositions on these two sites. Due to the variety of glycan structures on the peptide backbone, the intensity of glycopeptides with a given glycosylation structure are relatively low. Thus, a CID targeting approach was necessarily performed in the analysis of Asn-925 and Asn-928. In the Fig. 3-6, the MS/MS spectrums of glycopeptides at N925 and N928 with the highest intensity (3-6A: m/z 1465.34 (4+) in Pkd1+/+ cells and 3-6B: m/z 1461.09 (4+) in Pkd1−/− cells) are illustrated. In the CID spectrum of glycopeptides from Pkd1−/− cells (Fig. 3-6B), the ion present at m/z 657 is the marker ion for oligosaccharide containing GlcNAc-Gal-SA, and thus indicated the presence of sialic acid in the structure. The signals at m/z 818 and 948 Da are diagnostic ions for Man-GlcNAc-Gal-SA and GlcNAc-Gal-SA-SA connected to the same branch, leading to the deduction that two sialic acids are linked together and attached to the same glycan antenna. Additionally, the fragment ions illustrating the loss of 948 Da from a 4+ precursor mass also demonstrated the unusual di-sialylated containing structure in Pkd1−/− cells. On the contrary, in the CID-MS/MS spectrum of glycopeptides from Pkd1+/+ cells (Fig.3-6A), signals at high mass region showed the first loss of GlcNAc residue, followed by seven Hexose losses (Galatose or Mannose), indicating that the sialic acid residues were absent in the glycan structures on the Asn-925 and Asn-928 sites from Pkd1+/+ cells. MS/MS spectrums of other low-intensity glycoforms which were targeted in additional runs, also demonstrated the sialic-acid containing glycans were only present at these two sites in Pkd1−/− cells, rather than Pkd1+/+ cells (data not shown).
3.5 Discussion

We have utilized a powerful LC-MS approach to characterize α3 integrin both in the major two-chain form (including heavy chain and light chain) and minor single-chain form. We also characterized the glycoform heterogeneity present on the light chain of α3 integrin from both Pkd1+/+ and Pkd1−/− cells. Our approach is sufficiently sensitive to detect altered glycan structures of α3 integrin at a specific site in Pkd1−/− cells, compared to that in Pkd1+/+ cells. Table 2 shows the summary of the data for the major glycans at each site of α3 integrin light chain in both Pkd1+/+ and Pkd1−/− cells. So far, there is no clear mechanism reported for the roles of these site-specific glycans in the α3 integrin-mediated ligand binding or cell adhesion process, however, the glycosylation on integrins has been shown to maintain the correct 3-D conformation for the receptors and are strongly implicate in the modulation of their affinity to extracellular ligands[11, 12]. In addition, a point mutation (Ala-to-Ser substitution at position 349) on α3 integrin has been reported to introduce an extra N-linked glycosylation site and this additional point glycosylation prevented its heterodimerization with β1 unit, eventually leading to severe lung and kidney disease[13]. We therefore believe that it is of significance to measure detailed glycan heterogeneity on each individual site, which can explore in-depth the function of site-specific oligosaccharides in ligand binding or cell adhesion and thus provide mechanistic insights into cystic kidney disease.

Polysialic acid has been demonstrated to associate with many types of cancer, particularly in non-small cell lung cancer (NSCL), tumor progression has been shown to be related with the expression of polysiaic acids and the levels of the related glycosyltransferase, ST8SIA2[14, 15]. The ST8SIA2 gene has been shown to be capable of forming di-sialic acid structure [16]. In addition, the di-sialic acid glycans has been identified in the secreted form of epidermal growth
factor receptor (EGFR) [17], which has been utilized as a biomarker in lung, ovarian and breast cancers [18, 19]. Even though the function of such glycan structures on α3 integrin is unknown, but we can hypothesize that the integrin conformation could be altered upon the abnormal glycosylation pattern containing more negatively charged glycans, which further influence the receptor recognition as well as binding affinity for different ligands.

In the future, the same approach will be used to characterize the glycosylation pattern on α3 integrin heavy chain or other associated proteins as we believe the glycosylation defects may not be constrained to a specific protein or specific glycosylation site. In future studies we will examine changes in the glycosylation-biosynthetic enzymes, which may yield information about sub-phenotypes in Pkd1−/− cells with altered downstream signal pathways and related changes in kidney cyst formation.
Table 3-2 Major glycans at each site of α3 integrin light chain in Pkd1+/+ and Pkd1−/− cells

<table>
<thead>
<tr>
<th>Site</th>
<th>Glycan Type</th>
<th>Major Glycan Structure</th>
<th>[M+H]+</th>
<th>Pkd1+/+ (%)</th>
<th>Pkd1−/− (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn 937</td>
<td>High-mannose Type</td>
<td></td>
<td>3801.55</td>
<td>87.8±2.8</td>
<td>91.1±1.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3639.49</td>
<td>10.8±2.2</td>
<td>7.5±1.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3476.44</td>
<td>1.4±0.8</td>
<td>1.4±0.1</td>
</tr>
<tr>
<td>Asn 971</td>
<td>Complex Type</td>
<td></td>
<td>3178.33</td>
<td>3.6±2.7</td>
<td>9.9±1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3340.38</td>
<td>5.5±4.4</td>
<td>5.8±2.7</td>
</tr>
<tr>
<td>Asn 925 &amp; Asn 928</td>
<td>Hybrid-type (with SA)</td>
<td></td>
<td>5841.36</td>
<td>0</td>
<td>62.0±2.5</td>
</tr>
<tr>
<td></td>
<td>Hybrid-type (without SA)</td>
<td></td>
<td>5638.28</td>
<td>0</td>
<td>30.6±1.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5475.2</td>
<td>0</td>
<td>7.74±0.6</td>
</tr>
<tr>
<td>Asn 925 &amp; Asn 928</td>
<td>Hybrid-type (without SA)</td>
<td></td>
<td>6020.4</td>
<td>22.1±1.8</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5857.36</td>
<td>42.2±3.9</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5658.36</td>
<td>21.4±2.3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5495.28</td>
<td>13.9±0.7</td>
<td>0</td>
</tr>
</tbody>
</table>
3.6 Conclusions

In this study, we analyzed glycan structures linked to specific site on α3 integrin light chain from both $Pkd1^{+/+}$ and $Pkd1^{-/-}$ cells. Although the majority of glycan structures identified on multiple sites are quite similar, the LC-MS approach we developed was sufficiently sensitive to differentiate glycan structural changes in the disease state. The different glycan structures identified, with larger glycan mass in $Pkd1^{+/+}$ cells than in $Pkd1^{-/-}$ cells, can also explain the different mobility of α3 integrin light chain on SDS-PAGE electrophoresis. In addition, unusual di-sialic acid glycan structures were observed solely in $Pkd1^{-/-}$ cells. Thus, in this study, the LC-MS approach we utilized allowed us to characterize unique glycan structure specifically at an individual site, which can enable the discovery of site-specific disease markers.
Figure S3-1: Immunopurification of α3 integrin.

pkd1+/+ and pkd1−/− cell lysates, reduced α3 integrin IP extracts from pkd1+/+ and pkd1−/− cells and reduced polyclonal anti-α3 integrin antibody, were loaded in a mini-gel. Panel A shows an image of 1D SDS-PAGE gel stained with coomassie blue staining. B, shows western blot performed with anti-α3 integrin antibody. Bands at the expected molecular weight for 30KDa, 130KDa, and 150KDa from both pkd1+/+ and pkd1−/− cells are indicated by red squares. Interesting regions were cut for in-gel digestion and LC-MS proteomics analysis.
### Supplemental TableS3-1: Summary of the Identified peptide sequences for α3 integrin heavy chain and light chain with precursor mass, charge, identified m/z and X correction.

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>Amino acid No.</th>
<th>Identified peptides</th>
<th>Description</th>
<th>Sequence coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITGA3</td>
<td>842</td>
<td>41</td>
<td>α3 integrin heavy chain</td>
<td>87.1%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>m/z</th>
<th>Charge</th>
<th>MH+ [Da]</th>
<th>Sequence</th>
<th>Xcorr</th>
</tr>
</thead>
<tbody>
<tr>
<td>765.39</td>
<td>1</td>
<td>765.39</td>
<td>FNLDTR</td>
<td>*</td>
</tr>
<tr>
<td>908.97</td>
<td>2</td>
<td>1816.94</td>
<td>EAVNPGLFGYSLVALHR</td>
<td>4.19</td>
</tr>
<tr>
<td>487.29</td>
<td>2</td>
<td>973.58</td>
<td>YLLLAGAPR</td>
<td>3.50</td>
</tr>
<tr>
<td>620.27</td>
<td>2</td>
<td>1239.55</td>
<td>DLAVGDDYTNR</td>
<td>2.99</td>
</tr>
<tr>
<td>1053.49</td>
<td>2</td>
<td>2105.97</td>
<td>TGAVALCPLTAHKDCER</td>
<td>*</td>
</tr>
<tr>
<td>722.34</td>
<td>1</td>
<td>722.34</td>
<td>MDISEK</td>
<td>*</td>
</tr>
<tr>
<td>595.79</td>
<td>2</td>
<td>1190.58</td>
<td>KSDPDHHIE</td>
<td>2.72</td>
</tr>
<tr>
<td>822.91</td>
<td>2</td>
<td>1644.82</td>
<td>DMWLGVTVAASQGAPGR</td>
<td>3.68</td>
</tr>
<tr>
<td>797.45</td>
<td>1</td>
<td>797.45</td>
<td>VLVAHR</td>
<td>2.97</td>
</tr>
<tr>
<td>679.86</td>
<td>2</td>
<td>1358.72</td>
<td>VLSWGLEDQR</td>
<td>2.48</td>
</tr>
<tr>
<td>1058.94</td>
<td>2</td>
<td>2016.89</td>
<td>GNLDLQPDGDWQTYHE</td>
<td>4.71</td>
</tr>
<tr>
<td>1058.71</td>
<td>3</td>
<td>3015.52</td>
<td>VGGAVYVFMRGASFPQPSLSLHGPSR</td>
<td>6.91</td>
</tr>
<tr>
<td>1228.31</td>
<td>3</td>
<td>3682.92</td>
<td>GSEEQNLGYTVQVGNAILHPTDIITVTGAPR</td>
<td>*</td>
</tr>
<tr>
<td>513.26</td>
<td>3</td>
<td>1537.79</td>
<td>HQHMGAVFLLQYE</td>
<td>3.92</td>
</tr>
<tr>
<td>1005.84</td>
<td>3</td>
<td>3015.52</td>
<td>VGGAVVVFMRQGASFPQPSLSLHGPSR</td>
<td>6.91</td>
</tr>
<tr>
<td>726.38</td>
<td>2</td>
<td>1451.77</td>
<td>VVYHSSSGGLLR</td>
<td>3.95</td>
</tr>
<tr>
<td>653.55</td>
<td>4</td>
<td>2610.38</td>
<td>VVYHSSSGGLLRQPQQIHIYGEK</td>
<td>3.44</td>
</tr>
<tr>
<td>527.62</td>
<td>3</td>
<td>1580.87</td>
<td>LGIPGLATFGYSGLSR</td>
<td>4.98</td>
</tr>
<tr>
<td>909.47</td>
<td>3</td>
<td>2726.41</td>
<td>MDVDENLYPDVLVGLSDLHIVLLR</td>
<td>*</td>
</tr>
<tr>
<td>594.87</td>
<td>2</td>
<td>1188.74</td>
<td>ARPVINILHR</td>
<td>3.41</td>
</tr>
<tr>
<td>1236.14</td>
<td>2</td>
<td>2471.27</td>
<td>TLVAPAVNPDLCATTSBCVQVE</td>
<td>5.58</td>
</tr>
<tr>
<td>888.39</td>
<td>2</td>
<td>1775.78</td>
<td>LCFAYNQSAGNPNYRR</td>
<td>2.08</td>
</tr>
<tr>
<td>825.92</td>
<td>2</td>
<td>1650.84</td>
<td>NITLAYTLAEDDR</td>
<td>*</td>
</tr>
<tr>
<td>793.85</td>
<td>2</td>
<td>1586.70</td>
<td>SQSSVFGHFSMSPE</td>
<td>3.29</td>
</tr>
<tr>
<td>608.84</td>
<td>2</td>
<td>1216.67</td>
<td>LLMMDNVRDK</td>
<td>3.08</td>
</tr>
<tr>
<td>878.52</td>
<td>2</td>
<td>1756.03</td>
<td>LRPIVIAHNLQSLPLR</td>
<td>4.33</td>
</tr>
<tr>
<td>710.99</td>
<td>3</td>
<td>2130.98</td>
<td>SLDAYPVNNAQAMEHTE</td>
<td>4.37</td>
</tr>
<tr>
<td>912.38</td>
<td>2</td>
<td>1823.75</td>
<td>ECGPDNKCDNLSQM</td>
<td>*</td>
</tr>
<tr>
<td>Gene Name</td>
<td>Amino acid No.</td>
<td>Identified peptides</td>
<td>Description</td>
<td>Sequence coverage</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------</td>
<td>-------------------</td>
<td>-------------</td>
<td>------------------</td>
</tr>
<tr>
<td>ITGA3</td>
<td>179</td>
<td>41</td>
<td>α3 integrin light chain</td>
<td>61.45%</td>
</tr>
<tr>
<td>m/z</td>
<td>Charge</td>
<td>MH+ [Da]</td>
<td>Sequence</td>
<td>Xcorr</td>
</tr>
<tr>
<td>926.47</td>
<td>2</td>
<td>1851.95</td>
<td>QLDPGGDQSSPPVTLAAK</td>
<td></td>
</tr>
<tr>
<td>990.52</td>
<td>2</td>
<td>1980.04</td>
<td>QLDPGGDQSSPPVTLAAKK</td>
<td></td>
</tr>
<tr>
<td>711.85</td>
<td>2</td>
<td>1422.70</td>
<td>AKSETVLTCSNGR</td>
<td></td>
</tr>
<tr>
<td>825.42</td>
<td>2</td>
<td>1649.84</td>
<td>AKSETVLTCSNGRAR</td>
<td>*</td>
</tr>
<tr>
<td>1174.56</td>
<td>2</td>
<td>2348.12</td>
<td>CVWLECPLPDTSNITNVTK</td>
<td></td>
</tr>
<tr>
<td>967.95</td>
<td>2</td>
<td>1934.89</td>
<td>VWNSTFIEDYKDFDR</td>
<td></td>
</tr>
<tr>
<td>716.90</td>
<td>2</td>
<td>1432.81</td>
<td>VRVDGWATLFLR</td>
<td></td>
</tr>
<tr>
<td>589.31</td>
<td>2</td>
<td>1177.62</td>
<td>VDGWATLFLR</td>
<td></td>
</tr>
<tr>
<td>632.31</td>
<td>2</td>
<td>1263.62</td>
<td>TSIPTINMENK</td>
<td>*</td>
</tr>
<tr>
<td>694.38</td>
<td>1</td>
<td>694.38</td>
<td>ALYEAK</td>
<td></td>
</tr>
<tr>
<td>770.84</td>
<td>2</td>
<td>1540.68</td>
<td>SQPSETERLTDDY</td>
<td></td>
</tr>
</tbody>
</table>

Note

* Indicate the peptide is identified by manual inspection
C in black: Cys with carbamidomethylation
N in red: Asn is Deglycosylated and converted to Asp with PNGase F treatment
M in orange: Met with oxidation
Figure S3-2: Identification of glycosylated peptide backbone by using deglycosylation method (PNGase F treatment).

A, extracted ion chromatogram (XIC) of the deglycopeptide containing Asn-937 (without glycosylation); Accurate precursor mass was shown as inserted.

B, extracted ion chromatogram (XIC) of the glycopeptide containing Asn-937 (with glycosylation); Accurate precursor mass was shown as inserted.

C, CID-MS2 spectrum of deglycosylated peptide backbone (m/z = 645.96, z=3);

D, precursor mass of related glycan structures.
3.7 References


Chapter 4: Mass spectrometric analysis of leader sequence attached to N-terminus of recombinant human growth hormone produced from a yeast strain of, *Pichia pastoris*.

Co-authors’ work in this chapter: The MIT rhGH culture media was provided by Prof. Love’s lab in the chemical engineering department of MIT. The protein analysis by LC-MS/MS was done by Prof. Hancock’s group.

Anna (Fan) Zhang: experimental design and perform, data analysis, manuscript writing and revision; Kartik Shah: protein production and sample collection; Shiaw-lin Wu: idea contribution and experimental design; William Hancock: idea contribution and manuscript writing; J. Christopher Love: idea contribution and grant support.


4.1 Abstract

In this chapter, we have characterized the recombinant human growth hormone produced from *Pichia pastoris* by MIT. The sequence with N-terminal peptide variants as well as post-translational modifications (i.e. oxidation and deamidation) was analyzed by the state-of-the-art LC-MS approach. The extent of N-terminal end sequence variants as well as oxidation and deamidation were measured by the ratio of peak areas of the non-modified peptide vs. the sum of peak area of the non-modified and modified peptide in the same LC-MS analysis. We successfully identified two additional N-terminal variants with amino acids from leader sequence attached to N-terminus in MIT produced rhGH. Also, in the relative comparison with Genentech produced rhGH, a higher amount of oxidation but an equivalent amount of deamidation was found in MIT produced rhGH. The concentration of MIT produced rhGH was also measured by external standard Genentech rhGH with its calibration curve. The technique we applied can also
sensitively pinpoint the subtle but distinct difference due to a different manufacturing process. Our analytical results can provide critical feedbacks to the upstream manufacture groups, and then instruct them to improve the fermentation process.

4.2 Introduction

Leader sequence is a short peptide consisting of 5-30 amino acids, attached to the N-terminus of most newly synthesized proteins [1]. The leader sequence plays a critical role in leading the protein toward the destined secretary pathway and eventually either residing certain organelles (i.e. the endoplasmic reticulum, Golgi apparatus or Endosomes), inserted into cellular membranes or secreted from the cells[1]. Selection of a leader sequence for secretion of a particular recombinant protein is heavily dependent on the expression systems (i.e. Bacteria, yeast or CHO cells, etc) selected for production of recombinant proteins. In the yeast expression system, *Pichia pastoris* is one of the most popular yeast expression systems due to its efficient secretion of properly folded recombinant protein, which allows a direct capture of targeted product from the culture media and eliminates the need for expensive but less efficient cell disruption or refolding process. In *Pichia pastoris*, the *S. cerevisiae* α-mating factor leader sequence (α-MF) is commonly utilized and has been successful in many cases [2, 3]. Studies have shown that using α-MF as the signal peptide can even lead to secretion of higher amounts of a recombinant protein than using its native signal peptide [4].

Once the final destination of target proteins are reached, the signal peptides are cleaved off by signal peptidases, which correspondingly convert the proteins into their mature forms [5]. Kex2 protease (Kex2p) and Ste13 dipeptidyl aminopeptidase (Ste13p) are required in *Saccharomyces cerevisiae* for maturation of the alpha-mating factor in a late Golgi compartment [6]. Poor
efficiency of signal peptidases will directly introduce extra amino acids to N-terminal peptides and further affect the protein stability and in vivo half-life [7].

The recent achievement of analytical techniques, such as liquid chromatography coupled on-line with tandem mass spectrometry (LC-MS) has been demonstrated to be a powerful technology in analyzing amino acid sequence with its variance. For example, our laboratory has been applied this LC-MS approach to characterize a biosimilar version of beta interferon and its carrier protein, human serum albumin [8]. The similar LC-MS approach combined with multi-separation and multi-enzyme digestion strategy was also utilized to characterize the primary structure of a generic therapeutic drug, TNK-tPA, which is approved for treatments of acute myocardial infarction and ischemic stroke. The amino acid sequence of TNK-tPA was comprehensively characterized and three substitutions, which are different with its native tPA, were also identified [9]. In addition, comparability analysis of anti-CD20 commercial and RNAi-mediated fucosylated antibodies was performed by two LC-MS approaches, and two mutations, S258Y (fully mutated) and F174I/L (partially mutated), were detected in the production of the RNA interference (RNAi)-mediated molecule [10]. In this work, we used a similar LC-MS approach to extensively characterize the recombinant human growth hormone produced from *Pichia pastoris* by MIT (briefly: MIT rhGH), and compare it with Genentech rhGH (briefly: GNE rhGH) standard. As shown in the following, three N-terminal variants were observed in the MIT rhGH, of which, two variants unexpectedly contained amino acids from leader sequence. Our analysis could provide immediate and critical feedbacks to our upstream group and help them with their production process.
4.3 Materials and methods

4.3.1 Materials

MIT rhGH in our LC-MS analysis was provided by Love’s lab from chemical engineering department of MIT. (Cambridge, MA). The rhGH product is provided directly with culture media for early screening. Nutropin AQ (Genentech, South San Francisco, CA), a liquid formulation product, which consists of 10mg of recombinant human growth hormone (Somatropin, the active drug ingredient), 17.4 mg sodium chloride, 5mg phenol, 4mg polysorbate 20, and 10mM sodium citrate in 2mL sterile liquid. A total of 3 vials for each drug product were used for the following LC-MS analysis. Trypsin (sequencing grade) was purchased from Promega (Madison, WI). Ammonium bicarbonate, formic acid, dithiothreitol (DTT) and iodoacetamide (IAA) were purchased from Sigma-Aldrich (St. Louis, MO) Water at LC-MS grade was bought from J.T. Baker (Phillipsburg, NJ). HPLC grade acetonitrile was from Thermo Fisher Scientific (Fairlawn, NJ). NuPAGE® 4-12% Bis-Tris gels, Novex® sharp pre-stained protein standard, SimplyBlue™ safestain were bought from Invitrogen™ (Carlsbad, CA).

4.3.2 SDS-PAGE and in-gel digestion

A mini gel (8 cm x 8 cm and 4 to 20% Tris-glycine) stained with coomassie blue, was used to isolate the expected protein rhGH. The interesting band (22KDa) in several lanes, were cut-out and minced into small pieces (1mmx1mm). The gel slices were subjected to 2~3 cycles of gel dehydration with acetonitrile and rehydration with ammonium bicarbonate buffer (100mM, pH8.0), with shaking 15min in each buffer, until no visible blue color was observed. The remaining gel pieces were reduced with DTT (10mM in 100mM NH₄HCO₃) and incubated for 30min at 56 °C. The sample was then alkylated in IAA (55mM in 100mM NH₄HCO₃) and stay
in dark for 60 min at room temperature. An aliquot of 250 μL of trypsin digestion solution (10 ng/μL trypsin in 50 mM NH₄HCO₃, pH 8.0) was added to shrunked gel slice, and the sample was then incubated for 30-35 min at 4 °C. The incubated solution was then replaced with sufficient 50 mM NH₄HCO₃ to cover the gel pieces (50-100 μL) and then incubated overnight at 37 °C. The supernatant was collected. The remaining gel pieces were further extracted with 5% formic acid (100 μL) at 37 °C for 5 min, and shrunk by an equal amount of acetonitrile, in 2~3 cycles. The formic acid and acetonitrile solution, containing tryptic peptides, was combined with the previous collections and concentrated to ~10 μL. The concentrated peptide solution was subsequently used for LC-MS analysis.

4.3.3 LC-MS

The tryptic peptides were analyzed on an Orbitrap-Elite MS instrument (Thermo Fisher Scientific, San Jose, CA) equipped with a Nanospray ion source (New Objective, Woburn, MA). An Ultimate® 3000 nano LC pump (Dionex, Mountain View, CA) was on-line coupled to the mass spectrometer through a 75 μm i.d. x 10 cm C18 column, packed with Magic C18 beads (200Å pore and 5 μm particle size) (Michrom BioResources, Auburn, CA) Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1 % formic acid in acetonitrile. Before loading the samples, the C18 column was equivalent by using 2% mobile phase B at flow rate 200nL/min for at least 30 min. The tryptic peptides were separated by elution gradient, which was maintained at 200nL/min using 2% B to 40% B in 60 min, then from 40% to 80% B for 10 min. The mass spectrometry was operated in the data-dependent mode which alternatively switched between MS and MS2 acquisition. Full survey full-scan MS spectra was acquired in the FTICR cell with mass resolution of 100,000 at m/z 400 (after accumulation to a target value of
2x10^6 ions in the linear ion trap), followed by 8 sequential data-dependent MS/MS scans. Dynamic exclusion was utilized with exclusion duration of 30 sec and 2 repeat counts.

4.3.4 Protein identification

Peptide identification was initially searched against sequence of recombinant human growth hormone using the SEQUEST algorithm incorporated in Thermo Discovers software (version 1.3). Peptides are automatically identified if the peptide has an Xcorr score (+1 > 1.5; +2 > 2.0; +3 > 2.5) with trypsin cleavage at both ends. Orbitrap-Elite provides accurate mass measurement (<5 ppm) for confirmation of peptide identification. Thus, the Xcorr scores (+1 > 1.5; +2 > 2.0; +3 > 2.5) were used for the initial screening. Next, manual inspections were also performed for the MS/MS spectral of peptides with 1 or 2 miscleavages in trypsin digestion, to exclude the false positive or false negative results generated by software. Further, potential sequence variants and peptide modifications were examined manually from preferred fragmentation patterns in the observed CID-MS2.

4.4 Results and Discussion

Recombinant human growth hormone from Genentech, consisting of 191 amino acids, was utilized as a standard protein in our analysis. MIT produced rhGH should have identical amino acid sequence as it is the starting point to evaluate the successful production of any recombinant therapeutic drug. The primary structure of MIT rhGH was firstly examined by LC-MS analysis from trypsin digested peptides.

4.4.1 Protein and primary structure identification
The MIT rhGH was provided with culture media (without any concentration step) and was intended to initially examine the amino acid sequence, so the concentration of MIT rhGH is very low (approximately ng/µL), we performed in-gel digestion, instead of in-solution digestion, in our primary structure analysis. SDS-PAGE technique is useful for protein separation by molecular weight, which can separate the rhGH with other possible concomitant impurities in *Pichia pastoris*. It is also an efficient strategy for protein extraction from coomassie-blue stained gel. As shown in Fig4-1, MIT rhGH (in lane 2 and lane 3) and GNE rhGH with four different concentrations (in lane 4-7) were run side by side in reduced conditions (Disulfide bonds were reduced in this approach, because disulfide linkages were not important at this stage). Followed the detailed instructions in section 2, the interesting protein bands, from MIT and GNE respectively, were cut and subjected into in-gel trypsin digestion. The tryptic peptides were analyzed by online LC with Orbitrap-Elite MS.

As indicated in Fig 4-2, The C-terminal peptide T21 was identified with the accurate precursor mass and CID-MS2 fragmentation. In the Fig4-2, the C-terminal peptide was identified in both MIT and GNE rhGH, at 28.16 min (Fig. 4-2A) and 28.72 min (Fig. 4-2B), respectively. The accurate precursor mass of this peptide, m/z 842.3352 with 1+ charge, was shown in the Fig.4-2C, which is within 5ppm with its theoretical mass. The precursor ion (842.3352, 1+) was isolated using data-dependent acquisition mode and subjected to CID-MS2 fragmentation in the linear ion trap (Fig. 4-2D), the characteristic fragmentation pattern of CID was indicated as labels in the figure.

Other peptides of MIT rhGH, including N-terminal end peptide, were identified by using the same manner. The table 1 summarized identified tryptic peptides (without trypsin miscleavages) along with their molecular weight and retention time. In the table 4-1, T5 (EQK), T7 (EETQQK),
T14 (QTYSK), T17 (K) and T18 (DMDK) are too small to retain on the C18 column in the LC separation. However, they were further identified in the same the LC-MS run but considering trypsin digestion with 1 or 2 miscleavages. The very short peptides are combined with former or latter tryptic peptides and extracted manually from LC-MS data. And the identified peptide sequence, with precursor mass and retention time were summarized in the table 4-2.

In summary, only three amino acids (AHR) were not detected in tryptic peptide mapping. Further, we used Lys-C digestion and the missing region was also found (Data not shown). Thus, a total of 100% sequence coverage of MIT produced rhGH was achieved (indicated as Fig 4-3) which allowed us to monitor all sequence variants and other modifications in the next step.
FIG. 4-1. SDS-PAGE separation of recombinant human growth hormone produced from *Pichia pastoris* by MIT and GNE rhGH.

Lane 1: molecular weight standard.

Lane 2-3: recombinant human growth hormone produced from *Pichia pastoris* by MIT.

Lane 4-7: human growth hormone from Genentech, used as standard. Four different concentrations of GNE rhGH were loaded on the gel and used for relative quantitation.
FIG.4-2. LC-MS Identification of T21 peptide (C-terminal) from the tryptic digests of rhGH.

A: Extracted ion chromatography (XIC) of 842.3351 (1+) from MIT rhGH.

B: Extracted ion chromatography (XIC) of 842.3351 (1+) from GNE rhGH.

C: Precursor mass scan at 28.16 min using Orbitrap. For illustration purpose, only m/z 839-850 region is shown.

D: MS2 scan of the m/z 842.3351 (1+) ion. The peptide sequences with the observed fragment ions are shown in the insert.
Table 4-1. Summary of tryptic mapping of MIT produced rhGH (peptides identified without trypsin miscleavages)

<table>
<thead>
<tr>
<th>Tryptic#</th>
<th>AA#</th>
<th>Peptides</th>
<th>m/z</th>
<th>RT (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1-8</td>
<td>FPTIPLSR</td>
<td>465.77 (2+)</td>
<td>38.36</td>
</tr>
<tr>
<td>T2</td>
<td>9-16</td>
<td>LFDNAMLR</td>
<td>490.26 (2+)</td>
<td>36.25</td>
</tr>
<tr>
<td>T4</td>
<td>20-38</td>
<td>LHQLAFDYTEQEFEAYIPK</td>
<td>781.38 (3+)</td>
<td>47.14</td>
</tr>
<tr>
<td>T6</td>
<td>42-64</td>
<td>YSFLQNPQTSLCFSESIPNSR</td>
<td>891.76 (3+)</td>
<td>47</td>
</tr>
<tr>
<td>T8</td>
<td>71-77</td>
<td>SNLELLR</td>
<td>422.75 (2+)</td>
<td>32.36</td>
</tr>
<tr>
<td>T9</td>
<td>78-94</td>
<td>ISLILQSWLEPVQFLR</td>
<td>1028.11 (2+)</td>
<td>66.15</td>
</tr>
<tr>
<td>T10</td>
<td>95-115</td>
<td>SVFANSLVYGASDNSVYDLLK</td>
<td>1131.57 (2+)</td>
<td>50.85</td>
</tr>
<tr>
<td>T11</td>
<td>116-127</td>
<td>DLEEGIQTLMGR</td>
<td>681.34 (2+)</td>
<td>49.08</td>
</tr>
<tr>
<td>T12</td>
<td>128-134</td>
<td>LEDGSPR</td>
<td>773.38 (1+)</td>
<td>14.47</td>
</tr>
<tr>
<td>T13</td>
<td>135-140</td>
<td>TGGQIFK</td>
<td>693.39 (1+)</td>
<td>22.12</td>
</tr>
<tr>
<td>T15</td>
<td>146-158</td>
<td>FDTNSHNDALLK</td>
<td>745.35 (2+)</td>
<td>27.8</td>
</tr>
<tr>
<td>T16</td>
<td>159-167</td>
<td>NYGLLLYCFR</td>
<td>603.29 (2+)</td>
<td>44.11</td>
</tr>
<tr>
<td>T19</td>
<td>173-178</td>
<td>VETFLR</td>
<td>764.43 (1+)</td>
<td>29.79</td>
</tr>
<tr>
<td>T20</td>
<td>179-183</td>
<td>IVQCR</td>
<td>675.36 (1+)</td>
<td>10.55</td>
</tr>
<tr>
<td>T21</td>
<td>184-191</td>
<td>SVEGSCGF</td>
<td>842.33 (1+)</td>
<td>28.16</td>
</tr>
</tbody>
</table>
Table 4-2. Summary of tryptic mapping of MIT produced rhGH (peptides identified with 1 or 2 trypsin miscleavages)

<table>
<thead>
<tr>
<th>Tryptic#</th>
<th>AA#</th>
<th>Peptides</th>
<th>m/z</th>
<th>RT (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T4-T5</td>
<td>20-41</td>
<td>LHQLAFDTYQEFEEAYIPKEQK</td>
<td>909.78 (3+)</td>
<td>45.08</td>
</tr>
<tr>
<td>T6-T7</td>
<td>42-70</td>
<td>YSFLQNPQTSLCFSESITPSNREETQQK</td>
<td>1139.54 (3+)</td>
<td>44.11</td>
</tr>
<tr>
<td>T14-T15</td>
<td>141-158</td>
<td>QTYSKFDTNSSHNDALLK</td>
<td>1049.00 (2+)</td>
<td>29.24</td>
</tr>
<tr>
<td>T17-T19</td>
<td>168-178</td>
<td>KDMDKVETFLR</td>
<td>691.36 (2+)</td>
<td>31.26</td>
</tr>
<tr>
<td>T18-T19</td>
<td>169-178</td>
<td>DMDKVETFLR</td>
<td>627.31 (2+)</td>
<td>37.57</td>
</tr>
</tbody>
</table>

Fig.4-3. Summary of the identification of rhGH peptide sequence.

FPTIPLSRLF DNAMLRAHRL HQLAFDTYQE FEEAYIPKEQ KYSFLQNPQTSLCFSESITPSNREETQQK SNLELLRISLLIQSWLEPV QFLRSVFANS LVYGASDNSV YDLLKDLEEG IQTLMGRLED GSPRTGQIFK QTYSKFDTNSSHNDALLKNY GLLYCFRKDM DKVETFLRIV QCRSVEGSCG F

Note1: tryptic peptides identified (marked in red color) and not identified (in black color) by LC-MS analysis were indicated, respectively.
Note2: Amino acids with underline were identified in tryptic digested peptides with 1 or 2 miscleavage.
Note3: AHR, the very short peptide was not identified in trypsin digested sample, but was identified in the peptide of L1 (FPTIPLSRLF DNAMLRAHRL HQLAFDTYQE FEEAYIPKEQ KYSFLQNPQTSLCFSESITPSNREETQQK SNLELLRISLLIQSWLEPV QFLRSVFANS LVYGASDNSV YDLLKDLEEG IQTLMGRLED GSPRTGQIFK QTYSKFDTNSSHNDALLKNY GLLYCFRKDM DKVETFLRIV QCRSVEGSCG F) by Lys-C digested LC-MS analysis. (Data not shown)
4.4.2 Identification of N-terminal sequence variance

As shown in the Fig.4-4, the tryptic maps of rhGH from two different manufacturers (MIT and GNE) were compared and the two tryptic maps were quite similar except for the regions indicated by the circles. Relative to GNE rhGH tryptic map, two extra peaks, representing two peptides, were only observed in MIT rhGH, rather than GNE rhGH. Thus, we focused on the identification of these two peaks (peptides) as described in the following texts.

Fig.4-4. Comparison of LC-MS analysis of the tryptic map of rhGH from two different manufacturers.

A: Base peak ion chromatogram of the tryptic map of rhGH from MIT, novel strains: *Pichia pastoris*. B: Base peak ion chromatogram of the tryptic map of rhGH from the Genentech.
The two additional peaks (peptides) were analyzed by using the same approach as C-terminal end peptide T21 as shown in the Fig.4-2. Since we observed these two peaks are eluted right after the T1 peptide, and the intensity of T1 peptide decreased significantly (~10 fold) in MIT rhGH, compared to GNE rhGH, so it came to our attention that these two additional peaks, indicating two additional peptides, may be variants of T1 in MIT rhGH. Thus, we specially compared the accurate precursor mass and CID-MS2 spectrums of these two peptides, to the original N-terminal T1 peptide during our analysis.

As shown in Fig. 4-5, peak1 (T1 peptide), peak2 and peak3 were identified by accurate mass assignment of precursor ions and CID-MS2 spectrums of amino acid sequence. From the CID-MS2 spectrums, as labeled in the figure 4-5A, two characteristic ions, y4 (472.27) and y7 (783.40), explained in the figure 4-2, are highlighted in the red circles. Alike, same signals at 472 and 783 in the figure 4-5B and 4-5C, were deduced as same fragmentation ions as those labeled in T1 peptide (y4 and y7). Unique fragmentation ions in CID-MS2, such as 930.55 in both Fig 4-5B and Fig4-5C, 1001.66 and 1130.65 in Fig4-5C were deduced as labeled, demonstrating the extra amino acids EA and EAEA existed right before original T1 (FPTIPLSR). In addition, b ions at low mass regions (b2, b3 as labeled in Fig.4-5B and Fig.4-5C) also demonstrate the amino acid fragments with additional EA, EAEA on the left part of sequence. Thus, peptide sequences for Peak2 and Peak3 in MIT rhGH were confirmed as EAFPTIPLSR and EAEAFPTIPLSR, respectively. The peptide sequences were further confirmed by accurate precursor mass, 565.81(2+) for peak2 and 665.85 (2+) for peak3, as illustrated in the insert of Fig.5, which are within 5ppm of theoretical precursor mass.
Fig. 4-5. LC-MS analysis of the T1 and two additional peaks from the tryptic digests of MIT rhGH.

A: MS2 scan of the T1 peptide (Peak 1), m/z 465.77 (2+), with the Orbitrap MS measurement of the precursor ion in the insert.

B: MS2 scan of the Peak 2, m/z 565.81 (2+), with the Orbitrap MS measurement of the precursor ion in the insert.

C: MS2 scan of the Peak 3, m/z 665.85 (2+), with the Orbitrap MS measurement of the precursor ion in the insert.
As the two variants of T1 peptide were detected in the same LC-MS analysis along with the original T1 peptide, the relative quantitation for extent of variance can be easily performed as our previously studied[11, 12]. Fig.4-6 illustrated the extracted ion chromatogram (XIC) of the three N-terminal (T1 peptide) variants (m/z 465.77, 565.80 and 665.85) from MIT rhGH sample (4-6A) and GNE rhGH standard (4-6B). As shown, three N-terminal variants were eluted separately at retention time 38min (original T1), 39min (EA-T1) and 40min (EAEA-T1), however, only the original T1 peptide was detected in GNE rhGH standard. Besides, the percentage of each variant can be estimated by the peak area of the each peptide divided by the sum of the peak area of all three peptide variants. The equation is as followings: % of Peak 1 = Peak Area of [(Peak 1)/ (Peak1 + Peak2 + peak3)]. The percentage calculations in this equation were based on the assumption that all three variants had a similar response factor. Table 1 presents the percent of three N-terminal variants in MIT rhGH. The reported percentages were the average of 3 measurements with %CV less than ~10% in every measurement. As shown, the peptide variant involving EAEA for peak3, accounted for 63.65%, the peptide EA-T1 for peak2, accounted for 28.58%, while only 7.77% of original T1 was identified in MIT produced rhGH. These interesting observations were also further confirmed with our collaborator and the additional amino acids EA and EAEA come from signal peptides, which should be cleaved after rhGH secretion. In this manner, our analytical results provide immediate and critical feedbacks for the upstream production group.
C: Summary of three N-terminal variants in MIT rhGH

<table>
<thead>
<tr>
<th>Peak #</th>
<th>Amino acid sequence</th>
<th>m/z</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak 1 (T1)</td>
<td>FPTIPLSR</td>
<td>465.77 (2+)</td>
<td>7.77(±2.8)%</td>
</tr>
<tr>
<td>Peak 2</td>
<td>EAFPTIPLSR</td>
<td>565.81 (2+)</td>
<td>28.58(±0.2)%</td>
</tr>
<tr>
<td>Peak 3</td>
<td>EAEAFPTIPLSR</td>
<td>665.85 (2+)</td>
<td>63.65(±2.7)%</td>
</tr>
</tbody>
</table>

Fig.4-6. Extracted ion chromatography (XIC) of three N-terminal variants from tryptic digest of MIT rhGH and GNE rhGH

A: Base peak XIC of the ion m/z 465.77 (2+), m/z 565.81 (2+) and 665.85 (2+) from MIT rhGH.
B: Base peak XIC of the ion m/z 465.77 (2+), m/z 565.81 (2+) and 665.85 (2+) from GNE rhGH.
C: Summary of three N-terminal variants in MIT rhGH.
4.4.3 Oxidation

As well as post-translational modifications the presence of degradative reactions of therapeutic drugs are widely analyzed, because modifications could be introduced from manufacturing and formulation process and subsequently influence drug stability and efficacy [13, 14]. To evaluate the stability of protein drugs, oxidation and deamidation were commonly examined [15, 16]. In our study, oxidation of MIT rhGH, compared to GNE rhGH standard, were firstly analyzed. A higher oxidation percent in therapeutic drug would certainly cause FDA concerns and further affect the final approval.

Methionine is one of the amino acids that are easily susceptible to oxidation. Exposed Methionine is more readily oxidized by reactive oxygen species to Methionine Sulfoxide (MetO) and buried Methionine residues are relatively resistant to oxidation. Thus, the extent of oxidation on individual site can also be implicated with the protein 3-D conformations. In the analysis of rhGH products, three potential Methionine oxidation sites are located on T2, T11 and T18-T19 peptides.

As shown in Fig.4-7, both oxidized T2 peptide and its non-oxidized counterparts were identified by accurate precursor mass in the insert of Fig.4-7A and 4-7B. The oxidized peptide has an additional oxygen atom, and theoretically it should have an addition of 15.9944 Da for singlet charged ion and 7.9972 Da for doublet charged ion than the non-oxidized counterpart. Also, the oxidized peptide is more hydrophilic compared with its counterpart, so in the separation of reversed phase HPLC, the oxidized peptides should be eluted earlier than the non-oxidized peptides. As shown in the Fig.4-7A and Fig.4-7B, the Monoisotopic ion of T2 peptide (490.25, 2+) with its oxidation counterpart (498.25, 2+) were extracted from both MIT and GNE rhGH
and the difference of Monoisotopic ion (2+ charge) of oxidized T2 and non-oxidized T2 was 7.9973 Da (≈498.2516Da-490.2543Da), which matched the mass difference of doublet charged oxygen atom. The mass accuracy is within 2ppm of its theoretical mass. The position of oxidation is identified subsequently by CID fragmentation of the oxidized peptide, which was finally localized on the Methionine (M) residue. (Data not shown) In addition to T2 peptide, we also identified the oxidized T11 and T18-T19 peptides of rhGH by using the similar approach. (Data not shown)

Since both the oxidized and non-oxidized peptides were identified simultaneously in the same LC-MS run, the relative quantitation for the extent of oxidation on individual site can be easily performed. The extend of oxidation was estimated by the peak area of the oxidized peptide divided by the sum of peak area of the oxidized and non-oxidized peptides, the similar approach used for the sequence variants described above. The percents of oxidation for both MIT rhGH and GNE rhGH on T2 peptide were summarized in Table below as an example. The reported percentages were the average of three measurements. As a result, we observed the MIT rhGH have higher oxidation percent on T2 peptide, compared to GNE rhGH.

However, it was necessary to note that artificial oxidation was easily introduced from sample preparations. For example, the in gel digestion sample preparation need three days processing at room temperatures, so the MS results always have higher content of oxidation on Methionine residues, compared with the results from in-solution digestions. In addition, when the electrospray ion source was operated under atmospheric pressure, LC-MS analysis may generate some artificial oxidations on amino acid residues. However, the relative comparison between the analyzed therapeutic drug and standard was still valid, as very similar amount of artificial
oxidation was brought in from side by side sample handling and run to run mass spectrometry utilization.

![Graph A: Base peak XIC of the ion m/z 490.26 (2+) and 498.25 (2+) from GNE rhGH.

Graph B: Base peak XIC of the ion m/z 490.26 (2+) and 498.25 (2+) from MIT rhGH.

Graph C: Extend of Oxidation on T2 peptide.

<table>
<thead>
<tr>
<th>Tryptic #</th>
<th>AA #</th>
<th>Peptide</th>
<th>m/z</th>
<th>GNE rhGH</th>
<th>MIT rhGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>T2</td>
<td>9-16</td>
<td>LFDNAMLR</td>
<td>490.26 (2+)</td>
<td>48.94(±1.89)%</td>
<td>36.99(±0.29)%</td>
</tr>
<tr>
<td>T2 with Met Oxidation</td>
<td>9-16</td>
<td>LFDNAM*LR</td>
<td>498.25 (2+)</td>
<td>51.06(±1.89)%</td>
<td>63.01(±0.29)%</td>
</tr>
</tbody>
</table>

Fig. 4-7. Extracted ion chromatography (XIC) of T2 peptide with its oxidation counterparts from tryptic digests of MTI rhGH and GNE rhGH.

A: Base peak XIC of the ion m/z 490.26 (2+) and 498.25 (2+) from GNE rhGH.

B: Base peak XIC of the ion m/z 490.26 (2+) and 498.25 (2+) from MIT rhGH.

C: Extend of Oxidation on T2 peptide.
4.4.4 Deamidation

Deamidation of asparagines is another typical modification that can occur during manufacture and long-term storage. The hot spot for deamidation is the asparagines followed by glycine or serine [17]. Using the same approach for identification of oxidation, the deamidation of rhGH was also examined. The deamidated peptide should have the side chain of asparagines, NH2, be replaced by OH [16, 18], converting asparagines to aspartic acid. The concomitant mass difference was introduced with 0.9840Da for a 1+ charge ion or 0.4920Da for a 2+ ion (as in this case).

Similar to oxidation, as shown in the Fig.4-8A and Fig.4-8B, the Monoisotopic ion of T15 peptide (745.35, 2+) with its deamidation counterpart (745.85, 2+) were extracted from both MIT and GNE rhGH and the observed difference for the Monoisotopic ion (2+ charge) was 0.4921 (=745.8419 – 745.3498) between the deamidated and non-deamidated peptides, which matched the deamidation mass difference for a 2+ charge ion (within 2 ppm). The position of deamidation was further identified by the fragmentation (CID-MS2) observed for the deamidated peptide, Asparagine (N) at position 149. (Data not shown)

In addition, we did not observe any deamidation of the Asparagine at 152 positions in the same T15 peptide, and that could be explained by the next amino acid residue to N152 (not Glycine or Serine). After the characterization of the deamidation reaction, we used a similar approach as for monitoring oxidation to relatively quantitate the % deamidation of rhGH from both GNE rhGH and MIT rhGH and T15 peptide was used as an example in the Fig.4-8C. The reported percentages were the average of three measurements. As a result, we observed a similar percents of deamidation on T15 peptide from both the MIT rhGH and GNE rhGH. Other peptides with deamidation hot spots were also analyzed by using the similar approach. (Data not shown)
C: Extend of Deamidation on T15 peptide

<table>
<thead>
<tr>
<th>Tryptic #</th>
<th>AA #</th>
<th>Peptide</th>
<th>m/z</th>
<th>GNE rhGH</th>
<th>MIT rhGH</th>
</tr>
</thead>
<tbody>
<tr>
<td>T15 with Asn Deamidation</td>
<td>146-158</td>
<td>FDTNSHNDALLK</td>
<td>745.35 (2+)</td>
<td>6.95(±0.62)%</td>
<td>7.35(±1.12)%</td>
</tr>
<tr>
<td>T15</td>
<td>146-158</td>
<td>FDTDSHNDALLK</td>
<td>745.85 (2+)</td>
<td>93.12(±0.51)%</td>
<td>92.65(±1.12)%</td>
</tr>
</tbody>
</table>

Fig. 4-8. Extracted ion chromatography (XIC) of T15 peptide with its deamidation counterparts from tryptic digests of MTI rhGH and GNE rhGH.

A: Base peak XIC of the ion m/z 745.35 (2+) and 745.85 (2+) from GNE rhGH.

B: Base peak XIC of the ion m/z 745.35 (2+) and 745.85 (2+) from MIT rhGH.

C: Extend of Deamidation on T15 peptide.
4.4.5 Label-free quantitation by external standard with its calibration curve.

Since the concentration of MIT rhGH is unknown to us, we thus used the external GNE rhGH standards and its calibration curve to measure the concentration of MIT rhGH. Four different concentrations of GNE rhGH standard, as shown in the Fig.1 (0.5ug, 2ug, 5ug and 10ug) were loaded on the gel and processed side by side with MIT rhGH from SDS-PAGE, in-gel digestion to LC-MS detection. As samples were processed in parallel, variance caused by sample preparation and mass spectrometry can be minimized, which further enhance the data reliability.

As shown in Fig.4-9, the peak areas of two desired peptides were measured. The two peptides were selected based on the criteria that no variants or modifications occurred to those peptides. The peak areas from four different concentrations of GNE rhGH samples were displayed as a linear curve. As shown, both two desired peptides were linear from 0 to 10ug with r² greater than 0.99 (0.9914 for T4 and 0.9938 for T21). The concentration determined was quite similar by using either of the two peptides (slopes). As a result, the concentration of MIT rhGH was determined to be 6~7 ng/µL approximately.
The entire recombinant human growth hormone sequence with N-terminal peptide variants as well as post-translational modifications (oxidation and deamidation) has been extensively characterized by state-of-the-art LC-MS approach. Besides the correct N-terminal end peptide, two additional variants were successfully identified with amino acids from leader sequence attached to N-terminus in MIT produced rhGH. Both the N-terminal variants and modification sites for corresponding peptides of oxidation and deamidation were precisely determined by accurate mass measurement (Orbitrap MS) and peptide sequence assignment (MS/MS measurement). The extents of N-terminal variants, oxidation as well as deamidation were relatively quantitated by the ratio of peak areas of specific peptide (particular variant and modified peptide) on peak area of all related peptides. The concentration of MIT rhGH was also measured by external standard calibration.

**4.5 Conclusion**

![T4 standard curve](image1)

![T21 standard curve](image2)

*Fig.4-9. Linearity of quantitation of MIT rhGH by using external standards (T4 peptide and T21 peptide from GNE rhGH) with four different concentrations.*
In MIT rhGH, two additional variants, either two copies or one copy of Glu-Ala from leader sequence followed by rhGH amino acid sequence were identified and relatively quantitated. The original (correct) N-terminus peptide was found to account for only about 7.8% and the other two contaminated (incorrect) N-terminus peptides make up to 92.2% of all variants. As N-terminal amino acids exert critical influence on the efficacy of therapeutic drugs by mainly affecting the stability and serum half-life, a serious concern was raised up from the results that if the product is qualified.

Meanwhile, in relative comparison with GNE rhGH standard, MIT rhGH was identified with higher amount of oxidation on all three sites (Met14 in T2 peptide as an example in this chapter). And for the deamidation, no significant difference was observed between MIT rhGH and GNE rhGH (Asn149 in T15 peptide as an example in this chapter). To demonstrate the influence of subtle differences found in modifications, animal and human clinical studied are necessarily performed to alleviate the concerns.

In conclusion, we have demonstrated that the powerful LC-MS approach as described above enabled the extensive characterization and relative quantitation of recombinant human growth hormone produced by new manufacture process. This technique can also confidently and sensitively pinpoint the subtle but distinct differences due to different manufacturing process. Our analytical results provide critical and immediate feedbacks to the upstream manufacture groups and instruct them to improve the manufacture process. We believe our upstream group can benefit from our detailed analysis of their products as a valuable initial screening.
4.6 References


Supplemental to Chapter 4: Mass spectrometric analysis of N-terminal processing of recombinant viral protein produced from a CHO cell line.

Co-authors’ work in this chapter: The AAV5 protein was provided by Dr. Erno Pungor from BioMarin Pharmaceutical Inc. The protein analysis by LC-MS/MS was done by Prof. Hancock’s group.

Anna (Fan) Zhang: experimental design and perform, data analysis, manuscript writing and revision; Shiaw-lin Wu: idea contribution and experimental design; William Hancock: idea contribution and manuscript writing; Erno Pungor: grant support.

Industrial collaboration with BioMarin Pharmaceutical Inc.

S 4.1 Abstract

In the supplemental section of chapter 4, we have characterized the recombinant Adeno-Associated Virus protein (type 5) by our HPLC-mass spectrometric approach, the protein sequence has been characterized and sequence variants with unexpected N-terminal processing were discovered. The N-terminal amino acid sequence has been identified with the methionine residue cleaved and serine residue acetylated (for the major variant), as well as only the methionine residue cleaved (for the minor variant). The extent of N-terminal variants was relatively quantitated by the ratio of peak area of specific variant on the peak area of all related peptides. Also, to enhance the sequence coverage by single-enzyme digestion, the trypsin autolytic digestion was considered and matched against the database by no-enzyme search and two large peptides were identified by using this strategy. The analytical approach we developed in this study will provide an effective method to study the unexpected sequence variance during the production process and our analytical results can provide immediate and critical feedbacks to
our collaborators, assisting them in optimizing the production process with correct processing of the N-terminal sequence.

**S4.2 Introduction**

The biopharmaceutical industry, through the advent of recombinant therapeutic proteins, has revolutionized in the past decades [1]. A diversity of expression systems have been considered for the production of recombinant therapeutic proteins, including bacteria [2], yeast [3], and mammalian [4, 5]. Since the choice of an expression system has a profound impact on the maximum product yield as well as high product quality, choosing the expression system thereby has to be a very careful decision [6]. At the moment, biopharmaceutical manufacturing is mostly relying on mammalian cell expression system, as indicated by the pre-eminence of approved therapeutic produced in these cells [7-10]. Among the mammalian systems, the Chinese hamster ovary (CHO) cells, producing nearly 70% of recombinant protein therapeutics [6, 8], are the predominant choice in the current pharmaceutical market. Compared with other cell lines, the human-like post translational modifications produced, especially human-like glycosylation, are the most attractive advantages [5, 6, 9, 11]. Also, easy of transfection, extensive knowledge, as well as high product yield, are all benefits contributing to the extensive use of CHO cells [5, 6, 9, 11].

Two typical post-translational modifications occurred to protein N-terminus are N-terminal Methionine Cleavage (NMC) and N-terminal Acetylation (NTA). The NMC is a universally conserved activity and a highly specific mechanism across all life forms. Methionine is used to initiate protein synthesis for essentially all proteins. It is subsequently removed in a large of percentage of cases, either by cleavage of N-terminal signal peptide (as part of cellular
translation mechanisms or precursor activations) or by the action of specific Methionine Amino-peptidases (MetAPs). Approximately two-thirds of the proteins in any proteome are potential substrates for the latter N-terminal Methionine Cleavage. And the second amino acid in protein substrates is crucially important for NMC because MetAPs specifically mainly depends on the nature of this residue. These enzymes generally excise the N-terminal Met when the second residue is Gly, Ala, Ser, Thr, Cys, Pro, or Val [12-14] which are the amino acids smallest in size. The specificity of NME coincided with that of the N-end rule (NER) [15], an ubiquitin-dependent protein degradation process that is based on the recognition of N-terminal residues. The exposed P2-residues are all stabilizing residues, except for Methionine, introduced in the NER. The specificity of the MetAPs suggests an apparent connection between NMC and protein degradation. However, this connection has never been approved and thereby it is remains unclear whether exposing these stabilizing residues contributes to increasing protein half-life and thus represents a primary purpose of NMC.

The other N-terminal modification, N-terminal Acetylation (NTA), which involves the transfer of an acetyl group from Acetyl coenzyme A to the alpha amino group of the first amino acid residue of a protein, is catalyzed by N-terminal Acetyltransferases (NAT) [21]. Although the specificity of NAT is not as rigid as the MetAPs, the principal substrates in the stabilizing class are usually the four smallest residues (Gly, Ala, Ser and Thr). A second class of NATs can also modify the retained Met when the adjacent residues are Asp, Glu or Asn. N-terminal Acetylation has been suggested to be implicated with protein stability and degradation. Most examples indicated that blocking the N-terminus by NTA potentially prevents N-terminal Ubiquitination, and thus stabilizes the protein. But in contrast to the general idea, there are still a few examples showing that N-terminal acetylated amino acid sequence in certain proteins were found to be
involved in creating degradation signals and further marked with ubiquitin for destruction [21]. Thus, considering both NME and NTA, the influence of N-terminal processing on stability and degradation is shown to be protein-specific, without a general rule. It also probably that the N-terminal modifications may have other deleterious effects that are manifested in different ways.

In this work, a recombinant protein, adeno-associated virus (AAV) type 5, was expressed from a CHO cell strain line. Adeno-associated virus (AAV) proteins are always utilized as vectors of particular interest as they are capable of inducing transgenic expression in a broad range of tissues for a relatively long time without stimulation of a cell-mediated immune response [16]. Gene therapy using AAV vectors can infect both dividing and quiescent cells and persist in an extra chromosomal state without integrating into the genome of the host cell. These features make AAV proteins as very attractive candidates for creating viral vectors for gene therapy, recent human clinical trials using AAV vectors for gene therapy in the retina have shown promise[17]. AAV5, one of commonly used AAV vectors, is commonly used for transduction of protein of interest in the gene therapy. Native AAV5 is a single polypeptide (monomer) with 532 amino acids. The recombinant AAV5 expressed from CHO cell line has identical gene sequence and thus it should have the identical amino acid sequence with native AAV5. However, the protein AAV5 in our analysis was analyzed by RP-HPLC ESI-TOF system and the centroid mass obtained on the top of the peak was 59463.6 Da, which has -87.7Da (average) mass difference with the calculated isotopic average molecular mass. It is still not clear that where the difference between the predicted and measured molecular masses is located and why is caused during the expression, but it is highly possible to locate the exact difference by a detailed tryptic mapping. Thus, in this analysis, we used the LC-MS approach combined with single-enzyme digestion to characterize the recombinant AAV5 protein expressed from a CHO cell strain line, and the mass
difference was successfully localized on the N-terminal peptide and the major N-terminal variant was identified to be without N-terminal Methionine and with the second amino acid Serine acetylation.

**S4.3 Experimental Section**

**S4.3.1 Reagent and materials**

Adeno-Associated Virus type 5 (AAV5) proteins was obtained from BioMarin, Inc. (Novato, CA). Trypsin (Sequencing grade) was bought from Promega (Madison, WI). Ammonium bicarbonate, formic acid, dithiothreitol (DTT) and iodoacetamide (IAA) were purchased from Sigma-Aldrich (St. Louis, MO). Water at LC-MS grade was purchased from J.T. Baker (Philipsburg, NJ). HPLC grade acetonitrile was purchased from Thermo Fisher Scientific (Fairlawn, NJ). NuPAGE® 4-12% Bis-Tris gels, Novex® sharp pertained protein standard, SimplyBlue™ safestain were purchased from Invitrogen™ (Carlsbad, CA).

**S4.3.2 SDS-PAGE and In-gel digestion**

AAV5 protein was loaded on SDS-PAGE and separated with other impurities from CHO cell culture media. 2uL DTT and 4uL loading buffer were added to 10uL protein solution. The solution was heated at 70 °C for 10 min before gel electrophoresis. The gel band of protein AAV5 (about 60KDa) was cut out and minced into small pieces (approximately 1 mm²). The small gel pieces were dehydrated with acetonitrile and rehydrated with ammonium bicarbonate buffer (0.1M, pH 8.0) for 2~3 cycles, in order to remove the coomassie blue staining. The detained gel pieces was then reduced with DTT by addition of 200uL 10mM DTT in 0.1M NH4HCO3 and incubated at 56 °C for 30 min. The gel pieces were then alkylated with 200uL 55mM IAA in 0.1 M NH4HCO3 at room temperature for 60 min and this step is critical to
perform in the dark. After removal the liquid, the gel pieces were shrunked and rehydrated in the trypsin digestion buffer, which was prepared as 12.5 ng/µL trypsin in 50mM ammonium bicarbonate. Gel pieces with trypsin digestion buffer were firstly put at 4°C for 35 min and the trypsin buffer was maximally absorbed by gel pieces without digestion. Then the digestion buffer was placed by the buffer without trypsin, in case of self-digestion of excess trypsin and the protein solution was subsequently incubated at 37°C for an overnight. Then, the supernatant was collected on the second day; the peptides after enzymatic digestion were extracted from gel pieces by 25mM ammonium bicarbonate and acetonitrile. And the enzymatic digestion was quenched by 5% formic acid at 37°C for 15 min. Supernatants were collected in the same tube and concentrated to ~10uL by speed vacuum. The concentrated tryptic peptides were further subjected to LC-MS analysis.

**S4.3.3 LC-MS analysis**

The tryptic peptides were analyzed on an Orbitrap Elite MS instrument (Thermo Fisher Scientific, San Jose, CA) equipped with a nano spray ion source (New Objective, Woburn, MA). An Ultimate 3000 nano LC pump (Dionex, Mountain View, CA) was on-line coupled to the mass spectrometer through a 75 µm i.d. x 10 cm C18 column, packed with Magic C18 beads (200Å pore and 5 µm particle size) (Michrom BioResources, Auburn, CA) Mobile phase A was 0.1% formic acid in water, and mobile phase B was 0.1 % formic acid in acetonitrile. Before loading the samples, the C18 column was equivalent by using 2% mobile phase B at flow rate 200nL/min for at least 30 min. The tryptic peptides were separated by elution gradient, which was maintained at 200nL/min using 2% B to 40% B in 60 min, then from 40% to 80% B for 10 min. The mass spectrometry was operated in the data-dependent mode which alternatively switched between MS and MS2 acquisition. Full survey full-scan MS spectra was acquired in the
FTICR cell with mass resolution of 100,000 at m/z 400 (after accumulation to a target value of 2x10^6 ions in the linear ion trap), followed by 8 sequential data-dependent MS/MS scans.

Dynamic exclusion was utilized with exclusion duration of 30 sec and 2 repeat counts.

**S4.3.4 Protein identification**

Peptide identification was initially searched against sequence of Adeno-Associated Virus type 5 (AAV5) protein using the SEQUEST algorithm incorporated into Thermo Discovers software (version 1.3), Enzyme is set to trypsin with up to 3 miscleavages. Peptides are automatically identified if the peptide has an Xcorr score (+1 > 1.5; +2 > 2.0; +3> 2.5). Orbitrap-Elite provides accurate mass measurement (<5 ppm) for confirmation of peptide identification. Thus, the Xcorr scores (+1 > 1.5; +2 > 2.0; +3> 2.5) were used for the initial screening. Next, peptides which cannot be found by database search may be responsible for the precursor mass differences obtained from the Top-down analysis, which need to be manually extracted or searched again based on different criteria (i.e. no enzyme search or Methionine oxidation). Accurate precursor mass measured and CID-MS2 fragmentation pattern will provide a reliable result.

**S4.4 Results and Discussion**

Adeno-associated virus (AAV) based vectors are of particular interest as they are capable of inducing transgenic expression in a broad range of tissues for a relatively long time without stimulation of a cell-mediated immune response[16]. Gene therapy vectors using AAV can infect both dividing and quiescent cells and persist in an extra chromosomal state without integrating into the genome of the host cell. These features make AAV become very attractive candidates for creating viral vectors for gene therapy, recent human clinical trials using AAV for gene therapy in the retina have shown promise [17].
As mentioned in the introduction, the AAV5 analyzed in our analysis, has -87.7Da (average) mass difference with the calculated isotopic average molecular mass. It is still not clear that where the difference between the predicted and measured molecular masses is located and why is caused during the expression, but it is highly possible to locate the exact difference by a detailed tryptic mapping. Thus the primary structure of rAAV5 was extensively characterized by our LC-MS analysis, in the following section.

**S4.4.1 Identification of Primary structure**

In our analysis of primary structure, N-terminal T1 peptide was not found in the initial screening by using database search with traditional parameters, thus, we illustrated the identification of C-terminal end peptide of AAV5 in this section. As shown in the Fig.S4-1, the C-terminal peptide (T36) was identified at 32.53 min (Fig.S4-1A), with the accurate mass assignment for the peptide, m/z 762.45 with a single charge (Fig.S4-1B), the precursor ion (m/z 762.45) was isolated using the data-dependent acquisition mode and subjected to CID-MS2 fragmentation in the linear ion trap (Fig.S4-1C). The characteristic fragmentation pattern of CID was observed as shown in the figure.

Other peptides of rAAV5 were identified in a similar manner as shown in Fig.S4-1. The identification of tryptic peptides of rAAV5 were summarized in the table S4-1, along with the corresponding retention time and intensity. As shown in the table S4-1, all tryptic peptides were identified except for 3 small tryptic peptides (T10, T16 and T33), which cannot be retained on the C18 column, as well as 3 large tryptic peptides (T1, T13, and T25), which may be too hydrophobic to elute out. Or in another possibility, these unidentified peptides may have
unexpected sequence substitution or modification, which is responsible for the ~88 Da mass difference.

Fig. S4-1 LC-MS analysis of the C-terminal (T36) peptide from the tryptic digests of recombinant viral protein AAV5.

A: extracted ion chromatogram of m/z 762.45. B: Precursor mass scan at 32.53min using FTICR. For illustration purpose, only m/z 760-768 region is shown. C: MS2 scan of the m/z 762.45 (1+) ion.
<table>
<thead>
<tr>
<th>T#</th>
<th>AA#</th>
<th>Sequence</th>
<th>m/z</th>
<th>RT (min)</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>2-36</td>
<td>(-) MSAGGGGPLGD NNQGADGVGNASGDWHCDSTWMGDR(V)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2-T3</td>
<td>37-43</td>
<td>(R)JVTSTR(T)</td>
<td>396.74 (2+)</td>
<td>11.09</td>
<td>7.74E4</td>
</tr>
<tr>
<td>T4</td>
<td>44-56</td>
<td>(R)TWVLPSYNNHQR(E)</td>
<td>839.41 (2+)</td>
<td>41.01</td>
<td>5.37E7</td>
</tr>
<tr>
<td>T5-T6</td>
<td>57-59</td>
<td>(R)EIKSGVDGSAHANAIYFGYSTPGWGYDFDNRF(F)</td>
<td>1084.58 (3+)</td>
<td>53.19</td>
<td>6.74E5</td>
</tr>
<tr>
<td>T7</td>
<td>86-93</td>
<td>(R)FHSWSPR(D)</td>
<td>527.25 (2+)</td>
<td>24.81</td>
<td>6.76E7</td>
</tr>
<tr>
<td>T8</td>
<td>94-97</td>
<td>(R)DWQR(L)</td>
<td>604.28 (1+)</td>
<td>20.91</td>
<td>3.15E7</td>
</tr>
<tr>
<td>T9</td>
<td>98-108</td>
<td>(R)LINNYWFRP(R)</td>
<td>718.38 (2+)</td>
<td>43.1</td>
<td>8.08E7</td>
</tr>
<tr>
<td>T10</td>
<td>109-111</td>
<td>(R)SLR(V)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T11-T12</td>
<td>112-120</td>
<td>(R)VKIFNIOVK(E)</td>
<td>544.85 (2+)</td>
<td>42.48</td>
<td>3.10E7</td>
</tr>
<tr>
<td>T13</td>
<td>121-181</td>
<td>(K)ETVQDSTITANLTTSTVQVTDDQYLPY(V)</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>T14</td>
<td>182-190</td>
<td>(R)DNTENPTER(S)</td>
<td>538.23 (2+)</td>
<td>17.61</td>
<td>6.23E7</td>
</tr>
<tr>
<td>T15</td>
<td>191-202</td>
<td>(R)SSFCEYFPSK(M)</td>
<td>756.34 (2+)</td>
<td>53.37</td>
<td>7.65E7</td>
</tr>
<tr>
<td>T16</td>
<td>203-205</td>
<td>(K)MRL(T)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T17</td>
<td>206-233</td>
<td>(R)TGNFETYNFEEVFPFSSFAPSQNLK(L)</td>
<td>1100.17 (3+)</td>
<td>59.77</td>
<td>1.86E6</td>
</tr>
<tr>
<td>T18</td>
<td>234-245</td>
<td>(K)LANPLVDQYLRF(F)</td>
<td>488.93 (3+)</td>
<td>47.40</td>
<td>2.50E6</td>
</tr>
<tr>
<td>T19</td>
<td>246-259</td>
<td>(R)FVSTNNTGGQFQNNK(N)</td>
<td>756.87 (2+)</td>
<td>31.97</td>
<td>8.14E7</td>
</tr>
<tr>
<td>T20</td>
<td>260-264</td>
<td>(K)NLVRG(Y)</td>
<td>531.29 (1+)</td>
<td>16.27</td>
<td>6.36E5</td>
</tr>
<tr>
<td>T21</td>
<td>265-270</td>
<td>(R)YANTKY(K)</td>
<td>380.19 (1+)</td>
<td>18.63</td>
<td>3.28E6</td>
</tr>
<tr>
<td>T22</td>
<td>271-279</td>
<td>(K)NWFPPGPMGR(T)</td>
<td>531.25 (2+)</td>
<td>45.71</td>
<td>6.43E7</td>
</tr>
<tr>
<td>T23</td>
<td>280-291</td>
<td>(R)TQGWNLSQGNVRA</td>
<td>644.82 (2+)</td>
<td>36.12</td>
<td>1.60E8</td>
</tr>
<tr>
<td>T24</td>
<td>292-302</td>
<td>(R)ASVSAFTTNR(M)</td>
<td>562.78 (2+)</td>
<td>30.49</td>
<td>1.41E8</td>
</tr>
<tr>
<td>T25</td>
<td>303-368</td>
<td>(R)M*(oxidation)ELEGASYQVPQPQPGMTNNLNGSNTYALENTMIFSQAPNTATTLEYGNMLITSESETQVNR(V)</td>
<td>1173.33 (4+)</td>
<td>57.44</td>
<td>3.71E4</td>
</tr>
<tr>
<td>T26</td>
<td>369-407</td>
<td>(R)VAYNVGGQATNNQSTS TAPATGTYNQEIPPGSVWMER(D)</td>
<td>1036.49 (4+)</td>
<td>52.66</td>
<td>5.8E5</td>
</tr>
<tr>
<td>T27</td>
<td>408-418</td>
<td>(R)DYVLQGPIWAK(I)</td>
<td>645.34 (2+)</td>
<td>46.96</td>
<td>8.57E7</td>
</tr>
<tr>
<td>T28</td>
<td>419-438</td>
<td>(K)IPETGAHFPSAMPMDGGFKL(H)</td>
<td>684.34 (3+)</td>
<td>40.76</td>
<td>6.41E7</td>
</tr>
<tr>
<td>T29</td>
<td>439-447</td>
<td>(K)HPPPMMLK(N)</td>
<td>532.29 (2+)</td>
<td>39.67</td>
<td>1.33E8</td>
</tr>
<tr>
<td>T30-T31</td>
<td>448-485</td>
<td>(K)NTPVPGNITSDVVPQSFQTSYTGQQVTVQEMWELKKEKSR(R)</td>
<td>1173.33 (4+)</td>
<td>57.44</td>
<td>3.71E4</td>
</tr>
<tr>
<td>T32</td>
<td>490</td>
<td>(K)W(R)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T33</td>
<td>491-518</td>
<td>(R)WNPEIQYTNYNDP QFVDFAPDSTGDFR(T)</td>
<td>1127.49 (3+)</td>
<td>49.46</td>
<td>6.34E5</td>
</tr>
<tr>
<td>T34</td>
<td>519-526</td>
<td>(R)TRTPGTR(Y)</td>
<td>451.26 (2+)</td>
<td>16.60</td>
<td>1.92E8</td>
</tr>
<tr>
<td>T35</td>
<td>527-532</td>
<td>(R)YLTPR(-)</td>
<td>762.45 (1+)</td>
<td>33.84</td>
<td>4.75E7</td>
</tr>
</tbody>
</table>

Table S4-1: Summary of tryptic peptides of recombinant viral protein, AAV5
S4.4.2 Identification of N-terminal processing

After the initial screening of amino acid sequence by LC-MS approach, we then specifically look for the missing peptides, which may have potential substitutions or modifications. Asn-terminal end peptide always presents a higher chance to obtain variance (i.e. incomplete signal peptidase digestion or N-terminal protein processing), we primarily focused on the identification of possible N-terminal variants.

To look for the possible variants, the database was set up to no enzyme search in the enzyme specification; as a result, N-terminal peptide was automatically matched out with Met cleaved. The T1 peptide without Methionine residue, as shown in the Fig.S4-2, was identified at 40.93 min (Fig.S4-2A), with the accurate precursor mass measurement (m/z = 1159.1416, z=3) (Fig.S4-2B) and CID-MS2 of the precursor ion (Fig.S4-2C), the high abundant product ions, as a result of CID fragmentation, were indicated in the figure. However, if Met-cleaved T1 peptide was the only variant, the total molecular weight of protein should have 131 Da, instead of 88Da, less than the theoretical molecular weight. Also, the Met-cleaved T1 peptide was identified with very low intensity (~10E4). Thus, it is highly possible that the identified T1 peptide may only take a small portion of N-terminal variants.
Fig.S4-2 LC-MS analysis of the N-terminal peptide variant 1 (T1 peptide without Methionine) from the tryptic digest of recombinant viral protein AAV5.

A: Base peak ion chromatogram. B: Precursor mass scan at 40.93 min using FTICR. For illustration purpose, only m/z 1157-1163 region is shown. C: MS2 scan of the m/z 1159.14 (3+) ion.
For recombinant proteins expressed from CHO cells, the proteins may undergo many different types of chemical modifications within the cell. Particularly for N-terminal end peptide, N-terminal methionine cleavage and N-terminal acetylation are two typical modifications. It has been reported that the functional role of NMC may be primarily to expose Ala and Ser rather than other residues and the exposed Ala and Ser residues are tended to have NTA or other post-translational modifications [18]. And in our case, after methionine excision, the second exposed amino acid is just serine, which is highly possible to be acetylated.

Based on the assumption, we specifically extract the precursor mass of predicted T1 peptide. Interestingly, T1 peptide was also identified as expected, without methionine and with second amino acid serine acetylation, which is indicated in the Fig.S4-3, with the accurate precursor mass and CID-MS2 fragmentation pattern. As shown in the Fig.S4-3, the N-terminal T1 peptide, was identified at 42.80 min (Fig.S4-3A), the accurate precursor mass of this peptide, m/z 1173.81 with 3+ charge, was shown in Fig.S4-3B, which is within 5ppm with its theoretical mass. The precursor ion (1173.81, 3+) was isolated using data-dependent acquisition mode and subjected to CID-MS2 fragmentation in the linear ion trap (Fig.S4-3C); the characteristic fragmentation pattern of CID was indicated as labels in the figure.
Fig.S4-3 LC-MS analysis of the N-terminal peptide variant 2 (T1 peptide without Methionine but with Serine acetylation) from the tryptic digest of recombinant viral protein AAV5.

A: extracted ion chromatogram of m/z 1173.81. B: Precursor mass scan at 42.8min using FTICR. For illustration purpose, only m/z 1170-1178 region is shown. C: MS2 scan of the m/z 1173.81 (3+) ion.
In summary, two variants of T1 peptide in recombinant AAV5 protein were identified by our LC-MS approach. The major form is T1 peptide without Methionine but with Serine acetylation (S(acetyl)AGGGGPLGDNNQGADGVGNASGDWHCDSTWMGDR) and the minor form is T1 peptide without Methionine and also without Serine acetylation (SAGGGGPLGDNNQGADGVGNASGDWHCDSTWMGDR).

As the two variants of T1 peptide were detected in the same LC-MS analysis, the relative quantitation of these two variants can be easily performed as our previous studies [19-21]. Besides, the percentage of each variant can be estimated by the peak area of the each peptide divided by the sum of the peak area of both peptide variants based on an assumption that two variants had similar response factor. The T1 peptide with Serine acetylation accounted for 82.8% and T1 without serine acetylation accounted for 17.2%. The major form of T1 peptide without Methionine (-131Da) and with Serine acetylation (+42Da) has a mass difference of 89Da, which is very close to the average mass difference measured (88Da) in the top-down analysis.

**S4.4.3 Autolytic digestion of Trypsin Protease**

In the tryptic peptide mapping, two large peptides (T13 and T25) were not detected. Regularly multi-enzyme strategy is performed to further cut the peptide to proper size, which thereby can eluted out from C18 column and detected by MS. However, this approach needs a separate sample preparation. For samples which can only be processed once, we used no-enzyme search to identify non-specific digested peptides with accurate precursor mass and high confident MS2 spectrum, by considering the autolytic digestion of trypsin.

It is commonly known that trypsin specifically cleaves at the carboxylic side of Lysine and Arginine residues. However, trypsin also can also be subjected to autolysis, generating...
Pseudotrypsin, which exhibits a broadened specificity including a chymotrypsin-like activity, cleaving at the C-terminal of aromatic residues (Tyrosine, Tryptophan, and Phenylalanine). Even though the trypsin we purchased has been modified by reductive methylation, which was intended to yield a highly active and stable molecule and maximally resist to autolytic digestion, we still can observe some peptides, resulted from autolysis, in our results by software no-enzyme search. In the result of database search, T13 peptide was unexpectedly cleaved into two parts through amino acid Tyrosine 151. As shown in the Fig.S4-4, the first part of T13 peptide (EVTVQDSTTTIANNLTSTVQVFTDDYQLPY, 121-151) was identified at 63.23 min (Fig. S4-4A), with the accurate precursor mass measurement (m/z=1160.22, z=2) (Fig. S4-4B) and the CID-MS2 of the precursor ion (Fig.S4-4C), the precursor mass detected is less than 5ppm with its theoretical mass and the high-abundant product ions, generated from CID-MS2 were indicated in the figure as labeled. It has to be noted that trypsin autolysis only accounts for a very small amount of digestion, accordingly, the peptides resulted from autolytic digestion, were identified by mass spectrometry with reasonably low intensity.

Similarly, the second part of T13 peptide (VVNGTEGCLPAFPQVFTLPQGYATLNR, 152-181) was also identified with precursor mass less than 5ppm compared to the theoretical mass and high confident CID-MS2 with X-correction 5.65 for triplet charged precursor ion. (Fig. S4-5)
T13-01: EVTVQDSTTTIANLNLTSTVQFTDQDYPY (121-151)

Fig S4-4 LC-MS analysis of T13-part1 from the tryptic digest of recombinant viral protein AAV5.

A: Base peak ion chromatogram. B: Precursor mass scan at 63.23 min using FTICR. For illustration purpose, only m/z 1159-1163 region is shown. C: MS2 scan of the m/z 1160.90 (3+) ion.
T13-02: VVGNGTEGCLPAFPPQVFTLPQYGYATLNR (152-181)

Fig.S4-5 LC-MS analysis of T13-part2 from the tryptic digests of recombinant viral protein AAV5.

A: Base peak ion chromatogram. B: Precursor mass scan at 57.88 min using FTICR. For illustration purpose, only m/z 1087-1094 region is shown. C: MS2 scan of the m/z 1090.21 (3+) ion.
The same approach was also performed to T25 peptide, which was detected to be digested into three parts. The first part and third part amino acid sequence were also identified with reliable data (Data not shown).

In summary, the AAV5 protein was identified with about 95% sequence coverage with single-enzyme and one-time sample preparation. The mass difference was identified to N-terminal Methionine excision and with serine acetylation.

**S4.5 Conclusion**

The recombinant Adeno-Associated Virus protein (type 5) sequence with unexpected N-terminal processing has been extensively characterized by our LC-MS approach.

The sequence coverage can reach 95% only by single-enzyme (trypsin) digestion and one-time sample preparation, with consideration of trypsin autolytic digestion. The unexpected mass difference was localized at N-terminal end peptide. The major N-terminal variant was identified with N-terminal methionine cleavage and serine acetylation and the minor variant was T1 peptide also with Methionine excision but without serine acetylation. The extent of serine acetylation was relatively quantitated by the ratio of peak area of specific peptide (particular variant) on peak area of all related peptides. The T1 peptide with serine acetylation accounts for 83% and that without serine acetylation account for 17%. To enhance the sequence coverage by single-enzyme digestion and one-time sample preparation, the trypsin autolytic digestion was considered and matched against the database by no-enzyme search. Two large peptides, T13 digested into two parts, were completely identified and T25, cleaved into three parts, were partially identified by using this approach.
Even though the exact influence of N-terminal protein processing, including Methionine excision and serine acetylation, of AAV5 is still unknown, the roles of these N-terminal modifications has been suggested in acting as degradation signal and as a determining factor for preventing protein targeting to the secretory pathway in recent studies[21]. Thus, we assumed that N-terminal processing of AAV5, the vector of transgenic, may influence the transgenic stability or in-vivo half life time during transduction process, which need to be further tested in the related experiments. The analytical approach we developed in this study will also provide an effective method to study the unexpected sequence variance during the production process and our analytical results can also provide immediate and critical feedbacks to our collaborators, assisting them in trouble-shooting problems in the production process, such as factors inducing protein N-terminal processing.
S4.6 References


Dissertation Conclusion

In this dissertation, liquid chromatography- mass spectrometry based strategies have been extensively used to identify and characterize the protein glycosylation, including the glycosylation occupancy, glycopeptide and glycan analysis. Currently, the bottom-up MS-based approaches (started from glycopeptides and glycans) combined with different dissociation techniques (e.g. CID, ETD, or HCD) were mostly applied in the analysis of protein therapeutics while top-down MS-based approaches (started from intact glycoproteins) still have some problems due to constraints of MS instruments. With the improvements of MS instruments (e.g. resolution), the top-down based MS strategies will become a more convenient method to characterize glycoproteins, without the tedious time-consuming enzymatic digestion procedures. Also, the improvements of enrichment, purification and fractionation techniques will also be helpful to the glycosylation characterization. Besides the glycosylation analysis, the dissertation also demonstrated the successful application of the state-of-the-art LC-MS platform on other important post-translational modifications, such as oxidation, deamidation and N-terminal variants, which could provide the valuable information for better understanding of the biological pharmaceuticals.
Additional Research Projects and Publications

The work was done by senior students and my participation is part of data analysis and discussion.
The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome

To the Editor:

The Chromosome-Centric Human Proteome Project (C-HPP) aims to define the full set of proteins encoded in each chromosome through development of a standardized approach for analyzing the massive proteomic data sets currently being generated from dedicated efforts of national and international teams. The initial goal of the C-HPP is to identify at least one representative protein encoded by each of the approximately 20,300 human genes1,2. The proteins will be characterized for tissue localization and major isoforms, including post-translational modifications (PTMs), using quantitative mass spectrometry and antibody reagents. Our rationale is that effective integration of proteomics data into a genomic framework will lead to improved knowledge of complex biological systems and facilitate access to protein level data. Although the intent to engage in a C-HPP program has been noted1–3, our objective here is to define the goals and process for its development as a multinational program.

Over the past three years, the Human Proteome Organization (HUPO) has developed a strategy for the first phase of the Human Proteome Project (HPP; http://thehpp.org/; Supplementary Fig. 1). HPP1 goals will be achieved through cooperation with the C-HPP to characterize the human proteome on a chromosome-by-chromosome basis and with the biology- and disease-driven projects (B/D-HPP). Human genome studies, such as the 1000 Genomes Project and Encode, and transcriptome sequencing provide a basis for identification of protein isoforms generated by alternative splicing transcripts (ASTs) and by nonsynonymous single-nucleotide polymorphisms (nsSNPs; Supplementary Fig. 2). Additional protein forms will be identified through characterization of post-translational modifications. A basic premise of the HPP is that C-HPP data sets will have substantial utility for biological and disease studies. With development of new tools for in-depth characterization of the transcriptome and proteome, the HPP is well positioned to have a strategic role in addressing the complexity of human phenotypes. With this in mind, the HUPO has organized national chromosome teams that will collaborate with well-established laboratories building complementary proteotypic peptides, antibodies and informatics resources.

An important C-HPP goal is to encourage capture and open sharing of proteomic data sets from diverse samples to enhance a gene- and chromosome-centric display. This will display several layers of biological information on a common reference platform comparable to a genome browser. Such context will effectively integrate transcriptomics data such as RNA-Seq with proteomic data sets (Fig. 1).

Although the C-HPP program has some similarities to the Human Genome Project (HGP)4 in its quest for complete coverage across the genome, the C-HPP has the added challenge of characterizing protein expression at the tissue, cellular and subcellular levels, as well as PTMs, ASTs and protease-processed protein variants. An example of protein variation is shown for 6 selected genes on chromosome 13 (BRCA2, 3 ASTs and 54 SNPs in protein-coding regions (nsSNPs); RB1, 2 ASTs and 3 nsSNPs; and IRS2, 1 AST and 3 nsSNPs) and chromosome 17 (BRCA1, 24 ASTs and 24 nsSNPs; ERBB2, 6 ASTs and 13 nsSNPs; and TP53, 14 ASTs and 5 nsSNPs; Table 1 and Supplementary Table 1).

The C-HPP will build on the three HPP pillars that provide both technology and resources for mapping the human proteome: mass spectrometry–based SRMAtlas, antibody reagents in the Human Protein Atlas and bioinformatics knowledge linked by ProteomeXchange, specifically the proteomics identification database (PRIDE), Tranche, PeptideAtlas, the global proteome machine database (GPMDB), UniProt and nXtProt (Supplementary Fig. 3).

The C-HPP does not propose any alteration in the work flow of a typical proteomics laboratory; instead, it seeks more effective use of data encompassed in existing bioinformatics resources, which will be combined with targeted studies to generate a robust list of observed protein isoforms (Supplementary Fig. 3). A potential challenge to data collection from different laboratories is the diversity of instrument and bioinformatics platforms and quality criteria. The C-HPP will work closely with proteomics journals, and use existing data (GPMDB and PeptideAtlas), literature curation (Uniprot and nXtProt) and standardization programs (PSI, CPTAC, Unimod, ABRF and ASMS) to ensure that the data collection is efficient, with consistent quality assurance and quality control. Journal mandates for deposition of raw data upon publication will reinforce this process5. The C-HPP has already encouraged formation of chromosome-formatted databases (http://www.nextprot.org/; http://www.gpm.org/) in which new data sets are integrated with existing ones. In this manner the C-HPP will capture the protein evidence emerging from the hundreds of laboratories worldwide engaged in hypothesis-driven study.

Table 1 Features of salient genes on chromosomes 13 and 17

<table>
<thead>
<tr>
<th>Gene*</th>
<th>ASTs</th>
<th>nsSNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chromosome 13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA2</td>
<td>3</td>
<td>54</td>
</tr>
<tr>
<td>RB1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>IRS2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Chromosome 17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA1</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>ERBB2</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>TP53</td>
<td>14</td>
<td>5</td>
</tr>
</tbody>
</table>

*Ensembl protein and AST information can be found at http://www.ensembl.org/Homo_sapiens/. AST, alternative splicing transcript; nsSNP, nonsynonymous single-nucleotide polymorphism assembled from data from the 1000 Genomes Projects.
CORRESPONDENCE

![Diagram](image)

**Figure 1** Genomic, transcriptomic and protein information for the set of genes present in selected regions of chromosomes 13 and 17. (a,b) The information provides a comprehensive landscape with respect to protein evidence, quality of mass spectrometry–based protein identification, availability of antibody, disease relationship, and phosphorylation, acetylation, glycosylation and transcriptomic information. It shows the degree of protein annotation on two important regions on chromosomes 13 and 17. (c,d) The information provides a comprehensive landscape with respect to protein evidence, quality of mass spectrometry–based protein identification, availability of antibody, disease relationship, and phosphorylation, acetylation, glycosylation and transcriptomic information. It shows the degree of protein annotation on two important regions on chromosomes 13 and 17. (Fig. 1) summarizes the following extensive data sets on the basis of existing data compilations: protein evidence, mass spectrometry data, antibody availability, major PTMs, disease information and transcript level, including ASTs from three different samples in a format viewable for associations between data sets and information gaps in specific chromosome regions.

In phase 1 (~6 years), the C-HPP plans to map all proteins currently lacking high-quality mass spectrometry evidence, three major classes of PTMs, many representative AST products and many nsSNP sequence variants. The characterization will be followed by antibody-based detection in selected tissues and cell lines. In phase 2 (~4 years), identified proteins will be characterized and validated with additional proteomic and antibody measurements. Throughout this 10–year project, the C-HPP aims to generate information useful in the search for new biomarkers and drug targets and also in the study of disease gene families clustered in each chromosome (for example, the cytokerin family in chromosome 17). C-HPP outputs will be integrated with output from the parallel B/D-HPP project. The C-HPP has selected the UniProt protein list (based on Ensembl genome builds) as the starting point for identified proteins. Individual chromosome teams will use information collected in well-annotated databases (for example, GPMDB, PeptideAtlas and neXtProt) to develop a list of missing or poorly identified proteins for a particular chromosome. A plot of such data (for example, Fig. 1) can identify chromosomal regions with low amounts of data. For example, there is protein paucity for regions on chromosome 17 that contain olfactory receptors and keratin-binding proteins; this may be expressed in limited proteomic data sets for nasal epithelium and bone and hair samples, respectively (Fig. 1). The missing data can be obtained through collaborations with laboratories with expertise in such samples or by selection of new sample sets for protein identifications guided by transcriptomics measurements. To facilitate selection of samples suitable for mass spectrometry discovery of an individual missing protein, the C-HPP will collaborate with RNA-Seq laboratories to take advantage of specimens and transcriptomics data (Supplementary Fig. 2). We recognize that some proteins may not be suited for mass spectrometry measurements owing to their physical research or high-throughput proteome-wide studies.

Although chromosome-based protein data curation is a relatively new concept in proteomics, our justification is based in part on compatibility of this data format with the output of RNA-Seq. We think the search for yet-to-be-discovered protein products of genes can be informed by transcriptomics measurements of selected tissues and cell lines. The C-HPP will also prioritize specific tasks to laboratories with expertise in particular protein subsets (for example, membrane proteins), specific protein variations (PTMs, alternative splicing and protease-processed variants), deep profiling for low-abundance proteins and targeted subcellular localization studies. We recognize the popularity of other current bioinformatic methods used to organize complex data sets by functional classes; we will incorporate this information into the C-HPP browser. An example of such a global view for selected regions of chromosomes 13 and 17 (Fig. 1)
properties or lack of appropriate biological samples; other approaches such as generation of ribosomal DNA standards, antibody localization approaches and molecular biology tools will be used.

Given expected refinements in the human gene list, the C-HPP protein list will reflect updates in Uniprot that are captured in proteomic databases. To ensure consistent data quality across chromosome groups, the C-HPP will encourage prompt deposition of data. For antibody-based studies, the C-HPP will promote the use of cultured primary or transformed cells, including induced pluripotent stem cells, which can be maintained in perpetuity for reanalysis and for subcellular fractionation. Such cell-based studies will be augmented with tissue profiling, as in the Human Protein Atlas project. Enrichment for nuclear, mitochondrial and other subcellular organelles may be especially informative. The C-HPP will integrate antibody- and mass spectrometry–based measurements.

Another goal of the C-HPP is to procure high-quality reagents. To augment commercially available sources, the national teams will establish centralized antibody banks. This will be achieved through a close collaboration between each chromosome group and antibody resource groups or suppliers. In a similar manner, selected reaction monitoring peptide banks will be developed for quantitative mass spectrometry measurements.

The project will meet its aims when the comprehensive C-HPP database is 100% matched with the 20,300 protein-coding genes annotated on the human genome sequence, including at least one representative AST and nSSNP, tissue localization and three classes of PTMs in whole-chromosome sets (22 autosomal, X and Y; Supplementary Table 2).

The C-HPP is led by cochairs Young-Ki Paik (Korea), Bill Hancock (USA) and György Marko-Vargas (Sweden), an executive committee and a council of principal investigators of each of the chromosome teams (thus far, 15 investigators for 14 chromosomes; Supplementary Fig. 2). The initial C-HPP team emerged from an exploratory group in Korea that selected chromosome 13; it has several key metabolic disease genes (for example, IRS2, which is associated with diabetes, and CLF, which is associated with cholesterol metabolism). Diverse approaches have been pursued by other countries and teams. A US team has focused on breast cancer, selecting chromosome 17, which contains the oncogenes ERBB2 and BRCA1.

Similarly, the Australia–New Zealand team selected chromosome 7, with a focus on colon cancer and epidermal growth factor receptor. C-HPP guidelines now have been set for the assignment and progress review of chromosome-based teams and standardization of outputs (Y.-K. Paik, G.S. Omenn, M. Uhlen, S. Hanash, G. Marko-Varga et al., unpublished data). As of December 2011, based on their interests in a specific disease (for example, male infertility in Iran) or gene cluster (for example, liver-origin proteins in China), other international teams have chosen chromosomes 1 (China), 2 (Switzerland), 3 (Japan), 6 (Canada), 11 (Korea), 14 (France), 18 (Russia), 19 (Sweden and Germany, Norway, India, China and Spain), 21 (Canada), X (Japan) and Y (Iran). A Swedish team has published extensive findings for chromosome 21 (ref. 9). The C-HPP guidelines specify management of the project, data quality and data sharing metrics, reporting formats, and processes and criteria by which countries or researchers are designated to take the lead for a specific chromosome (Y.-K. Paik et al., unpublished data).

In conclusion, we envision that effective integration of transcriptomics and proteomics data will provide insights through a more complete ‘parts list’ and enhance a comprehensive understanding of human biology. The HPP and the C-HPP represent an even larger endeavor than the HGP. This challenge has led the EUPO to promote an efficient approach of recruiting national teams with clear areas of responsibility and effective collaborations among leading proteomic laboratories in the HPP consortium. Recognizing the complexity of the human proteome, we have set 10-year goals for characterizing the major forms of the complete set of proteins. The C-HPP will provide a global open Web interface for data collection, curation and presentation of the proteome parts list and will stimulate availability of high-quality protein capture and signature peptide reagents (Supplementary Table 2). Importantly, the C-HPP will work with governmental funding bodies to address major gaps in proteomics infrastructure, such as secure archiving of large data sets.

Note: Supplementary information is available on the Nature Biotechnology website.

AUTHOR CONTRIBUTIONS

ACKNOWLEDGMENTS
We thank R. Beavis for his critical comments and support for this work. Y.-K.P thanks the Korean Human Proteome Organization HPP planning committee members for their contribution to the project in the early phase. Work involving chromosomes 13 and 17 in this paper was supported in part by the World Class University program funded by the Korean Ministry of Education, Science and Technology (to Y.-K.P and W.S.H.), a grant from the Korean Ministry of Health and Welfare (to Y.-K.P), and grants from the US National Cancer Institute (to M.S. and W.S.H.).

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Young-Ki Paik1, Seul-Ki Jeong1, Gilbert S Omenn2,3, Mathias Uhlen4, Samir Hanash5, Sang Yun Choi6,7, HyoYoung-Joo Lee8, Keun Na1, Eun-Young Choi1, Fangfei Yan9, Fan Zhang9, Yue Zhang9, Michael Snyder7, Yong Cheng7, Rui Chen7, György Marko-Varga8, Eric W Deutsch9, Houghen Kim9, Ja-Young Kwon9, Ruedi Aebersold10, Amos Bairoch11, Allen D Taylor4, Kwang Youl Kim1, Eun-Young Lee1, Denis Hochstrasser11, Pierre Legrain12 and William S. Hancock1,5

1Yonsei Proteome Research Center, Yonsei University, Seoul, Korea. 2Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA. 3Institute for Systems Biology, Seattle, Washington, USA. 4Royal Institute of Technology, Stockholm, Sweden. 5Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. 6Northeastern University, Boston, Massachusetts, USA. 7Stanford University, Palo Alto, California, USA. 8Lund University, Lund, Sweden. 9Yonsei University College of Medicine, Seoul, Korea. 10Department of Biology, Institute of Molecular Systems Biology, Eidgenössische Technische Hochschule, Zürich, Switzerland, and Faculty of Science, University of Zurich, Zurich, Switzerland. 11Swiss Institute of Bioinformatics and University of Geneva, Geneva, Switzerland. 12Ecole Polytechnique, Palaiseau, France. 13Present address: Korean National Institute of Health, Ongso, Korea. e-mail: Y.-K.P (pailyk@yonsei.ac.kr) or W.S.H. (w.hancock@neu.edu)

ABSTRACT: We report progress assembling the parts list for chromosome 17 and illustrate the various processes that we have developed to integrate available data from diverse genomic and proteomic knowledge bases. As primary resources, we have used GPMDB, neXtProt, PeptideAtlas, Human Protein Atlas (HPA), and GeneCards. All sites share the common resource of Ensembl for the genome modeling information. We have defined the chromosome 17 parts list with the following information: 1169 protein-coding genes, the numbers of proteins confidently identified by various experimental approaches as documented in GPMDB, neXtProt, PeptideAtlas, and HPA, examples of typical data sets obtained by RNASeq and proteomic studies of epithelial derived tumor cell lines (disease proteome) and a normal proteome (peripheral mononuclear cells), reported evidence of post-translational modifications, and examples of alternative splice variants (ASVs). We have constructed a list of the 59 "missing" proteins as well as 201 proteins that have inconclusive mass spectrometric (MS) identifications. In this report we have defined a process to establish a baseline for the incorporation of new evidence on protein identification and characterization as well as related information from transcriptome analyses. This initial list of "missing" proteins that will guide the selection of appropriate samples for discovery studies as well as antibody reagents. Also we have illustrated the significant diversity of protein...
INTRODUCTION

A new scientific initiative, the Chromosome-Centric Human Proteome Project (C-HPP) of the Human Proteome Organization, has a 10 year goal of characterizing the “parts list” of the entire human proteome encoded by the approximately 20300 human protein-coding genes.1,2 We believe that integration of proteomics data into a genomic framework will promote a better understanding of the relationship of the transcriptome to the proteome and facilitate international collaborations with different national teams volunteering for an individual chromosome. In this manner, a group of primarily US-based scientists have decided to study chromosome 17 and to characterize the full set of proteins coded by this chromosome as well as identify the major variants. The reason for selection of this chromosome was based on the presence of the driver oncogene, ERBB2 as well as the close association of a significant number of genes present on chromosome 17 with cancer. In addition, our team has developed a close association with the Australian and New Zealand scientists who are studying chromosome 7 which contains the oncogene EGFR, which together with ERBB2 forms a heterodimer complex which results in receptor kinase activation and oncogenic signaling. We will, therefore, report in this publication on the current status of the proteogenomic parts list of chromosome 17 and discuss future steps in our part of the C-HPP initiative.3

The DNA sequence of chromosome 17 was most recently defined in 2006 and chromosome 17 contains 78 839 971 bases or 2.8% of the euchromatic genome. In RNA-sequencing studies it was noted that there is an average of 5 distinct transcripts per gene locus and approximately 75% with at least two transcripts, as well as 274 pseudogenes.3 Chromosome 17 was also gene locus and approximately 75% with at least two transcripts, or 2.8% of the euchromatic genome. In RNA-sequencing studies in various diseases.4

RESULTS AND DISCUSSION

1. Background to Chromosome 17 and its Unique Properties, Especially Related to Cancer Biology

Chromosome 17 has a strong association with cancers and is extensively rearranged in at least 30% of breast cancer tumors. Whereas the short arm undergoes frequent losses, the long arm has complex combinations of overlapping gains and losses.5 Studies of the transcriptome map revealed regions (17p11, p13, q11, q12, q21, q23 and q25) with higher expression levels in specific chromosomal regions in 10 tumor tissue types.5 An increase in gene copy numbers for the region from 17q22 to 17q24 was observed in a large set of cancer cell lines and primary tumors by comparative genomic hybridization and cDNA arrays4 and was related to amplification of ERBB2 and collinear genes. In fact, chromosome 17 contains many regions with cancer-associated genes (see Table 1), including such prominent genes as TP53 (DNA damage response/usually called tumor suppressor), BRCA1 (breast cancer), NF1 (neurofibromatosis) and ERBB2 (breast cancer). To construct a more complete list of cancer associated genes we have integrated information from several Web sites (primarily Sanger, Waldman and GeneCards, see legend to Table 1) and have identified 44 such genes on chromosome 17. In this table we also explore the tendency of genes that are strongly associated with cancer to originate in transcriptionally active regions (high gene density) and to be clustered with other cancer related genes.4,8 In chromosome 17 these associations are indeed observed; only 8 of the 44 oncogenes occur in a region with a gene density of less than 30 genes per Mb and all oncogenes had >5 other cancer associated genes in proximity. In addition, regions identified with high expression in tumor studies6 contained significant numbers of cancer genes listed in this Table: p11 (2 including FLCN), p13 (9, TP53), q11 (NF1), q12 (4, ERBB2), q21 (13, BRCA1), q23 (4, DDX5) and q25 (9, GRB2).

2. List of Protein Coding Genes as Baseline for C-HPP and Corresponding Transcriptomic and Proteomic Information

Our current knowledge of the proteogenome of chromosome 17 was aggregated by information stored in the following resources: neXtProt,7 Uniprot,9 Unipep,11,16 PeptideAtlas,11,16 Genecards,15 Oncomine13 and Human Protein Atlas (see this issue). The data sets of GPMDB (release 2012/07/01) and PeptideAtlas (release 2012/09) are based on an aggregation of curated protein identifications by mass spectrometry that have been deposited in the public domain (PRIDE, PeptideAtlas, Tranche) and then undergone standardized reanalysis of the mass spectra. Direct deposition at neXtProt (Release 2012/09/11) is based on the annotation resources of Swiss-Prot and Uniprot,10 including literature curation. Thus the different major databases represent alternative snapshots of the information flow into proteomics ranging from experimental data sets to reviewed publications to epitope-based antibodies and immunohistochemistry at HPA(release 2012/09/12). The designation of the status of protein-coding genes is based on the Ensembl genome browser (current version 68). For chromosome 17 there are 1169 such genes with the following information in Uniprot: 861, 269, and 40 with the designation of protein or transcript evidence or with uncertain status (green, yellow, and red, respectively in Figures 1 and 3). For the HPP, our baseline accepts only those
protein characterizations with the highest-grade identifications, not just those inferred from transcripts, and recognizes the propensity to false discovery and lack of confirmation of many reported findings. The number of such highest-grade identifications for this chromosome is as follows: 601 (51%, neXtProt gold), 824 (70%, GPMDB, green) and 745 (64%, PeptideAtlas, 1% FDR for proteins) of the number of protein coding genes, respectively. As examples of the state of knowledge at the level of proteomics, the numbers of protein identifications with lower status listed in GPMDB and PeptideAtlas are, respectively, medium (49 and 43), low probability (201 and 190) and “missing” or black as 95 in GPMDB. In Protein Atlas there are antibodies corresponding to 725 genes on chromosome 17 of which 503 are of high quality (HPA score medium or high) and in addition the number of available polyclonal and monoclonal antibodies listed in antibodypedia and Labome (see Figure 3 legend) are currently 900 and 416, respectively.

One goal of the C-HPP initiative is to promote the integrated analysis of multiple ‘omics platforms, starting with use of RNA-Seq studies to guide deeper proteomics. The HPP initiative has the overall goal of both defining the protein parts list and establishing the biology/disease context of such information; we wish to capture in this discussion the significant amount of ‘parts’ information that could be fed into biology with such an integrated approach. As a starting basis, with common ‘shotgun’ proteomics approaches one could expect only about 50 out of a total of 1169 predicted proteins to be identified on chromosome 17 in a proteomic study with 1000 identifications (4.5% of the total genome). The number of identifications of proteins coded by genes present in chromosome 17 was 140 in the disease study

Table 1. Cancer Gene List for Chromosome 17

<table>
<thead>
<tr>
<th>Sanger List</th>
<th>GeneCards</th>
<th>Waldman</th>
<th>CancerIndex</th>
<th>Adjacent Cancer Related Genes</th>
<th>Gene Density</th>
</tr>
</thead>
<tbody>
<tr>
<td>17p13.3</td>
<td>ABAR</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>YWHAe</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>CRK</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>HIC1</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td>17p13.2</td>
<td>USP6</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>ALOX12</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>1+</td>
</tr>
<tr>
<td></td>
<td>TP53*</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>PER1</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>RCVN</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>1+</td>
</tr>
<tr>
<td>17p11.2</td>
<td>MAP2K4</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>1+</td>
</tr>
<tr>
<td></td>
<td>FLCN</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>TRAF4</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>1+</td>
</tr>
<tr>
<td></td>
<td>NF1*</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>TAF15</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>LASP1</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>ERBB2*</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>3+</td>
</tr>
<tr>
<td>17q21.1</td>
<td>THRA</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>RARA*</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>TOP2A</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>CCR7</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>3+</td>
</tr>
<tr>
<td>17q21.31</td>
<td>BRCAl*</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>NBR1</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>ETV4</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>WNT3</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>PHB</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>NGFR</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>ABCC3</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>NME1</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>NME2</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td>17q22.1</td>
<td>CTC</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td>17q22.2</td>
<td>BRI1</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td>17q23.3</td>
<td>CD79B</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>2+</td>
</tr>
<tr>
<td></td>
<td>DDX5</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>2+</td>
</tr>
<tr>
<td>17q24.1</td>
<td>AKIN2</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>2+</td>
</tr>
<tr>
<td>17q24.2</td>
<td>PRKAR1A</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>×</td>
</tr>
<tr>
<td></td>
<td>SST2</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>1+</td>
</tr>
<tr>
<td></td>
<td>GRB2</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>1+</td>
</tr>
<tr>
<td></td>
<td>ITGB4</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>1+</td>
</tr>
<tr>
<td>17q25.2</td>
<td>SEPT6</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>2+</td>
</tr>
<tr>
<td>17q25.3</td>
<td>TK1</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>BRC5</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>TIMP2</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>LGALS3BP</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>CANT1</td>
<td>×</td>
<td>√</td>
<td>×</td>
<td>3+</td>
</tr>
<tr>
<td></td>
<td>MAFG</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>1+</td>
</tr>
</tbody>
</table>

“Web sites that list oncogene information are: http://www.sanger.ac.uk/genetics/CGP/Census/, http://www.genecards.org/, http://waldman.ucsf.edu/GENES/completechroms.html, and http://www.cancerindex.org/geneweb/genes_d.htm. Information derived from GeneCards. As a way to assess degree of cancer associations we have calculated a score as follows: oncogene designation +5 points, direct literature association with cancer +3 points, present in cancer data sets +1 point. The score scale was set as follows: ≥25, 3+; scores 10–25, 2+; scores <10: 1+. The gene density was derived from GeneCards. The solid bars represent the regions with a gene density above the average on chromosome 17, that is 30.3. Recognized as driver oncogene.”
(two cell lines, triplicates, LTQ-FTMS) and 237 in normal sample (peripheral mononuclear cells)\textsuperscript{17} a deep study with 164 serial analyses, orbitrap vs 601 and 730 transcripts (disease and normal respectively). In proteomic studies the number of observed proteins will be related to the approach used (sample fractionation steps, type of mass spectrometer) as well as the number of replicates and size of the sample set but in both cases the depth of the transcriptomic finding was much deeper than the proteomics and underlines the importance of promoting the adoption of improved analytical procedures in proteomics.\textsuperscript{18} The disease implications of this comparison will be the subject of a separate manuscript in which we discuss the selection of appropriate control or normal samples for a disease biomarker study.

To further illustrate the data sets generated by a combined proteomic/transcriptomics study and compare with information listed in Uniprot and GPMDB, Figure 1 shows the region around the important and well-studied oncogene ERBB2. Again the RNASeq data sets are more extensive (11 and 12 transcripts for the 20 genes flanking ERBB2, respectively) vs proteomic coverage (5 and 6 respectively). Uniprot has annotated 5 genes in this region only at the level of transcript (STAC2, NEUROD2, PGAP3, ZPB2, and LRR3C) and one of these was not observed at the transcript level in our studies (LRR3C). In the aggregated data of GPM one protein from the gene STAC2 was observed with medium level probability (yellow) and 4 with low probability (red) while one had no evidence (LRR3C).

3. The Location of Protein Families on Chromosome 17

As is shown in Figure 2 there are several regions on the chromosome with clusters of genes in protein families. Genes in human families can exhibit close clustering on a single chromosome and presumably arose by tandem gene duplication,\textsuperscript{19} such as the growth hormone family (5 genes) at \textsuperscript{17q23}, CD300 (7 genes, 6 clustered at \textsuperscript{17q25.1}) and Schlafen family of 5 growth regulatory genes at \textsuperscript{17q12}. Alternatively, gene families may be of a complex type and dispersed and consist of multiple gene clusters at different chromosomal locations, such as the olfactory receptors at \textsuperscript{17p13.2}, \textsuperscript{13.3} where 12 of a large family of 398 members are present. In this case the evolution process may have involved a mixture of gene and genome duplication events and the sequences may be divergent. Other examples of gene families substantially clustered on chromosome 17 include cytokines, chemokine ligands, keratins, keratin - associated proteins, homeobox and chromobox proteins. The number of protein coding genes in each band is listed in red in the figure and interestingly band \textsuperscript{17q21.2} contains 109 genes with 50\% comprised of either keratin or keratin-associated proteins (28 and 25, respectively). In the future we plan to explore the challenge of characterizing closely homologous proteins that may be very difficult to identify unambiguously by proteomics, given limited sequence coverage that is obtained in a proteomics experiment.

4. What are the “missing or black proteins”?\textsuperscript{20}

To guide the search for "rare" tissue and cell lines samples and experimental protocols we have developed a list of the 59 “missing” proteins from chromosome 17. The list is limited to only those proteins that do not have any proteomic identifications in neXtProt, PeptideAtlas or GPM. The last column in Figure 3 shows the availability of antibody evidence for protein expression from HPA and only 4 genes have good antibody evidence (MYH4, HOXB1, CD300C, CD300LD). There is, in addition, significant number of proteins with only preliminary evidence in the databases. For example, the number of low probability identifications listed in GPMDB and PeptideAtlas for chromosome 17 are 201 and 190 respectively. A further example of incomplete information is provided by the important region of 20 genes around ERBB2, where 1 gene product has not been identified and 4 only with low levels of evidence (Figure 1). We will concentrate on the “missing” proteins initially, but in the future we will perform targeted studies on suitable tissues that are rich sources for poorly characterized proteins and deposit additional data sets in the public databases to improve the confidence of these identifications. As shown in Figure 3 tissues for targeted analyses can be identified through transcript expression data. The development of a list of “missing” proteins

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Parts list for the genomic region around ERBB2 (10 protein coding genes on each side). The color coding used the following: green, protein level evidence in Uniprot, RNA-Seq RPKM ≥15, proteomic data for the two studies and in GPMDB log(e) < -10; yellow, transcript level evidence, RNA-Seq RPKM 3-15, proteomics, log(e) = -5-10; red, uncertain protein evidence, RNA-Seq RPKM 1-3, proteomics log(e) = -1-5; black, no information. The cancer data were from two breast cancer cell lines: SKBR3 and SUM190 and normal sample consisted of peripheral mononuclear cells collected over serial time points from an individual. The samples were trypsinized and analyzed by nanoLC–MS/MS using a FTMS (cancer) or orbitrap (normal) linear ion-trap mass spectrometer. Strand-specific RNA-Seq libraries were prepared and sequenced on a lane of the Illumina HiSeq 2000 instrument per sample to obtain transcript data.\textsuperscript{15} All RNA-Seq data are available at Short Read Archive (SRS366582, SRS366583, SRS366584, SRS366609, SRS366610, SRS366611). For proteomics data, see: (http://gpmdb.thepgm.org/data/keyw...)
\end{figure}
will also allow the identification of situations where the lack of identification is due to a technical issue (poor enzymatic cleavage steps or unsuitable protein/peptide physical properties) and also identify cases where the gene model itself is problematic or the protein nomenclature is confounded due to synonyms.

The process we used to refine the list of “missing” proteins was as follows: (1) Eliminate any faulty gene annotation of non-protein coding genes (2 were eliminated from the chromosome 17 list and both were also identified as red in the Uniprot evidence list). (2) Identify “missing” proteins that had abundant RNASeq evidence (and in some cases proteomic data) from our experimental data sets (epithelial cancer cell lines, peripheral mononuclear cells). In some cases we identified nomenclature problems where a “missing” proteins was in fact reported in the proteomic databases under alternate gene symbols with good quality identifications. While our cell line data shown in Figure 3 did not show any highly credible protein IDs some did show significant levels of transcript (see Figure 3 yellow or green transcript levels for genes RNF43, STRADA, ARL16). Also one can use gene and transcript data from other sources (obtained from GeneCards in this figure legend) to identify target tissues for follow up RNASeq and proteomic studies. For example, we have initiated a study on nasal epithelial cells to identify missing olfactory receptors. In addition, brain and hair cortex look to be promising samples for additional studies. Brain has long been known to express a much higher proportion of single-copy DNA expressed in mRNA than in liver, kidney, and spleen. (3) Obtain suitable antibodies specific for the “missing” proteins: in the case of chromosome 17 there are 5 “missing” proteins with no antibody availability. Our next step is to prepare suitable monoclonal antibodies in collaboration with the antibody resources of the initiative. These antibodies will be used to confirm the proteomic identifications in Western blots as well as for affinity isolation steps to facilitate the identification of protein isoforms.

5. Splice Variant Analysis of Chromosome 17 Genes based on RNA-Seq Data from six ErbB2+ Cancer Cell Lines

Alternative splicing and post-translational modifications in higher eukaryotes increase the diversity of protein products derived from a single genetic locus and enable regulation of cellular and developmental processes. Across various cancers, examples of aberrant splicing events include alternative splice sites, alternative promoters, exon skipping, retained introns, and inclusion of presumed 5′ or 3′ untranslated regions (UTRs). The translated protein products of the alternatively spliced transcripts of a gene may play different roles in cancer mechanisms as there are various studies showing distinct opposite functions for the variants from a single gene.21,22

Activation of the ErbB2 receptor signaling pathways has been shown to increase cancer metastasis.23 We studied alternatively spliced transcripts (ASTs) expressed as distinct RNA-Seq reads from transcript-specific exons in the following six ErbB2+ cancer cell lines: two colorectal (LIM2405, LIM1899), two gastric (KATOIII, SNU16) and two breast (SUM149, SUM190). Across these 6 cancer cell lines, we identified 195 distinct ASTs from 144 genes; 46 of the 144 genes had more than one alternative transcript expressed (Table 2). We compared the differential expression of the transcripts in these cell lines based on the FPKM (fragments per kilobase of exon per million fragments mapped). Interesting observations include the following:

1. Seven distinct ASTs of septin 9 (SEPT9) were identified, with strikingly different expression levels of these ASTs across
the six cell lines (Figure 4a); sept9 epsilon is the isoform most expressed, by far. Altered expression of sept9 is observed in several carcinomas.24 (2) We identified reads from exons specific for the variant of ERBB2 which translated to the shorter protein variant (ENSP00000385185) (Figure 4b). (3) The longest AST variant of each of three genes (CDK12, FBXL20, GRB7) that are located quite near the ERBB2 gene on chromosome 17 was identified (Figure 4c). (4) Overexpression of the shorter variant

<table>
<thead>
<tr>
<th>Protein</th>
<th>Evidence</th>
<th>RPKM</th>
<th>Tissue</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOC2B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BHLM09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR1D5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR1G1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR1A2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR1A1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR2A2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C17orf107</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INCA1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C17orf100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SLC36A11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPDY64</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RCVN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MYH4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDR15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMEM99</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRTAP1-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRTAP1-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRTAP4-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KRTAP17-1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAMP2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RND2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NBR2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PYY</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HGD1B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STH</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPIML</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOX1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MPRC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAC4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TMEM92</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORK1D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ORK2D</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RNF43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>YPEL2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C17orf82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>STRADA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KCN16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD300C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CD300LD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OTOP3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AANAT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRCO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARL16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NP8</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. Uncharacterized gene products on chromosome 17. The gene names used are those listed in Ensembl. The following RNaseq data is illustrated (RPKM: reads per kilobase per million mapped reads): SKBR3, A431, SNU16, KATOIII, H9, LIM1899, LIM1215, LIM2405. The proteomics data are listed: SUM149, SUM190, SKBR3, Stem cell, ERBB2+ GI control, ERBB2+ GI tumor, ERBB2- GI control, ERBB2- GI tumor. The color coding of HPA evidence are used as following rules: green (high), yellow (medium), low (red) and black (very low or NULL). The color coding of the rest of the table are the same as Figure 1 (see Figure 1 legend). Antibody information is shown as: NA, not available; A, AntibodyPedia, polyclonal; B, AntibodyPedia, monoclonal; C, Labome (http://www.antibodypedia.com, http://www.labome.com). The potential tissue sources information was obtained from GeneCards15.
of PPP1R1b was observed in the KATOIII colorectal cancer cell line (Figure 4d).

6. Additional Protein variants resulting from Alternative Splice and Single Nucleotide Variants and Post-translational Modifications (PTMs)

We illustrate the diversity of protein isoforms for three regions on chromosome 17 that are adjacent to three important oncogenes, ERBB2, TP53 and BRCA1 (see Figure 5 and in Supporting Information Figures 1 and 2, 10 genes on either side). In this figure we list the number of alternative splice variants (ASVs, red circle) and single nucleotide variants (SNVs, purple circle) resulting from the transcription of missense SNPs. Examples of genes with a significant number of variants in the region around the four important oncogenes include ERBB2 with 4 and 88, IKZF3 (15, 7), TP53 (9, 1394), and BRCA1 (6, 320) ASVs and SNVs, respectively. In this figure we also list the four major PTMs (nexProt data) in boxes with separate listing for phosphorylated serine and threonine vs tyrosine and for N- vs O-glycosylation (see figure legend). At the level of the entire chromosome 17 with 1169 protein coding genes, the number of proteins with at least one known post-translational modification is 526 (approximately 45%), with the following major categories (426 phosphorylations; 185 acetylations; 51 glycosylations; 10 methylations; 8 palmitoylations; 2 myristoylations; many proteins with several identifications). In the cases

Table 2. Forty-six Genes Identified with More than One Transcript from the Six ErbB2+ Cancer Cell Lines: Two Colorectal (LIM2405, LIM1899), Two Gastric (KATOIII, SNU16) and Two Breast (SUM149, SUM190)

<table>
<thead>
<tr>
<th>Gene</th>
<th>Transcript 1</th>
<th>Transcript 2</th>
<th>Transcript 3</th>
<th>Transcript 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AANAT</td>
<td>CARD14</td>
<td>GPS1</td>
<td>RENCL5</td>
<td></td>
</tr>
<tr>
<td>ABR</td>
<td>CASKIN2</td>
<td>GRB2</td>
<td>RNP213</td>
<td></td>
</tr>
<tr>
<td>ACBD4</td>
<td>CEB1X</td>
<td>HIC1</td>
<td>Septin9</td>
<td></td>
</tr>
<tr>
<td>AFM1D</td>
<td>CDK12</td>
<td>LLGL2</td>
<td>SKA2</td>
<td></td>
</tr>
<tr>
<td>ALOXN3</td>
<td>CEP112</td>
<td>MAPK7</td>
<td>SMG6</td>
<td></td>
</tr>
<tr>
<td>ANAPC11</td>
<td>CDP</td>
<td>MS12</td>
<td>SPHK1</td>
<td></td>
</tr>
<tr>
<td>B4GALNT2</td>
<td>CTNS</td>
<td>MXRA7</td>
<td>TADA2A</td>
<td></td>
</tr>
<tr>
<td>BIAP2</td>
<td>DLX4</td>
<td>NME2</td>
<td>THRA</td>
<td></td>
</tr>
<tr>
<td>C17orf57</td>
<td>EFCA3</td>
<td>NXN</td>
<td>TTYH2</td>
<td></td>
</tr>
<tr>
<td>C17orf81</td>
<td>EIF5A</td>
<td>PIDK2</td>
<td>UBTF</td>
<td></td>
</tr>
<tr>
<td>CACNB1</td>
<td>GAS7</td>
<td>PEMT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAMKK1</td>
<td>GO5R2</td>
<td>PHF12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. (a) Septin 9 (SEPT9) transcript expression in the six ERBB2+ cell lines. (b) Relative abundance of the shorter variant of ERBB2 (ENSP00000385185) across the six cancer cell lines. (c) Relative expression levels of ASTs of three genes around ERBB2 on chromosome 17 in six ERBB2+ cell lines. Only the longest AST of each of these 3 genes was expressed. (d) Relative expression levels of the shorter transcript variant of Ppp1r1b in the six ErbB2 + cancer cell lines.

Figure 5. PTM information for genes around ERBB2. 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/nexProt; Purple circles, number of variants in nexProt; Hexagon shape in yellow, phosphorylation at Ser/Thr; Hexagon shape in orange, phosphorylation at Tyr; Rectangular shape acetylation; Triangle, N-glycosylation; 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.
Figure 6. Set of protein coding genes comprising the ERBB2 amplicon which presents transcriptomic and proteomic information determined for ERBB2-expressing breast cancer cell lines (SKBR3 and SUM190). The color coding is listed in Figure 1 legend. ERBB2 is denoted with a star. The presence of a gene with cancer associations is denoted with a number and the following information, which was collected from GeneCards as well as cited references. (1) TIAF1, antiapoptotic factor, induced by TGFβ1, functionally interacts with p53 in regulating apoptosis;46 (2) TRAF4, commonly overexpressed in a wide range of tumors, adaptor protein and signal transducer, links members of the tumor necrosis factor receptor (TNFR) family to signaling pathways, regulation of apoptosis; (3) PCGF2, transcriptional and tumor suppressor, a diagnostic marker for poor prognosis in breast and prostate cancer patients; (4) PSMB3, proteasome subunit, beta type, 3, ubiquitin-proteasome system is an important regulator of cell growth and apoptosis;45 (5) LASP1, regulation of dynamic actin-based, cytoskeletal activities and zyxin localization in breast carcinomas, tumor growth and migration in cancer, enhances proliferation of ovarian and colorectal cancer; (6) MED1 or PPAR binding protein (PPARBP), nuclear coactivator, activates the transcription of vitamin D receptor, retinoic acid receptor and estrogen receptor, involved in cell growth, differentiation and amplified in a subset of breast tumors, regulates p53-dependent apoptosis; (7) CDK12, deletions within 17q12 region leading to CDK12-ERBB2 fusion protein, related to gastric cancer; (8) PPP1R1B, (protein phosphatase 1, regulatory inhibitor, subunit 1B) signaling member of wnt pathways that is frequently overexpressed in breast, prostate, colon, and stomach carcinomas; (9) STARD3, overexpression in cancer cells increases steroid hormone production, promoting growth of hormone-responsive tumors such as breast cancer; (10) PNMT, phenylethanolamine N-methyltransferase which converts noradrenaline to adrenaline, observed by increased gene copy number in breast cancer; (11) PGAP3, lipid remodeling steps of GPI-anchor maturation, key step in lipid raft assembly of ERBB2 heterodimers; (12) MIEN1, migration and invasion enhancer, overexpression in various breast and prostate cancer, enhances migration and invasion of tumor cells, regulation of apoptosis; (13) GRB7, growth factor receptor-bound protein 7 binds to tyrosine phosphorylated HER2, promotes activation of HRAS, associated with invasive breast, ovarian, gastric prostate and esophageal carcinomas; (14) IKZF3, IKAROS family zinc finger 3, transcription factor, major tumor suppressor involved in human B-cell acute lymphoblastic leukemia, interacts with HRAS; (15) Gasdermin-like (GSDML), linked to cancer development and progression, involved in secretory pathways; (16) ORMDL3, negative regulator of sphingolipid synthesis, may indirectly regulate endoplasmic reticulum-mediated Ca^(+2) signaling; (17) PSMD3, proteasome 26S subunit, non-ATPase, 3 (P58), ubiquitin-proteasome system is an important regulator of cell growth and apoptosis;45 (18) MED24, component of the Mediator complex, a coactivator of RNA polymerase II-dependent genes, mediates growth of breast carcinoma cells;46 (19) NR1D1, nuclear receptor subfamily 1, group D, required for energy production in ERBB2 expressing breast cancer cells;47 (20) CASC3, cancer susceptibility candidate 3, component of mRNA splicing-dependent multiprotein exon junction complex (EJC), overexpressed in breast and gastric cancer; (21) CDC6, cell division cycle 6 homologue, regulator at the early steps of DNA replication, transcription regulated by mitogenic signals, overexpressed or associated with prostate, squamous, cervical, lung and liver cancer; (22) RARA, retinoic acid receptor, alpha, mediates embryogenesis, differentiation and growth arrest, some ER-negative breast cancer cell lines (SKBR3) express high levels of RAR alpha protein and RARA has been observed as part of the ERBB2 amplicon; (23) TOP2A, topoisomerase IIa, markedly upregulated in breast, prostate, gastric, ovarian and lung cancer, catalyzing the ATP dependent breakage and rejoining of double strand of DNA, shown to be amplified in a subset of breast tumors with ERBB2 amplification.48
residues and the heterogeneity of N-glycosylated structures; this summary does not attempt to capture this level of data complexity.

An important future goal of proteomics, as well as functional genomics, is to identify which of the potential protein variants that are identified at the genome or transcriptome level are observed at the level of the proteome. An example of this process is the characterization of the numbers of ASVs of ERBB2 observed in the proteome that is listed as 6 in Ensembl (Release 2012/07) and 4 in neXtProt. There is more complexity to be revealed, however: the total number of ASTs listed in Ensembl is 14 with the following amino acid lengths: 1255, 1240, 1225 (3 forms), 1055, 979, 603 (2 forms), and <252 (5 forms). The shorter variants have not been observed at the protein level and only 4 ASVs are listed in the consensus CDS protein set (an NCBI collaborative effort to identify the well annotated core set of human protein coding regions); these variants are listed with proteomics information in neXtProt (amino acid lengths of 1255, 1225 (3 forms)). A larger set of 6 protein forms is listed in GeneCards (1255, 1240, 1225 (2 forms) and 979) and same set of 6 is listed in GPM with MS identifications.

One could also expect that the site and nature of PTMs could be altered in a protein variant. Since there is little information on the MS characterization of SNVs in the human proteome, we have used data from the NCBI dbSNP database to capture potential polymorphism sites through EnSEMBL BioMart. Again using ERBB2 as an example, we examined the PTM information for 5 ASVs listed for ERBB2 in GMDB. For ENSP000000269571, phosphorylation sites are T701, Y735, T862, Y877, S998, I1023, S1051 (S/Y for SNV), S1054, S1073, S1083, S1107, S1115, T1166, S1174, S1214, T1240, Y1248; for ENSP000000385185, phosphorylation sites are T671, Y705, T832, Y847, S996, Y993, S1021 (S/Y for SNV), S1024, S1043, S1053, S1077, Y1109, T1136, S1144, T1210, Y1218; for ENSP000000443562, sites are T671, Y705, S968, Y993, S1024, S1048, S1121, S1144, T1210; for ENSP000000446466, sites are T686, Y720, S983, Y1008, S1039, S1063, S1136, S1159, T1225; for ENSP000000404047, sites are T425, Y459, T586, Y601, S722, Y747, S775 (S/Y for SNV), S778, S797, S807, S831, Y863, T890, S898, T964, Y972. NeXtProt lists 4 phosphorysine residues in 3 of the 4 ASVs while there are 7 N-linked glycan sites reported for the major ASV but only 1 structure for other ASVs. To denote an amino acid change in an ASV relative to ENSP0000044047, we underlined the residue. In the cases where there are lesser numbers of PTMs listed for lower abundance ASVs, this observation may be due to lack of experimental evidence rather than absence of PTMs. At this stage of characterization of the proteogenome there is little experimental evidence for PTMs in protein variants produced by missense SNPs but the polymorphism of serine to tyrosine (residue 1051 in ENSP269571) would be expected to affect that site of modification.

The C-HPP initiative believes that the management of such a complex data set as illustrated here (and with regular updates) is best served by an informatics system and associated interfaces that can integrate such information from a diversity of information sources. This new type of informatics system will be the subject of a separate report in this issue.  

7. Cancer-associated Studies
We have chosen as an example the important oncogene, ERBB2 (HER2), which resides on chromosome 17 and illustrates the value of considering proteomic observations in the context of the environment of the chromosome region in which the corresponding gene resides. ERBB2 encodes for a 185 kDa transmembrane glycoprotein that belongs to the family of epidermal growth factor receptors (EGFRs). Other members of this family include EGFR (ERBB1/HER1), ERBB3 (HER3), and ERBB4 (HER4). Ligand binding to the extracellular domain of these receptors induces homodimer (e.g., EGFR–EGFR or heterodimer (e.g., EGFR–ERBB2) formation leading to the activation of the intracellular tyrosine kinase domain and subsequent signaling cascade. Several ligands capable of activating the ERBB receptors have been identified, EGF-like ligands binding
to EGFR and neuregulins (NRG1 (ERBB2), NRG2 (ERBB3,4), NRG3 and 4 (ERBB4)). ERBB2 has been shown to be a preferable interaction partner for all other ERBB receptors and such heterodimers are long-lived and have a particularly high signaling potency.

The term amplicon is used to define gene amplification or the selective increase in the copy number of an oncogene and adjacent genes that can occur in development of solid tumors and should not be confused with elevated gene expression. ERBB2 is amplifiable in many tumor types, typically over a 1.5 Mb region and covers multiple genes in the region 17q12-q21 of chromosome 17. The amplicon may have an impact on the phenotype and clinical characteristics of ERBB2-amplified tumors as amplified genes may contribute to disease progression. Genes in close proximity to ERBB2 and observed in tumor studies include the following: TIAF1, TRAF4, PSMB3, LASP1, MED1, CDK12, PPP1R1B, STAR3D, PNMT, PGAP3, ERBB2, MIEN1, GRB7, IKZF3, ORMDL3, GSDMB, PSMD3, MED24, NR1D1, CASC3, CDC6, RARA, TOP2A, listed in order from the centromeric and telomeric ends of the amplicon (17q11.2 to q21.2). The amplicon can range from the minimal set of ERBB2 - GRB7 to the genes listed above. Moreover, increased expression of this set of genes was directly linked to gene amplification via copy number analysis of ERBB2 and adjacent genes. However, not all genes in this region are amplified in tumor samples and thus not included in the above list, for example the expression levels of PNMT were generally rather low, while expression of TCAP and NEUROD2 was either very weak or absent in these tumor samples.

The existence of the ERBB2 amplicon is a strong example of the value of the integration of proteomics with transcriptomic data with the potential for discovery of additional protein features of disease interest in a given data set. In Figure 6, we plot the data for two breast cancer cell lines that express high levels of ERBB2, namely SKBR3 and SUM190 (breast cancer). The figure shows the value of integrating RNASeq measurement with the corresponding proteomic data, for example, PGAP3, MED1, CDK12, MED24 and CDC6 are not observed by proteomics but are part of the ERBB2 amplicon. Conversely, the members of the amplicon were observed by proteomics as well as RNASeq measurement: LASP1, ERBB2, MIEN1, GRB7, GSDMB, ORMDL3, PSMD3 and TOP2A which can give insights into the phenotypically important parts of the amplicon, namely actin cytoskeletal reorganization, regulation of apoptosis, stabilization of the phosphorylated form of ERBB2, endoplasmic reticulum-mediated Ca(2+) signaling and DNA synthesis.

In a follow up study we explored the ERBB2 amplicon in Oncomine, which contains the largest collection of curated cancer microarray data. Its integrated data-mining platform facilitates extensive meta-analyses across large numbers of mRNA data sets. We performed differential meta-analyses on human breast cancer microarray data sets and studied the commonly overexpressed chromosome 17 genes from ERBB2 positive, Estrogen receptor positive and Triple negative (ERBB2/ER/PR-negative) breast cancer subtypes. The top 20 overexpressed genes in each of the 3 breast cancer subtypes are given in Table 3. Remarkably, we found no overlap between the top 20 genes expressed in these 3 different major breast cancer types.

In a separate analysis, we identified the 20 most overexpressed genes from all chromosomes across the ten ERBB2+ breast cancer microarray data sets in Oncomine. Remarkably, 13 of the top 20 genes were from chromosome 17 (Figure 7), with several clustered very near ERBB2. The genes (transcripts) are ranked by their association with ERBB2+ breast cancers; to compare these data with the consensus list of genes for the ERBB2 amplicon, we have repeated the amplicon set identified in previous studies and underlined the genes overexpressed in Oncomine and from our cancer cell line studies (underline and bold denote presence in both studies): TRAF4, TIAF1, PCGF2, PSMB3, LASP1, MED1, CDK12, PPP1R1B, STAR3D, PNMT, PGAP3, ERBB2, MIEN1, GRB7, IKZF3, ORMDL3, GSDMB, PSMD3, MED24, NR1D1, CASC3, CDC6, TOP2A. While...
some genes in Oncome are expressed at low levels in our study (e.g., TCAP, FBXL20) and are not included in this comparison, others are present on chromosome 17 but outside the amplicon e.g. CYB561 (see Table 3). It is noteworthy that most of the genes in close proximity to ERBB2 are highly ranked in Oncome and observed in our studies. Thus the combined view of experimental data generated in our laboratories together with curated literature information strengthen the case for the integration of transcripomic and proteomic data in the study of this amplicon and will be explored further with studies of ERBB2+ breast, gastric and colon cancer patient samples. Such a comparison is particularly germane since there is clinical evidence that Herceptin does not seem to yield benefit in ERBB2+ colorectal cancer patients and may have mixed results in such gastric cancer patients.36

■ CONCLUDING REMARKS

We have demonstrated an organized approach to defining the baseline of what is currently known about protein products of the protein-coding genes on Chromosome 17, utilizing and comparing multiple valuable data resources. Post-translational modifications, sequence polymorphisms, and splice variants have been documented, features which must be studied at the protein level and to this end we have integrated of transcriptomics, mass spectrometry and antibody protein capture approaches, which is a model for integrated analyses of additional platforms like epigenomic and metabolomics data. We have highlighted the cancer-associated genes on this chromosome and especially the epigenomic and metabolomics data. We have demonstrated the value of a chromosome-centric approach to even such a complex amplicon from 17q12 to 17q23 demonstrates the value of a chromosome-centric approach to even such a complex phenotype as a particular subtype of human breast cancers.

■ ASSOCIATED CONTENT

4 Supporting Information

Supplemental figures. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: wi.hancock@neu.edu. Telephone: 617-373-4881. Fax: 617-373-8795.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the following for valuable discussions: Dan Rhodes, Compendia Bioscience, Inc, Ann Arbor, Michigan (Oncomine analysis) and Mark Fink for the insight that PGAP was part of the ERBB2 amplicon. This work was supported by the following research grants: (Korea) The World Class University program through the National Research Foundation of Korea (2012R1A1A3010623) and the Ministry of Education, Science and Technology (R31-2008-000-10086-0 (W.S.H and Y.K.P.)), National Project for the Personalized Genomic Medicine A111218-11-CP01 (to Y.K.P.) from the Korean Ministry of Health and Welfare: (USA) The National Institutes of Health grants, U01-CA128427 to W.S.H., U54DA021519, UL1 RR024986, RM-08-029, and U54ES017885 to G.S.O.; R41 GM103362 to D.F., NIH grant (M.P.S. and H.I.); Texas State Rider for the Morgan Welch inflammatory Breast Cancer Program and the G. Morris Dorrance Jr. Chair in Medical Oncology (M.C., Z.M.); NHGRI grants R01 GM087221, 2PS0 GM076547 (E.D.) (Europe) EU FP7 grant “ProteomeXchange” [grant number 260558], and from the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg (E.W.D.); SIB Swiss Institute of Bioinformatics; Genebio SA; the Swiss Confederation’s Commission for Technology and Innovation (C.T.I., grant 10214.1 PFLS-LS). Eurostars grant 6715 (A.B., P.G.); (Sweden) the Knut and Alice Wallenberg foundation for funding the HPA project (E.L., M.U.). Contribution number 1035 from the Barnett Institute.

■ REFERENCES


Genome Wide Proteomics of ERBB2 and EGFR and Other Oncogenic Pathways in Inflammatory Breast Cancer

Emma Yue Zhang,† Massimo Cristofanilli,‡§ Fredika Robertson,‖ James M. Reuben,§‖ Zhaomei Mu,‡ Ronald C. Beavis,†‡ Hogune Im,‡ Michael Snyder,‡† Matan Hofree,‡ Trey Ideker,‡ Gilbert S. Omenn,‡ Susan Fanayan,‡ Seul-Ki Jeong,◆ Young-ki Paik,◆ Anna Fan Zhang,† Shiaw-Lin Wu,† and William S. Hancock*†

†Barnett Institute and Department of Chemistry and Chemical Biology, Northeastern University, 360 Huntington Avenue, Boston, Massachusetts 02115, United States
‡Kimmel Cancer Center, Thomas Jefferson University, Philadelphia, Pennsylvania 19107, United States
§The Morgan Welch Inflammatory Breast Cancer Research Program and Clinic, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, United States
‖Departments of Experimental Therapeutics and Hematopathology, The University of Texas, MD Anderson Cancer Center, Houston, Texas 77030, United States
⊥Department of Biochemistry and Medical Genetics, George and Faye Yee Centre for Healthcare Innovation, University of Manitoba, Winnipeg, MB R3T 2N2, Canada
∇Department of Genetics, Stanford University, Stanford, California 94305, United States
◆Department of Bioengineering, University of California, San Diego, La Jolla, California 92093, United States
¶National Center for Integrative Biomedical Informatics, University of Michigan, Ann Arbor, Michigan 48109, United States
○Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, New South Wales 2109, Australia
◆Yonsei Proteome Research Center, Yonsei University, Seoul, Korea

ABSTRACT: In this study we selected three breast cancer cell lines (SKBR3, SUM149 and SUM190) with different oncogene expression levels involved in ERBB2 and EGFR signaling pathways as a model system for the evaluation of selective integration of subsets of transcriptomic and proteomic data. We assessed the oncogene status with reads per kilobase per million mapped reads (RPKM) values for ERBB2 (14.4, 400, and 300 for SUM149, SUM190, and SKBR3, respectively) and for EGFR (60.1, not detected, and 1.4 for the same 3 cell lines). We then used RNA-Seq data to identify those oncogenes with significant transcript levels in these cell lines (total 31) and interrogated the corresponding proteomics data sets for proteins with significant interaction values with these oncogenes. The number of observed interactors for each oncogene showed a significant range, e.g., 4.2% (JAK1) to 27.3% (MYC). The percentage is measured as a fraction of the total protein interactions in a given data set vs total interactors for that oncogene in STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, version 9.0) and I2D (Interologous Interaction Database, version 1.95). This approach allowed us to focus on 4 main oncogenes, ERBB2, EGFR, MYC, and GRB2, for pathway analysis. We used bioinformatics sites GeneGo, PathwayCommons and NCI receptor signaling networks to identify pathways that contained the four main oncogenes and had good coverage in the transcriptomic and proteomic data sets as well as a significant number of oncogene interactors. The four pathways identified were ERBB signaling, EGFR1 signaling, integrin outside-in signaling, and validated targets of C-MYC transcriptional activation. The greater dynamic range of the RNA-Seq values allowed the use of transcript ratios to correlate observed protein values with the relative levels of the ERBB2 and EGFR transcripts in each of the four pathways. This provided us with potential proteomic signatures for the SUM149 and 190 cell lines, growth factor receptor-bound protein 7 (GRB7), Crk-like protein (CRKL) and Catenin delta-1 (CTNND1) for ERBB signaling; caveolin 1 (CAV1), plectin (PLEC) for EGFR signaling; filamin A (FLNA) and...
actinin alpha1 (ACTN1) (associated with high levels of EGFR transcript) for integrin signalings; branched chain amino-acid transaminase 1 (BCAT1), carbamoyl-phosphate synthetase (CAD), nucleolin (NCL) (high levels of EGFR transcript); transferrin receptor (TFRC), metadherin (MTDH) (high levels of ERBB2 transcript) for MYC signaling; S100-A2 protein (S100A2), caveolin 1 (CAV1), Serpin B5 (SERPINB5), straffin (SEF), PYD and CARD domain containing (PYCARD), and EPH receptor A2 (EPHA2) for PI3K signaling, p53 subpathway. Future studies of inflammatory breast cancer (IBC), from which the cell lines were derived, will be used to explore the significance of these observations.

KEYWORDS: ERBB2, EGFR, Inflammatory breast cancer, Chromosome-centric Human Proteome Project, Inflammary breast cancer, Chromosome-centric Human Proteome Project

INTRODUCTION

Breast cancer is a major health problem with over 40 000 deaths each year in the United States. We have previously studied proteomics and glycoproteomics in samples collected from breast cancer patients2−4 as potential markers for the early detection of breast cancer. As an extension of these studies, we report in this manuscript on a study of protein expression as measured by both RNA-Seq and proteomics of two cell lines established from primary inflammatory breast cancer (IBC) tumors,5 namely, SUM149 and SUM190, which are ER (−) and PR (−), as well as the well-studied cell line SKBR3, which is known to express high levels of ERBB2 and is ER (−) and PR (−).

EGFR and ERBB2 are members of the epidermal growth factor receptor (EGFR) family, one of 20 subfamilies of human receptor tyrosine kinases (RTK).6 The EGFR family is one of the best studied growth factor receptor systems, often overexpressed in human tumors.7−9 Several small molecule inhibitors and protein drugs have been developed to modulate disorders in the EGFR family.10,11 Moreover, determination of ERBB2 status by immunohistochemistry (IHC) or immunofluorescence in situ hybridization (ISH) has been recommended by the American Society of Clinical Oncology (ASCO) as a marker for diagnosis and evaluation in primary invasive breast cancer.12 Initially we will describe the analysis of the RNA-Seq data to determine the presence or absence of oncogenes typically associated with breast cancer as well as the levels of the target oncogenes ERBB2 and EGFR. These studies demonstrated the importance of EGFR and ERBB family members in the cell lines, as well as other oncogenes such as TP53, CRKL, EZR and MYC. We then explored different approaches to integrate the proteomic information with the transcriptome data and compared the proteomic levels as measured by spectral count with the transcript level as well as interaction values of the observed proteins with the panel of oncogenes. These comparisons highlighted the 4 oncogenes, namely, EGFR, ERBB2, MYC and GRB2, and allowed the identification of protein-based subpathways of interest for the different cell lines.

MATERIALS AND METHODS

Cell lines, Cell Lysis, and In-Gel Digestion

Cell Lines SKBR3, SUM149 and SUM190. The human breast cancer cell lines SKBR3 (ER/PR−, HER2+), metastatic pleural effusion), was obtained from the American Type Culture Collection (Manassas, VA) and maintained in culture with DMEM/F-12 medium supplemented with 10% FBS (Tissue Culture Biologicals, Seal Beach, CA) and 1% of Antibiotic-Antimycotic 100X (Gibco, Carlsbad, CA).

SUM149 and SUM190 cells were obtained from Dr. Stephen Ethier (Kramanos Institute, MI, USA) and are commercially available (Asterand, Detroit, MI). SUM149 cells are ER/PR−, HER2− (triple receptors negative), and the SUM190 cells are ER/PR−, HER2+. Both human IBC cell lines were maintained in culture with Ham’s/F-12 medium supplemented with 10% FBS (Tissue Culture Biologicals, Seal Beach, CA), 5 μg/mL of insulin, 1 μg/mL of hydrocortisone and 1% of Antibiotic-Antimycotic 100X (Gibco, Carlsbad, CA).

Twenty microliters of lysis buffer (2% SDS in 50 mM NH4CO3) was added to 10 μL of cell lysate. Cells were solubilized by sonication using 20 s bursts, followed by cooling on ice for 20 s, in a process that was repeated for 10 times. The entire extract was concentrated down to 15 μL in a speed vacuum and loaded onto a gel (SDS-PAGE, 4−12% gradient) to separate proteins by molecular weight. After staining with Coomassie blue, each gel lane was cut into five individual slices as shown in Figure S1 (Supporting Information).

Each slice was further minced into smaller pieces (approximately 0.5 mm3). The gel slices were washed with 600 μL of water for 15 min and centrifuged, supernatant was removed, and 50% ACN was added (1 mL), followed by shaking until no visible Coomassie stain remained. Proteins were then reduced with dithiothreitol (DTT) by adding 250 μL of 10 mM DTT in 0.1 M NH4CO3, and incubated for 30 min at 56 °C. Samples were subsequently alkylated at room temperature and in the dark for 80 min with 250 μL of 55 mM iodoacetamide (IAA) in 0.1 M NH4CO3. Trypsin digestion reagent (200 μL; 10 ng/mL of trypsin in 50 mM NH4CO3, pH 8.0) was added, and samples were incubated for 30 min at 4 °C. The trypsin concentration was based upon an estimate of approximately 0.1−0.5 mg of protein per gel slice and adjusted as necessary. The solution was then replaced with 50 mM NH4CO3 to cover the gel pieces (50 μL) and incubated overnight at 37 °C to elute peptides from the gel. Following this step, supernatant was removed and stored. Gel pieces were further extracted with 5% formic acid (30 μL) and acetonitrile (ACN) (200 μL) at 37 °C for 10 min and then twice with 5% formic acid (30 μL) and ACN (200 μL). The formic acid solution containing tryptic peptides was combined with the previous supernatant and concentrated to 5−10 μL. The concentrated solution (trypsin-digested peptides) was subjected to LC−MS analysis.

LTQ-FT MS

The in-gel digested peptides were analyzed by online LC using a linear IT coupled to a Fourier transform mass spectrometer (LTQ-FT MS, Thermo Electron, San Jose, CA) with a Dionex nano-LC instrument (Ultimate 3000, Sunnyvale, CA) and a 75 mm i.d. × 15 cm C-18 capillary column packed with Magic C18 (3 mm, 200 Å pore size) (Michrom Bioresources, Auburn, CA). The LTQ-FT mass spectrometer was operated in the data-dependent mode to switch automatically between MS and MS/MS acquisition. Survey full-scan MS spectra with two microscans (m/z 400−2000) were acquired in the Fourier transform ion cyclotron resonance cell with a mass resolution of 100 000 at m/z 400 (after accumulation to a target value of 2 × 106 ions in the linear IT), followed by ten sequential LTQ-MS/MS scans throughout the 90 min separation. The analytical separation was carried out using a three-step linear gradient, starting from 2% B to 40% B in 40 min (A: water with 0.1% formic acid; B: ACN with 100X (Gibco, Carlsbad, CA).
0.1% formic acid), increased to 60% B in 10 min, and then to 80% B in 5 min. The column flow rate was maintained at 200 nL/min.

### Protein Identification

Peptide sequences were identified using Thermo Proteome Discoverer 1.3 from a human database SP.human.56.5 with full trypsin specificity and up to three internal missed cleavages. The tolerance was 50 ppm for precursor ions and 0.8 Da for product ions. Dynamic modifications were deamidation of asparagine, and static modification was carboxymethylation for cysteine. Peptides were identified with Xcorr scores above the following thresholds: ≥3.8 for 3+ and higher charge state ions, ≥2.2 for 2+ ions, and ≥1.9 for 1+ ions.

We used the spectral count approach to measure relative abundance of protein samples as reported by Choi et al.13 We have selected several housekeeping proteins, glyceraldehyde-3-phosphate dehydrogenase (GAPDH), b-actin (ACTB), b-tubulins (2A, 2B, 2C, 3 and 5),14 which are ubiquitously expressed in a wide range of tissues and cell types, as internal standards for relative quantification. In order to minimize the amount of samples loaded on the 1D SDS-PAGE gel. These proteins met the required criteria of high abundance and consistent ratios across the 3 cell lines, as measured by peptide counts and extracted ions in the same gel section between the different cell lines.

The protein list also was submitted to the Gene A La Cart (provided by www.genecards.com, uploaded to Gene A La Cart for analysis in August, 2011) to acquire data for bioinformatics analysis, including gene symbols and other genomic information.

### RNA-Seq Measurement

Strand-specific RNA-Seq libraries were prepared and sequenced on a lane of the Illumina HiSeq 2000 instrument per sample to obtain transcript data.16 All RNA-Seq data are available at Short Read Archive (SRS366582, SRS366583, SRS366584, SRS366609, SRS366610, SRS366611). From total RNA, strand-specific RNA-Seq libraries were prepared according to Illumina TruSeq standard procedures and sequenced at both ends (paired-end RNA-sequencing) on Illumina HiSeq 2000. Tophat embedded with Bowtie was used to align the sequence reads to human genome (hg19). Using Cufflinks, the alignments were assembled into gene transcripts (NCBI build 37.2), and their relative abundances (RPKM) were calculated.

## RESULTS AND DISCUSSION

We have previously studied on the role of two driver oncogenes, EGFR, ERBB2, in epithelial cancers5,16 and have investigated the changes in their glycosylation patterns.7–4 To further expand on our previous observations, we have performed a comparative study to explore the total lysate proteome of a well-established epithelial breast cancer cell line, SKBR3, which overexpresses ERBB2 and two primary cell lines (SUM149 and SUM190) isolated from patients with inflammatory breast cancer.7 We have employed a traditional proteomic analysis of the data and compared these results with an alternative format, namely genome-wide proteomics using the chromosome format (C-HPF260), which is being developed as part of the HUPO human proteome initiative. One benefit of such approach is the facile integration of proteomic and transcriptomics data as well as allowing for the identification of genomic regions in which a driver oncogene may affect gene transcription of adjacent genes.

### Analysis of Cell Lines SKBR3, SUM149, and SUM190

Each cell line was analyzed in triplicate, and relative quantitation was achieved with spectral counts using a correction factor based on housekeeping proteins. With the availability of a deep measurement of the transcriptome, by RNA-Seq (100 million reads), it is common to measure 10 000–11 000 transcripts in a cell line study. In contrast, a proteomic study comparable to what is reported here will sample only approximately 10% of the expressed the expressed set of proteins. While the transcriptome can enhance the proteomic measurement, the opposite is also true as a medium level protein study can be used to explore the major phenotypic patterns observed in a study of disease versus normal cell lines and patient tissue. In addition, there are examples of a protein being identified in the absence of a measurable transcript level.21

In the proteomic analysis we used a conservative protocol for identifying proteins in replicate analysis, which included high protein confidence and high peptide rank (Proteome Discoverer) and with a FDR of less than 1%. We identified a total of 1444, 1396, and 964 proteins (numbers of proteins with 2 or more peptides) in the SKBR3, SUM149 and 190 cell line samples, respectively (numbers of proteins with 2 or more peptides were 1071, 1134, and 686 for SKBR3, SUM149 and SUM190, respectively). In addition, selected proteins identified by one single peptide were further analyzed using additional criteria such as high mass accuracy, fragmentation spectra and observation of the corresponding transcript (see Table 1). In the cell line studies a comparison of the SKBR3 with SUM190, SKBR3 with SUM149, and SUM190 with SUM149 proteome contents identified 751, 934, and 695 common proteins, respectively.

### Characterization of EGFR and ERBB2

EGFR was identified in SUM149 and SKBR3 cell lysates, while ERBB2 was identified in SKBR3 and SUM190 cell lysate preparations, consistent with IHC results in previous studies.5,22 As shown in Table S1 (Supporting Information), EGFR and ERBB2 were identified with 11 and 13 peptides for cell lines SUM149 and SKBR3, respectively. This table employs data from GPMDB (Global Proteome Machine database)23 to assess the quality of peptides observed for the two proteins. The peptides observed in our study have been frequently reported in the literature, e.g., rank 1–5 and 1–4 for the most frequently observed, as well as other peptides for EGFR and ERBB2, respectively. The MS/MS data for a diagnostic peptide for EGFR and ERBB2 in shown in Figure S2 (Supporting Information). Both EGFR and ERBB2 were detected with good sequence coverage (15.5 and 15.8%), although peptides derived from the N-terminal domain of ERBB2 were not observed. The identification of ERBB2 was confirmed by immunoprecipitation with the monoclonal antibody trastuzumab (Herceptin) and subsequent analysis on 1D SDS-PAGE and detected at an approximate molecular weight (MW) of 110 000 (theoretical 138 kD, data not shown).

### Protein Observations with RNA-Seq Data and Expressed in a Genome Wide Format (Chromosomes)

Besides proteomic analysis, we have also discovered potential proteins of interest by comparing proteomics data with the corresponding transcriptomic data in a chromosome format (see Tables S3 and S4, Supporting Information, for the RNA-Seq results for SKBR3, SUM149 and SUM190). We collected the genomic information from the Gene A La Cart tool provided by www.genescards.org. In doing so, UniProt accession numbers for
their corresponding proteins were first extracted from the search result files in Proteome Discoverer, prior to submission to Gene a la Cart as identifiers to retrieve their genomic information, including gene symbols, genomic locations (chromosome number, base pair location of gene start and end, and gene size), and Ensemble cytobands. This would allow the protein list to be organized by their locations on different chromosomes. The resulting data sets for the three cell lines are shown in Table S3 (Supporting Information).

**Use of RNA-Seq Data to Explore ERBB2 Signaling Pathways**

As a first step we generated a list of 33 oncogenes associated with breast cancer from the Sanger, Genecards databases, and literatures, which had either measurable transcript level (RPKM >1) and in some cases proteomic data (see Table 1). The RNA-Seq values showed that the cell line SUM149 had a high level of transcript for EGFR (RPKM = 60) and a relatively low value for ERBB2 (RPKM =14). Conversely the cell line SUM190 had values of 400 and ND for ERBB2 and EGFR, respectively. The immortalized cell line SKBR3 expressed a high level of ERBB2 and a low level of EGFR (RPKM = 300 and 1.4, respectively). Other oncogenes with a high level of transcript (RPKM > 40) were TP53, MYC (SUM149); GRB7, CRKL (SUM190) and EZR, TOP2A (SKBR3). As described in a later section we also explored reported interactions between the group of 33 oncogenes and the proteins observed in the SUM149 and 190 proteomic results.

Figure 1A and B compares the ERBB2 signaling pathway in two IBC cell lines, SUM149 (high levels of EGFR transcript) and SUM190 (high levels of ERBB2 transcript) with the ERBB2 pathway derived from the KEGG database. SUM190 presents an interesting situation with high transcript levels of ERBB2 and ERBB3 (RPKM = 400 and 23, respectively) and a low level for ERBB4 (RPKM = 3), without detectable transcript levels of EGFR (ERBB1) and a low RNA-Seq value (4.91) for amphiregulin (AR in Figure 1), one of the EGFR ligands. ERBB2 is a special member in the ERBB family in that there has been no ligand discovered for ERBB2 and signaling largely depends on heterodimer formation with either EGFR, ERBB3 or other ERBB family members.

### Table 1. List of Oncogenes Associated with Breast Cancer with Associated Proteomic and Transcriptomic Data

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>SKBR3</th>
<th>SUM190</th>
<th>SUM149</th>
<th>SKBR3</th>
<th>SUM190</th>
<th>SUM149</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACOT8</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>34.4</td>
<td>2.2</td>
<td>9.4</td>
</tr>
<tr>
<td>APC</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>3.4</td>
<td>1.2</td>
</tr>
<tr>
<td>BRCA1</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>1.0</td>
<td>28.3</td>
<td>1.9</td>
</tr>
<tr>
<td>CDKN2A</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>6.4</td>
<td>3.3</td>
<td>ND</td>
</tr>
<tr>
<td>CEACAM6</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>16.2</td>
<td>ND</td>
</tr>
<tr>
<td>CRKL</td>
<td>ND</td>
<td>9</td>
<td>ND</td>
<td>3.3</td>
<td>56.3</td>
<td>12.3</td>
</tr>
<tr>
<td>DEK</td>
<td>5</td>
<td>ND</td>
<td>ND</td>
<td>22.4</td>
<td>12.5</td>
<td>26.0</td>
</tr>
<tr>
<td>EGFR</td>
<td>12</td>
<td>ND</td>
<td>22</td>
<td>1.4</td>
<td>ND</td>
<td>60.1</td>
</tr>
<tr>
<td>ERBB2</td>
<td>46</td>
<td>19</td>
<td>ND</td>
<td>300.1</td>
<td>399.7</td>
<td>14.4</td>
</tr>
<tr>
<td>ERBB2IP</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>1.1</td>
<td>15.2</td>
<td>8.3</td>
</tr>
<tr>
<td>ERBB3</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>6.6</td>
<td>23.1</td>
<td>6.6</td>
</tr>
<tr>
<td>ERBB4</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>3.0</td>
<td>ND</td>
</tr>
<tr>
<td>EZR</td>
<td>92</td>
<td>40</td>
<td>46</td>
<td>54.0</td>
<td>10.9</td>
<td>26.4</td>
</tr>
<tr>
<td>GRB2</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>9.0</td>
<td>7.7</td>
<td>10.0</td>
</tr>
<tr>
<td>GRB7</td>
<td>2</td>
<td>2</td>
<td>ND</td>
<td>30.3</td>
<td>57.5</td>
<td>5.8</td>
</tr>
<tr>
<td>JAK1</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>1.8</td>
<td>12.0</td>
<td>7.9</td>
</tr>
<tr>
<td>KRAS</td>
<td>ND</td>
<td>ND</td>
<td>2</td>
<td>1.5</td>
<td>8.5</td>
<td>6.9</td>
</tr>
<tr>
<td>MET</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>5.6</td>
<td>9.3</td>
</tr>
<tr>
<td>MLH1</td>
<td>2</td>
<td>ND</td>
<td>ND</td>
<td>3.5</td>
<td>3.6</td>
<td>6.3</td>
</tr>
<tr>
<td>MSH2</td>
<td>ND</td>
<td>ND</td>
<td>3</td>
<td>2.8</td>
<td>2.9</td>
<td>10.6</td>
</tr>
<tr>
<td>MTOR</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>1.7</td>
<td>7.2</td>
<td>16.2</td>
</tr>
<tr>
<td>MYC</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>19.0</td>
<td>9.6</td>
<td>79.1</td>
</tr>
<tr>
<td>PIK3R1</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>16.3</td>
<td>3.3</td>
</tr>
<tr>
<td>PIK3R2</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>7.4</td>
<td>1.7</td>
<td>16.0</td>
</tr>
<tr>
<td>PIK3R3</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>4.5</td>
<td>28.1</td>
<td>1.5</td>
</tr>
<tr>
<td>PPP1R1B</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>1.3</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>PTEN</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>0.8</td>
<td>6.1</td>
<td>0.8</td>
</tr>
<tr>
<td>RB1</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>1.2</td>
<td>2.0</td>
<td>1.7</td>
</tr>
<tr>
<td>RHOC</td>
<td>ND</td>
<td>ND</td>
<td>16</td>
<td>53.6</td>
<td>24.7</td>
<td>47.8</td>
</tr>
<tr>
<td>SRC</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>6.2</td>
<td>ND</td>
<td>7.5</td>
</tr>
<tr>
<td>STAT1</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>1.5</td>
<td>9.7</td>
<td>11.3</td>
</tr>
<tr>
<td>TOP2A</td>
<td>ND</td>
<td>58</td>
<td>ND</td>
<td>50.2</td>
<td>5.3</td>
<td>25.9</td>
</tr>
<tr>
<td>TP53</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
<td>14.8</td>
<td>1.3</td>
<td>40.6</td>
</tr>
</tbody>
</table>

— Gene symbols are from Genecards. Spectral counts. RPKM values. ND = not detected. Oncogenes used for pathway analysis are highlighted by box. Identifications of single peptide proteins are shown in Figure S3 (Supporting Information).
ERBB4. However, a high level of ERBB3 is found in the SUM190 transcript, and it has been reported that the ERBB2/ERBB3 heterodimer is active in cell proliferation in breast tumor cells (see highlighted blue lines in Figure 1A). Conversely, as is shown in Figure 1B (highlighted blue lines) with the observed transcript values in the SUM149 cell line for the EGFR family signaling pathway there are several possibilities for signaling with involvement of EGFR dimers, ERBB2 heterodimers with EGFR or ERBB3. Since ERBB4 is not detected at either the transcript or protein level, it is presumably not part of the signaling cascade. Thus RNA-Seq studies identified potential differences between the two cell lines and thus set the stage for a proteomic investigation. Another advantage of the RNA-Seq studies was the greater dynamic range than the proteomic measurement; one
important example was identification of high levels of the transcript for the MYC oncogene in SUM149 and 190 (19 and 10, respectively) in the absence of a proteomics signal. The importance of this oncogene is consistent with the importance of the MEK/ERK pathway in carcinogenesis (see arrow in Figure 1) and is supported by the large number of MYC interactors identified in the proteomics study (see Figure 2 and discussion later).

To further explore the difference between EGFR and ERBB2 signaling in SUM190 and 149 transcriptome, we used the ratio of the RPKM values to interrogate the NCI Erbb receptor signaling network and visualized the data by assigning different colors based on the ratio values. First, EGFR and ERBB2 are the most differentially expressed genes in this network. As can be seen in Table 2 increased levels of EGFR transcript are associated with increased levels of the ligands amphiregulin (AR), epiregulin (EPR) and transforming growth factor, alpha (TGFA) for SUM149 vs 190, while the transcript levels for ERBB2 and associated receptors/ligands HBEGF, ERBB3 and 4 are increased in SUM190 vs 149. Amphiregulin is identified as a ligand of EGFR and acts as an effective mitogen for epithelial cells. Epiregulin is another EGFR ligand that binds directly to EGFR and regulates tyrosine phosphorylation of EGFR. On the other hand, ERBB3 and ERBB4 are reported to be part of ERBB2 heterodimer in ERBB2 signaling, and ERBB3 has been reported to be necessary for tumor cell proliferation in breast cancer.

**Proteomic Analysis of SKBR3, SUM149, and 190 Cell Lines**

With the importance of ERBB2 and EGFR signaling indicated by the RNA-Seq data, we then examined the correlation between our proteomics data, transcript levels and chromosome location. In Table S2 (Supporting Information) we have ranked the 20 most abundant proteins as measured by spectral count in the SKBR3 cell line (highest number of protein observations) and compared these values with the corresponding RNA-Seq levels as well as the proteomic values for SUM149 and 190 cell lines. As has been reported elsewhere there is a general correlation between the levels of a transcript and the corresponding proteins, although relative differences in transcript and protein stability as well as temporal events can result in exceptions to this rule. The genes TUBB, ACTB and GAPDH, which were selected as housekeeping proteins for normalization of the proteomic data, were indeed observed at high levels (spectral count rank 17, 8, and 7, respectively). Conversely, the genes HIST1H4A, EPPK1, ENO3 and FLNA offer examples of poor correlation with a rank of 15, 9, 10, 11 in the proteomic data and a RPKM of only 3, 4, 2, and 6, respectively. While the selection of 20 examples in Table S2 (Supporting Information) as a representative protein set is arbitrary, it is of interest to note that 11 of the 20 proteins are

![Figure 2](http://example.com/image.png)

**Figure 2.** A composite of SUM149 (A) and SUM190 (B) transcriptomic, proteomic, and interaction data for significant oncogenes observed in SUM149 and SUM190. The following notations are used. Line length: Interaction score (shorter line, stronger interaction with ERBB2). Circle size: RPKM value (largest: RPKM > 15, medium: RPKM between 3 and 15, small: RPKM between 1 and 3, spot: RPKM <1). Black circle: if observed in proteomic experiments. Percentage: percentage of proteins identified in SUM149 or 190 with specific oncogene interactions as listed by STRING or I2D in Genecards.org.

### Table 2. Erbb Receptor Signaling Networka with RNA-Seq Ratios (SUM149 vs SUM190)b

<table>
<thead>
<tr>
<th>Gene</th>
<th>RPKM-SUM149</th>
<th>RPKM-SUM190</th>
<th>Ratio (149/190)</th>
<th>ERBB2 interact</th>
<th>Novoseek tumor hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR</td>
<td>60.1</td>
<td>0.6</td>
<td>120</td>
<td>+</td>
<td>1513</td>
</tr>
<tr>
<td>AREG (AR)</td>
<td>73.7</td>
<td>4.9</td>
<td>37</td>
<td>+</td>
<td>27</td>
</tr>
<tr>
<td>EREG (EPR)</td>
<td>7.9</td>
<td>0.5</td>
<td>3.2</td>
<td>+</td>
<td>926</td>
</tr>
<tr>
<td>TGFα</td>
<td>6.5</td>
<td>0.1</td>
<td>-4</td>
<td>+</td>
<td>169</td>
</tr>
<tr>
<td>HSCC</td>
<td>1.5</td>
<td>4.7</td>
<td>-1.2</td>
<td>+</td>
<td>103</td>
</tr>
<tr>
<td>ERBB3</td>
<td>6.6</td>
<td>23.1</td>
<td>+1.7</td>
<td>+</td>
<td>5807</td>
</tr>
<tr>
<td>ERBB4</td>
<td>6.0</td>
<td>3.0</td>
<td>-2.0</td>
<td>+</td>
<td>5807</td>
</tr>
</tbody>
</table>


**b**RPKM values are used to show the expression differences in two IBC cell lines, and the values in the ratio column are calculated as follows: Ratio (149/190) = log((RPKM SUM149 + 1)/(RPKM SUM190 + 1)). By adding 1 to RPKM values artificially, the ratio could still be calculated even if RPKM value is 0.
located on just 3 chromosomes: 6, 12, and 17. The possible significance of this observation will be discussed in the next section. One of the proteins coded by the gene MYH9 is a known oncogene,\textsuperscript{32} and such a high level of expression is of potential interest.

It has been reported there is a relationship between levels of gene expression and gene density in a chromosome region.\textsuperscript{9} Figure 3 shows the number of proteins identified in the SKBR3 study reported for each chromosome together with the % observed (number of protein observations divided by the number of protein coding genes on the chromosome). It is not surprising that the highest number of protein observations occurs for chromosome 1 and the lowest for chromosome 13 (largest chromosome and chromosome with lowest number of protein coding gene density, respectively). The highest % values were observed for genes 17, 12, 20, and 22, and while there is some correlation with reported gene densities on each chromosome (order of gene density is 19, 17, 20 and 22, high to low) it is relevant to note that chromosome 12 had 5 of the 20 most abundant proteins in Table S2 (Supporting Information), followed by chromosome 17 (3). Another factor is that chromosome 17 contains the highly expressed oncogene ERBB2 that can amplify a set of genes colocated near this oncogene (9).

Comparison of Proteomic Observations between Cell Lines

One of the challenges of studies with cancer cell lines compared with patient derived tumor samples is the lack of a suitable control samples. We chose the levels of ERBB2 as the comparator and compared the relative abundance of proteins in the two ERBB2 expressing cell lines (SUM190 and SKBR3, RPKM = 400 and 300) with SUM149 (RPKM = 14) in terms of unique proteins and for proteins with a 10-fold higher expression (see Tables S3, Supporting Information). Examples of proteins observed with this approach include the RAS associated proteins that are commonly activated in tumors in which ERBB2 is over-expressed.\textsuperscript{31,33} RAS-related proteins were preferentially observed in SUM190 and SKBR3 (ERBB2+) in that of the 24 different types of RAS-related proteins identified, SUM190 and SKBR3 accounted for 15 and 20, respectively, while only 6 were shared by all three cell lines. In addition, there are 5 RAS proteins with relative abundance 2-fold higher in SUM190 and SKBR3 compared to SUM149. Another example is cathepsin D which was elevated 6\times and 10\times more in SUM190 and SKBR3 compared to SUM149 and has previously been associated with Her2 amplification\textsuperscript{23} and is a marker of breast cancer marker.\textsuperscript{12} While this type of data analysis did detect some proteins with cancer associations it did not lead to pathway discoveries similar to that observed with the RNA-Seq analysis, and thus we explored alternative approaches.

Mapping of Oncogene Interactions with Proteomic Observations

With the use of interaction scores provided by Genecards (String, I2D) we recorded the values for interactions between the proteins identified in the proteomic studies of the two IBC cell lines and 21 oncogenes listed in Table 1. The large data set is given in Tables S3 and S4 (Supporting Information), and a summary is given in Figure 2 with the proteomic and transcriptomic experimental data as well as number of interacting proteins. First, Figure 2 shows oncogenes that are known to interact with ERBB2, and the oncogenes that show a high degree of interaction (EGFR, ERBB3, ERBB2IP, GRB2, GRB7, KRAS) are denoted by a shorter line. A relatively high RNA-Seq measurement is shown by the size of the circle, e.g., ERBB2, GRB7 and MYC, and those oncogenes with a proteomics value are shown with a black outline, e.g., ERBB2, GRB7, CRKL, TOP2A (see Table 1 for numerical values). For each oncogene, the number of interactions with proteins observed in the proteomic studies of either SUM149 or 190 is given in the circle as a percentage of the total oncogene interactions. As shown in Figure 2 the top 3 oncogenes with the greatest number of interactions with observed proteins are MYC, GRB2 and EGFR with 268, 235, and 143 interactions,
respectively, for SUM149. The basis for this approach has been used by others in the development of bioinformatic processes for prioritizing cancer associated genes with gene expression data combined with protein–protein interaction network information, as well as the observation that proteomic data when combined with genomic information can add further discrimination to pathway analysis. Thus in our approach we have combined mapping of oncogenes with RNA-Seq levels and identification of interacting proteins in the proteomic data set, and we will now use this data to search for additional pathways of interest in breast cancer.

**Identification of Pathways That Contain ERBB2, EGFR, GRB2 and MYC Interactors**

As an example of our process we describe the selection process for ERBB2 interactors. From the proteomic data set 35 proteins were found to be interacted with ERBB2 on the basis of I2D and STRING databases. We then selected a subset of 14 proteins according to levels of protein expression (spectral counts) and RNA-Seq values in the two IBC cell lines, SUM149, 190 and the model cell line SKBR3 (see Table S3, Supporting Information).

Table S3 (Supporting Information) also lists the chromosome locations of the interacting proteins, and it is noteworthy that many of the genes in these pathways are located on cytoband 17q12, which is the site of the ERBB2 amplicon. Of this group of chromosome 17 genes, ERBB2, GRB7, STAT3 and KRT17 are located in the same chromosome region (17q12 to q21.2) and have the following Novoseek tumor associations based on literature text-mining (Genecards): 5, 22, 693, and 24. The next stage in our process was to select disease relevant pathways based on our integration of transcriptomic, proteomic and interaction data. Our goal was to find at least one pathway for each of the 4 oncogenes that were well represented by the proteins listed in Table S3 (Supporting Information) and we used Cytoscape and Pathway Commons in this search. The pathways that we have selected are ERBB2, MYC, and PI3K signaling pathways from NCI Pathway Interaction Database, EGFR from the Cancer Cell Map and Integrin Signaling (GRB2) from GenGo.

In Table 3 we have listed all the proteins identified in EGFR1 signaling pathway as well as oncogenes (including those only observed with significant levels of transcript) in order of the ratio of SUM149/190 RNA-Seq values. This approach allows us to take advantage of the much greater dynamic range for RNA-Seq vs proteomics to compare differences between the two cell lines. We then compared these ratios with the proteomics data obtained for these two cell lines. The control cell line SKBR3 expresses high levels of ERBB2 transcript (300) and lower levels of EGFR (1.4) and shows proteomic values that are mostly intermediate between SUM149 and 190. In Table 3 we highlighted in yellow the proteins with higher expression in SUM149 and in yellow the proteins with higher expression in SUM149 and in...
blue those with higher levels in SUM190. In general there was

good agreement between RNA-Seq and proteomic values, e.g.,

CAV1, PLEC for higher ratios of EGFR vs ERBB2 and GRB7,

CRKL and CTNND1 for higher ratios of ERBB2 vs EGFR.

CRKL has been shown to associate with lamellipodia formation
in breast carcinoma, and coactivation of CRKL and estrogen

receptor alpha has been shown to be a promoter of tumori-
genesis. These observations are supported by literature reports,
such as CTNND1 was genomically correlated to breast cancer
and cell proliferation in ERBB2 positive breast cancer cell

lines. The overexpression of caveolin-1 (CAV1) is frequently
related to breast cancer and has been reported to be

\[ \text{Ratio} (149/190) = \log_2 \left( \frac{\text{RPKM SUM149} + 1}{\text{RPKM SUM190} + 1} \right) \]

Proteins with higher expression in SUM149 are highlighted in yellow and in blue those with higher levels in SUM190.

\[ \text{Ratio} (149/190) = \frac{\text{SUM149} - \text{SUM190}}{\text{SUM190}} \]

Proteins with higher expression in SUM149 are highlighted in yellow and in blue those with higher levels in SUM190.

Table 4. Integrin Outside-In Signaling

<table>
<thead>
<tr>
<th>Gene</th>
<th>SKBR3</th>
<th>SUM190</th>
<th>SUM149</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNI</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>FLNA</td>
<td>420</td>
<td>5</td>
<td>524</td>
</tr>
<tr>
<td>ACTN1</td>
<td>75</td>
<td>54</td>
<td>154</td>
</tr>
<tr>
<td>S100</td>
<td>7</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ALCAM</td>
<td>158</td>
<td>62</td>
<td>103</td>
</tr>
<tr>
<td>ACTB</td>
<td>430</td>
<td>545</td>
<td>478</td>
</tr>
<tr>
<td>VCL</td>
<td>100</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>ITGB1</td>
<td>0</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>LN1</td>
<td>72</td>
<td>0</td>
<td>46</td>
</tr>
<tr>
<td>CNTN1</td>
<td>0</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>FGFR2</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>VTN</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ITGA9</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ACTR3</td>
<td>28</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td>MAP3K1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>RAC1</td>
<td>15</td>
<td>14</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 5. Validated Targets of C-MYC Transcriptional Activation (A Subpathway of c-MYC Pathway)

<table>
<thead>
<tr>
<th>Gene</th>
<th>SKBR3</th>
<th>SUM190</th>
<th>SUM149</th>
<th>SKBR3</th>
<th>SUM190</th>
<th>SUM149</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAC1</td>
<td>2</td>
<td>0.1</td>
<td>0.0</td>
<td>28.6</td>
<td>4.8</td>
<td>0</td>
</tr>
<tr>
<td>HMGA1</td>
<td>4</td>
<td>35.6</td>
<td>3.5</td>
<td>119.0</td>
<td>4.7</td>
<td>99</td>
</tr>
<tr>
<td>ICA1</td>
<td>1</td>
<td>4.0</td>
<td>0.0</td>
<td>11.1</td>
<td>3.6</td>
<td>0</td>
</tr>
<tr>
<td>CAD</td>
<td>43</td>
<td>2.5</td>
<td>3.2</td>
<td>22.0</td>
<td>2.5</td>
<td>2</td>
</tr>
<tr>
<td>RUVBL2</td>
<td>2</td>
<td>4</td>
<td>73.0</td>
<td>10.4</td>
<td>53.9</td>
<td>2.3</td>
</tr>
<tr>
<td>NCL</td>
<td>1</td>
<td>60</td>
<td>45</td>
<td>142</td>
<td>118.1</td>
<td>14.9</td>
</tr>
<tr>
<td>GAPDH</td>
<td>4</td>
<td>522</td>
<td>495</td>
<td>539</td>
<td>2966.0</td>
<td>205.2</td>
</tr>
<tr>
<td>BAX</td>
<td>2</td>
<td>2</td>
<td>12.8</td>
<td>2.8</td>
<td>13.3</td>
<td>1.9</td>
</tr>
<tr>
<td>CCNB1</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>15.2</td>
<td>6.3</td>
<td>26.3</td>
</tr>
<tr>
<td>PFKM</td>
<td>5</td>
<td>7</td>
<td>5.5</td>
<td>19.3</td>
<td>1.6</td>
<td>0</td>
</tr>
<tr>
<td>E4F1</td>
<td>4</td>
<td>49</td>
<td>96</td>
<td>77</td>
<td>128.0</td>
<td>21.6</td>
</tr>
<tr>
<td>ENO1</td>
<td>5</td>
<td>758</td>
<td>1073</td>
<td>661</td>
<td>481.8</td>
<td>169.7</td>
</tr>
<tr>
<td>LCHAPF</td>
<td>3</td>
<td>1</td>
<td>12</td>
<td>3.3</td>
<td>10.7</td>
<td>1.4</td>
</tr>
<tr>
<td>LDIA</td>
<td>1</td>
<td>179</td>
<td>115</td>
<td>215</td>
<td>213.6</td>
<td>19.2</td>
</tr>
<tr>
<td>TK1</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>120.2</td>
<td>11.0</td>
<td>28.3</td>
</tr>
<tr>
<td>SLC2A1</td>
<td>5</td>
<td>22.1</td>
<td>7</td>
<td>19.7</td>
<td>1.3</td>
<td>451</td>
</tr>
<tr>
<td>HUWE1</td>
<td>15</td>
<td>15</td>
<td>57</td>
<td>3.3</td>
<td>16.1</td>
<td>36.0</td>
</tr>
<tr>
<td>HSPDI</td>
<td>216</td>
<td>117</td>
<td>187</td>
<td>73.6</td>
<td>14.5</td>
<td>31.8</td>
</tr>
<tr>
<td>PRDX5</td>
<td>21</td>
<td>7</td>
<td>7</td>
<td>25.6</td>
<td>14.6</td>
<td>31.2</td>
</tr>
<tr>
<td>ACTL6A</td>
<td>4</td>
<td>12</td>
<td>12</td>
<td>12.1</td>
<td>12.6</td>
<td>24.1</td>
</tr>
<tr>
<td>EIF4G1</td>
<td>25</td>
<td>12</td>
<td>45</td>
<td>35.8</td>
<td>37.8</td>
<td>69.9</td>
</tr>
<tr>
<td>NME1</td>
<td>29</td>
<td>23</td>
<td>20</td>
<td>0.1</td>
<td>26.4</td>
<td>36.0</td>
</tr>
<tr>
<td>PTMA</td>
<td>6</td>
<td>9</td>
<td>3</td>
<td>550.7</td>
<td>77.5</td>
<td>85.0</td>
</tr>
<tr>
<td>EIF2S1</td>
<td>6</td>
<td>9</td>
<td>6</td>
<td>5.9</td>
<td>13.1</td>
<td>12.9</td>
</tr>
<tr>
<td>DDX5</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6.9</td>
<td>6.7</td>
<td>6.5</td>
</tr>
<tr>
<td>HSPA4</td>
<td>26</td>
<td>31</td>
<td>36</td>
<td>8.0</td>
<td>23.0</td>
<td>21.6</td>
</tr>
<tr>
<td>HSP90AI</td>
<td>260</td>
<td>326</td>
<td>287</td>
<td>254.6</td>
<td>58.8</td>
<td>42.3</td>
</tr>
<tr>
<td>EIF4E</td>
<td>3</td>
<td>7.5</td>
<td>3.1</td>
<td>1.8</td>
<td>0.5</td>
<td>181</td>
</tr>
<tr>
<td>MT1DH</td>
<td>41</td>
<td>21</td>
<td>12</td>
<td>23.2</td>
<td>48.2</td>
<td>14.3</td>
</tr>
</tbody>
</table>

This pathway was retrieved from GeneGo in January, 2012. Spectral counts. RPKM values. RPKM values are used to show the expression differences in two IBC cell lines, and the values in the ratio column are calculated as follows: Ratio (149/190) = log_2 \left( \frac{\text{RPKM SUM149} + 1}{\text{RPKM SUM190} + 1} \right). Proteins with higher expression in SUM149 are highlighted in yellow and in blue those with higher levels in SUM190.
associated with EGFR activation. Interestingly, the over-expression of both CAV1 and CAV2 has been discovered in triple negative (TN) invasive breast cancer. In our study, SUM149 is only the TN cell line, and CAV1 is only identified by proteomics in this cell line and a RPKM value (33.6) that is much higher than for two ERBB2+ cell lines, i.e., SUM190 (0.2) and SKBR3 (0.4). At the other extreme of Table 3, higher levels of ERBB2 transcript are associated with the proteomic measurement of GRB7, growth factor receptor-bound protein 7, which is part of the ERBB2 amplicon in breast cancer. In addition most of the proteins in Table 3 have been reported to interact with EGFR (30/34) and had literature associations with cancer (27/34).

Table 4 shows a similar analysis of the Integrin outside-in signaling pathway, which was selected as an example of the oncogene GRB2, and shows an elevation of filamin A (FN1), actinin alpha1 (ACTN1) in both the transcriptome and proteome of SUM149 vs 190 cell lines. Of interest, Filamin A phosphorylation has been shown to mediate the effects of caveolin-1 on cancer cell migration. For the c-MYC pathway (Table 5) the higher ratios of EGFR transcript were associated with increased proteomic levels of branched chain amino-acid transaminase 1 (BCAT1), cytosolic, carbamoyl-phosphate synthetase 2 (CAD) and nucleolin (NCL), while higher ERBB2 ratios are associated with transferrin receptor (TFRC) and metadherin (MTDH).

Table 6. p53 Pathway (A Subpathway of Class I PI3K Signaling Events Mediated by Akt)

<table>
<thead>
<tr>
<th>Gene</th>
<th>SKBR3</th>
<th>SUM190</th>
<th>SUM149</th>
<th>SKBR3</th>
<th>SUM190</th>
<th>SUM149</th>
<th>Ratio 149/190</th>
<th>PTEN interact</th>
<th>Novoseek tumor kits</th>
</tr>
</thead>
<tbody>
<tr>
<td>S100A2 d</td>
<td>14</td>
<td>1.6</td>
<td>0.4</td>
<td>53.8</td>
<td>5.3</td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EGFR d</td>
<td>12</td>
<td>0.4</td>
<td>0.2</td>
<td>53.6</td>
<td>5.3</td>
<td>66</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAV1 d</td>
<td>4</td>
<td>0.0</td>
<td>0.4</td>
<td>23.5</td>
<td>4.2</td>
<td>482</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TP53 d</td>
<td>14</td>
<td>0.0</td>
<td>0.4</td>
<td>23.5</td>
<td>4.2</td>
<td>482</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SERP BM5</td>
<td>6</td>
<td>15.2</td>
<td>2.8</td>
<td>64.2</td>
<td>4.1</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRMT1 d</td>
<td>1</td>
<td>0.4</td>
<td>0.2</td>
<td>33.6</td>
<td>4.9</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NDRG1 d</td>
<td>20</td>
<td>0.8</td>
<td>0.2</td>
<td>33.6</td>
<td>4.9</td>
<td>43</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PYCARD d</td>
<td>1</td>
<td>0.1</td>
<td>0.0</td>
<td>9.5</td>
<td>3.4</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPIX d</td>
<td>4</td>
<td>0.1</td>
<td>0.0</td>
<td>9.5</td>
<td>3.4</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPHA2 d</td>
<td>4</td>
<td>0.1</td>
<td>0.0</td>
<td>9.5</td>
<td>3.4</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRKCD d</td>
<td>3</td>
<td>0.1</td>
<td>0.0</td>
<td>9.5</td>
<td>3.4</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSEIL d</td>
<td>2</td>
<td>0.1</td>
<td>0.0</td>
<td>9.5</td>
<td>3.4</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BAN d</td>
<td>12</td>
<td>0.1</td>
<td>0.0</td>
<td>9.5</td>
<td>3.4</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRIM25 d</td>
<td>5</td>
<td>0.1</td>
<td>0.0</td>
<td>9.5</td>
<td>3.4</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IT1T d</td>
<td>1</td>
<td>0.1</td>
<td>0.0</td>
<td>9.5</td>
<td>3.4</td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMARCA4 d</td>
<td>5</td>
<td>21.7</td>
<td>8.4</td>
<td>31.4</td>
<td>1.8</td>
<td>61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCNB1 d</td>
<td>1</td>
<td>17.3</td>
<td>8.8</td>
<td>32.3</td>
<td>1.8</td>
<td>61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSHE2 d</td>
<td>2</td>
<td>2.8</td>
<td>0.9</td>
<td>30.6</td>
<td>1.6</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCNA d</td>
<td>25</td>
<td>16</td>
<td>20</td>
<td>47.2</td>
<td>5.0</td>
<td>2016</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRRAP d</td>
<td>2</td>
<td>1</td>
<td>0.2</td>
<td>1.2</td>
<td>1.5</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BID d</td>
<td>2</td>
<td>0.2</td>
<td>0.1</td>
<td>1.2</td>
<td>1.5</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PRMT5 d</td>
<td>4</td>
<td>0.3</td>
<td>0.1</td>
<td>1.2</td>
<td>1.5</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HUWE1 d</td>
<td>15</td>
<td>1</td>
<td>0.2</td>
<td>1.2</td>
<td>1.5</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLH1 d</td>
<td>52</td>
<td>31</td>
<td>5</td>
<td>345</td>
<td>21.9</td>
<td>430</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MET d</td>
<td>5</td>
<td>3.4</td>
<td>2.1</td>
<td>36.6</td>
<td>1.7</td>
<td>1123</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RPL5 d</td>
<td>34</td>
<td>35</td>
<td>29</td>
<td>658.1</td>
<td>145.3</td>
<td>379</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NEDD8 d</td>
<td>9</td>
<td>0.6</td>
<td>0.3</td>
<td>5.3</td>
<td>0.4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USP7 d</td>
<td>2</td>
<td>0.1</td>
<td>0.0</td>
<td>1.2</td>
<td>0.1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RB1 d</td>
<td>5</td>
<td>0.0</td>
<td>0.0</td>
<td>1.2</td>
<td>0.1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PPPCA d</td>
<td>5</td>
<td>0.0</td>
<td>0.0</td>
<td>1.2</td>
<td>0.1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DDX5 d</td>
<td>6</td>
<td>0.0</td>
<td>0.0</td>
<td>1.2</td>
<td>0.1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AP2M1 d</td>
<td>2</td>
<td>0.0</td>
<td>0.0</td>
<td>1.2</td>
<td>0.1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PTEN d</td>
<td>2</td>
<td>0.0</td>
<td>0.0</td>
<td>1.2</td>
<td>0.1</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Examples of the significance of these proteins include the observation that nucleolin colocalizes with BRCA1 in breast carcinoma tissue, and metadherin is a valuable marker of breast cancer progression, and high expression may play a role in tumorigenesis of breast cancer. As was observed for the EGFR pathway, most of the proteins in Table 4 (GRB2) and Table 5 (MYC) contained a significant number of interactors (12/17 and 28/31) and literature associations with cancer (16/17 and 24/31) respectively.

A similar analysis of the p53 pathway is shown in Table 6. This pathway is a subpathway of Class I PI3K signaling events mediated by Akt and was selected as an example of the oncogene PTEN (phosphatase and tensin homologue). Tumor suppressor PTEN has been observed to be deleted in TN breast cancer, which shown related to resistance of EGFR targeting therapy. In our data set, SUM149, which is a classic TN breast cancer, has a very low level transcript expression of PTEN (0.8), compared to SUM190 (6.1). Interestingly, another tumor suppressor, SERPINB5 (Serpin B5), which has been reported to be negatively correlated with both ER and PGR genes in a quantitative DNA analysis, was only observed in SUM149 (proteomics and transcriptomics), which is the only TN cell line in the study. Likewise, amplification of S100A2 (Protein S100-A2) was observed in both proteomics and transcriptomics experiments.
This protein, as one of S100 families, has been reported to be upregulated in mRNA expression in ER-negative breast cancer patients and potentially promote cancer metastasis.\(^{52}\) SFN (14-3-3 protein sigma), which acts as p53-regulated inhibitor of G2/M progression, has been reported to be silencing due to DNA hypermethylation in breast cancer.\(^{33,34}\) A similar silencing due to methylation for PYCARD (or TMS1) has been observed in breast cancer cells.\(^{35}\) However, both overexpression of SFN and PYCARD in transcript and proteomic level was detected in SUM149, which could provide a potential diagnostic marker for TN breast cancer. Similarly, EPH2, which overexpresses in more than 60% of breast cancer patients,\(^{36}\) has been listed as potential clinical target in TN breast cancer.\(^{44}\) Expression of EPH2 has been observed to be stimulated by the activation of EGFR.\(^{37}\) This is consistent with the EPH2 expression in our experiment, in which EPH2 was only identified in SUM149 (proteomics) and greatly amplified in transcriptomic level.

\section*{CONCLUSION}

In view of the importance of EGFR/ERBB2 heterodimer signaling in breast cancer, it is of interest to explore the transcriptomic and proteomic analysis of two primary cell lines isolated from inflammatory breast cancer patients, one (SUM149) that expresses high levels of EGFR transcript with much lower levels of ERBB2 (1/4), while the other expresses very high levels of ERBB2 transcript (SUM190) and no detectable EGFR transcript. As a control we used a SKBR3 cell line that expressed high levels of ERBB2 transcript and low levels of EGFR. Analysis of the transcript levels indicated that the most likely signaling pathway for SUM190 involved the ERBB2/ERBB3 heterodimer, while SUM149 had several possibilities with involvement of EGFR dimers, ERBB2 heterodimers with EGFR and ERBB2 or ERBB3. We then explored the proteome of the two cell lines in terms of correlations between the transcriptome and proteome measurements, identification of a panel of 21 oncogenes expressed in the two cell lines, interaction analysis of the observed proteins with this panel of oncogenes and selection of relevant cancer pathways. The analysis resulted in 4 pathways in addition to ERBB2 signaling (EGFR, integrin, MYC signaling, and PI3K signaling), see Tables 4–6 that contained many of the oncogene interacting proteins. In general there was reasonable agreement between the RNA-Seq and proteomic values shown in these tables except for some housekeeping proteins (see Discussion section). We list here those proteins that were correlated with higher levels of EGFR or ERBB2 transcript, respectively. EGFR signaling: caveolin 1 (CAV1), pleckin (PLEC) (EGFR); growth factor receptor-bound protein 7 (GRB7), Crk-like protein (CRKL) and Catenin delta-1 (CTNND1) (ERBB2). Integrin signaling: filamin A (FN1) and actinin alphal (ACTN1) (EGFR), MYC signaling: branched chain amino-acid transaminase 1 (BCAT1), carboxyl-phosphate synthetase (CAD), nucleolin (NCL) (EGFR); transferrin receptor (TFRC), metalderin (MTDH) (ERBB2). p53 signaling: S100-A2 protein (S100A2), caveolin 1 (CAV1), Serpin BS (SERPINB6), stratatin (SFN), PYD and CARD domain containing (PYCARD), and EPH receptor A2 (EPHA2) (EGFR). While the depth of the proteomic analysis was limited partly because of technical issues with analysis of the primary cell lines, this study was designed to use proteomics to identify higher level protein expressions that correlated with the transcriptome study. In this study we have demonstrated that one of the goals of the chromosome-centric human proteome project (C-HPP), which is to integrate RNA-Seq with proteomics measurement, is of value. We plan in a future study to explore the potential of the proteins identified in this study as markers of ERBB2 and EGFR signaling as well as activation of the oncogenes MYC and GRB2 in a study of breast cancer tumor samples.

\section*{ASSOCIATED CONTENT}

\section*{Supporting Information}

Supplemental figures and tables. This material is available free of charge via the Internet at http://pubs.acs.org.

\section*{AUTHOR INFORMATION}

Corresponding Author

*E-mail: wi.hancock@neu.edu. Tel: 617-373-4881. Fax: 617-373-8795.

Notes

The authors declare no competing financial interest.

\section*{ACKNOWLEDGMENTS}

This work was supported by following research grants: (Korea) The World Class University program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10086-0 (W.S.H. and Y.-K.P.), National Project for the Personalized Genomic Medicine A111218-11-CP01 (to Y.-K.P.) from the Korean Ministry of Health and Welfare; (USA) The National Institutes of Health Grants, U01-CA128427 to W.S.H., U54DA021519, UL1 RR024986, RM-08-029, and U54ES017885 to G.S.O.; NIH grant (M.P.S. and H.I.); Texas State Rider for the Morgan Welsh Inflammatory Breast Cancer Program and the G. Morris Dorrance Jr. Chair in Medical Oncology (M.C., Z.M.).

\section*{REFERENCES}


overexpressed and in a tight complex with HER2 in breast cancer. EMBO J. 1994, 13 (6), 1331–40.


